18

Assessment for Licensing and Certification

Melissa J. Margolis
NBME

Rebecca S. Lipner
American Board of Internal Medicine

Chad W. Buckendahl
ACS Ventures

Margolis, M. J., Lipner, R. S., & Buckendahl, C. W. (2025). Assessment for licensing and certification. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 1265–1338). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0018

Regulation, credentialing, licensure, certification: These terms refer to aspects of occupational oversight that may occur at the federal level, the state level, the intra-professional level, or a combination of all three. Regardless of the responsible entity or entities, this oversight serves the primary purpose of ensuring that practitioners demonstrate a specific level of competence to protect themselves, the profession, and, most important, the public from harm.

Regulation tends to refer more generally to higher level oversight of different aspects of a profession. In essence, professional regulation can refer to "any activity that is intended to promote and protect the public interest by reducing, suppressing, mitigating or eliminating harms or potential harms stemming from the practice of a profession" (Balthazard, 2017). In some domains, professional regulation is, relatively speaking, straightforward: Professional engineering practice, for example, is governed by individual states that have specified educational, internship, and examination licensure requirements. The work of the state licensing boards is supported by a regulatory body, the National Society of Professional Engineers, which promotes strong licensure laws intended to protect "public health, safety, and welfare" (2023, Our Values section). Other domains, such as medicine, have a multifaceted and complex regulatory process involving numerous required or elected entities that have interconnected levels of oversight throughout a given physician's professional career (White, 2014).

Credentialing is a term that can refer to the process of verifying eligibility for participating in a profession. Though such verification may occur at multiple points along the continuum of education and training (e.g., verifying completion of the educational requirements for taking a required examination, verifying successful completion of required examinations), the ultimate credentialing activity requires verification of *all* credentials that are necessary for performing the required job. Examples of credentials include licensing and certification, as well as other types of designations such as professional certificates, badges, or microcredentials. For the purposes of this chapter, the examples focus primarily on professional licensing and certification rather than on other types of credentials that have more recently gained traction in the professional arena. When relevant to the discussion, these other designations will be considered.

Licensure and *certification* typically refer to processes that are intended to provide the public with evidence that certain professional standards have been met; through these two processes, members of a profession are granted certain privileges that are associated with that profession (Raymond, 2015). A historic, yet somewhat tenuous, distinction between the two credentials relates to whether obtaining the credential is mandatory or voluntary. With licensing, a governmental agency grants permission to engage in an occupation after the specified requirements have been met; these requirements usually relate to educational activities, experience working in a supervised setting, and one or more examinations that a candidate must complete to demonstrate minimum competence with respect to the knowledge, skills, and abilities that are required for practice. For those occupations that require it, licensing is mandatory; without a license, one is not permitted to engage in the occupation even if all other requirements have been

met. Alternatively, certification typically is *not* a requirement for occupational practice. Instead, it traditionally has been viewed as an activity that is designed to demonstrate competence for entering an advanced level of practice; as such, it is more a reflection of professional oversight rather than the governmental oversight that applies to licensing. To further complicate matters, certification in business and industry may be used to indicate mastery of knowledge and skills associated with products or services, rather than as a gateway into a field (e.g., technology hardware, software, medical devices). Though the historical approach to distinguishing between licensure and certification based on the mandatory versus voluntary nature of the process has been popular and in many cases is accurate, there are exceptions to this general rule. In some professions, such as teaching, certification is mandatory and acts as a de facto license (Raymond, 2015).[1] There are professions in which both credentialing practices are relevant, and the distinction between them is clear. In the licensing of allopathic physicians, for example, licensure is governed by the state medical board; candidates must meet the educational, training, and assessment requirements for licensure and then apply to the state for the undifferentiated license to practice medicine. Should a licensed physician desire specialty certification, they must complete additional required training in that discipline and then pass one or more certification examinations administered by the specialty board; passing the examination(s) allows that diplomate to use the associated specialty designation (e.g., a board-certified plastic surgeon). In other professions, the distinction between licensing and certification is less clear. The medical example leads to consideration of an additional distinction that often is used to differentiate the two credentials: Licensing tends to be conceptualized as more of a hurdle to enter a given profession, whereas certification tends to reflect a more advanced level of practice (Kane et al., 1999). There are additional complexities associated with variability in credentialing requirements within a given field. In psychology, for example, health service (clinical, counseling, and school) psychologists typically require a license to practice, whereas the same credential may not be required in other applied psychological domains (DeMers et al., 2014; Elchert, 2016). With few exceptions, because licensing in the United States is a function of individual states/territories, there is variability across jurisdictions regarding which professions require licenses and what is required to maintain them.

Despite differences in terminology or specific requirements, the overarching goal of both credentialing processes is the same: to ensure that the practitioner has attained at least a minimum level of competence in the profession to protect the welfare of the public. As we have mentioned, professional credentialing organizations typically specify at least three types of requirements for receiving the credential: educational, experiential, and examination. This is not universal, however, and many certifications do not have education or experience as eligibility requirements to sit for an examination. The specific educational and experiential requirements for licensure and certification vary greatly across professions and jurisdictions; considering the third requirement—completion of one or more examinations—allows for focusing on the similarities in requirements that exist across these high-stakes contexts.

Each year, hundreds of thousands of test takers complete examinations that are designed to provide information about the extent to which they have met certain professional standards (Raymond, 2015). Because licensing and certification examinations are intended to inform decisions about a candidate's ability to perform effectively at the relevant level of practice (Kane et al., 1999), developing these tests focuses specifically on identifying the necessary requirements for professional practice at the level of interest (e.g., entry level for licensing, a more advanced level for certification); this allows for greater confidence in the link between performance on the assessment and effective job performance. Additional important considerations with respect to these examinations relate to topics such as legal issues associated with using tests to make decisions about the ability to enter practice (or practice using a specific title), the importance of methodological rigor throughout the item and test development process, the extent to which relevant stakeholders are included in the process, elements of standard setting that are unique to or warrant special attention in credentialing contexts, and other factors that are relevant to settings in which results of the assessment process have significant professional implications. By outlining these critical areas, the chapter aims to support the development of more effective, equitable, and evidence-based assessment practices in high-stakes professional contexts.

As should be clear from the preceding pages, we present a broad review of considerations that are applicable to professional licensing and certification contexts. Other chapters in this volume delve into many of these issues in substantial detail, and the interested reader is referred to these chapters for an in-depth discussion of relevant content within a given topic area. We intentionally limit our review to the specific context in which tests are used to make decisions about whether a practitioner has demonstrated competence to practice in a given profession or to use a specific practice-related designation. As such, the content of the chapter is presented with the intent of meeting the following goals:

- providing a clear presentation of the issues that historically have been and currently are most relevant for defining use(s) for developing, scoring, evaluating, and defending scores from licensing and certification tests;
- describing the legal precedents that have continued to shape standards for licensing and certification tests;
- presenting a detailed explanation of the foundational measurement issues that guide development and use of tests for licensing and certification (e.g., considerations of fairness, reliability, validity, establishing performance standards); and
- discussing related factors including policy considerations, stakeholder perceptions, and antitesting sentiments.

We conclude our introduction with an overview of the major sections of the chapter: "Rationale and Historical Perspective," "Fairness and Legal Issues," "Test Development," "Reliability," "Validity," "Standard Setting," and "The Future of Assessment for Licensing and Certification."

1. "Rationale and Historical Perspective": In this section, we begin by describing the historical context for licensing and certification assessments and explore how perspectives on these types of assessments have developed and changed over time. Content represented in this historical background includes a discussion of different types of tests and their uses, with a focus on licensing and certification, an examination of how licensing and certification practices differ in different contexts, and a focus on how assessment for licensing and certification has changed since the publication of the prior *Educational Measurement* volume in 2006.
2. "Fairness and Legal Issues": This section reviews important considerations with respect to fairness and legal issues in credentialing contexts. The primary focus is on the legal bases that would allow for challenging testing practices and/or outcomes in both licensing and certification contexts (e.g., Title VI and VII of the Civil Rights Act of 1964; Equal Employment Opportunity Commission, 1985). We also consider fairness from the perspectives of mitigating risk for bias during test development, testing accommodation policies, and the implications of the outcomes of accommodated and nonaccommodated test scores.
3. "Test Development": This section describes test development considerations that are particularly relevant to licensing and certification assessments. These include a discussion of the importance of clearly specifying the purpose of the test and focusing on test development as a data-driven process informed by an understanding of the knowledge and skills that one must demonstrate in the context relevant to the specific credential. Recent research on advanced test development approaches such as automated item generation and automated test assembly are included, as is a discussion of innovative item types and issues surrounding the provision of performance feedback.
4. "Reliability": This section addresses relevant issues that are unique to or require special emphasis in licensure and certification contexts. Content includes discussion of several key topics, including classification consistency and accuracy, with a particular focus on false-positive and false-negative errors and the impact of these factors on policy decisions, as well as a review of measures of classification accuracy in the contexts of both traditional and cognitive diagnostic assessments. We also discuss the special issues in classification accuracy associated with allowing failing test takers to retake an examination.
5. "Validity": In this section, we describe validity issues that are unique to or need special emphasis within credentialing contexts. We highlight Kane's (2006) validity framework to discuss the importance of a multifaceted and thorough approach to collecting validity evidence in professional testing environments, discuss the particular importance of the extrapolation and interpretation components with respect to evidence for performance in practice, reflect on the theoretical underpinnings of why assessment may be important in the professions, and discuss threats to validity such as security-related issues.

6. "Standard Setting": This section addresses issues that are central to standard setting in assessments used for licensing and certification, including a brief discussion of the content-based methods typically used in credentialing contexts, considerations related to the variety of activities that comprise the overall standard-setting process, and a discussion of the role of judgment in standard setting.
7. "The Future of Assessment for Licensing and Certification": In the final section, we discuss what is to come in the domain of assessment for licensing and certification. This includes a review of major milestones that have influenced the practice of assessment for licensing and certification, such as the availability of new technology, reconceptualization of how competence is developed and demonstrated, and the rise of antitesting and antiregulation sentiments.

RATIONALE AND HISTORICAL PERSPECTIVE

Licensing and certification are based on a foundation that continues to evolve as professions seek to distinguish themselves through designation processes for both the profession and individuals. This history goes back hundreds of years. Some of the earliest records of licensure requirements are presented by Schmitt (1995), who described when tariffs were imposed on medical practitioners as part of an effort to regulate practice; these early requirements date to approximately 2000 B.C.E. (and are analogous to the requirement that exists in many jurisdictions of registering and paying a fee for a business license). Some of these fees were levied by governmental agencies, but as noted in the following paragraphs, professions also had an interest in controlling access to the profession itself. During this same period, the Chinese civil service began periodic testing to evaluate fitness for remaining in designated positions (see Dubois, 1970, and Clauser et al., this volume). Evidence of these efforts to regulate or monitor professions were related to healthcare, which contributed to the concern for protection of the health and welfare of the public. However, motivations for developing these systems were neither entirely altruistic nor fully driven by governmental or regulatory bodies.

The guild systems that emerged in Europe between the 1200s and the 1400s established rules for membership, expectations for education or training through apprenticeships, and member-determined expectations for entry (Schmitt, 1995). In this regard, these systems may appear analogous to contemporary professional associations that sponsor programs, but the original purpose of these systems was more about control of the profession and commerce than about ensuring minimum competency of individuals entering the profession. From an economic perspective, this control over the profession also would have influenced prices and wages. This initial shift from government-driven oversight to profession-driven monitoring represents an early phase in the movement toward more professional self-regulation. As with other policies that have used testing as part of oversight or accountability, licensing and certification programs have experienced periods of greater or lesser support for their efforts.

One reason for the reduction of oversight of certain professional licensing requirements in the mid-1800s was that there was a perception in some fields—such as medicine and law—that the schools training these practitioners were of a higher quality than had been observed at the outset of the licensing legislation. Not surprisingly, the growing recognition of supply and demand principles led to the opening of more schools to meet the increasing demand for training. However, by the late 1800s, the expectation that candidates from all training programs were equivalent had waned, as did the quality of practitioners (due to the increase in the population; Schmitt, 1995). Since then, the number of professions regulated by states or monitored by professional associations has continued to grow as professions and roles have continued to evolve.

Discussions among professional training or educational programs, credentialing bodies, practitioners, and the public have continued for decades, along with considerations for why some licenses or certifications exist. Administrators and faculty at many professional training programs often take the position that graduation from an accredited training program should be sufficient for candidates to claim that they have achieved minimum competence in the knowledge, skills, and abilities required for the credential. There have been criticisms of this perspective, however, including the fact that training programs may have incentives to retain students that are not based on competence. Instead, they are perceived to be based on factors such as the need to maintain tuition and related revenue, along with program accreditation requirements that might rely on test information to inform outcomes assessment. Concurrently, there are practitioners and policy makers (e.g., legislators, licensing board members, certification board members) who would claim that these conflicts of interest are precisely why an independent demonstration of competence is needed. Criticisms of the practitioner stakeholders often are consistent with those levied against the historic guild system, where there may be perceptions that the licensure or certification process is used as a barrier to entry of a profession and that it limits opportunities for otherwise qualified candidates, particularly those from underserved communities. Because the influence of licensing and certification testing impacts millions if not billions of people, concerns related to these perspectives are important considerations for the measurement community.

Focusing on licensure to illustrate the point, Shimberg (1982) noted that more than 800 professions require licenses; significant expansion in licensing requirements began in the early 1900s with the proliferation of licensing agencies established through jurisdictional legislation. Morath (2015) subsequently estimated that the number of licensed professions had increased to more than 1,100. When applied to the United States workforce, this represents approximately 30% of workers (Kearney et al., 2015). The increase is driven partly by the creation of new professions and roles that did not previously exist. In addition, professions have sought to increase credibility and add a layer of protection through legislative intervention. The public protection goal of professional licensing and certification serves as a powerful reference point for professions seeking both visibility and a modicum of protection from practitioners who do not meet a standard of minimum competence for professional practice.

Considering the now thousands of professions and specializations that are licensed or certified, the meaning of some of these credentials is at risk of being devalued. As an example, the plaintiff in a legal case in Utah challenged the Barber, Cosmetology/Barbering, Esthetics, Electrology and Nail Technology Licensing Board for the right to provide African hair braiding services without a license after being ordered to cease and desist from offering the service (Romboy, 2012). The court in this case noted that the specific hair braiding service was beyond the scope of the public protection interests of the licensing board and supported the plaintiff's right to be able to practice without a license. Similarly, there are instances where licensing and certification programs have been eliminated. For example, in 2019, Texas removed licensing requirements and related laws for plumbers (Ura, 2019). This decision was related to an effort to reduce barriers to entry into a trade, but unintended consequences for the public may result from making such a decision without consideration of the need for additional independent verification of the competence of individuals offering these services. These examples illustrate instances where licensing boards have potentially exceeded the scope of their public protection charge or where support for a licensing requirement has eroded (affirmatively ending a program) or become inactive (allowing a program to sunset through lack of renewal or reauthorization).

Licensure and Certification Processes

Licensing and certification are the two most prevalent examples of formal recognition associated with an occupation, role, or profession. Although we discuss them collectively, the requirements, interpretation, and use for these and other professional credentials are unique to the defined eligibility, demonstration of competence, and expectations for maintaining the credential. A first step in designing and developing a testing program involves beginning with a clear statement of purpose. As will be discussed, terminology associated with licensing and certification testing can quickly expand beyond the historical definitions. For the sake of consistency in this chapter, we will use licensing and certification terminology, but additional types of professional credentials will be discussed when relevant, along with how they connect to these more common forms.

Just as we recognize a range of achievement levels in education and a variety of roles within employment settings, it stands to reason that licensure and certification would reflect a similar diversity. This variation may be seen not only in the reasons for establishing specific licenses or certifications, but also in how scores are used and interpreted in relation to those purposes. For credentialing programs, the primary claim relates to public protection: Candidates who receive the credential meet the minimum standard for entry to the profession or ability to use the designation(s) afforded by the credential. This claim can be distinguished from a claim for uses of certification credentials in a work setting that might be used for employment selection or retention. The concept of public protection stems from the desire—of judiciary, legislative, or regulatory authorities, as well as professional associations—to help members of the public distinguish

between qualified and unqualified practitioners. For licensed professions that are directly associated with potential risk to the public (e.g., architecture, aviation, healthcare, law), actions of an incompetent practitioner can produce lasting harm.

In a more historical context, the concept of public protection often is closely associated with professional licensing and certification. However, particularly in certification, the purposes have expanded beyond this core purpose. For example, some certification programs were created to recognize individuals who have demonstrated specialized knowledge, skills, and abilities in a particular domain or subdomain within a profession (e.g., specialty certification, subspecialty certification). Other certifications may function more like educational assessments and are aligned with curriculum, instruction, training, or experience and designed to serve as evidence of knowledge or skills that would be needed or desired by employers (e.g., assessment-based certificates). Even for industry certifications that are based on experience or lack formal education eligibility requirements, the underlying emphasis remains on learning and the acquisition of knowledge and skills needed to demonstrate job-related competencies. Additional certification programs have expanded more into areas such as stackable credentials, microcredentials, and badging. Although these types of designations have become more prevalent and are increasingly visible, there is little consensus on quality standards and how they are to be designed, developed, evaluated/validated, or maintained. Therefore, though we note that this is an emerging area to continue to monitor, the breadth of licensing and certification testing continues to be guided by industry standards for quality.

Distinctions between goals of public perception and market recognition are important to remember when thinking about the value of these processes for stakeholders. From a candidate's perspective, the purpose of seeking a license or certification may be to demonstrate sufficient competence to be eligible to practice in or use a specific designation within a profession. Beyond minimal requirements for a role or profession, there often are market-based incentives for candidates to demonstrate deeper or more specialized competence that help them to distinguish themselves from colleagues. In addition, certain credentials may be used as part of employment eligibility requirements; this has implications for necessary validity evidence as well as legal expectations.

Link to Employment Testing

Because results from credentialing examinations may be required as part of employment eligibility processes, the question of whether these examinations could potentially be interpreted as employment tests that support selection purposes becomes relevant. Phillips (2017) discussed these implications based on the *Gulino v. Board of Education* (2002) case in New York. In that case, the court viewed the licensing requirement for teachers as a de facto employment test because teachers who had been employed without certification were either not promoted, not hired for full-time positions, or terminated based on their certification test performance. Although the primary

purpose of a credentialing examination may be to differentiate between candidates who are minimally competent and those who are not, integration with the employment selection process indicates a clear—though perhaps not universal—secondary use. The *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA] et al., 2014) distinguish between tests used for selection, placement, and promotion in employment contexts and those used for licensing and certification, which are designed to determine whether individuals meet established criteria for professional practice. Selection tests within employment settings may be used to rank order or categorize candidates in the hiring process. This is particularly useful if the number of available positions is limited. To accomplish this goal, the test needs to yield scores that support those types of decisions. This distinction between rank ordering candidates for selection, promotion, or retention purposes versus a broader claim of minimally qualified or not has implications for test design, development, and validation, as well as the evidence needed to support legal defensibility of the resulting scores (see Ercikan & Solano-Flores, this volume).

Because licensing and certification programs often have educational components as part of their eligibility criteria, the intersection of expectations for testing in these areas along with employment considerations requires clear distinctions among purposes and heightens the importance of transparency with respect to associated limitations. Confusion among stakeholders increases when programs' purposes are not clearly delineated. The expansion of purposes can raise challenges for licensing and certification programs in terms of both validation evidence and defensibility. For example, within student achievement testing in the United States, we have observed policies that incorporate results for purposes of both school and educator accountability. In higher education, results from licensing and certification exams often are used as an indicator of outcomes for purposes of program accreditation. Within licensing, provisional or conditional licenses may be granted in emergency situations that then require successful examination performance as a condition of continued employment. Often seen as intuitive by policy makers or sometimes even by programs seeking to provide evidence for or expand credibility of its credential, additional interpretations or use cases should be critically evaluated prior to implementation.

Program Sponsors

Licensure and certification programs have a wide range of sponsors. Agencies, associations, or organizations that grant the credential may assume responsibility for development and validation of the associated tests (e.g., American Board of Internal Medicine, National Commission on Certification of Physician Assistants). At the same time, some professions have formed associations or collaboratives where one or more intermediary agencies assume responsibility for development and validation of the tests (e.g., National Council of Architectural Registration Boards [NCARB], National Board of Medical Examiners, American Board of Dental Examiners, Federation of State Boards of Physical Therapy). In these instances, test users (e.g., state licensing boards,

regulatory agencies) may contribute to the process and then agree to accept or recognize the results of the tests to support consistency and portability.

