JOURNAL OF MENTAL HEALTH AND CLINICAL PSYCHOLOGY
www.mentalhealthjournal.org

Review article Open Access

Mental health clinical exams' evident adherence to industry standards for testing

Benjamin E. Caldwell
California State University Northridge, CA, USA

Article Info

Article Notes
Received: August 03, 2023
Accepted: September 25, 2023

***Correspondence:**
*Dr. Benjamin E. Caldwell, PsyD, California State University Northridge, CA, USA. Email: drbencaldwell@gmail.com

Keywords:
Licensing exam
Examination
Clinical exam
Testing standards
Mental health
Licensure

Abstract

The developers of clinical exams for US mental health licensure have faced significant recent criticism and calls for their exams to be paused or discontinued.[1,2] Critics cite concerns over exams lacking evidence of validity, while they demonstrate strong evidence of racial and ethnic bias. Developers, in turn, argue that their exams are developed using accepted methods that conform with industry standards, specifically, the *Standards for Educational and Psychological Testing*.[3]

This manuscript challenges that assertion. Based on external research as well as developers' own statements and publications, clinical exams for US mental health care licensure appear to deviate in important ways from both the letter and the spirit of the *Standards*. Clinical exams should be paused unless and until they are shown to be fair, equitable, valid, and more fully consistent with industry norms.

Introduction

For each of the four major licensed mental health professions in the United States, licensure is typically conditioned upon the passing of a clinical exam. Psychologists must pass the Examination for Professional Practice in Psychology (EPPP), and soon, the EPPP Part 2.[4] Counselors, depending on their jurisdiction, must pass either the National Counseling Exam or the National Clinical Mental Health Counselor Exam.[5] Clinical social workers must pass the ASWB Clinical Exam.[6] Marriage and family therapists (MFTs) in all states except California must pass the National MFT Exam.[7] California MFTs must pass a similar exam developed by the state.[8]

These exams have faced significant criticism. Concerns over the ASWB Clinical Exam have intensified following the 2022 release of 10 years of pass rate data.[9] This data showed that Black examinees were more than three times as likely to fail the exam on their first attempt compared to white examinees.[9] Additional disparities in pass rates were shown to exist on the basis of age (pass rates declined with age, with those 50 and older being significantly less likely to pass than those 18-39) and primary language (those whose primary language was not English were significantly less likely to pass). Psychology's EPPP has shown similar disparities.[10-14] Criticism of the EPPP has increased as the ASPPB has pursued implementation of the EPPP Part 2, which some have argued is unnecessary and lacks evidence of validity.[15] And in marriage and family therapy, a pattern of racial disparity in exam outcomes similar to the ASWB Clinical Exam and the EPPP has been observed, with race being a particularly strong statistical predictor of whether one passes their clinical exam.[16] Across mental health fields,

clinical exams for licensure have similar structures and development processes, leading critics to argue that they are also likely to share the same significant flaws.[17]

State licensing boards and legislatures have taken note of these equity and validity concerns. As the EPPP Part 2 was in development, several states expressed resistance to implementing the new test in light of concerns about necessity, cost, and validity.[18,19] In response to ASWB's 2022 report, several states proposed legislation that would create alternate pathways to licensure for social workers at various levels.[20] These alternate pathways would allow new professionals to achieve licensure without first passing a licensing exam. Illinois ultimately created such a pathway.[21]

In choosing to adopt clinical exams and then to continue using them, licensing boards rely on specific assurances. Exam developers assert to boards that these exams are developed in accordance with industry standards.[9,22] State boards, typically not comprised of testing experts, rely on these assurances in determining that a specific exam is valid, appropriate, and legally defensible. This manuscript critically examines such assertions.

AERA Standards

The *Standards for Educational and Psychological Testing* (from here, "AERA standards" or simply *"Standards"*) were jointly developed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, each of which formally adopted the final version.[3] As such, the *Standards* are considered the defining industry standards for measurement in education and psychology in the US. When any exam developer claims that their exam was developed consistent with "industry standards," the AERA standards are the appropriate point of reference. In some cases, exam developers and reviewers specifically note that these are the standards being considered.[9,23]

Limitations

A number of factors limit the scope of this review. First, this review relies solely on publicly available information. That information is shared in publicly available exam handbooks, developers' web sites, other statements and assertions made by test developers in public venues, and published research. This review makes no presumptions about the information that developers may or may not have shared with licensing boards behind closed doors.