For some credentialing contexts, membership associations may serve as the sponsor for the program (e.g., American Physical Therapy Association for physical therapy specialties, National Strength and Conditioning Association for strength and personal training). In some professions, this may be a single professional association; in others, there may be multiple membership associations or programs that compete for credibility and market share among members of the eligible population. In addition, some credentialing programs are sponsored by organizations that then seek to certify or confirm qualifications of individuals to perform specific roles within divisions of the parent organization (e.g., Armed Services, Federal Bureau of Investigation, National Security Agency within the U.S. Department of Defense). However, these credentials sometimes begin to have value in the workforce or labor market beyond just the requirements of the sponsoring organization. For example, industry certifications developed by companies in diverse economic sectors such as finance, healthcare, or technology are valued by a range of employers, internally and externally, that use products and services from these companies. Within employment settings, companies and governmental agencies may develop internal credentials that are used to support nonselective promotion or wage decisions.

The distinctions among government-regulated, profession-driven consortia or combinations of these processes are discussed in the next section. As noted previously, the primary differentiator is whether holding the license or certification is mandatory or voluntary for practice, with these differences creating more confusion among stakeholder groups.

Mandatory Professional Credentials

A license likely is the most frequently cited instance of a mandatory professional credential. Well-known professions such as architecture, accounting, law, medicine, nursing, and clinical psychology require practitioners to obtain a license to practice or represent themselves in their profession. What may be less well known, however, is that many jurisdictions also require professionals such as auctioneers, barbers, contractors, cosmetologists, real estate agents, and security guards to obtain a license to practice.

Licensure generally is a jurisdictional legislative responsibility. In the United States, there are some exceptions, including the Federal Aviation Administration's oversight of private pilot licenses and the judiciary's responsibility for admission to the bar for lawyers. As such, there may be jurisdiction-specific requirements for eligibility, demonstration of competence, and maintenance of the license. As described earlier, the focus of licensure is public protection and is based on the rationale that a governmental agency can serve as a dispassionate intermediary between the candidate and the public in evaluating minimum competence. However, even these agencies will necessarily rely on or seek input from members of the profession to help define expectations for entry-level practice.

Although having a license may be necessary to practice in a given field, it does not ensure employment or even employability because of other factors that may go into selection. Further, there are exceptions even to the *mandatory* requirement. For example, some emergency situations (e.g., severe shortages of practitioners, pandemics, natural disasters) may necessitate a temporary relaxation of requirements to grant a provisional or conditional license so that more practitioners can be mobilized. A credentialing examination may serve as an eligibility criterion within the hiring process, similar to the expectation that a valid driver's license is a prerequisite for applicants for a delivery driver position. However, scores from these examinations are not designed to predict successful performance on the job or distinguish among applicants who meet minimum eligibility requirements. In contrast to claims made about scores for employment selection tests, validity evidence for a licensing or certification exam should be sufficient to support the inference that a candidate has met the standard for minimum competence. In addition to this initial demonstration, some credential holders are expected to maintain the credential in good standing; this may involve activities such as continuing education, retesting, or contributions to the profession. The purpose of this practice is to ensure that licensees have a level of continued competence to maintain currency in the profession. This maintenance component often is what distinguishes professional credentials from certificates or industry certifications of products or services that are intended for specific skill sets (e.g., technology certifications, trade certifications, course-based certificates). However, it is important to note that some of these industry certifications do have a maintenance or continuing education requirement.

Voluntary Professional Credentials

In instances where someone can practice in a field without a license, a certification can serve as a point of distinction for the public. Because a voluntary certification may be sponsored by an association, a private organization, or the profession itself, there are fewer barriers to creating a certification; this can be offset by the challenge of building credibility for the value of the credential. Licensed physicians can work as general practitioners, but it is common to seek one of the many specialty/subspecialty certifications offered by American Board of Medical Specialties (ABMS) member boards (e.g., Emergency Medicine, Family Medicine, Internal Medicine, Obstetrics/Gynecology, Pediatrics). Within some jurisdictions, lawyers may seek specialty certifications in criminal law, estates and trusts, family law, or intellectual property. Individuals in the financial services sector may strive to obtain recognition through programs like the Chartered Financial Analyst Institute or as a Certified Financial Planner. These examples illustrate the range of options available to individuals seeking opportunities to voluntarily demonstrate competence within a specific professional domain or subdomain.

In contrast to the jurisdictional oversight of licensure programs, certification programs could be developed by a professional association or a company that offers a product or service on which someone desires to be certified. As a result, there is a wider

range of program designs and methods for implementing them in practice. There may be some level of competition within the certification program space where multiple providers offer credentials (e.g., crane operators, personal trainers, food protection). In these instances, it can be more difficult to ensure independence and mitigate conflicts of interests in governance policies with respect to program development and implementation.

In terms of test development and validation, certification programs are like licensing programs in that the primary claim relates to competence for practice. In addition to eligibility or prerequisite requirements, the demonstration of competence is based on the knowledge, skills, and abilities related to the type of practice implied by the certification. In industry, this may be defined in terms of a level of mastery; in other professions, it may relate to competence in some area of specialization. Earning a voluntary certification does not ensure employment but may be used in the employment process as part of preferred qualifications of applicants. In contrast, if a certification is used as an employment eligibility requirement, it could be interpreted as shifting the certification from voluntary to mandatory.

Although not universal, many professional certification programs have maintenance of competence requirements like those seen in licensure. These requirements will be unique to the program but may include such activities as continuing education, contributions to the profession, or retesting. In a separate section of this chapter, we further explore how some healthcare certification programs offer maintenance of competency assessments that blend learning and assessment. In addition, some certification programs have depended on assessment to inform continuing education recommendations for certified individuals. Related to licensure, there is some debate about whether maintenance of competence requirements should include a demonstration of *continued* or *enhanced* competence. From a measurement perspective, the question is related to the intended interpretation and use of the evidence from the maintenance of licensure process.

Program Stakeholders

Stakeholders for licensing and certification programs are varied, and each has its own perspectives and interests associated with the credentialing process. The public often is noted as the primary stakeholder because public protection tends to be the principal reason for the existence of these types of programs. Clearly, it is important to protect the public from an incompetent physician, teacher, attorney, engineer, or crane operator. In general, when there is a jurisdictional license involved, there has been a decision that incompetent practice could harm the public and that market forces may not offer sufficient protection. This may be particularly true for more technical or specialized domains where members of the public would not be able to differentiate competent from incompetent performance. Beyond initial licensure, another public interest may be served by recognizing individuals who have demonstrated more specialized competence within their field to distinguish themselves from others who have not achieved an additional credential.

Another key stakeholder group comprises the candidates or applicants seeking the credential. These individuals have a personal interest in achieving certification and a legitimate expectation of fair access to professional practice. Their perspective may shift after becoming certified, as the same standards that once posed a barrier now serve to uphold the value and credibility of the credential they hold. In many certification programs, candidates also are connected to professional associations or organizations that define standards for the profession. These groups represent another type of stakeholder with a dual responsibility: to protect the public interest and to serve their members. At the same time, they may have an organizational self-interest in maintaining the perceived value of the credential and sustaining membership, which supports their continued existence.

The training programs that prepare candidates for practice comprise yet another important stakeholder group. The nature of these programs varies across professions. In the case of medicine, the stakeholders in this group are medical schools. For certification, the stakeholders are residency programs. In both cases, access to the credential is important both because individuals are attracted to the program based on their belief that they will succeed in receiving the credential and because accreditation of the training programs may depend on satisfactory performance using this same metric.

Within the healthcare community, insurance providers are another stakeholder group. They may require medical professionals to have certain credentials to provide treatment and may have different pricing or reimbursement tiers for different types of credentials. Possession of the credential also may serve as a form of risk management by having third-party agencies or organizations recognize someone as having a particular skill set. In this instance, it distributes the risk across the insurance company, the provider, and the credential-granting entity.

For employers, credentialing adds a layer of quality control and external attestation of competence that may not be able to be discerned on one's own. Employers in the public and private sectors are stakeholders of credentialing programs and sometimes sponsors of these programs. For example, employers may require a given credential as an expectation for consideration of employment. In these instances, employers are placing trust in the credentialing body and the credential itself as being able to support associated claims about competence. Similarly, labor unions may want to use credentials not only for helping to establish quality standards for their members but also as part of a negotiating strategy to support increased leverage in advocating for the competence of their members.

Additional Evolving Practices

Whereas the fourth edition of *Educational Measurement* highlighted some of the measurement approaches in computerized scoring of performance assessments (e.g., NCARB), challenges associated with staying current with respect to changes in the technologies used in practice have altered the strategy for some of these programs.

For example, NCARB removed performance items (i.e., vignettes) that required candidates to use a generalized but simplified design software to respond to tasks. This decision was driven both by changes in practice regarding the roles responsible for these activities and by the job relatedness of the software that was being used. In response to those concerns, they opted for a more flexible strategy that permitted continued measurement of the important content and cognitive processes associated with being an entry-level architect and provided better alignment of the measurement with the task (see NCARB, 2023a; NCARB, 2023b). This approach to designing an examination—which is less dependent on a specific technology and more responsive to the knowledge and skills required by the profession—is becoming more prevalent in practice.

Similarly, as noted in the fourth edition, a content-neutral demonstration of professional skills has been part of many jurisdiction bar examinations for years. In this approach, candidates (or applicants) are provided with a library of resources and are expected to complete a task that requires them to engage with these materials and respond in a job-relevant manner. The emphasis on skills has taken on new meaning as the number of practice areas has continued to expand. Based on national job analysis research (National Conference of Bar Examiners, 2020), more than 25 different practice areas were noted by survey respondents, suggesting that the scope of practice is diverse. An additional unique feature of the bar examination is that, rather than being overseen by the legislative branch of government as is typical for other professions, it is the only licensure process overseen by the judicial branch of government (within jurisdictions).

As we hope has been demonstrated, there is a long and complex history behind the field of assessment for licensing and certification. Many advances have been made with respect to understanding the complexities of credentialing contexts and how to advance both theory and practice, yet many challenges still exist. The next section reviews topics relating to fairness and other legal considerations that can provide more insight into some of the issues and challenges that have faced this field in recent years.

FAIRNESS AND LEGAL ISSUES

The high-stakes nature of decisions made based on scores from credentialing examinations places additional responsibilities on programs that develop and maintain these examinations. Specifically, in addition to the professional expectations articulated in the *Standards* (AERA et al., 2014), programs and practitioners should be aware of the legal frameworks and case law that may interact with theory and practices within the measurement community. Phillips (2017) organized her comprehensive discussion of legal challenges that have occurred in licensure and certification using the following framework:

1. *Protecting the public*: Conditions under which it is fair to require candidates to demonstrate minimum competence for defined knowledge, skills, and abilities.
2. *Testing accommodations policies*: Defining criteria for disabilities and reasonable accommodations and ensuring that the meaning of the license or certification is not changed through modification of the construct.
3. *Test security policies*: Procedures for maintaining confidentiality of test content, ensuring valid score interpretation through detection and enforcement of test security violations.
4. *Test construction procedures*: Options for licensing and certification programs pressured to adopt item selection criteria that minimize differential performance between majority and minority candidates.

We discuss some of the recent case law associated with the first three of these areas and refer interested readers to Phillips (2017) for discussion of some of the historical background associated with the fourth area as it relates to legal precedent that conflicts with what often are considered professional expectations and best practices within measurement. Though this section largely focuses on case law covering issues related to fairness, the topic of fairness also is explicitly discussed relevant to test accommodation and score interpretation and is included as central to test development, reliability, and validity considerations. For a broader discussion of issues related to fairness, the interested reader is referred to Zwick, and Rodriguez and Thurlow, both this volume.

With a focus on U.S. jurisdictions, federal law forms the basis for most of the challenges described in this section. Most legal challenges to licensing and certification programs point to alleged violations of the 14th Amendment to the U.S. Constitution, Title VII of the Civil Rights Act (1964), Section 504 of the Rehabilitation Act (1973), the Copyright Act (1976), and the Americans With Disabilities Act (1990). The courts have relied on the *Uniform Guidelines on Employment Selection Procedures* (Equal Employment Opportunity Commission, 1985) for interpreting expectations for licensing and certification litigation. These legal expectations generally are intended to guard against discrimination of protected classes that include the demographic characteristics of age, disability, ethnicity, race, or sex. It is important to note that these legal expectations are not universal and some of the case law discussed in this section may not warrant action or be part of legislation in other countries. As such, readers should consider the jurisdiction-specific expectations applied to their licensing, certification, or equivalent program (e.g., qualification, accreditation, registration).

As noted earlier, there are documented examples of licensing boards overreaching in terms of their charge to protect the public (e.g., the plaintiff offering African hair braiding services). Similarly, decisions like the one against the North Carolina Board of Dental Examiners (2015) indicated that the courts recognize that there is a limit to how a licensing board defines its need for oversight and public protection. This case began with a complaint filed with the Federal Trade Commission by owners of teeth-whitening kiosks after the state board of dental examiners claimed exclusive domain under their practice

act and told them to cease and desist. The Federal Trade Commission concluded that the interpretation was an overreach of the practice act and ruled against the board, citing that, even within a dentist's office, nonlicensed staff members administered and monitored teeth-whitening procedures, thereby suggesting that the risk to the public was limited. Within Phillips's (2017) organizing framework, this illustration would fall within challenges related to assertion of public protection. Challenges to licensing and certification examinations may not come through judicial processes such as lawsuits because legislative (e.g., House, Assembly, Senate) and policy bodies (e.g., licensing or certification boards) seeking to sunset (Drew, 2016) or remove (Taylor, 2017) requirements may be as effective as litigation.

Because the treatment of the historical foundation for legal challenges for licensing and certification programs is articulated in greater detail by Phillips (2017), B. E. Clauser et al. (2006), Phillips and Camara (2006), and Ercikan and Solano-Flores, and Zwick (both this volume), we have limited our discussion to key decisions that have occurred within each of these areas since the publication of the fourth edition of *Educational Measurement*. This is not a comprehensive review of all decisions but rather illustrates the types of challenges that have occurred in licensing and certification.

Protecting the Public

Because of the potentially broader implications for licensing and certification testing programs, the *Gulino v. Board of Education of the City School District of the City of New York* (2002) case is of particular interest for the measurement community. In the Gulino case, the court interpreted a teacher certification examination as having properties of an employment test because of how the state permitted provisional or temporary licenses that allowed teachers to teach. This temporary license included an expectation for completion of licensure examination requirements within a specific time frame. Because passing the examination became a condition of continued employment in a given role, the court viewed this as a de facto employment test. When examinations are mandatory for employment eligibility, as is the case for licensing and many certification examinations, the Gulino ruling could be extended to the broader family of licensing and certification examination programs in which passing the examination is a condition of initial or continuing employment. Litigation on the case has been ongoing for more than 2 decades.

The original action was filed in 1996, when a group of African American and Hispanic teacher candidates alleged that the teacher certification test discriminated against racial minority candidates because passing rates for both racial/ethnic groups were below the 80% rule that defined disparate impact under the *Uniform Guidelines*. The 80% rule means that a protected class has an observed passing rate that is less than 80% of the passing rate of the majority group. The lawsuit was certified as a class action in 2001. As it relates to test development and validation, the court's interpretation of necessary evidence to support the test's continued use was important. The court ruled that five sources of evidence were needed to support the job relatedness of an employment test.

Specifically, these sources of evidence were (a) a suitable job analysis, (b) competent test development, (c) test content related to job content, (d) tested content representative of the job, and (e) an appropriate performance standard (i.e., passing score) for selecting competent applicants. The court held that the new test was not properly validated because the vendor did not conduct a rigorous job analysis, did not competently develop the test, tested content unrelated to and unrepresentative of the job of teaching, and applied a passing standard unrelated to teaching competence (*Gulino*, 2002).

More specifically, the court found that the vendor's test framework—developed by reviewing teacher education materials, consulting with education experts, and surveying teachers and teacher education faculty—was flawed because the vendor: (a) did not start with a task analysis of teaching; (b) had not documented the reference materials or persons interviewed; (c) had pilot tested items on college students rather than on working teachers; (d) had failed to link the tested content to minimum and representative content required for teaching competence; and (e) set the passing standard based on a small subset of items with no definition of minimum competence or data relating test scores to student outcomes. Thus, the court concluded that the board violated Title VII when it required teachers to pass the new test to obtain permanent licenses. The decision was affirmed on appeal.

Several lessons emerged from this case. An important lesson was that even though the *Standards* (AERA et al., 2014) is intended to serve as a primary guide for the testing profession, the court in this case did not recognize it as such. Further, because the court appointed a neutral expert who prioritized the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003) along with the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1985), the value of the *Standards* in this proceeding was further diluted. Although not intended to be prescriptive, if guidance from the *Standards* does not align with legal expectations, practitioners will be exposed to risk. Although the *Standards* includes a caveat about its utility in legal proceedings that may supersede the profession's expectations, these legal standards often are not widely known, nor are they necessarily consistent. Different jurisdictions may have different expectations, so beyond some core expectations that would be articulated in federal law (e.g., U.S. Constitution) or regulations (e.g., Civil Rights Act of 1964 as amended in 1991), the legal standard may not become fully evident until it is settled by the Supreme Court. The court's rationale for interpreting the state's credentialing exam as an employment test was that it was used as a condition for continued employment following issuance and potential extension of a temporary or provisional license, not just eligibility for employment. Therefore, without passing the examination, an applicant would not have been eligible for a permanent license. The focus of the proceedings then centered on the job analysis evidence and whether there was a clear connection between knowledge, skills, and abilities that an educator would need to know in practice.

Although evidence from the state and most evidence collected by the vendor from practicing educators and faculty who train educators suggested that there was a connection, the court's expert disagreed and offered an opinion that suggested a specific methodological preference for what would constitute an appropriate job analysis study. The perspective offered by the court's expert focused on the incremental validity evidence provided by the test in question and whether all educators should be required to demonstrate this level of minimum competence. During the hearing, the plaintiff's attorneys and the judge questioned whether the range of content represented on the test was job related and important for all educators. The intuitive perspective of the judge was that content areas such as fine arts would not apply to what a science teacher would need to know and be able to do (*Gulino*, 2015a).

Concurrently, the state had been in the process of redesigning and developing the examination with a focus on making changes to the content specifications. The revised examination recharacterized the construct as inclusive of literacy and writing abilities rather than the range of liberal arts skills that were defined in the prior version. This new exam also was challenged as part of the proceedings in the case as having adverse impact and not being properly validated. In this instance, the state's test vendor contracted with an external organization to conduct a practice analysis. The extensive study evaluated the knowledge, skills, and abilities of educators to determine whether the content of the examination was consistent with job-related expectations.

Again, the court's appointed expert concluded that the examination was not properly validated for reasons almost identical to those advanced for the prior iteration of the examination in the broader program. However, in this instance, the court's ruling for *Gulino* (2015b) yielded a different outcome in that the court accepted the content evidence for the program. Although the methodology for evaluating content evidence for these examinations was different, they were both within a common set of validation activities conducted for credentialing examinations. It is possible that the court's acceptance of the evidence of examination content for the new examination was based on an intuitive understanding of content that educators would need to know and be able to use. In other words, it was more difficult to argue that the literacy skills of reading and writing content areas were not important for all educators in practice, as opposed to the diversity of content represented in the prior version. However, the opinion of the court's expert likely leaves room for additional litigation on this issue because the implications for the testing program and its users are substantial. The ongoing litigation costs and questions regarding how to interpret the court's rulings in this case have continued to have ripple effects for other educator certification programs as well as for the licensing and certification community more broadly.

Testing Accommodations Policies

The Americans With Disabilities Act (ADA, 1990) requires that licensing and certification programs make decisions about accommodations for candidates with disabilities (see Rodriguez & Thurlow, this volume, for further discussion of fairness

issues related to test takers with disabilities). To support a violation of the ADA, plaintiffs need to provide evidence that there is a qualifying disability and that a reasonable accommodation was denied (*Cox v. Ala. State Bar*, 2004). The three elements of a qualified disability are (a) an impairment (b) that substantially limits (c) a major life activity. As it directly relates to some of the knowledge and skills measured in licensing and certification tests, it includes many prerequisite abilities such as literacy and communication skills that generalize across professional domains.

Accommodation, Not Modification

It is important to differentiate between what is intended with an accommodation and something that would modify or change the construct that is being measured. In the instance that providing an accommodation would fundamentally change how the job-related knowledge or skills are demonstrated, a candidate does not need to be permitted to receive the accommodation under the ADA. It then becomes important for a licensing or certification program to define the critical knowledge, skills, and abilities that are part of the intended construct. Phillips (2017) suggested that licensing and certification programs that use simulation or more real-world measurement approaches may not be able to offer accommodations as easily as examination programs that focus more on knowledge or application of information without potentially modifying the intended construct.

Although accommodations often are interpreted within the measurement community as facilitating opportunities for candidates with disabilities to demonstrate their skills, there is an important distinction in stakeholder groups' perceptions. Phillips (2012) noted that advocates for disabled candidates may interpret increased access as applying to opportunities within a profession or occupation; this is notably different from attempting to facilitate greater access to the examination. She further makes the point that if the accommodation is construct relevant, the interpretation of the candidate's job-related abilities may be called into question as a result of the shift in the meaning of the score and resulting decision.

The interesting court ruling in *Palmer College of Chiropractic v. Davenport Civil Rights Commission* (2014) raises the question of how job relatedness is interpreted, and it extended the question about what types of accommodations are reasonable. In this case, the court ruled that the college was required to provide a visually impaired student with a sighted assistant for reading and interpreting radiographs (i.e., X-rays). The question of the job relatedness of being able to read, interpret, and diagnose radiographs for the purpose of developing and implementing a treatment plan is an important one. The court's ruling suggests that it is a skill that in practice can be delegated to another role in support of the chiropractor role. If this is the case, then it is a reasonable interpretation. If not, however, then the court in this instance has appeared to define expectations that may run counter to the legislative practice act that defines the scope of practice for the profession in the jurisdiction.

Evaluating Reasonableness

Part of the ADA's expectations for accommodations rests in the interpretation of what constitutes a *reasonable* accommodation, recognizing that there are some limits to the interpretation. In *Kelly v. W. Va. Bd. of Law Exam'rs* (2010), the plaintiff (Kelly) was diagnosed with a reading disability after being accepted to law school and received a time-and-a-half accommodation for exams. He then transferred to a new law school, where he received a double-time accommodation. He did not receive any accommodations during college or for his admission exams.

For the West Virginia bar examination, Kelly requested double time but was granted a time-and-a-half accommodation. Although it was not part of the bar exam, Kelly also passed the required professional responsibility exam within the licensure process for lawyers without an extended time accommodation. A member of the licensing board testified that it was a job-related skill for a lawyer to be able to work under time constraints and that most applicants used the full testing time, with not everyone finishing. The plaintiff failed the West Virginia exam twice with the time-and-a-half accommodation and then successfully appealed to Kentucky's Board of Bar Examiners to provide a double-time accommodation for its bar exam.

In evaluating the differing opinions of expert witnesses for the plaintiff and the board, the court held that the board's position was more credible given the plaintiff's prior education experience and related test score performance that were provided as evidence of the need for the accommodation. The court further held that providing more time than was necessary would shift the impact of the decision from being a disability-based accommodation to being an advantage over other applicants. As part of its deliberation, the court considered the historical pattern of accommodations for the plaintiff: specifically, that there were no accommodations provided until law school. As a result, the court interpreted that the request was unreasonable and not required under the ADA.

Currier v. NBME (2012) was another case that focused on the definition of a reasonable accommodation. The context for this case was Step 2 of the United States Medical Licensing Examination (USMLE); the two-part accommodation request was for (a) extended time for diagnosed learning disabilities and (b) additional break time for the candidate, a nursing mother, to pump breast milk. In response, the National Board of Medical Examiners (NBME) offered to provide a double-time accommodation for the learning disabilities (dyslexia, attention deficit hyperactivity disorder), administering the exam over 2 days, in a separate room. To respond to the candidate's request for the 60 extra minutes each day to pump breast milk, the NBME offered to provide its standard break time of 45 minutes each day, along with a power outlet and the ability to bring food into the separate room that was previously offered for the double-time accommodation.

The plaintiff sued to get the extra break time to be able to pump. NBME's position was that the break times needed to be uniform for all candidates, and the combination of accommodations—double time and extra break time—made

the administration logistics untenable. The trial court sided with NBME, but on appeal, the court then sided with the plaintiff and ruled that the extra-time accommodation should be provided. With that additional court-ordered accommodation, the plaintiff failed the examination; she then passed the examination after retesting without the accommodation. As the legal proceedings continued, the court decided in favor of NBME, with the plaintiff appealing again. The appeals court ruled that nursing mothers were covered under the *state* equal rights act, but not under the ADA, based on sex discrimination. It is important to note the distinction between state and federal jurisdictions and how they may not always overlap with respect to interpretation.

Test Security Policies

Many licensing and certification testing programs have transitioned to computer-based testing. This occurs in testing centers, through in-person events, or with remote proctoring. Although early proponents of computer-based testing argued that it would increase test security because the test material could be encrypted, concerns about security of test content and validity of scores have continued to be a significant problem. Not surprisingly, this has led to litigation related to test security and administration. Here we highlight two instances of these types of cases.

In *ETS v. Hildebrant* (2007), a candidate who was taking a principal certification test had signed a candidate agreement about confidentiality, nondisclosure, and the testing conditions included in the candidate manual for the program. According to the test proctor, the plaintiff twice refused to stop writing when time was called. Following an initial warning, the proctor completed a test irregularity report to document the nonstandard condition and noncompliance of the candidate to the administration timing conditions. The candidate insisted that she had not violated the administration conditions. After an investigation, the vendor concluded that there was a violation and canceled her test score. During litigation, the court held that the candidate's unsworn, general denial was insufficient evidence to create doubt that the vendor acted in good faith. Further, the plaintiff was unable to convince the court that the proctor had a motive to lie about the violation of the administration protocol. The court held that the vendor complied with the terms of its own policies and procedures in relying on the proctor's observations and documentation of the violation. As such, the court agreed with the vendor's position that the misconduct justified score cancelation.

A second example, *NCBE v. Multistate Legal Studies* (2006), deals specifically with the security of test material. In this case, the defendant was the Preliminary Multistate Bar Review, a test preparation provider that offered review courses for the Multistate Bar Exam across multiple sites. Detection of an attempt to steal test content occurred after a proctor caught an applicant—who was an employee of the provider—attempting to remove test content on scratch paper. Following an investigation that uncovered approximately 100 stolen items, the National Conference of Bar Examiners (NCBE) brought a lawsuit against the provider for copyright infringement. In the court's

discussion of their decision in support of a ruling for NCBE, they found substantial overlap between items that were offered by the test preparation provider and the actual items from NCBE's examination.

Discussion of legal considerations that are relevant to licensing and certification tests provides some insight into areas that should be a central focus of test developers and score users because of the associated risk for legal action. In the next sections, we review three main topics that are critical in high-stakes credentialing contexts because of their direct impact on the interpretation and use of test scores: test development, reliability, and validity.

TEST DEVELOPMENT

In most, if not all, credentialing contexts, candidates seeking the credential will be required to complete at least one examination to demonstrate their competence. A passing score on this examination will serve as an indication that the test taker has the requisite knowledge and skills to competently and safely perform the duties of the profession. As described in the *Standards*, a number of specific considerations and recommendations apply to design and development of tests used for educational, psychological, or credentialing purposes (AERA et al., 2014). This section presents test development considerations that warrant particular attention in the context of licensing and certification; for a comprehensive review of test design and development more broadly, we refer the interested reader to Huff et al., this volume.

Specifying the Purpose

The general purpose of credentialing examinations is clear: to identify individuals who have the minimum knowledge and experience to perform required tasks competently and safely (Chinn & Hertz, 2010). As noted previously, the focus in licensure is on the need to demonstrate that an individual is qualified to practice in the profession; in certification, the focus may be on specific specialization within the profession or on a more advanced level of performance.

The importance of developing and publicizing a clear statement about the purpose of the overall test—or, when relevant, the main components comprising a test—cannot be overstated. Doing so serves several functions. First, it guides subsequent test development efforts by specifying what is in and out of scope for the examination. It also informs stakeholders about the nature of the examination. For example, the NCARB has publicized the following purpose statement about the Architect Registration Examination (ARE):

> The ARE is designed to assess aspects of architectural practice related to health, safety, and welfare. Specifically, the ARE focuses on areas that affect the integrity, soundness, and health impact of a building, as well as an architect's responsibilities within firms, such as managing projects and coordinating the work of other professionals. (NCARB, 2023b)

Although the specifics of the examination content are not provided, the reason for administering the test and general test content are clear to test developers and other stakeholders, including candidates who will be taking the examination. According to Cizek et al. (2011), "The test development process begins and is guided by clear, explicit statements regarding the purpose of the test and the inferences that are intended to be made from the test scores" (p. 6).

One of the challenges inherent in constructing a purpose statement is developing a clear understanding of (a) what *should be* assessed with respect to the critical knowledge and skills needed by practitioners, as well as a realistic idea of (b) what *can be* assessed given the inherent constraints associated with assessment. Careful consideration should be given to the requirements of practitioners at the specific level at which the examination is targeted. This need highlights the importance of a systematic evaluation of the profession that will yield information to guide decisions about examination content (AERA et al., 2014; Chinn & Hertz, 2010; Cizek et al., 2011; Raymond, 2001, 2002) and, by extension, development of a purpose statement.

Identifying Appropriate Content

As we have mentioned, one of the central requirements for credentialing in professional disciplines is successful completion of an examination or series of examinations that is/are intended to provide insight into whether the test taker has the necessary knowledge and skills to practice in the profession safely and competently. Regardless of the discipline or the type of credentialing, the common thread among these examinations is the need to determine the content that will provide the evidence for minimally acceptable performance. According to the *Standards*, best practice for determining the content of credentialing examinations is surveying members of the profession about the knowledge, skills, and abilities that are required of practitioners to identify the actual tasks that practitioners must be able to perform competently and safely. If done properly, this *job* or *practice analysis* will collect information about what activities are required for practice and thus will inform decisions about the content that is relevant for the examination (A. L. Clauser & Raymond, 2017). It is important to note that the practice analysis does not dictate what will be included on the test, but instead provides test developers with information about the requirements of the profession; it is the responsibility of the test developer to carefully review the results to identify the content that can be assessed within the constraints of the examination format.

A license indicates that the practitioner has the knowledge, skills, and abilities required for competent and safe practice. It is not a prediction that the individual will develop those skills over time and, in most cases, it is not restricted to a set of entry-level tasks.[2] To use an example from outside the professions, a driver's license allows an individual to operate a motor vehicle. It may be ill-advised for the newly licensed driver to operate a vehicle in a highly congested area, but the license does not include such a restriction. Similarly, when a state licenses someone to practice medicine, the license does not come with prescribed limitations. There are, however, circumstances where these lines are blurred.

For example, some states require passing the first two steps in the three-step USMLE to enter a residency program and practice under supervision. Similarly, residency programs typically require successful completion of those components of the examination, making those components of the test a de facto license for practice under supervision. In many professions, there is a presumption that practitioners will improve their skills after being licensed; advanced certification documents that improvement, but licensing tests typically are a stand-alone evaluation. It is not uncommon for language about "entry level" or "readiness to enter the profession" to be used when discussing licensure. Insofar as this characterization implies a trajectory, we believe that it may be misleading.

It is worth mentioning that practice analyses done for the purposes of occupational licensing and certification may be different than those conducted for other purposes. For credentialing purposes, the focus is on those areas that are observable and are related to "impact on public health, safety, and welfare" (Chinn & Hertz, 2010); consideration of those areas that provide insight into other professional activities, such as those that allow for personal achievement within a profession, for example, would not be relevant. This focus on the safety of the public is a thread that carries through the entire test development process.

Much has been written about practice analysis design and the relationship between the design and the ability to extrapolate data to inform test specifications (and, ultimately, examination content; Chinn & Hertz, 2010; Raymond, 2001, 2002). The following paragraphs provide a brief overview of practice analysis in credentialing as well as recommendations and cautions associated with its use (see Raymond, 2001, 2002; and Raymond & Neustel, 2006, for a more detailed review of relevant considerations).

The decision to conduct a job analysis is only the first step in the process; a series of additional decisions about the specifics of the process must follow. These decisions generally relate to the scope of the project, the overall methodology that will be used, and the rating scales that will allow for collecting the actual practice data. Subsequent decisions about use of the resulting data to inform test specifications also must be made, but careful attention to the initial design considerations will help to ensure a more seamless process of applying the data to the next phase of test development.

Numerous approaches to practice analysis are available to practitioners, and much has been written about the different methodologies and their associated strengths and weaknesses in particular contexts and for particular uses (Raymond, 2001, 2002; Raymond & Neustel, 2006). That being said, of particular focus in credentialing contexts has been the task inventory approach and, subsequently the professional practice model (Raymond, 2002).

With the task inventory approach, a list is created that outlines the tasks or activities that are performed by practitioners in a profession. The list typically would be created by a group of subject matter experts, and the results then would be included in a survey that asks a representative sample of people in the field to respond to rating scales about different aspects of the included tasks (e.g., the importance of the task for

practice or the frequency with which the task is performed). The primary benefit of this approach is efficiency: A lot of information can be collected in a short time. The breadth of information that can be collected is especially important in credentialing contexts (where readiness for performing a variety of tasks needs to be specified across a range of settings), and the data translate well to the test specification phase (Chinn & Hertz, 2010; Raymond, 2002). A disadvantage, however, is that the focus is on individual tasks rather than higher level skills such as critical thinking and problem-solving approaches. Task inventories that emphasize discrete, observable tasks may overlook the cognitive nature of many professions (LaDuca, 1994).

This approach, regarded as an overarching framework within which practice information can be organized rather than a distinct method (Chinn & Hertz, 2010), focuses on including multiple dimensions that are important to the practice. Doing so allows for creating a matrix in which the practice analysis content is included at the intersection of the two dimensions. In medicine, this might be represented as a setting-by-competency design in which the practice settings (such as inpatient wards, emergency department, ambulatory clinic) are crossed with critical competencies such as medical knowledge or communication skills. The tasks that rely on a given competency within a setting then are included in a given cell. This approach is more comprehensive than a strict task inventory and likely is more appropriate for the complex domain of professional credentialing. The framework also is consistent with the model of practice analysis proposed by Kane (1997).

The professional practice model was developed as a way to address these limitations; it focuses on the problems that practitioners will need to solve within the relevant contexts in which they occur (Chinn & Hertz, 2010; LaDuca, 1994; LaDuca et al., 1995; Poniatowski et al., 2019; Raymond, 2002).

Though a variety of acceptable options for data collection exist, research findings indicate that a more comprehensive approach may be preferable to any single identified or named method. A comprehensive practice analysis therefore might include the following elements. First, instead of looking at discrete tasks or specific knowledge, skills, or abilities, it would focus on the types of problems professionals solve and the tools and ideas they work with to solve them. Second, it would be model driven. The model could be used for specifying a preliminary domain of practice, developing questionnaires, and establishing content weights. The goal of the practice analysis would be to confirm, modify, and refine the model. Third, the information would be obtained from multiple qualified sources, depending on the type of information being sought (e.g., a large sample of practitioners for some types of information, panels of subject matter experts [SMEs] for others). For example, the study of dietetic practice conducted by D'Costa (1986) contained elements of functional job analysis (Fine, 1986), emphasized the social and environmental context of dietetic practice, and described practice in terms of a multifaceted model. It also specified critical scenarios, which provided the basis for test items. Other examples of integrating multiple methods into a comprehensive practice analysis questionnaire include studies of nurse anesthetists (Zaglaniczny, 1993) and psychologists (Rosenfeld et al.,

1983). Future projects might follow suit by going beyond the traditional task inventory as the basis for describing occupations and professions.

Specifically designing the data collection approach to answer the question(s) of interest is extremely important. If task importance is the most critical determinant of examination content, contributors should respond to questions about task importance. Similarly, if knowing the frequency with which a particular activity or competency is required, questions about frequency should be asked explicitly. It also is critical to design the practice analysis so that the results can readily inform the development of test specifications (Raymond, 2015). See Huff et al., this volume, for more information about designing test specifications that are used to guide examination development.

It is important for the entire process of conducting a practice analysis to be viewed as an empirical activity that must be performed carefully and systematically. Done correctly, a practice analysis is likely to be resource intensive. It also is a process that should be repeated on a regular basis (e.g., every 7–10 years) so that any changes in professional practice can be identified and examination content coverage can be revised. Though potentially daunting, the effort is critical to the credibility of the examination program. As was described earlier, the legal implications are potentially significant for a credentialing program that has not done its due diligence in aligning examination content with practice requirements.

Adapting the Assessment to Changes in the Profession's Practice

In professions such as medicine and law, the potential for frequent changes in practice necessitates ongoing attention to evaluating practice data and making necessary adjustments to test content. New science changes medicine and new judicial decisions change law. This means that in addition to periodically updating the practice analysis, item pools must be reviewed and updated frequently. In some cases, specific events may trigger such a review; for example, the change in guidelines for the use of hormone replacement therapy that occurred in the early 2000s led to a review of related items in the USMLE item pools.

In addition to changes in practice specific to test content, technology has created more pervasive changes in how professions work. An important example of one of these technological changes is the availability of electronic resources that provide easy access to a wealth of information. For example, Myers and Bashkov (2020) reported that, on average, physicians consult resources for approximately 30% of the patients they see. A central role of credentialing organizations is to steer away from questions that assess factual recall of test content and instead assess performance related to constructs that are needed for practice (e.g., judgment on case law). This has led to several professions deciding to include online resources as part of their assessments. Aligning the assessment conditions with practice activities has the potential to support the validity of the resulting inferences that are made based on the test scores. At the same time, such changes lead to a question of whether access to electronic external resources during an assessment may unintentionally change the construct that is being assessed from actual professional knowledge, skills, or abilities in a particular area to the ability to retrieve pertinent information. Lipner et al.

(2017) provided an example of research designed to answer this question. They conducted a randomized experiment using case scenarios and multiple-choice questions to assess physicians' clinical judgment skills under timed testing conditions. They found that the inclusion of an electronic resource did not adversely affect assessment performance and did not change the construct targeted by the assessment. Due to testing time limitations, the external resource functioned as a tool much like a calculator used on the SAT exams. If, however, there were no time limit, they posit that the construct might change significantly such that what was being measured was the test taker's ability to find the correct answer through information retrieval skills. This change could make it possible for a nonphysician to pass the assessment, which clearly undermines test validity.

Automated Item Generation

Recent advances in testing and technology have allowed researchers to explore the use of model-based approaches to content development that will improve efficiency in the production of test material and support the design of assessments to enhance both the quality of the resulting questions (Gorin & Embretson, 2012; Leighton, 2012) and the validity of associated score interpretations. Mislevy and Haertel (2006) stated that "evidence-centered assessment design (ECD) provides language, concepts, and knowledge representations for designing and delivering educational assessments, all organized around the evidentiary argument an assessment is meant to embody (p. 6)"; this conceptualization meshes well with Kane's popular validity framework (2006, 2013).

Automated item generation (AIG) is an advanced test development approach that attempts to address the need for fast, efficient production of domain-specific test items by employing a generative process that uses models and computer technology to create new test items. Where traditional item development is a resource- and labor-intensive process requiring subject matter experts (SMEs) to develop individual items one at a time, AIG automates and streamlines the process by using technology to rapidly generate hundreds of test item iterations. Gierl and Lai (2015) described a model-based approach to content development using cognitive task analysis (Clark, 2014). At present, relatively little is known either about the cognitive processes used by test takers in responding to test items or about the characteristics of those items that impact their psychometric performance, although one study indicates that AIG items perform similarly to other test items (Gierl et al., 2016). Although AIG continues to be studied, the technology is not currently in widespread operational use. More research is needed to evaluate whether the anticipated benefits of AIG will be realized for the approach to be adopted more broadly.

Innovative Item Types

Since the 1960s, multiple-choice items have been the most commonly used item format in credentialing contexts (M. McDonald, 2014).[3] This format was introduced in 1915 as a way to (all but) eliminate the scoring errors that were rampant in educational assessments, and it has additional benefits in that it both supports broad sampling of test content in a fixed amount of testing time and reduces the effort required for scoring.

The primary disadvantage, particularly in credentialing contexts, is that test takers select a response from a predetermined list rather than constructing a response based entirely on their own knowledge and skills. At more advanced levels of credentialing, the concern is that this does not necessarily provide clear evidence that test performance will generalize to performance in the real-world practice environment. In some contexts, this disadvantage may be trivial. If, for example, it is important for Certified Public Accounts to know/be able to identify the fraudulent activities that could be perpetrated because of a lack of effective internal controls in the revenue cycle, (a) that knowledge can be readily assessed using multiple-choice questions and (b) doing so is completely relevant to what is necessary for performance in practice. In other contexts, competent practice necessitates not only content knowledge but also demonstration of how that foundational knowledge is to be applied; in some of these situations, multiple-choice items may be less appropriate. It is these situations in which there has been a focus on developing and administering *innovative* or *alternative* item types.

Innovative (or alternative) items are those presented in a format other than the traditional multiple-choice format (Association of Test Publishers & Institute for Credentialing Excellence, 2017; Parshall et al., 2002; Wendt & Harmes, 2009). Many varieties of innovative items exist, and we provide a brief overview here both to provide insight into why these item types increasingly have been used for credentialing purposes and to explain why this trend is important and relevant in credentialing contexts. (See Huff et al. and Bennett et al., both in this volume, for further information about item types.)

The hallmark of alternative item types is that at least one of the following aspects of the items is different from traditional multiple-choice items: item presentation, the test-taker task, the response method, or item scoring (Association of Test Publishers & Institute for Credentialing Excellence, 2017). Though many types of innovative items exist, several categories tend to be used most commonly: items that include audio or video as stimulus material, multiple choice multiple response (two or more keys, often more than four options), hot spot (answer in the form of clicking on part of graphic), drag and drop/place (clicking and dragging text or graphics to match, sort, or rank), written response (an open response to a prompt such as an essay or short-answer question), and audio/video prompt (an audio or video clip is part of the item stem; Cadle & Parshall, 2015; International Test Commission & Association of Test Publishers, 2022; Parshall & Harmes, 2008).

The decision to include innovative item types should be motivated by a need and/or desire to create assessment tasks that capture important construct-relevant information about test takers, rather than simply because it is possible and exciting to develop something new and different. In credentialing contexts, where the assessment is intended to provide some level of insight about a candidate's ability to practice competently and safely, the focus may be on including more realistic assessment tasks that provide a clearer link between test performance and performance in the profession. The reason for developing these items also may be the need and/or desire to present

tasks that require the same types of cognitive processes that are necessary in practice. The typical guiding principle behind including these items in the assessment is greater fidelity; unfortunately, greater fidelity does not guarantee validity (Parshall & Harmes, 2008). When developing these novel assessment tasks, it is important to be sure that the constructs of interest are the focus of the assessment and that the item type is selected because it adds something important to the measurement of the construct. If, for example, it is important for a medical student to be able to interpret heart sounds, one complex assessment option might be a performance-based examination in which students examine standardized patients using an electronic stethoscope that presents simulated heart sounds. An alternative approach would be to incorporate audio into a text-based multiple-choice item. The latter option is a more efficient approach that likely allows for equal, if not better, measurement of the construct of interest.

Increasingly sophisticated technology permeates most aspects of daily life, and the domain of testing is no exception. A desire for increased authenticity has continued to influence the trend toward technologically innovative assessment solutions in which test takers engage with and demonstrate competence through activities that more closely approximate real-world practice. Computer simulations have gained in popularity, though the complexity, fidelity, and interactivity of these approaches can vary significantly (Association of Test Publishers & Institute for Credentialing Excellence, 2017; Margolis et al., 2002). A low-fidelity example might present four pictures of dermatologic conditions and have the test taker click on the one that is appropriately treated with cryotherapy (Association of Test Publishers & Institute for Credentialing Excellence, 2017). A high-fidelity example might be using a patient mannequin and simulated catheterization laboratory to assess interventional cardiologists' procedural skills. This type of simulation could yield information supporting the claim that candidates know how to do the procedure instead of just knowing the right thing to do without actually performing the task under pressure (Lipner et al., 2010). As noted previously, the evolving nature of professional practice has prompted innovations in assessment design. A notable example—aimed at more closely mirroring real-world conditions by allowing access to key resources during testing—is the American Board of Internal Medicine's decision to incorporate commonly used point-of-care tools into its certification exams (Lipner et al., 2017). Similarly, the Uniform Certified Public Accountant examination uses task-based simulations that typically require test takers to use authoritative literature provided in the examination (https://www.aicpa.org/becomeacpa/cpaexam/downloadabledocumentes/cpa-exam-digital-brochure.pdf). Other examples of new innovations include drag and drop/place as well as hot spot items in the ARE exam (https://www.ncarb.org/blog/a-deep-dive-are-5-item-types), drag and drop/place items used by the National Council of Examiners for Engineering and Surveying (https://ncees.org/exams/cbt/), and the computer-based patient management simulations used in Step 3 of the USMLE (https://www/usmle.org/step-3-test-question-formats/computer-based-case-simulations).

Feedback

Test takers have increasingly demanded that testing organizations enhance the educational value of assessment by providing feedback about strengths and weaknesses. This is especially true for continuing certification programs where a new focus on learning and improvement is intertwined with the assessment (American Board of Medical Specialties, 2019). Metacognition has been shown to be important in self-regulated lifelong learning, from having the ability to adjust one's learning to becoming a better professional (e.g., a better clinician who makes fewer medical errors; Medina et al., 2017). Metacognition includes planning, monitoring, and evaluating one's own learning process, and the idea is to make this process more automatic so that errors are not repeated. For example, in the Doctor of Pharmacy program, a key goal of ensuring the reliability and validity of examination scores is to enhance student learning and achievement of the program goals; this is done by providing meaningful feedback for metacognitive learning after an examination (Ray et al., 2018). To enhance metacognition, the authors suggest an exercise after the test where cohort performance data by topic are shared with students and they engage in a three-step process of describing their understanding of concepts when they took the test, reviewing faculty feedback and identifying essential concepts, and reflecting on their knowledge gaps. Despite the clear value of feedback from the perspective of the test taker, testing organizations must contend with the natural tension between the risks and benefits of providing such information. Ideally, once the assessment is complete, individuals would benefit from receiving feedback in the form of the question, the correct answer, and an explanation of why the correct answer is right and the incorrect answers are wrong. From the perspective of the test taker, the benefits of receiving this information are clear. For the testing company, however, the associated risks of exposing examination content may not be acceptable.[4]

Much has been written about score reporting guidelines and feedback (Zenisky & Hambleton, 2015) and the influence of the context of the testing experience on what makes a report useful and understandable (see also Zenisky et al., this volume). Klesch (2010) complemented this literature with detailed work on score reporting for teacher certification, specifically identifying ways to use technology to provide better and more engaging feedback to teachers on their performance. She also noted that raw scores, percentage scores, or even narrative performance descriptions were more desired than scaled scores and that use of statistical terms or abbreviations was not recommended. Likewise, information about confidence intervals (i.e., measurement error) was not useful for those preparing to take the test again. Klesch found that providing information in many different ways (e.g., contextual, statistical, visual) was important for different learning styles.

Short of giving the full question/answer/rationale, some medical testing organizations have provided feedback in the form of subscores (performance scores for subdomains of an exam) to provide more diagnostic performance information (e.g., areas of relative strength and areas in need of improvement). An example of this type of feedback can be seen in the Internal Medicine Certification exam in which

subdomain scores are presented to test takers following the examination (American Board of Internal Medicine, 2023). Presenting subscores introduces an array of measurement concerns because assessments are designed to measure the overall construct of the discipline such that the total score is the most appropriate measure of the test taker's knowledge in the field (e.g., Feinberg & Jurich, 2017; Feinberg & von Davier, 2020; Feinberg & Wainer, 2014a, 2014b; Sinharay, 2010; Zapata-Rivera, 2018). Despite this caution, test takers desire performance information because it helps them to understand their areas of strength as well as areas that are in need of improvement. Particularly in credentialing contexts, test takers who are not successful often request additional information specifically to aid them in understanding where they fell short so that they can address those weaknesses in a focused way. Subscore use has increased over the years, and this makes it even more important for testing organizations to ensure that reported subscores are reliable and valid. Augmentation of subscores with information obtained from other parts of the assessment sometimes is done to achieve this goal, especially when the subscores are correlated (de la Torre & Patz, 2005; de la Torre et al., 2011; Edwards & Vevea, 2006; Sinharay et al., 2011; Wainer et al., 2001). Puhan et al. (2010) showed that reporting subscores may be most beneficial (or at least not harmful) at an institutional level (e.g., programs where 30 or more students were trained) but that they typically do not add much value over the total score. However, augmenting subscores using information from the total score did result in increased subscore reliability. Similarly, augmenting subscores using even just one prior testing occasion can improve subscore reliability, especially when the number of domains is limited and the overall test length is relatively short (Smiley, 2019). Feinberg and von Davier (2020) described a method for identifying unexpectedly high or low subscores that could stand out for test takers and be more actionable. This may help to limit misinterpretation of subscore reporting.

Another approach to enhancing the educational value of an assessment is to provide feedback at the topic and task level such that the intended measurement objectives of the incorrectly answered questions are revealed without exposing the exact item content. This enables test takers to understand their knowledge gaps at a more detailed level than does providing simple subdomain scores. The development of this approach is described in detail elsewhere (Brossman, 2018), and an example of the final presentation of the score report feedback is presented by Lipner et al. (2019). The obvious concern about this approach is low subscore reliability; there are few questions in any single subdomain because the assessment is a sampling of the entire domain. Focusing on areas with low reliability may give test takers a wrong signal of where their weaknesses lie. When there is significant variability in the number of questions represented by a subscore, there also may be instances in which studying the topic area with more questions is a better strategy than studying the topic area in which the test taker had relatively poorer performance. That said, providing more specificity does help test takers better understand what they got wrong on the test

and, as long as they are aware of the limitations, it helps them focus on the topics where test questions were missed.

A slightly different perspective on the provision of feedback relates to contexts that employ a hybrid assessment approach intended to satisfy both formative and summative assessment needs. Formative feedback is provided with the goal of helping learners identify areas of strength and areas in need of improvement so that they can focus their efforts on those areas for the summative component. This practice can lead to challenges in interpreting test results and in evaluating the overall validity of the inferences made from those scores. This topic is further discussed in the "Validity" section, specifically from the perspective of consideration of threats to test validity. In the next section, we address aspects of reliability that either are unique to or warrant special consideration in credentialing contexts.

RELIABILITY

Reliability relates to the estimation of error in measurement that might arise from different sources. The measurement community has developed several formulas and indices to estimate the error of scores, scorers, decisions, and classifications (Haertel, 2006; Lee & Harris, this volume) and, more specifically, has provided practical insight into the applications of these measures in credentialing contexts (B. E. Clauser et al., 2006). The expected reliability for licensing and certification examinations will depend on multiple factors. For example, the types of items, scoring methods, decisions, and decision rules that are used for the examination will influence the selection of methods to estimate reliability. Practitioners are encouraged to use methods that focus on the greatest potential sources of error. To illustrate this point, for a human-scored performance task that has subjective features of a rubric, one important source of error likely is in inconsistent application of scoring criteria to the performance; when the assessment is based on relatively few tasks, the sampling of tasks (the person-by-task interaction in generalizability theory terms) also may make a substantial contribution to error.

Kane (1996) discussed another concept that is particularly important to credentialing tests: error tolerance. The high-stakes nature of any test used for classification and decision-making increases the need for precision around the passing (cut) score. The conditional standard error of measurement (CSEM) around the cut score provides users with an estimate of the uncertainty of the decision: The larger the CSEM, the greater the uncertainty. The location of the cut score in the score distribution also influences the confidence and connects with Kane's recommendation to consider the error tolerance in describing it. As users think about Type I (false-positive/pass) and Type II (false-negative/fail) errors in terms of decisions, the tolerance for these types of errors will be dependent on the risk of making each one. In the instance of a Type I error, policy makers will need to evaluate the risk to stakeholders and the public of someone entering a profession whose true score may not meet the minimum standard for competence (assuming all other eligibility and licensure or certification requirements are

met). This is where the interpretation and use of the credential is important to consider. One can make a reasoned argument that for members of the healthcare community, the risk of an unqualified candidate can have devastating, if not deadly, consequences. The risks of false-positive outcomes in other professions (e.g., architects, lawyers, educators) can be similarly costly.

Conversely, the risk of a Type II error has additional consequences, such as potentially limiting the pool of qualified practitioners available to the public. The challenge is that programs are unable to accurately determine which candidates are the results of Type I or Type II errors; it is only possible to know that there are some candidates within that range of uncertainty that may be misclassified. Because of these risks and the associated limitations, it is important for testing programs to review information about classification and discuss the levels of risk tolerance associated with different decision points. It is in these instances where the combination of validity, reliability, fairness, legal risk, and testing policy intersect.

Methods

To reduce redundancy with the discussion of methods provided by Lee and Harris (this volume), we have focused our discussion on measures of classification consistency and accuracy. The *Standards* (AERA et al., 2014) noted that licensing and certification tests should prioritize these sources of reliability evidence. It is important to note that methods for estimating reliability of scores also are applicable for credentialing testing. However, per Standard 11.14, estimates of decision consistency also should be provided, and "the consistency of decisions on whether to certify is of primary importance" (AERA et al., 2014, p. 182). As applied to licensing and certification tests, Popham and Husek (1969) discussed the issue of calculating reliability for criterion-referenced tests in an educational context when there was an expectation of limited score variability and scores well above the passing score. This led to the subsequent development of methods that could respond to estimates of classification or decision consistency (see Brennan & Kane, 1977; Brennan & Prediger, 1981; Huynh, 1976; Livingston, 1972; Subkoviak, 1976). These initial efforts represented an expansion of the concept of reliability and, by extension, an expansion of some of the evidence that would be valuable for licensing and certification programs.

An additional wave of methods emerged in the late 1980s and 1990s, driven by efforts to clarify the interpretation of these indices (S. A. Livingston, personal communication, January 18, 2002) and by advances in computing power that made it easier to implement more sophisticated techniques. Approaches described by Hanson and Brennan (1990), Breyer and Lewis (1994), and Livingston and Lewis (1995) sought to respond to these questions. Further research by Brennan and Wan (2004) described the application of bootstrap techniques. More recently, work by Cui et al. (2012) using cognitive diagnostic models provided additional value to this still growing line of research. The application of cognitive diagnostic models along with application of measurement decision theory methods (see Rudner & Gao, 2011) has seen greater utilization in licensing and certification, particularly in maintenance of certification and the use of longitudinal

assessments where test/assessment items may be used more as assessment for learning to evaluate and promote continued competence. These longitudinal approaches are currently being used (e.g., American Board of Anesthesiology, American Board of Pediatrics, American Board of Internal Medicine) and explored (e.g., National Commission for Certification of Physician Assistants) as new forms of assessment for maintenance of certification that are not event-based test-delivery experiences. In these longitudinal assessments, programs often will administer a smaller number of questions (e.g., from 1–5 to 20–30) every month or two as a strategy to facilitate ongoing engagement in professional literature and to maintain currency in the profession.

Implications for Practice

Although there is no single approach that is appropriate for all situations, it is important to understand the assumptions and limitations of available methods in order to make informed decisions about which approach to use in a given context. As it relates to the passing score, a CSEM can be interpreted, but with CSEMs often yielding relatively standard values, the use of a classification consistency index that is reported on a metric between 0 and 1 may be easier to communicate to more diverse stakeholder audiences. This means that when communicating with policy bodies or lay stakeholders, stating that a program has 90% or 95% confidence in the accuracy of its pass/fail decisions is often more interpretable than sharing a CSEM value. When these statistics are reported, it also is appropriate to note limitations or cautions with respect to the stability of the estimate. The practicality of using classification consistency estimates illustrates the interrelated nature of dependent characteristics of scores that are produced by licensing and certification exams; these include sample sizes, first-time takers versus retest candidates, location of the passing score, and related risk management considerations.

Something unique to credentialing contexts that is not frequently discussed in methods of reliability relates to smaller volume programs. The data dependency of most methods (bootstrapping or simulation being exceptions) means that estimating reliability or classification accuracy for smaller volume programs remains a meaningful challenge for which neither historical nor modern methods offer satisfactory solutions. As with the initial nudge from mastery testing within educational assessment to explore single administration estimates of decision consistency, there may be an opportunity to again look to K–12 education for consideration of additional methods or concepts that could be applicable solutions for credentialing programs.

Statistical estimates of reliability are often unstable with small sample sizes. Similarly, estimates of the conditional standard error of measurement at the passing score also may be imprecise, though they can still offer some empirical insight. However, in response to the challenge of providing evidence of reliability for the many smaller volume credentialing programs, it seems necessary to move beyond statistical calculations/estimates only. If we conceptually define reliability as the replicability of observations of samples of candidates' abilities that are judged to align with a given construct, then extending

the interpretation of reliability may be appropriate. To that point, the ideas suggested by J. K. Smith (2003) in a reconceptualization of reliability for classroom assessments may be worth exploring in greater detail. Classroom assessments often suffer from some of the same challenges as smaller volume credentialing programs: There is a small sample of test takers and the expected range of abilities for those test takers may be limited. The combination of these factors often yields unstable and potentially suppressed estimates of score reliability. Corrections for attenuation may not be enough to provide a reasonable estimate.

In rethinking the characterization of reliability, J. K. Smith (2003) suggested using a criterion of sufficiency as a proxy for traditional measures of reliability. In other words, psychometricians often recommend having a minimum number of measurement opportunities (e.g., items, questions, score points) for each important concept or domain of interest. Using that approach, and absent robust data sets to conduct traditional analyses, we may be able to make an argument that the object of measurement is being sufficiently represented by multiple observations in alignment with other sources of recommended best practices. For licensing and certification programs, the concept of sufficiency could be applied as an interim step until data can be collected to apply more traditional methods. This approach also may provide a solution for an important source of evidence for credentialing programs where the current practice may be to simply ignore consideration "due to low sample sizes." Professional standards suggest we can do better, yet currently there is a dearth of evidence-centered approaches to guide such improvement. In 2017, the Institute for Credentialing Excellence published a revised version of the National Commission for Certifying Agencies' Standards for the Accreditation of Certification Programs. In this document, Standard 20 addresses reliability concerns and provides specific commentary associated with estimating reliability for small-volume programs:

> When candidate volumes are so small or there are other factors which lead to reliability estimates that are not meaningful, programs should describe the procedures used to demonstrate that the decisions made on the basis of scores are reasonable and fair. (Institute for Credentialing Excellence, 2014)

This comment clearly acknowledges limitations associated with providing meaningful reliability estimates in certain contexts, but limiting solutions for small-volume credentialing programs to provision of a thoughtful narrative about their procedures provides some insight into the fact that there still is a long way to go. Newer approaches, such as bootstrapping techniques, may hold some promise in advancing the assessment of reliability with small samples (Amalnerkar et al., 2020), but more work is needed to evaluate the application of these procedures to credentialing contexts.

Another unique aspect of licensing and certification tests is the opportunity for failing candidates to retest with different policy conditions governing these opportunities. One of the outcomes of this approach is that it results in inflated false-positive/passer (Type I) error rates. B. E. Clauser et al. (2006) illustrated that administration conditions with only repeat candidates lead to inflated Type I error rates. Attempts to

control this characteristic were discussed by B. E. Clauser and Nungester (2001) and included such decision rules as using average scores, most recent scores, or best scores within a defined period. Another strategy would be to potentially increase the passing score for subsequent attempts. However, this may lead to equal protection concerns if the passing expectation is differentially applied to candidates, raising questions about the extent to which any given administration is an independent observation of the candidate's ability that is not dependent on prior information or test-taking performance.

Similarly, there is an important relationship between estimates of classification consistency and the location of the passing score. This relationship extends in multiple directions. First, depending on the choice of classification consistency method, the location of the passing score relative to the distribution of candidate scores can potentially inflate or suppress the estimate. For example, this risk is highlighted if a method is chosen that is more influenced by the location of the passing score relative to the mean of the distribution. Second, as discussed in the section on legal challenges to licensing and certification tests, the location of the passing score also may be influenced by the impact on different demographic groups if there is a goal to reduce the potential effects of adverse or disparate impact.

In this section, we discussed the unique aspects of reliability as they relate to licensing and certification tests. As noted, concepts and methods are generally applicable across a range of testing applications. The interpretations and uses of scores from licensing and certification tests suggest additional and integrated considerations that are noted here as they relate to the larger validity framework articulated in the next section.

VALIDITY

Validity is a fundamental concept in educational measurement that is relevant to all aspects of the testing process, spanning test development and delivery, the analysis of testing data, and the interpretation and use of resulting scores. Lane and Marion (this volume) provide a comprehensive review of validity and validation; as such, the focus of this section will be on the aspects of validity that are most applicable to licensing and certification assessment programs. We use Kane's validity and interpretative argument as our guiding framework for this discussion.

Kane's Validity Framework/Interpretative Argument

In credentialing contexts, Kane's framework espouses employing a multifaceted approach to collecting validity evidence to maintain a program of high integrity that is fair to test takers and defensible to stakeholders. It calls for "a clear statement of the proposed interpretations and uses and a critical evaluation of the interpretations and uses" (Kane, 2013, p. 1). Four types of inferences, or claims, are important within Kane's approach: scoring, generalization, extrapolation, and interpretation/use. For assessments with scorable item types, the assigned score becomes a reflection of performance in a test setting where the assessment is taken under standardized conditions and test takers have no prior

knowledge of the specific content that is included on the test. The *scoring* component includes considerations of adherence to the defined administration procedures as well as appropriate use of raw scores (if used), item response theory, field testing, scaling, and equating. *Generalization* involves making inferences from observed scores based on the specific sample of questions on the test or the sample of observations scored by raters to the universe of questions/observations from which they were sampled. Within classical test theory, this is represented by the relationship between the observed score and the true score. Generalization is impacted by errors of measurement, which typically are random; systematic errors also may exist, and when they do, they may be both difficult to identify and damaging to score interpretations. *Extrapolation* relates to the ability to make inferences about real-world performance based on the observed scores. The fourth component of the validity argument, *interpretation/use*, refers to using the test scores to draw a conclusion about the test taker or to inform a decision such as the granting of a license.

The more ambitious the claims one seeks to make based on assessment outcomes, the more validity evidence is required. Yet the more ambitious claims typically are harder to validate because *all* inferences in the interpretive argument must hold for the interpretation to be considered valid. In licensing and certification, the concept of validity has expanded beyond the question of whether the test measures what it was intended to measure. It also emphasizes the idea that the testing organization must take broad responsibility for the consequences of the testing outcomes. A claim might be that those who pass the assessment and are deemed competent should perform competently in real-world settings. This is an ambitious claim and one that is difficult to validate, because assessments typically focus on constructs that rarely can be isolated in real life. For instance, medical assessments often test knowledge/clinical judgment in a discipline, but patient care clearly reflects more than just discipline-specific knowledge. Patient-care decisions are influenced by a myriad of factors including patient compliance, professionalism, communication skills, and systems factors such as teamwork. Although good performance on an assessment of knowledge may relate to higher quality patient care, it is difficult to control all other factors that influence patient care in a real-world setting when trying to determine the strength of that relationship. In addition, knowing the right thing to do on a multiple-choice test taken in an artificial setting is not a guarantee that the test taker would do the same thing in practice. Although some credentialing programs still use oral examinations to test practice-relevant skills, many are transitioning to sophisticated multiple-choice items that evaluate clinical judgment and the ability to synthesize information. Innovative item types also have been introduced in an attempt to address these challenges, and though initial results are encouraging, additional empirical evidence is needed to evaluate whether they are better than sophisticated multiple-choice items for assessing the desired competencies (Swanson & Hawkins, 2017). Similarly, in medical contexts, assessment approaches based on case records or logs (e.g., patient charts from electronic health records) eventually may supplement or replace multiple-choice questions as metrics based on actual practice become more universally accessible.

Importance of Extrapolation and Interpretation

All claims from Kane's framework are important in establishing the chain of evidence. For professionally developed credentialing assessments, evidence to support the scoring and generalization aspects of Kane's framework typically is much easier to produce. In most testing organizations, great effort is placed on test construction procedures: Practice analysis helps to ensure relevance to practice, standardization of testing procedures helps to ensure fairness, and scoring processes help to ensure accuracy. In addition, generalizability of test scores is demonstrated both with longer assessments covering the breadth of the field and by controlling for content and statistical specifications included in the blueprint. These are well-understood processes that have undergone significant specification and refinement over time.

Providing evidence to support the appropriateness of both *extrapolation* from test scores to real-life situations and score *interpretations/uses* is much more challenging. The high-stakes nature of licensure and certification makes the extrapolation and interpretation components critical with respect to validity. If claims are being made about the assessment that go beyond the actual test setting, an evaluation of whether these claims hold true should be done by conducting a program of research designed to generate a preponderance of evidence that either supports or disproves the claims. The higher the stakes and the claims made about the assessment results, the stronger the research program will need to be to support validity arguments. In medical licensing and certification, there is a growing body of peer-reviewed literature that provides evidence related to extrapolation to real-life situations; unfortunately, this work remains limited because definitive outcome measures are limited.

The gold standard for validity evidence in credentialing contexts is the ability to demonstrate that those who pass the assessment perform competently in practice and those who do not pass do not perform competently. In licensure contexts, however, this evidence is impossible to obtain because those who fail the examination(s) never receive a license. This type of evidence is feasible for certification exams, because certification typically is not mandated for practice. For example, a considerable amount of evidence exists to support the claim that being certified or performing better on a certification exam is linked to fewer disciplinary actions by state medical licensing boards. In internal medicine, a board-certified physician is five times less likely to have a disciplinary action than a noncertified physician (Lipner et al., 2016), and the higher the physician's score on the certification exam, the less likely they are to have a disciplinary action (Papadakis et al., 2008). This relationship has been shown in many studies comprising numerous specialties in medicine (Jones et al., 2018, 2020; Kinney et al., 2019, 2020; Kocher et al., 2008; Kopp et al., 2020; Lipner et al., 2016; F. McDonald et al., 2018; Nelson et al., 2019; Papadakis et al., 2008; Peabody et al., 2019; Zhou et al., 2017, 2018, 2019).

Other studies have sought to provide evidence to support or refute extrapolation claims by examining the relationship between measures of performance on the assessment and measures of performance in practice. One study showed that, for patients with acute myocardial infarction, treatment by a board-certified internist

or cardiologist was associated with a 19% reduction in mortality compared with treatment by a noncertified internist or cardiologist (Norcini et al., 2002). Another study showed that receiving care from an internist who scored in the top quartile of their exam was associated with a 17% increase in odds of *guideline-compliant* diabetes care compared to those who scored in the bottom quartile (Holmboe et al., 2008). A third study showed that physicians who maintain certification at 20 years after initial certification have better care management on the Healthcare Effectiveness Data and Information Set performance measures (Gray, Vandergrift, Landon, et al., 2018)—including diabetes care, mammography screening, and coronary artery disease—compared to similar physicians who did not maintain certification. In Quebec's universal healthcare system, scores achieved on licensing and certification examinations show a persistent relationship, over 4 to 7 years, with specific measures of preventive care and acute and chronic disease management in primary care practice (Tamblyn et al., 2002). Additionally, research has found a strong link between a physician's diagnostic knowledge and likelihood of patient death, emergency department visits, or hospitalizations (Gray, Vandergrift, McCoy, et al., 2021), that physicians who score higher on certification exams are less likely to prescribe inappropriate medication to geriatric patients (Vandergrift et al., 2021), and that physicians who do not keep medical knowledge current are more likely to overprescribe opioids for back pain (Gray, Vandergrift, Weng, et al., 2021).

These examples represent just a few of the studies linking performance in practice with performance on credentialing exams. This type of research is extremely challenging, and controversy exists over whether correlational studies like these are sufficient or whether studies must demonstrate a more direct link between program/exam performance and outcomes as a necessary condition for supporting the extrapolation inference. Showing a causal relationship in medical education research is extremely difficult because randomized controlled trials are rare. Other studies using quasi-experimental designs have the potential to get closer to the randomized controlled trials design; two studies that did so (possibly due to a policy change that occurred within a particular time frame) revealed that physicians required to maintain certification save the healthcare system $167 per Medicare patient, per physician, per year without sacrificing quality (compared to those not required to maintain certification). This result translates into roughly $5 billion per year in Medicare cost savings (Gray et al., 2014). In addition, women who had not received prior mammography screening were 8.5% more likely to get screened when seen by a general internist required to maintain certification (Gray, Vandergrift, & Lipner, 2018). These studies are closer to a randomized controlled trial design than correlational studies because they compare physicians who are very close in age and demographics and who have very similar patient populations before and after a policy change. With a preponderance of correlational studies showing positive relationships between status in a program or performance on an assessment and patient care outcomes, a convincing argument can be made that assessment performance matters to patient care. At a minimum, the

evidence is beneficial to patients who, when selecting a provider, should know which physicians "know enough" to practice in their specialty.

The other concern about procuring evidence to support or refute extrapolation claims is that, even in simulated scenarios, examinations typically test knowledge and judgment but not other important behaviors (e.g., professionalism, communication skills, or ability to function effectively within their current system) that are needed by practitioners. These other behaviors undoubtedly impact practice performance differently, such that poor performance in practice could be associated with high assessment scores. However, in the studies cited previously, this seems to be the exception and not the norm. With appropriate designs, attempting to control for many of these other behaviors is possible although far from perfect. For instance, a study can attempt to control for types of systems by identifying practice type (e.g., academic or community setting) or for communication skills using ratings provided by program directors during residency and fellowship training. Solid research designs that use matched samples, nesting of patients within physicians (generalized estimating equations), and patient risk adjustments are critical to studying the relationships between performance on examinations and practice setting outcomes.

As a final comment on correlational studies, it is worth noting that Kane (2013) does not give this type of study much consideration in his discussion of validity. He argues that demonstrating correlations with another measure simply creates the need for a new validity argument for the criterion measure.

Threats to Validity

This section will cover a variety of topics that have a potential impact on the validity of testing processes and score interpretations. These are especially important in the context of high-stakes credentialing decisions.

Test Security

Test security plays a critical role in the validity of interpretations based on test scores. For large-scale, high-stakes examinations, the ability of testing agencies to offer frequent—if not continuous—testing leads to increased item exposure. This presents a significant security challenge because item exposure can provide test takers with advanced knowledge of the questions, leading to an unfair advantage on the test and undermining the validity of score interpretations. Item exposure is an ongoing issue for high-stakes testing (Way et al., 2014), and contemporary testing innovations—including frequent administrations, larger items banks, and communication technologies—have added to the security risk (Foster, 2016).

A number of technology-based advancements that use modeling approaches to content development such as evidence-centered assessment design (Mislevy & Haertel, 2006), assessment engineering (Luecht, 2012), and AIG (e.g., Gierl & Lai, 2015) are intended to improve efficiency in item production. Several studies have demonstrated AIG's ability to generate hundreds or thousands of test items in very short periods of time, often in just seconds (Gierl et al., 2012; Gierl & Alves, 2011, as cited in Morrison

& Embretson, 2018). As long as the generated items are psychometrically equivalent and meet other quality assurance requirements, these items can be used to populate the item bank and expand the pool of active items from which tests can be assembled. This, in turn, prevents test takers from becoming familiar with item content in advance of the examination.

Automated test assembly provides a platform to build item exposure controls into the test development process through heuristics and algorithms. This can help reduce the overexposure of individual items and promote broader utilization of the full test bank. For example, in computer adaptive testing, item exposure is a frequent problem because of the tendency to select highly discriminating items more frequently than others (Hau & Chang, 2001). Exposure controls—such as limiting the number of times any one item from a bank can be used—can mitigate this tendency. Notably, while embedded exposure controls theoretically have been available for quite some time, the technological improvements that have allowed for them to be operationalized have only been made available more recently (van der Linden & Li, 2016).

It should be noted that our comments regarding AIG reflect on the potential promise of these methods. Actual results have lagged behind that promise. There are few (if any) published results showing that these items can be used in high-stakes testing without careful selection and postproduction editing. Additionally, the available evidence suggests that it will not be practical to use these items without pretesting because evaluations of the current technology for predicting item parameters based on item characteristics—or from the parameters that have been estimated for other items from the same model—are not encouraging. Finally, there are few, if any, studies that provide evidence to support the contention that items from the same model are sufficiently different, so that advanced access to one (or more) item(s) from the model will not improve the probability of a correct answer to the next item from that same model. Nevertheless, with continued technological developments, including significant advances in machine learning and artificial intelligence, realization of the promise of some of these methods may not be far in the future.

PREKNOWLEDGE AND UNETHICAL BEHAVIOR Preknowledge of examination content is a threat to validity that can unfairly change the construct that the test is intended to measure. Many review or test-preparation courses exist to prepare test takers for credentialing examinations, and these courses may *teach to the tes*t (or even teach the content of the actual test) instead of focusing on the construct the test is intended to measure. If these courses truly are construct focused, there is no inherent problem; if, however, the focus is on teaching tactics to pass the test, this raises concerns about the inferences that can be made based on test performance. Medical schools and training programs capitalize on using test blueprints to guide curricula. This practice could be viewed as positive in that these curricula improve professional education by covering a complete spectrum of the field. Alternatively, if the curriculum neglects important aspects of practice that are not on the test because the focus is solely on passing the test, the approach is problematic.

As the consequences of passing an examination become more important for a test taker's career, the likelihood of test takers attempting to engage in unethical behaviors increases. Security challenges associated with high-stakes assessments have changed along with advances in technology for computer delivery and remote testing (see Shermis et al. and Bennett et al., this volume). These testing improprieties threaten the validity of the inferences that can be made based on the resulting scores (Cizek & Wollack, 2017). This is critically important in terms of being able to attest that the professional whose name is on a credential did in fact *ethically* master the test. Without this validation, the credential would hold little meaning for stakeholders—especially, and, and perhaps most importantly, the public. Testing organizations may employ numerous strategies to prevent potential security threats. Test takers typically are required to sign a code of conduct that clearly specifies the policies of the examination program and the consequences for not adhering to the code of conduct. Verifying test-taker identity is a critical step to help prevent proxy test takers from subverting the system. Prevention methods in test centers have improved in that test takers often do not sit near one other and questions can be easily scrambled or different forms of the test can be used. Finally, surveillance by test proctors, video recording of test takers while testing, documentation of unscheduled break time, and records of response time data collected routinely for both live and pretest items are all part of the detection and data forensic processes.

CONTINUOUS TESTING A further development in the continuous certification arena is a testing paradigm in which test takers answer questions on a year-round basis in a combined formative/summative assessment process (Continuing Board Certification, 2019). As previously mentioned, testing organizations have been under significant pressure to loosen security measures and increase convenience and flexibility; as a result, at-home, unproctored testing at the test taker's convenience is becoming the norm. In addition, questions and answers along with rationales and references are provided immediately after the question is answered to serve formative assessment purposes. To keep the assessment secure, new questions are written and field tested every year because there is no longer the traditional concept of a large, stored item bank with questions that can be reused numerous times. Field-tested questions that survive then can be used as live scored questions. Although (generous) time limits typically are imposed per question, test delivery procedures allow access to external resources during the test. This new testing paradigm intensifies security issues, because, in addition to the demand for a large number of test questions and the test being taken at home at any time of the day, the message that the program is both for learning *and* for assessment may be confusing. Learning typically allows for working in groups or consulting with peers, but in this situation it does not; security breaches are obviously harder to prevent and detect within such a paradigm. Watermarking and encoding test material have been introduced to trace information back to any individual who tries to subvert the system. Codes of conduct need to be enforced and strengthened, and more sophisticated web-patrolling programs and data forensics techniques must be implemented.

Guidelines have been written about security measures in online testing (Foster, 2013). However, there are unique challenges associated with at-home, online, unproctored tests, and more research is needed to help make informed decisions about the most effective way(s) to incorporate these requests without increased security breaches that may go undetected and ultimately affect the validity of score interpretations. The increasing demand for more flexible and convenient test delivery models has made the move away from brick-and-mortar centers to at-home delivery a reality that will necessitate significant enhancements in prevention and detection techniques, possibly taking advantage of the explosion of artificial intelligence capabilities.

Construct-Irrelevant Variance

Unintended factors that influence test performance pose additional threats to the validity of inferences drawn from examination results. One such factor is content bias, which can advantage certain subgroups over others. Predictive bias is concerning if the exam is meant to measure a construct for the population but is found to give different predictions for subgroups of that population who should be equivalent on that construct (e.g., females and males, Asians and Caucasians). When bias exists, we are measuring construct-irrelevant variance and the test may be seen as unfair.

Use of a statistical method known as differential item functioning is a common initial analytic step for identifying biased items. After running the differential item functioning analysis, items displaying differential item functioning are presented to experts who carefully review them and determine whether the content is truly biased (e.g., Holland & Wainer, 1993). Eliminating questions containing bias is an extremely important step in maintaining the fairness of an examination. In licensing and certification, differences in performance between groups often can be explained by other relevant factors, rather than being indicative of true bias (O'Neill et al., 2022). For instance, medical oncologists who are female more often focus on breast cancer, whereas more male oncologists focus on prostate cancer. These practice tendencies may help to explain a difference in performance that is *not* considered bias.

With computer-based testing, the automatic collection of metadata such as response time and answer changes has contributed to exponential growth in the study of aberrant behaviors (e.g., Lu & Sireci, 2007; Margolis & Feinberg, 2020). Examination time limits are important from a practical, administrative viewpoint and might be important to the construct if the testing organization believes that answering a question within a realistic time frame is important. This issue becomes especially relevant when testing is in the open-book format because typically the construct is not about whether a test taker can look up the correct answer, but rather whether they can synthesize information in a timely manner to arrive at the correct answer. If a layperson can pass a licensing examination by looking up answers over an unlimited time period, the interpretation of any performance on the examination would be meaningless. Time limits are necessary in most contexts, but those that impose unrealistic constraints introduce construct-irrelevant variance into measurement of the construct.

Other Threats to Validity

Less-obvious threats to validity in the credentialing arena can arise from disciplinary pressure to create more individualized certifications—even in well-established areas—resulting in highly specialized credentials for very small populations. (Continuing Board Certification, 2019). Certificate programs have been established in disciplines that may have fewer than 50 test takers a year. The small number of test takers in these populations presents challenges in following best practices in measurement, including field-testing questions, estimating item parameters (calibration), and scoring and equating across forms (Finch & French, 2019). Sample size affects the accuracy and efficiency of model parameter estimates that are used in making decisions about individuals. When high-stakes decisions—such as admission to training programs or employment—are based on exam performance, small sample sizes can undermine the validity of the resulting inferences. In fact, *Applied Measurement in Education* devoted an entire journal issue to studies that seek solutions to the limitations associated with small sample sizes (2020, Volume 33, Issue 1).

A final threat to validity relates to situations in which credentialing programs permit test takers to narrow the focus or scope of their assessment. This approach questions the general construct that is being tested and how that credential is being represented to the public. For instance, if test takers whose primary clinical focus is breast cancer take an assessment that focuses on breast cancer but then are provided with a certificate in general medical oncology, there are consequences for patients who might not be aware of this narrower examination focus. This begs the question of who the certificate is for: the physician, the employer, or the patient population (Vandergrift et al., 2020). Confusion over what exactly the certificate means could raise threats to the validity of the inferences based on test performance. However, asking questions that are irrelevant to those physicians that narrow the scope of their practice also can be seen as a threat to validity in that it is not measuring exactly what they are doing in practice but more what the typical physician practices in the discipline. More narrow scopes and credential designations have the potential to address this controversial issue.

Validity considerations are central to any testing program. In high-stakes contexts, where credentialing examination scores influence critical decisions about individuals, collecting robust evidence to support score interpretation and use becomes even more essential. A validity argument is only as strong as the weakest link in the chain of evidence; attention to the strength of the argument is paramount.

STANDARD SETTING

In educational testing, standard setting refers to the process of determining "how much is enough" with respect to test performance (Kane, 2017). In credentialing contexts, the question more specifically becomes, What level of competence is enough to consider a candidate minimally qualified to enter practice or to use a specific practice

designation? In these contexts, standard setting begins with the assumption that any given discipline can be defined by an underlying body of knowledge, skills, and abilities that are required for practice (Way & Gialluca, 2017). Proficiency in a domain can be measured along a continuum, and there is a theoretical point at which the difference between competence and lack of competence can be distinguished. The intent of the standard-setting process is to first describe that body of knowledge, skills, and abilities and then to estimate the point along the continuum that best distinguishes between those candidates who have enough of what is needed for safe and competent practice and those who do not (Way & Gialluca, 2017).

The above logic indicates that standard setting involves specifying two types of standards: content standards and performance standards. *Content standards* are "collections of statements that describe specific learning outcomes or objectives" (Cizek, 2012); they relate to the underlying domain of knowledge, skills, and abilities in a given discipline and the need to review the domain to identify what must be mastered for a particular purpose. In credentialing, the domain review typically focuses on what is necessary to be considered competent versus not competent to receive the credential. *Performance standards* are defined by the "distinctions between two adjacent categories of performance (e.g., between acceptable performance and unacceptable performance" (Kane, 1998, p. 130). Setting performance standards therefore refers to specifying the level of performance that indicates that the content standards have been met. Once content and performance standards have been specified, there is yet another step: determining the *passing score* (the point on the score scale that corresponds to the identified performance standard; Kane, 1998). As Kane (2017) explained, a performance standard is a *minimally adequate level of performance* for the given purpose, and a passing score is *a point on a score scale* that divides test takers by the level of performance. Passing scores therefore provide an objective way to answer the central policy question about the adequacy of a given test performance.

Though standard-setting considerations are largely similar across educational and credentialing contexts, a number of areas specific to credentialing are addressed in the *Standards* (AERA et al., 2014). In the following section, we touch on some areas that we feel warrant explicit consideration in the context of credentialing examinations. For a comprehensive review of standard setting in educational and credentialing contexts, including a detailed review of standard-setting methods and activities, the interested reader is referred to Cizek (2012), Cizek and Bunch (2007), Hambleton and Pitoniak (2006), and Ferrara et al. (this volume).

Standard-Setting Methods

Standard setting for credentialing examinations typically involves using one (or more) of many judgment-based processes to collect information from content experts about the level of performance that is necessary to meet the content standards; this information is used to determine the examination score that is associated with the minimum acceptable

level of test performance (Institute for Credentialing Excellence, 2018). Perhaps the most important consideration with respect to the selection of a standard-setting methodology is the extent to which that methodology aligns with the purpose of the testing program and its associated models of achievement (Kane, 1998).

There are two overarching categories of standard-setting approaches: test centered and test taker centered. Test-centered approaches are those in which panelists review test items or tasks and, for each one, make judgments about the expected level of performance for minimally qualified test takers (those just meeting the performance standard); item-level judgments then are aggregated to produce a passing score. In test-taker-centered approaches, panelists provide judgments about test-taker performance based on an external criterion or overall test performance; the passing score is set by identifying a point on the score scale that is most consistent with the judgments (Kane, 1998; Kane et al., 1999). Test-centered approaches are used most commonly in licensing and certification; this likely is because of the preponderance of credentialing programs that use objective (rather than performance-based) tests in the credentialing process (Buckendahl & Davis-Becker, 2012; Institute for Credentialing Excellence, 2018; Kane, 1998; I. L. Smith & Springer, 2009). (To the extent that licensing and certification examinations move toward more performance-based testing and away from multiple-choice testing, this balance could shift.) Despite the existence of numerous test-centered methods that promote different approaches to data collection, relatively few are used consistently in licensing and certification, with the Angoff method and its derivative (modified Angoff) procedures being used most frequently (Buckendahl & Davis-Becker, 2012; Hurtz & Auerbach, 2003; Institute for Credentialing Excellence, 2018; Way & Gialluca, 2017).

As typically implemented, modified Angoff procedures require panelists to review test items and provide estimates of the proportion of minimally qualified candidates that would answer each item correctly; the average of the sums of the probabilities from all experts across the set of items becomes the panel's final passing score. This procedure is attractive in its simplicity and in its flexible applications, such as being able to adapt it to a variety of item types (Buckendahl & Davis-Becker, 2012; Plake & Cizek, 2012). Despite these general strengths, these procedures have several associated limitations, perhaps most notable of which is that the task (of estimating the probability of a minimally qualified candidate answering questions correctly) tends to be a difficult one. For example, panelists—particularly those for whom more time has passed since initial licensure or certification—may struggle with appropriate expectations for a minimally qualified candidate, which results in passing scores that the testing program would consider unreasonable (Professional Testing Corporation, 2019). That is, unless panelists have been licensed or certified relatively recently, they may not have a clear sense of the level of knowledge and experience of a test taker seeking the credential, and this may lead them to overestimate the performance of a minimally qualified candidate. Another concern is that panelists may not have a clear understanding of the

extent to which test takers prepare for credentialing examinations, and therefore they may underestimate the performance of the minimally qualified candidate. These concerns lead to questions about the meaningfulness of the resulting data (Busch & Jaeger, 1990; B. E. Clauser et al., 2002; B.E. Clauser, Harik et al., 2009; B. E. Clauser, Mee, et al., 2009; B. E. Clauser et al., 2013; Mee et al., 2013).

To address these concerns, supplementary methods can be used alongside the Angoff process to offer panelists alternative perspectives on this critical task. For example, using a compromise approach such as the Hofstee method (Cizek, 2006; Ferrara et al., this volume; Thompson, 2018) and presenting those results to panelists along with their Angoff judgments allows them to see the outcomes and impact of their item-level Angoff judgments (i.e., passing scores, percentage of items answered correctly, associated fail rates), compare those results with their normative expectations for fail rates and content mastery (i.e., minimum and maximum failure rates and minimum and maximum percentages of items answered correctly), and make any necessary/desired adjustments to their judgments. (If these data are not presented to panelists for review and modification, they can be shared with the decision-makers as they consider all relevant information.)

Though criterion-referenced standard-setting approaches are most appropriate for credentialing examinations (AERA et al., 2014), the reality is that incorporating some consideration of the performance of the population of test takers can provide panelists with valuable context and allow them to make more meaningful judgments (Bowers & Shindoll, 1989). The preceding example illustrates tensions that may exist between making criterion-referenced versus norm-referenced judgments and is intended to convey one aspect of the complexity associated with standard setting in credentialing contexts. There are certain standard-setting methods that may be more or less appropriate for credentialing examinations, but that does not mean that there is a single approach or procedure that is best for all credentialing contexts. The specific features of each context must be evaluated, the benefits and challenges of relevant approaches should be reviewed, and a method or combination of methods should be selected that is most likely to yield meaningful and, perhaps most important, defensible results.

Standard-Setting Activities

Content-based standard setting involves numerous steps that include, but certainly are not limited to, recruiting content experts to serve as standard-setting panelists (or judges) and provide the required judgments, facilitating the standard-setting meeting(s), and collecting or assembling data from additional sources to aid policy makers in making a passing score decision. These general activities tend to be similar across credentialing (and many other) contexts; the differences are in the details and in the way(s) that specific organizations approach and prioritize those details. In the sections that follow, we briefly describe aspects of the above activities that we believe warrant explicit consideration for credentialing examinations.

Panel Selection

Because the judgments of panelists directly affect passing scores and the overall credibility of the standard-setting process, panel composition is of critical importance to the process (Plake et al., 1991). Raymond and Reid (2001) asserted that panelists should meet two critical requirements:

1. Individually, participants should have extensive knowledge of and experience with both the subject matter and the test-taker population; and,
2. As a group, participants should be representative of all stakeholder groups.

An additional requirement, which may be assumed but should in fact be explicit, is that participants must have a commitment to protecting the security of the testing materials (Nancy Tippins, personal communication, September 10, 2023).

When identifying relevant stakeholder groups, Buckendahl and Davis-Becker (2012) argued that stakeholders differ between licensure and certification programs. In licensure testing, the primary stakeholder is the public, which is interested in protection from harm caused by unqualified practitioners. Industry professionals, who have an interest in protecting the integrity and reputation of the profession, and educators, who may use testing information to evaluate and inform curricula, instruction, and accreditation, also are relevant licensing stakeholders. Finally, though candidates themselves are stakeholders, their *opinions* are not of primary importance because they likely have a strong self-interest with respect to entering the profession. That said, the candidates' interest in ensuring that entrance into the profession is not unduly restrictive is completely legitimate and should be reflected in the final decision.

Conversely, certification typically is a voluntary pursuit where the candidate seeks to demonstrate greater knowledge, skills, or abilities than is required of entry-level practitioners (Buckendahl & Davis-Becker, 2012). Therefore, relevant stakeholders for certification exams differ from stakeholders in licensure exams. The primary stakeholders in certification may be the candidates themselves, who seek to enhance their credentials for a variety of reasons. Employers—many of whom consider certification status when determining eligibility for a position or for making hiring decisions—also may be considered a primary stakeholder in the certification process (Buckendahl & Davis-Becker, 2012). (In some professions, the public will be the employer and so will have a clear interest in the integrity of the certification.) The professional community, which has an interest in maintaining the integrity of the profession, also is likely to be an interested stakeholder. Patients sometimes are considered stakeholders, as certification can be a process to let the public know about candidates' competency in specialized areas. Finally, credential-sponsoring organizations are stakeholders in that they have an interest in promoting awareness and use of their products and/or services.

Any standard-setting process that influences public policy will inevitably impact stakeholders (Cizek, 1993; Cronbach, 1982). As such, researchers agree that participants should adequately represent stakeholders in the standard-setting process (Hambleton & Pitoniak, 2006; Loomis, 2012; Raymond & Reid, 2001). Buckendahl and

Davis-Becker (2012) explained that in the context of licensure, recently credentialed practitioners offer a thorough understanding of the content and expectations of professionals with the target credential. Experienced practitioners may have more specialized practices and so may have more advanced skills and knowledge in some areas; they also may lack proficiency in some areas in which they no longer practice. Educators have a strong knowledge of curricula and are most familiar with the target population, but they may lack a thorough understanding of what is required in practice.

Increasing stakeholder representation also can present a number of challenges. Buckendahl and Davis-Becker (2012) noted that broad panel composition may be complicated by stakeholders with competing interests. Policy decisions require the integration of content-based knowledge and stakeholder values. However, stakeholders may have differences in opinions and motivations that are not easily reconciled (Kane, 1994b). For example, experienced practitioners may desire higher passing standards if they result in a less-saturated professional market. In contrast, trainers and educators who use student performance as evidence of programmatic success may desire lower passing standards. Recently licensed professionals also may have diverse interests: Some may push for higher standards to preserve the level of rigor they were faced with; others may represent student-level groups that actively advocate for the elimination of licensure exams in favor of diploma privilege. In an effort to balance conflicts of interest, Buckendahl and Davis-Becker (2012) recommended equivalent participation by experienced practitioners, recently licensed or certified practitioners, and educators.

In credentialing examinations, there is a debate over the inclusion of panelists who lack expert-level knowledge given the focus on increased representation from diverse stakeholders. Certification and licensing exams typically cover a wide range of topics. However, it is the nature of an expert to have specialized knowledge (Raymond & Reid, 2001). Content expertise has been regarded as the most important qualification for judges (Buckendahl & Davis-Becker, 2012; Downing et al., 2006), and a lack of expertise could undermine the credibility of the judges and the soundness of the standard-setting procedure (J. C. Clauser et al., 2017; Norcini & Shea, 1997). Kane (1994b) suggested that panelists who lack the requisite expertise to knowledgably rate all items should only respond to items that they are qualified to address. However, additional data analysis would be required to account for the missing information. There are, of course, standard-setting contexts in which important stakeholders are likely to completely lack the knowledge and experience to make the judgments required of panel members; for example, members of the public cannot make the probability judgments required of the Angoff procedure for items from a medical licensing test. In this circumstance, the interests of that stakeholder group can be incorporated into the process as part of the instructions to the panelists. In addition, in situations in which the estimated passing score from the standard-setting exercise is reviewed and approved by examination governance, the interests of otherwise unrepresented stakeholders can be reflected in their decision process.

Nonetheless, it remains critically important to consider stakeholder perspectives in the standard-setting process because they represent the interests of groups with bona fide stakes in the outcome of a testing program (Hambleton & Powell, 1983). As noted by Kane (1994b), stakeholder input is a time-honored tradition and part of the observed democratic processes for establishing public policy, thereby providing evidence of the reasonableness of policy decisions related to a particular standard.

Procedural Factors: Training and Practice

We cannot underscore enough the importance of providing ample training and practice opportunities for panelists during the content-based standard-setting activity. As previously mentioned, panelists are recruited because of their content expertise and their experience with the relevant population of test takers. There is no reason to believe that they have special skills in content-based standard setting; in fact, in some contexts there is an explicit requirement that panelists *do not* have such experience. As such, it is wise to consider the training and practice aspects of the process to be as important as, if not even more important than, the process of collecting the operational judgments.

This training and practice can come in many forms. The first relates to providing specific information about the examination of interest. A thorough review of the examination purpose and structure is a good way to ensure that all panelists are beginning the process with a reasonable understanding of the test they are about to review. Similarly, it often is beneficial to have panelists participate in a session during which they complete a set of items much like a test taker would; this provides a general sense of the test-taker experience, which is directly relevant to their standard-setting judgments. The judgment task is to make decisions about an item based on how a minimally qualified candidate *would* answer the item; focus on the *would* (rather than the *should*) represents consideration of many factors, including the reality of the testing situation. Having the panelists take the test as the test takers do, for example, including a set of questions (a) without access to the correct answers and (b) with the same time limits that are used for the operational test, can provide valuable insight into the test-taker experience, particularly for panelists who are farther away in time from the credentialing process. Engaging in several rounds of judgments has been another feature of standard-setting activities that benefits the process by allowing panelists to practice with the overall task—and its component parts—prior to providing judgments for the data set that will be used to derive the cut score. Many different approaches to this iterative process have been described in the literature, and there is no single approach that has proven to be best. Please refer to Ferrara et al. (this volume) for additional information about workshop design and execution.

A final recommendation relates to replications of the content-based standard-setting process: Programs that have the resources to do so should consider repeating the process with different panels on multiple occasions. This provides examination governance with important information about the precision of the estimated passing score.

Decision-Making: A Policy Perspective

Although standard setting is typically viewed as a statistical and procedural process, it also involves elements of policy-based judgment (Cizek, 2006). As the *Standards* noted, the process of standard setting incorporates both value judgment and empirical considerations (AERA et al., 2014). Results from content-based standard-setting panels are one of the most important factors in establishing a passing score, but that source of data is not the only one used to inform a final passing score decision. Results typically are combined with additional sources of diverse, relevant, and often policy-related information to arrive at the most appropriate recommendation for a passing score in a given credentialing context (Geisinger & McCormick, 2010).

Proposed test score interpretations are not valid unless they are supported by evidence (Kane, 1994a; Kane 1994b). That is, the standard-setting process must be supported by evidence to validate a performance standard as an interpretation of a passing score. Specifically, Kane argued that validity of a pass/fail decision depends on the plausibility of two assumptions supporting the inference that a test taker's achievement is adequate if and only if their observed score is greater than or equal to the passing score. The first assumption, termed the *descriptive assumption*, is that the passing score corresponds with a particular performance standard. The second assumption, the *policy assumption*, is that the performance standard is appropriate for the purpose of the decision. The descriptive assumption, which relates to the content-based aspect of the standard, can be empirically evaluated. The policy assumption is more complex. Policy decisions regarding "how much is enough" (e.g., how much knowledge is required to demonstrate competence) are based on assumptions about the consequences of various choices and the values they reflect. Some assumptions about the consequences of choices that frame policy decisions can be empirically confirmed, but assumptions about values necessarily imply judgment. Therefore, to validate the policy aspect of the standard-setting process, evidence must demonstrate that the passing score and associated performance standard are reasonable. The following sections briefly describe several sources of evidence that we feel are of particular importance in credentialing contexts.

STANDARD ERROR Evidence supporting the policy assumption can be drawn from a number of different sources. The standard errors of various methods, for example, can provide procedural evidence of the reasonableness of the process (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Cross et al., 1984; R. L. Smith & Smith, 1988). The standard error of a passing score refers to the likely variation in passing scores that would result if the standard-setting activity were repeated with different items and/or panels (Kane, 1994b).

PSYCHOMETRIC INFORMATION Presenting pass rates and other psychometric information can promote consistency in how judges approach the standard-setting activity and also provides useful procedural evidence supporting the validity of the

policy portion of the standard-setting process (Hambleton & Powell, 1983; Jaeger, 1989; Kane, 1994b; Linn, 1978; Shepard, 1980). Statistical data on the performance of test-taker groups, including pass rates, can aid judges in setting cut scores at realistic levels. As Loomis (2012) described, in many standard-setting methods judges make initial recommendations and are subsequently provided with estimated pass rates associated with their recommendation. They then are given the opportunity to adjust their recommendations in light of the data.

In many credentialing organizations, the results of the standard-setting exercise are not used directly to establish a cut score but instead are considered one source of data intended to inform the governance group that is responsible for establishing the final cut score. That group has the authority to accept or modify the recommended cut score, and they may adjust the recommendation to ensure that different pass rates withstand public scrutiny and meet the demands of the profession (Geisinger & McCormick, 2010). For example, a decision-making body may want to compare expected pass rates with the pass rates of previous years because a substantial disparity would undermine the public confidence in and validity of the cut score. However, decision-makers should avoid adjusting cut scores to conform to a desired passing rate, especially if doing so is contrary to the goals of the standard-setting process (AERA et al., 2014; Geisinger & McCormick, 2010). In licensing, the goal of standard setting is to determine a cut score that identifies which candidates have the competence required for safe and effective practice. As such, pass rates cannot justify alterations that move the cut score outside the range of minimal competence.

CLASSIFICATION ACCURACY Data on classification accuracy is a critical source of evidence that may be used to support the policy portion of standard setting. False positives occur when a candidate's true score is below the cut score and their observed score is at or above the cut score; false negatives occur when a candidate's true score is above the cut score and their observed score is at or below the cut score (B. E. Clauser et al., 2006). In licensing and certification, it often is the case that classification errors are not equally important. For example, passing an incompetent physician may be deemed more problematic than failing a competent one. In such situations, it is reasonable for normative data to impact the recommended or adopted standards (Hambleton & Powell, 1983).

In credentialing contexts, the individual test takers bear the cost of false negatives and society bears the cost of false positives. These costs are difficult to approximate and balance. Nonetheless, balancing classification accuracies is an important tool for policy-related decision-making because it accounts for the societal costs of various decisions (Geisinger & McCormick, 2010).

Some sources of evidence are subject to stronger critique than others. For example, while the nature of the professional market is often a supplemental factor in determining cut scores for employment testing, the use of this information in credential-

ing assessments is far more dubious. The *Standards* explicitly state that this is not appropriate in the credentialing context (AERA et al., 2014). However, Millman (1989) found that cut scores on licensing tests often were lowered in situations where the supply of professionals was not able to meet market demand (Geisinger & McCormick, 2010).

Providing opportunities to retake a failed exam also has the potential to influence policy decisions because these opportunities can change the consequence of cut scores. For example, decision-makers may raise cut scores to allow for a higher initial failure rate (Geisinger & McCormick, 2010) but a lower false-positive rate across test administrations. Some have even suggested raising the cut score for subsequent examinations for test takers who have failed in a previous administration (B. E. Clauser & Nungester, 1999; Millman, 1989). However, any policy decision that incorporates retesting as a factor should first consider whether the standard-setting panel was instructed to consider retesting opportunities; if they were, further adjustment may be inappropriate.

In credentialing, this distinction is critical because achieving an examination score above the cut score is interpreted as an indication that the test taker has the requisite knowledge to receive the credential (or is a step toward receiving the credential if multiple examinations are required). Standard setting in credentialing contexts therefore has direct and potentially significant consequences on the test taker and the public with respect to permitting or denying either entry to a profession or the ability to use a specific designation to practice.

The Role of Judgment

Standards (and cut scores) for a given examination are fundamentally the product of human judgment (Kane, 1994b) and represent the intersection between a complex set of policy decisions informed by empirical data and the interests of a diverse group of stakeholders. As such, the inherent subjectivity of the standard-setting process has led to criticism about the "arbitrariness" of cut scores (Glass, 1978). Cut scores typically result from a mathematical transformation of the judgments made by a group of standard-setting panelists onto a score scale; the resulting cut score then may be presented to the policy makers so that they can make the final decision about where on the score scale the cut score should be located. These activities represent a string of judgment-based processes that understandably might lead to the perceived arbitrary nature of cut scores: panelists make judgments about test material or test-taker performance, the recommendation of a single cut score based on the panelists' judgments often proposes only one score from a range of legitimate scores, and the credentialing bodies have the authority to use their judgment to accept or modify cut score recommendations (Kane, 1994b). The inclusion of human judgment is unavoidable, however, because there is no statistical answer to the question, How much is enough? (Kane, 1998, 2002). Although aspects of the process may be perceived as arbitrary because of its foundations in human judgment, it is important to note that an arbitrary standard is not necessarily a capricious standard (Hambleton, 1978; Hambleton & Pitoniak,

2006). Capricious standards may undermine the validity and integrity of the process (Geisinger, 1991; Kane, 1994b), but arbitrary standards may escape the perception of capriciousness if they demonstrate that the standards align with the goals of the testing program. This demonstration requires methodical design and documentation of standard-setting processes that yield cut scores with a data-informed, empirical basis (AERA et al., 2014; Kane, 2017; Hambleton & Powell, 1983). The importance of this process is perhaps most critical in credentialing contexts where the risks of making decisions that are *not* based on sound methodology have implications for the safety of the public and the careers of the candidates.

As such, it is important for practitioners involved in standard-setting activities to (a) be aware of the myriad factors that have the potential to influence standard-setting judgments and (b) do their best to mitigate the associated risks. Ferrara et al. (this volume) provide a detailed description of hypotheses about factors that influence standard-setting judgments. These include task features, cognitive processes, judgmental heuristics, and cognitive biases. We strongly recommend review and consideration of this information for those seeking to implement standard setting in credentialing contexts. The authors also present a framework to guide both the design of standard-setting workshops and a structured approach to gathering evidence to support the validity of cut score interpretations (see Ferrara et al., Table 12.3, this volume). Careful attention to the guidance provided here and by Ferrara et al. will help ensure that standard setting for credentialing examinations is intentional, aligned with best practices, and results in defensible cut scores.

Major Influences on the Practice of Assessment for Licensing and Certification

Substantive changes in assessment for licensing and certification have been influenced by four markers. These include: (a) the rapid evolution of information and advances of technology and tools including more sophisticated psychometric approaches (Norcini et al., 2013); (b) the rise of competency-based education (Hauer et al., 2016); (c) an increased focus on the need for integrating learning and assessment over a professional's career (Baron, 2016; Continuing Board Certification, 2019); and (d) the rise of antitesting and antiregulation sentiments (Flier et al., 2016; Teirstein, 2015).

Rapid Evolution of Information and Advances of Technology and Tools

Recent decades have seen an explosion of information and, though keeping pace with it is essential, the task is increasingly challenging. Advances in cognitive psychology have deepened our understanding of how people use knowledge and information. At the same time, technological innovations are transforming assessment practices; for example, the Dental Interactive Simulation Corporation is doing important work using simulations to assess problem solving in dental hygiene (Williamson et al., 2006).

The challenge is making sense of the complex data resulting from this type of novel assessment format.

Technological advances abound, many of which directly support healthcare professions: these include the mandate of electronic health systems for all physician practices in the United States, increased use of point-of-care resources, precision medicine, the genome project, health wearables, 3D printing, robotic surgery, and decision support systems (Waldren et al., 2017). New assessment approaches are trying to incorporate some of these advances into the testing experience to make it more authentic to actual practice, such as allowing access to online external resources (e.g., UpToDate, an electronic clinical tool) during the testing experience and using simulation technology to assess performance more directly (Levine et al., 2012; Lipner et al., 2017). However, it is important to balance the cost and feasibility of these approaches against the additional value they provide in evaluating the desired construct(s). As mentioned previously, the technology itself inadvertently may introduce construct-irrelevant variance into the measurement. As such, design and impact of the new assessment should be researched thoroughly before it is used for making high-stakes decisions about test takers.

Item response theory (IRT) has become the standard psychometric model used in large-scale testing programs (see Cai et al., this volume, for discussion of IRT and other models). Its use has expanded in recent years in part due to sophisticated technology like computerized adaptive testing, as well as because of extensions of IRT for handling polytomous items and measuring multiple traits. There also has been increased interest in assessments that reflect how people practice their work. This has led to the development of more complex assessments that use advanced psychometric techniques such as computational psychometrics to assess both cognitive and noncognitive skills (von Davier, 2017). Computational psychometrics is quite sophisticated, combining mathematical models, machine learning, and data mining approaches as well as theory-driven psychometric approaches like evidence-centered design to measure latent abilities. The psychometric models that deal with complex data generally are IRT and Bayesian networks (Levy & Mislevy, 2016). These more complex models are important in credentialing settings because they allow for emulating real-world situations, such as using the Internal Revenue Service tax code on the Association of International Certified Public Accountants (AICPA) exams. Psychometric approaches also have become more sophisticated, particularly in areas such as data forensics. This field involves statistical analysis of the rich information automatically collected during computer-based testing—such as test taker responses, keystroke logs, and timing data—and prior performance to detect aberrant behavior and test fraud (de Klerk et al., 2019).

Rise of Competency-Based Education

In 1999, the Accreditation Council for Graduate Medical Education and the ABMS introduced six core competencies to define the relevant skills, knowledge, and attitudes that a trainee or physician should possess to provide quality health care

(Accreditation Council for Graduate Medical Education, 2020). Soon after, the competency-based education movement arose in residency and fellowship training programs; this movement resulted in defining subcompetencies and milestones for each of the six competencies to track a trainee's developmental progress within each competency and provide a structured approach to providing developmental feedback (Nasca et al., 2012). Instead of the traditional paradigm in which time spent in a training program is the primary indicator of a trainee's progression, mastery of knowledge and skills is the key to assessing progression. An outcomes-based approach is used to evaluate a trainee's progress in the medical program using the milestones framework. Assessment is both formative (giving feedback to enhance learning) and summative (making a decision about whether a trainee has achieved a certain level of competence). Research on the validity of the inferences based on milestone ratings is growing in the medical community, and undoubtedly the milestones will continue to be refined as the community learns more about the effectiveness of the framework (Hauer et al., 2018).

With the rise of competency-based education, the concept of programmatic assessment also has evolved. This approach combines various assessment instruments to provide a more holistic evaluation of a trainee (Schuwirth & van der Vleuten, 2012; Uijtdehaage & Schuwirth, 2018). The notion is based on the fact that each assessment instrument has its own inherent advantages and disadvantages, and it focuses on the meaningful triangulation across instruments, the stakes involved in the decision-making, and how the validity evidence is collected as the keys to success. The approach appears to be most beneficial in controlled environments such as training programs where it is largely associated with lower-stakes use cases (i.e., in the assessment *for* learning paradigm). It is less clear how this approach could be implemented for professionals in practice where there are few controlled environments. Furthermore, combining programs of assessment *for* and *of* learning (formative and summative) must be done with careful planning, especially with respect to validity considerations. Although there is appeal to approaching individual assessment from a programmatic perspective, and there is some evidence that it can work, more research is needed (Bok et al., 2018).

Increased Focus on Integrating Learning and Assessment Over a Professional's Career

One of the most significant changes in assessment relates to efforts to combine different assessment methods (e.g., formative and summative) into a cohesive, longitudinal program that supports a professional's development throughout their career. Much controversy has existed around programs for recertification (i.e., the maintenance of certification and continuing certification for medical professionals). As a result, a commission was formed that included public members as well as representatives from physician-led organizations. *The Continuing Board Certification: Vision for the Future Commission Final Report* (American Board of Medical Specialties, 2019),

which resulted from this effort, suggests that continuing certification programs for physicians should emphasize ongoing professionalism, focusing on learning and improvement in addition to assessment of current competence, with the ultimate goal of providing high-quality patient care. Initial certification programs, representing a culmination of training and traditional point-in-time assessments, sufficed for this purpose. Maintenance of certification, however, should be more continuous in nature. The American Board of Anesthesiology was the first of the ABMS boards to restructure its maintenance of certification program to a longitudinal program, called MOCA Minute, which claims to enable physicians to continuously assess themselves and identify knowledge gaps using an online portal (American Board of Anesthesiology, n.d.). Other ABMS boards soon followed. This is not just the case in medicine: Nursing and physician assistant credentials all have certification renewal programs (American Nursing Credentialing Center, n.d.; National Commission on Certification of Physician Assistants, n.d.).

From a measurement perspective, the challenge for these programs is being able to a) deliver enough quality questions in a nonsecure environment and to b) provide good educational feedback that includes identifying and remediating areas of weakness while c) maintaining a reliable and valid summative decision after 5 years (i.e., the time span for most of the medical education programs). There are many obstacles to doing this well; some of these include threats to security (because the questions and answers typically are exposed immediately after a question is answered), the difficulty of field testing enough questions with small sample sizes, maintaining the same standard over a period of time (e.g., 5 years), and choosing a scoring model that can adjust over time to allow for improvement. Strong research agendas are critical to evaluate the success of these longitudinal programs in terms of acceptability, feasibility, and meaningful learning and measurement.

The Rise of Antitesting and Antiregulation Sentiments

There has been a growing sentiment against testing and regulation across all of testing. For instance, in 2020 the University of California system announced that the SAT and ACT would no longer be required for undergraduate admissions (Hubler, 2021; see also Camara et al., this volume). For licensure and certification, most of the controversy has centered on healthcare, particularly regarding continuing certification programs and the Step 2 Clinical Skills examination for U.S. medical students (Flier et al., 2016; Teirstein, 2015; Teirstein & Topol, 2015). In 2018, the Institute for Credentialing Excellence, a leading developer for standards for certification programs, and the American Society of Association Executives formed a Professional Certification Coalition (Professional Certification Coalition, 2023). This coalition aims to counteract the negative impact of attempts to enact state legislation that undermines certification programs designed to provide data-driven, external validation of practitioners' competence by nongovernment private certification organizations. This legislation is directly working to weaken professional standards in the healthcare industry.

Weakened criteria for certificates could have serious consequences for patient care, particularly if the certificate is meaningless due to being obtainable solely through monetary exchange. Although the issue of weakened credentialing criteria is not the same as the proliferation of fake organizations offering phony certificates (Marcus, 2021), it still threaten the value of programs that maintain defensible standards and have conducted legitimate research to demonstrate that the assessments are meaningful and relevant.

CONCLUSION

Assessment for licensing and certification represents a segment of testing focused on protecting the public from unqualified practitioners. As this chapter has demonstrated, though assessment in this context shares similarities with other testing contexts, it is the differences that set it apart from tests used for other purposes. These differences are the focus of this chapter.

From a practical standpoint, the primary issues for credentialing bodies relate to test development and measurement considerations that impact the resulting examination and, by extension, examination outcomes. Validity evidence for the procedures used to develop, administer, score, and set passing scores for the test is central to the inferences made about candidates based on the test results. The stakes associated with licensing and certification examinations—including the risks of granting credentials to unqualified candidates and creating barriers to employment, thereby limiting the public's access to practitioners—make these considerations even more critical.

ACKNOWLEDGMENTS

First and foremost, thank you to our esteemed editors Linda Cook and Mary Pitoniak for inviting us to contribute to this prestigious volume and for providing unwavering support and patience throughout the entire production process. Even a global pandemic could not stop this dynamic duo! Thank you to our reviewer collaborators, Dave Swanson, Leslie Keng, and Nancy Tippins, for their insightful recommendations that yielded a much more concise (believe it or not!) and informative chapter. Finally, thank you to Benjamin Chesluk, Brian Clauser, Joanna Gorin, Michelle Johnson, Maxamillia Moroni, and Siddhartha Reddy, whose hard work and thoughtful contributions to our manuscript were invaluable!

REFERENCES

Accreditation Council for Graduate Medical Education. (2020). *The milestones guidebook*. https://www.acgme.org/globalassets/milestonesguidebook.pdf

Amalnerkar, E., Lee, T. H., & Lim, W. (2020). Reliability analysis using bootstrap information criterion for small sample size response functions. *Structural Multidisciplinary Optimization*, *62*, 2901–2913.

American Board of Anesthesiology. (n.d.). *MOCA minute*. Retrieved October 1, 2023, from https://www.theaba.org/maintain-certification/moca-minute/

American Board of Internal Medicine. (2023). *Internal medicine blueprint*. https://www.abim.org/~/media/ABIM%20Public/Files/pdf/exam-blueprints/certification/internal-medicine.pdf

American Board of Medical Specialties. (2019). *The continuing board certification: Vision for the Future Commission*. Retrieved June 9, 2025 from https://www.abms.org/wp-content/uploads/2020/11/commission_final_report_20190212.pdf

American Educational Research Association, American Psychological Association, & National Council on Educational Measurement. (2014). *Standards for educational and psychological testing*.

American Nursing Credentialing Center. (n.d.). *Renew your certification*. Retrieved October 1, 2023, from https://www.nursingworld.org/certification/renewals

Americans With Disabilities Act of 1990, 42 U.S.C. § 12101 *et seq*.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36*(1), 45–50.

Association of Test Publishers and the Institute for Credentialing Excellence. (2017). *Innovative item types: A white paper and portfolio*. Beyond MCQ: A Showcase of Examples. https://atpu.memberclicks.net/assets/innovative%20item%20types%20w.%20appendix%20copy.pdf

Balthazard, C. (2017, September 26). *#39 Defining professional regulation*. https://www.linkedin.com/pulse/39-defining-professional-regulation-claude/

Baron, R. J., & Braddock, C. H., III. (2016, December 29). Knowing what we don't know—Improving maintenance of certification. *New England Journal of Medicine*, 375(26), 2516–2517. https://doi.org/10.1056/nejmp1612106

Bok, H. G. J., de Jong, L. H., O'Neill, T., Maxey, C., & Hecker, K. G. (2018). Validity evidence for programmatic assessment in competency-based education. *Perspectives on Medical Education*, 7(6), 350–351.

Bowers, J. J., & Shindoll, R. R. (1989). *A comparison of the Angoff, Beuk, and Hofstee methods for setting a passing score* (ACT Research Report Series 89-2). ACT.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277–289.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4*(2), 219–240.

Brennan, R. L., & Prediger, D. J. (1981) Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687–699.

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). University of Iowa.

Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94-39). ETS.

Brossman, B. (2018, April 12–16). *Score reports: A collaborative design between measurement, communications, and subject matter experts* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.

Buckendahl, C. W., & Davis-Becker, S. L. (2012). Setting passing standards for credentialing programs. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 485–502). Routledge.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*, 145–163.

Cadle, A., & Parshall, C. G. (2015). *Innovative item types: Strengths and weaknesses*. http://www.proftesting.com/blog/2015/05/11/2015511innovative-item-types-strengths-and-weakness/

Chinn, R., & Hertz, N. (2010). *Job analysis: A guide for credentialing organizations* (CLEAR Research Brief). CLEAR.

Civil Rights Act of 1964, Pub. L. 88-352, 78 Stat. 241 (1964).

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93–106.

Cizek, G. J. (2006). Standard setting. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (pp. 225–260). Routledge.

Cizek, G. J. (2012). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3–14). Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications.

Cizek, G. J., Germuth, A. A., & Schmid, L. A. (2011). *A checklist for evaluating credentialing testing programs*. The Evaluation Center, Western Michigan University. http://www.wmich.edu/evalctr/checklists/

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge.

Clark, R. E. (2014). Cognitive task analysis for expert-based instruction in healthcare. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 541–551). Springer.

Clauser, A. L., & Raymond, M. (2017). Specifying the content of credentialing examinations. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policy and practice* (pp. 64–84). Routledge.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education, 22*(1), 1–21.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701–732). Praeger.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement, 46*(4), 390–407.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. *International Journal of Testing, 13*, 65–85.

Clauser, B. E., & Nungester, R. J. (1999). Considerations in adjusting cut-scores for certification and licensure decisions. *CLEAR Exam Review, 10*(2), 18–23.

Clauser, B. E., & Nungester, R. J. (2001). Classification accuracy for tests that allow re-takes. *Academic Medicine, 76*(Suppl. 10), S108–S110.

Clauser, B. E., Swanson, D. B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39*, 269–290.

Clauser, J. C., Hambleton, R. K., & Baldwin, P. (2017). The effect of rating unfamiliar items on Angoff passing scores. *Educational and Psychological Measurement, 77*(6), 901–916. https://doi.org/10.1177/0013164416670981

Copyright Act, 17 U.S.C. § 101 *et seq*. (1976).

Cox v. Ala. State Bar, 330 F. Supp.2d 1265 (M.D. Ala. 2004).

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey–Bass.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods of establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21*(2), 113–129.

Cui, Y., Gierl, M. J., & Chang, H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19–38.

Currier vs NBME, Case No. 07-J-434 (Mass. Ct. App. 2007), *vac. & rem.*, 965 N.E.2d 829 (2012).

D'Costa, A. G. (1986). The validity of credentialing examinations. *Evaluation & the Health Professions, 9*(2), 137–169. https://doi.org/10.1177/016327878600900202

de Klerk, S., van Noord, S., & van Ommering, C. J. (2019). The theory and practice of educational data forensics. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 381–399). Springer Cham.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295–311.

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35*(4), 296–316.

DeMers, S. T., Webb, C., & Horn, J. B. (2014). Psychology licensure and credentialing in the United States and Canada. In W. B. Johnson & N. J. Kaslow (Eds.), *The Oxford*

handbook of education and training in professional psychology (pp. 201–213). Oxford University Press.

Downing, S. M., Tekian, A., & Yudkowsky, R. (2006). Research methodology: Procedures for establishing defensible absolute passing scores on performance examinations in health professions. *Teaching and Learning in Medicine, 18*(1), 50–57.

Drew, J. (2016, April 4). NC may eliminate licensing for a dozen professions. *The Charlotte Observer*. http://www.charlotteobserver.com/news/politics-government/article69815452.html

Dubois, P. H. (1970). *A history of psychological testing*. Allyn & Bacon.

Edwards, M. C., & Vevea, J. K. (2006). An empirical Bayes approach to subscore augmentation. *Journal of Educational and Behavioral Statistics, 31*(3), 241–245.

Elchert, D. M. (2016). *Educating students about professional licensure in health service psychology*. https://teachpsych.org/resources/Documents/otrp/resources/Educating%20Students.pdf. University of Iowa: Office of Teacher Resources in Psychology.

ETS v. Hildebrant, 923 A.2d 34 (Md. 2007).

Equal Employment Opportunity Commission, Uniform guidelines on employee selection procedures (Title VII Regulations), 29 C.F.R. § 1607.2B (1985).

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Journal of Educational Measurement, 36*(1), 5–13.

Feinberg, R. A., & von Davier, M. (2020). Conditional subscore reporting using iterated discrete convolutions. *Journal of Educational and Behavioral Statistics, 45*(5), 515–533.

Feinberg, R. A., & Wainer, H. (2014a). A simple equation to predict a subscore's value. *Journal of Educational Measurement, 33*, 55–56.

Feinberg, R. A., & Wainer, H. (2014b). When can we improve subscores by making them shorter? The case against subscores with overlapping items. *Journal of Educational Measurement, 33*, 47–54.

Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education, 32*(2), 77–96.

Fine, S. A. (1986). Job analysis. In R. A. Berk (Ed.), *Performance assessment: Methods & applications* (pp. 53–81). Johns Hopkins University Press.

Flier, L. A., Henderson, C. R, & Treasure, C. L. (2016). Time to eliminate the Step 2 Clinical Skills Examination for US medical graduates. *JAMA Internal Medicine, 176*(9):1245–1246. doi:10.1001/jamainternmed.2016.3753

Foster, D. (2013). Security issues in technology-based testing. In J. Wollack & J. Fremer (Eds.), *Handbook of test security* (pp. 39–83). Routledge.

Foster, D. (2016). Testing technology and its effects on test security. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 235–253). Routledge.

Geisinger, K. F. (1991). Using standard-setting data to establish cut off scores. *Educational Measurement: Issues and Practice, 10*(2), 17–22.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*(1), 38–44.

Gierl, M. J., & Lai, H. (2015). Automatic item generation. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 410–430). Routledge.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, *46*(8), 757–765.

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education*, *29*(3), 196–210.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*(4), 237–261.

Gorin, J. S., & Embretson, S. E. (2012). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 23–35). Routledge.

Gray, B. M., Vandergrift, J. L., Johnston, M. M., Reschovsky, J. D., Lynn, L. A., Holmboe, E. S., McCullough, J. S., & Lipner, R. S. (2014). Association between imposition of a maintenance of certification requirement and ambulatory care–sensitive hospitalizations and health care costs. *Journal of the American Medical Association*, *312*(22), 2348–2357.

Gray, B. M., Vandergrift, J. L., Landon, B., Reschovsky, J. D., & Lipner, R. S. (2018). Association between American Board of Internal Medicine maintenance of certification status and performance on a set of healthcare effectiveness data and information set process measures. *Annals of Internal Medicine*, *169*(2), 97–105.

Gray, B. M., Vandergrift, J. L., & Lipner, R. S. (2018). Association between the American Board of Internal Medicine's general internist's maintenance of certification requirement and mammography screening for Medicare beneficiaries. *Women's Health Issues*, *28*(1), 35–41.

Gray, B. M., Vandergrift, J. L., McCoy, R. G., Lipner, R. S., & Landon, B. E. (2021). Association between primary care physician diagnostic knowledge and death, hospitalization and emergency department visits following an outpatient visit at risk for diagnostic error: A retrospective cohort study using Medicare claims. *BMJ Open*, *11*(4), Article e041817.

Gray, B. M., Vandergrift, J. L., Weng, W., Lipner, R. S., & Barnett, M. L. (2021). Clinical knowledge and trends in physicians' prescribing of opioids for new onset back pain, 2009–2017. *JAMA Network Open*, *4*(7), Article e2115328.

Gulino v. Board of Education of the City School District of the City of New York. (2015a, August 7), U.S. District Court, S.D. New York.

Gulino v. Board of Education of the City School District of the City of New York. (2015b, June 5), U.S. District Court, S.D. New York.

Gulino v. Board of Education of the City School District of the City of New York, 236 F. Supp. 2d 314 (S.D. N.Y. 2002), *aff'd in part*, *rev'd in part*, 460 F.3d 361 (2nd Cir. 2006), *on rem.*, Opinion & Order, Case 1:96-cv-08414-KMW (S.D. N.Y. 2012), *aff'd*, Summary Order, No. 1301001-cv (2nd Cir. 2014).

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Praeger.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15*(4), 277–290.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Praeger.

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard-setting. *Evaluation & the Health Professions, 6*(1), 3–24.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345–359.

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*(3), 249–266.

Hauer, K. E., Vandergrift, J., Hess, B., Lipner, R. S., Holmboe, E. S., Hood, S., Iobst, W., Hamstra, S. J., & McDonald, F. S. (2016). Correlations between ratings on the resident annual evaluation summary and the internal medicine milestones and association with ABIM certification examination scores among U.S. internal medicine residents, 2013–2014. *Journal of the American Medical Association, 316*(21), 2253–2262.

Hauer, K. E., Vandergrift, J., Lipner, R. S., Holmboe, E. S., Hood, S., & McDonald, F. S. (2018). National internal medicine milestone ratings: Validity evidence from longitudinal three-year follow-up. *Academic Medicine, 93*(8), 1189–1204.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates.

Holmboe, E. S., Wang, Y., Meehan, T. P., Tate, J. P., Ho, S., Starkey, K. S., & Lipner, R. S. (2008). Association between maintenance of certification examination scores and quality of care for Medicare beneficiaries. *Archives of Internal Medicine, 168*(13), 1396–1403.

Hubler, S. (2021, May 15). University of California will end use of SAT and ACT in admissions. *The New York Times*. https://www.nytimes.com/2020/05/21/us/university-california-sat-act.html

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*(4), 584–601. https://doi.org/10.1177/0013164403251284

Huynh, H. (1976). On the reliability of decisions in domain-referenced tests. *Journal of Educational Measurement, 13*, 253–264.

Institute for Credentialing Excellence. (2014). *National Commission for Certifying Agencies Standards for the Accreditation of Certification Programs*. https://www.credentialingexcellence.org/Portals/0/NCCA%20Standards%202021%20DRAFT%20REVISIONS_Sept%202021.pdf

Institute for Credentialing Excellence. (2018). *Standard setting overview for credentialing programs*. https://www.credentialingexcellence.org/blog/rsrch18-standard-setting-overview-for-credentialing-program

International Test Commission & Association of Test Publishers. (2022). *Guidelines for technology-based assessment*.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Macmillan.

Jones, A. T., Kopp, J. P., & Malangoni, M. A. (2018). Association between maintaining certification in general surgery and loss-of-license actions. *JAMA*, *320*(11), 1195–1196.

Jones, A. T., Kopp, J. P., & Malangoni, M. A. (2020). Recertification exam performance in general surgery is associated with subsequent loss of license actions. *Annals of Surgery*, *276*(6), 1020–1024.

Kane, M. T. (1994a). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions*, *17*, 133–159.

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

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, *9*, 355–379.

Kane, M. T. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education*, *10*, 5–18.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment*, *5*(3), 129–145.

Kane, M. T. (2002). Practice-based standard setting. *The Bar Examiner*, https://thebarexaminer.ncbex.org/wp-content/uploads/2018/10/710302-kane.pdf

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. T. (2017). Using empirical results to validate performance standards. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective* (pp. 11–30). Springer.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education*, *4*, 195–207.

Kearney, M. S., Hershbein, B., Boddy, D., Jácome, E., & Nantz, G. (2015). *Three targeted approaches to expand employment opportunities*. Brookings. https://www.brookings.edu/wp-content/uploads/2016/06/THP_three_approaches_expand_employment_framing.pdf.

Kelly v. W.Va. Bd. of Law Exam'rs, Case No. 2:08-00933 (S.D. W.Va. 2010).

Kinney, C. L., Raddatz, M. M., Sliwa, J. A., Clark, G. S., & Robinson, L. R. (2019). Does performance on the American Board of Physical Medicine and Rehabilitation initial certification examinations predict future physician disciplinary actions? *American Journal of Physical Medicine & Rehabilitation*, *98*(12), 1079–1083.

Kinney, C. L., Raddatz, M. M., Sliwa, J. A., Driscoll, S. W., & Robinson, L. R. (2020). Association of participation in the American Board of Physical Medicine

and Rehabilitation Maintenance of Certification program and physician disciplinary actions. *American Journal of Physical Medicine & Rehabilitation, 99*(4), 325–329.

Klesch, H. S. (2010). *Score reporting in teacher certification testing: A review, design, and interview/focus group study* (Publication No. AAI3409610) [Doctoral dissertation, University of Massachusetts Amherst]. Scholarworks. https://scholarworks.umass.edu/dissertations/AAI3409610.

Knapp, J., Anderson, L., & Wild, C. (Eds.), (2015). *Certification: The ICE handbook* (2nd ed.). Institute for Credentialing Excellence.

Kocher, M. S., Dichtel, L., Kasser, J. R., Gebhardt, M. C., & Katz, J. N. (2008). Orthopedic Board certification and physician performance: An analysis of medical malpractice, hospital disciplinary action, and state medical board disciplinary action rates. *The American Journal of Orthopedics, 37*(2), 73–75.

Kopp, J. P., Ibáñez, B., Jones, A. T., Pei, X., Young, A., Arnhart, K., Rizzo, A., & Buyske, J. (2020). Association between American Board of Surgery initial certification and risk of receiving severe disciplinary actions against medical licenses. *JAMA Surgery, 155*(5), Article e200093.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation & the Health Professions, 17*(2), 178–197. https://doi.org/10.1177/016327879401700204

LaDuca, A., Downing, S., & Henzel, T. (1995). Systematic item writing and test construction. In J. C. Impara, (Ed.), *Licensure testing: Purposes, procedures and practices* (pp. 117–148). Buros Institute of Mental Measurements.

Leighton, J. P. (2012). Learning sciences, cognitive models, and automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 23–35). Routledge.

Levine, A. I., Schwartz, A. D., Bryson, E. O., & DeMaria, S., Jr. (2012). Role of simulation in U.S. physician licensure and certification. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine, 79*(1), 140–153.

Levy, R. & Mislevy, R. (2016). *Bayesian psychometric modeling*. Chapman and Hall/CRC.

Linn, R. L. (1978). Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement, 15*(4), 301–308.

Lipner, R. S., Brossman, B. G., & Grosso, L. J. (2019). Changes to the American Board of Internal Medicine's maintenance of certification program: Preserving core assessment values. *Journal of Applied Testing Technology, 20*(S2), 11–20.

Lipner, R. S., Brossman, B. G., Samonte, K. M., & Durning, S. J. (2017). Effect of access to an electronic medical resource on performance characteristics of a certification examination. *Annals of Internal Medicine, 167*(5), 302–310.

Lipner, R. S., Messenger, J. C., Kangilaski, R., Baim, D. S., Holmes, D. R., Jr., Williams, D. O., & King, S. B., III. (2010). A technical and cognitive skills evaluation of performance in interventional cardiology procedures using medical simulation. *Simulation in Healthcare, 5*(2), 65–74.

Lipner, R. S., Young, A., Chaudhry, H. J., Duhigg, L. M., & Papadakis, M. A. (2016). Specialty certification status, performance ratings, and disciplinary actions of internal medicine residents. *Academic Medicine, 91*(3), 376–381.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9*, 13–26.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

Loomis, S. C. (2012). Selecting and training standard setting participants: State of the art policies and procedures. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). Routledge.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29–37.

Luecht, R. M. (2012). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59–76). Routledge.

Marcus, J. (2021, December 26). Growing "maze" of education credentials is confusing consumers, employees. *Washington Post.*

Margolis, M. J., Clauser, B. E., Harik, P., & Guernsey, M. J. (2002). Examining subgroup differences on the computer-based case-simulation component of USMLE Step 3. *Academic Medicine, 77*(10), S83–S85.

Margolis, M. J., & Feinberg, R. (Eds.) (2020). *Integrating timing considerations to improve testing practice*. Routledge.

McDonald, F., Duhigg, L., Arnold, G., Hafer, R., & Lipner, R. L. (2018). The American Board of Internal Medicine maintenance of certification examination and state medical board disciplinary actions: A population cohort study. *Journal of General Internal Medicine, 33*(8), 1292–1298.

McDonald, M. (2014). *The nurse educator's guide to assessing learning outcomes* (3rd ed.). Jones & Bartlett Publishers.

Medina, M. S., Castleberry, A. N., & Persky, A. M. (2017). Strategies for improving learner metacognition in health professional education. *American Journal of Pharmaceutical Education, 81*(4), Article 78.

Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. *Educational Measurement: Issues and Practice, 32*(3), 27–35.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*(6), 5–9.

Mislevy R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues & Practice, 25*(4), 6–20.

Morath, E. (2015, November 14–15). License law is nixed in D.C. *Wall Street Journal.*

Morrison, K. M., & Embretson, S. (2018). Item generation. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test development* (Vol. 1, pp. 75–94). Wiley.

Myers, A. J., & Bashkov, B. M. (2020). *Evaluating use of an online open-book resource in a high-stakes credentialing exam* [Unpublished manuscript]. American Board of Internal Medicine.

Nasca, T. J., Philibert, I., Brigham, T., & Flynn, T. C. (2012). The next GME accreditation system—rationale and benefits. *New England Journal of Medicine, 366*(11), 1051–1056.

National Commission on Certification of Physician Assistants. (n.d.). *Maintaining certification*. Retrieved October 1, 2023, from https://www.nccpa.net/maintain-certification/

National Conference of Bar Examiners. (2020, November). *Phase 3 report of the Testing Task Force*. National Council of Bar Examiners. https://nextgenbarexam.ncbex.org/reports/phase-3-report/

National Council of Architectural Registration Boards. (2023a, October). *ARE 5.0 format*. National Council of Architectural Registration Boards. https://www.ncarb.org/pass-the-are/prepare/are-5-0-format

National Council of Architectural Registration Boards. (2023b). *Pass the ARE*. https://www.ncarb.org/pass-the-are

National Council of Bar Examiners v. Multistate Legal Studies, 458 F. Supp. 2d 252 (E.D. Pa. 2006).

National Society of Professional Engineers. (2023). *NSPE: Who we are and what we do*. https://www.nspe.org/membership/nspe-who-we-are-and-what-we-do

Nelson, L. S., Duhigg, L. M., Arnold, G. K., Lipner, R. S., Harvey, A. L., & Reisdorff, E. J. (2019). The association between maintaining American Board of Emergency Medicine certification and state medical board disciplinary actions. *Journal of Emergency Medicine, 57*(6), 772–779.

Norcini, J. J., Lipner, R. S., & Grosso, L. J. (2013). Assessment in the context of licensure and certification. *Teaching and Learning in Medicine, 25*(Suppl. 1), S62–S67.

Norcini, J. J., Lipner, R. S., & Kimball, H. R. (2002). Certifying examination performance and patient outcomes following acute myocardial infarction. *Medical Education, 36*(9), 853–859.

Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10*(1), 39–59.

North Carolina Board of Dental Examiners s. FTC 135 S. Ct 1101 (2015).

O'Neill, T. R., Wang, T., & Newton, W. P. (2022). The American Board of Family Medicine's 8 years of experience with differential item functioning. *Journal of the American Board of Family Medicine, 35*(1) 18–25. https://doi.org/10.3122/jabfm.2022.01.210208

Palmer College of Chiropractic s. Davenport Civil Rights Commission, No. 12-0924 (2014). https://caselaw.findlaw.com/court/ia-supreme-court/1671310.html

Papadakis, M. A., Arnold, G. K., Blank, L. L., Holmboe, E. S., & Lipner, R. S. (2008). Performance during internal medicine residency training and subsequent disciplinary action by state licensing boards. *Annals of Internal Medicine, 148*(11), 869–876.

Parshall, C. G., & Harmes, J. C. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. *CLEAR Exam Review*, *19*(2), 18–25.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Issues in innovative item types. *Practical considerations in computer-based testing* (pp. 70–91). https://doi.org/10.1007/978-1-4613-0083-0_5

Peabody, M. R., Young, A., Peterson, L. E., O'Neill, T. R., Pei, X., Arnhart, K., Chaudhry, H. J., & Puffer, J. C. (2019). The relationship between board certification and disciplinary actions against board-eligible family physicians. *Academic Medicine*, *94*(6), 847–852.

Phillips, S. E. (2012). Legal issues for standard setting in K–12 educational contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 535–569). Routledge.

Phillips, S. E. (2017). Legal issue for credentialing examination programs. In S. Davis-Becker & C. Buckendahl (Eds.), *Testing in the occupations: Credentialing policies and practice* (pp. 228–276). Routledge.

Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–755). Praeger.

Plake, B. A., & Cizek, C. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 485–502). Routledge.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing interjudge consistency during standard setting. *Educational Measurement: Issues and Practice*, *10*(2), 15–16, 22, 25–26.

Poniatowski, P. A., Dugosh, J. W., Baranowski, R. A., Arnold, G., Lipner, R. S., Dec, G. W., Jr., & Green, M. M. (2019). Incorporating physician input into a maintenance of certification examination: A content validity tool. *Academic Medicine, 94*(9), 1369–1375.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, *6*, 1–9.

Professional Certification Coalition. (2023, October). *About PCC*. https://www.profcertcoalition.org/about-pcc

Professional Testing Corporation. (2019). *Standard setting: How to build a panel of subject matter experts*. https://ptcny.com/standard-setting-build-panel-subject-matter-experts/

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*(3), 266–285.

Ray, M. E., Daugherty, K. K., Lebovitz, L., Rudolph, M. J., Shuford, V. P., & DiVall, M. V. (2018). Best practices on examination construction, administration, and feedback. *American Journal of Pharmaceutical Education*, *82*(10), 7066. https://doi.org/10.5688/ajpe7066

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14*(4), 369–415.

Raymond, M. R. (2002). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues and Practice, 21*(3), 25–37.

Raymond, M. R. (2015) Job analysis, practice analysis, and the content of credentialing examinations. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 144–164). Routledge.

Raymond, M. R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 181–223). Lawrence Erlbaum Associates.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Standard setting: Concepts methods, and perspectives* (pp. 119–157). Lawrence Erlbaum Associates.

Romboy, D. (2012, August 10). *Court sides with woman in hair braiding case*. KSL Broadcasting. https://www.ksl.com/article/21638994/court-sides-with-woman-in-hair-braiding-case

Rosenfeld, M., Thornton, R., & Shimberg, B. (1983). Job analysis of licensed psychologists. *Professional Practice of Psychology, 4*(2), 17–24.

Rudner, L. M., & Gao, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *Journal of Applied Testing Technology, 12*, 1–12.

Schmitt, K. (1995). What is licensure? In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 3–32). Buros Institute of Mental Measurements. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1004&context=buroslicensure

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education, 46*(1), 38–48.

Section 504 of the Rehabilitation Act [Section 504], 29 U.S.C. § 701 *et seq*. (1973).

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*(4), 447–467.

Shimberg, B. (1982). *Occupational licensing: A public perspective*. ETS.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*, 150–174.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instruction module on subscores. *Education Measurement: Issues and Practice, 30*(3), 29–40.

Smiley, W. (2019, April). *An examination of classification accuracy in a continuous testing assessment framework* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Toronto, ON, Canada.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice, 22*(4), 26–33.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25*(4), 259–285.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.).

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*, 265–276.

Swanson, D. B., & Hawkins R. E. (2017). Using written exams to assess medical knowledge and its application. In E. S. Holmboe, S. J. Durning, & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (2nd ed., pp. 42–59). Mosby Elsevier.

Taylor, K. (2017, March 13). Regents drop teacher literacy test seen as discriminatory. *New York Times*. https://www.nytimes.com/2017/03/13/nyregion/ny-regents-teacher-exams-alst.html?mwrsm&_r=0

Tamblyn, R., Abrahamowicz, M., Dauphinee, W. D., Hanley, J. A., Norcini, J., Girard, N., Grand'Maison, P., & Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care. *Journal of the American Medical Association, 288*(23), 3019–3026.

Teirstein, P. S. (2015). Boarded to death–Why maintenance of certification is bad for doctors and patients. *New England Journal of Medicine, 372*(2), 106–108.

Teirstein, P. S., & Topol, E. J. (2015). The role of maintenance of certification programs in governance and professionalism. *Journal of the American Medical Association, 313*(18), 1809–1810.

Thompson, N. (2018, May 23). *What is the Hofstee method for setting cutscores?* Assessment Systems Corporation. https://assess.com/what-is-the-Hofstee-method-for-setting-cutscores/

Title VII of the Civil Rights Act, 42 U.S.C. § 2000e *et seq*. (1964).

Uijtdehaage, S., & Schuwirth, L. W. T. (2018). Assuring the quality of programmatic assessment: Moving beyond psychometrics. *Perspectives on Medical Education, 7*(6), 350–351.

Ura, A. (2019, May 27). Texas plumbing board and laws abolished after legislative strife. *Texas Tribune*. https://www.texastribune.org/2019/05/27/texas-plumbing-board-and-laws-abolished-after-legislative-strife/

Vandergrift, J. L., Gray, B. M., Barnhart, B. J., Lynn, L. A., & Lipner, R. S. (2020). Opportunities for maintenance of certification to better reflect scope of practice among medical oncologists. *JCO Oncology Practice, 16*(8), 641–648.

Vandergrift, J. L., Weng, W., & Gray, B. M. (2021). The association between physician knowledge and inappropriate medications for older populations. *Journal of the American Geriatric Society, 69*(12), 3584–3594.

van der Linden, W. J., & Li, J. (2016). Comment on three-element item selection procedures for multiple forms assembly: An item matching approach. *Applied Psychological Measurement, 40*(8), 641–649.

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement, 54*(1), 3–11.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K., & Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores

based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Routledge.

Waldren, S. E., Agresta, T., & Wilkes, T. (2017). Technology tools and trends for better patient care: beyond the EHR. *Family Practice Management*, *24*(5), 28–32.

Way, W. D., Steffen, M., & Anderson, G. A. (2014). Developing, maintaining, and renewing the item inventory to support CBT. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (3rd ed., pp. 143–164). Routledge.

Way, W. D., & Gialluca, K. A. (2017). Interpreting the meaning of test scores. In C. Buckendahl & S. Davis-Becker (Eds.), *Testing in the professions* (pp. 105–122). Taylor & Francis.

Wendt, A., & Harmes, J. C. (2009). Evaluating innovative items for the NCLEX, Part 1: Usability and pilot testing. *Nurse Educator*, *34*(2), 56–59.

White, W. D. (2014). Professional self-regulation in medicine. *Virtual Mentor*, *16*(4), 275–278.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Psychology Press.

Zaglaniczny, K. L. (1993). Council on certification professional practice analysis. *Journal of the American Academy of Nurse Anesthetists*, *61*(3), 241–255.

Zapata-Rivera, D. (Ed.). (2018). *Score reporting research and applications*. Routledge.

Zenisky, A. L., & Hambleton, R. K. (2015). A model and good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed, pp. 585–602). Routledge.

Zhou, Y., Sun, H., Culley, D. J., Young, A., Harman, A. E., & Warner, D. O. (2017). Effectiveness of written and oral specialty certification examinations to predict actions against the medical licenses of anesthesiologists. *Anesthesiology*, *126*(6), 1171–1179.

Zhou, Y., Sun, H., Macario, A., Keegan, M. T., Patterson, A. J., Minhaj, M. M., Wang, T., Harman, A. E., & Warner, D. O. (2018). Association between performance in a maintenance of certification program and disciplinary actions against the medical licenses of anesthesiologists. *Anesthesiology*, *129*(4), 812–820.

Zhou, Y., Sun, H., Macario, A., Keegan, M. T., Patterson, A. J., Minhaj, M. M., & Warner, D. O. (2019). Association between participation and performance in MOCA Minute and actions against the medical licenses of anesthesiologists. *Anesthesia & Analgesia*, *129*(5), 1401–1407.

NOTES

1. Teacher certification represents an interesting example of ambiguity between licensure and certification. States require certification for individuals teaching in public schools, but there is often no similar restriction for teaching outside the public system. This is similar to physician certification in the United Kingdom, where

physicians can practice without being a member of one of the Royal Colleges but cannot receive reimbursement from the National Health Service.

2. As with many issues related to licensure and certification, there are exceptions—or at least gray areas. For example, the USMLE comprises three steps. Successful completion of Steps 1 and 2 typically is required to enter residency training, and in some states individuals receive a formal training license. Because of this special use, these first two steps focus on what might be viewed as entry-level skills. A full unrestricted license is, however, only awarded after completing the third step in the examination process—the focus of that step is not constrained in the same way.
3. Work samples are another common item type in personnel assessments for certification.
4. In addition, it is important that decisions about feedback be made at the test development phase (i.e., as an intentional part of the overall testing process) rather than as an additional consideration after the test has been developed and/or administered.