Second, as the *Standards* themselves note, conformity with those standards is not based simply on a checklist. Individual standards should not be considered in isolation, and satisfying the literal language of a standard may be less important than meeting its intent.[3] For example, minor changes were required for the administration of many common tests in the early stages of the COVID-19 pandemic. These changes may have amounted to minor deviations from the *Standards* as written, which call for test administrators to carefully adhere to the standardized guidelines published by the test developer. But these changes often were made in the interest of examinee or administrator safety, with appropriate cautions to the examinee and to those making use of test results, and with attention to preserving fairness and objectivity in testing to the greatest degree possible.

The discussion in this review is limited to a handful of critical areas where it appears that clinical exams in mental health care fall short of both the letter *and* the spirit of the AERA standards.

Finally, at some points, this review notes the absence of evidence for compliance with specific standards. (While some frame "adherence" as more proactive, and "compliance" as more passively doing what one has been told,[24] both terms refer to behavior that is consistent with external directions. They are used interchangeably here.) Based on the AERA standards, exam developers have an affirmative responsibility to validate their constructs and claims. However, a lack of evidence should not be construed to mean that those constructs or claims are necessarily invalid – such a conclusion would be a logical fallacy. Rather, this review asserts only that the constructs or claims at issue require supporting evidence, and that evidence has not been offered.

The issues discussed here do not constitute a comprehensive accounting of potential concerns related to clinical exams' adherence to the *Standards*. Rather, this review highlights some of the most significant areas of immediate concern.

Construct Clarity

The very first of the specific AERA standards demands that "the construct or constructs that the test is intended to assess should be described clearly" (p. 23).[3] Standard 8.0, labeled an "overarching" standard related to examinees' rights, goes farther, asserting that examinees "have the right to adequate information to help them properly prepare for a test[,] so that the test results accurately reflect their standing on the construct being assessed and lead to fair and accurate score interpretations" (p. 133).[3]

Clinical exam developers generally describe their constructs using language related to the knowledge base necessary for minimally competent, entry-level practice of the profession. This terminology is quite broad, and rather than using external referents for the construct of "minimal competence," the construct appears to exist wholly and exclusively as the achievement of a passing score on the exam. Developers note that they determine a threshold of minimum competence using a committee process that is a component of exam development.

Developers' non-specific definitions of this broad construct make it difficult for examinees to know how they can effectively prepare. While exam developers provide general lists of *topics* covered on their exams through specific exam handbooks, these lists are not specific enough to allow examinees the level of clarity the AERA standards appear to demand.

Consider, for example, questions related to professional ethics. For those preparing for the California MFT Clinical Exam, both the California Association of Marriage and Family Therapists and the American Association for Marriage and Family Therapy offer relevant ethics codes.[25,26] The two codes have more than two dozen areas of substantive disagreement, where behavior that would be considered ethical under one code may not be considered ethical under the other. The California Office of Professional Examination Services (OPES), the exam developer, offers *no references or sourcing at all* for its exam development work,[27] so examinees are left to fend for themselves in determining which code to study. OPES has, to date, never publicly specified which code it considers the right one. It did so privately via email (the CAMFT code turns out to be the "correct" one) only in response to a specific request challenging their lack of disclosure as a potential legal defensibility issue.[28] Less severely, while the guidebook for the NCMHCE[29] does list the American Counseling Association's Code of Ethics[30] as a "useful reference" for examinees, it does not clarify whether the ACA code, or the somewhat different code offered by the American Mental Health Counseling Association (AMHCA),[31] is the one actually used in exam development.

Similarly, exam developers do not typically provide guidance as to whose writing is considered the "source of truth" for each specific treatment model covered on the exam. As a result, examinees have no way of knowing whose version of a model is considered the "right" one for the purposes of their exam. Models evolve over time, and even those who contributed to the development of a model can significantly disagree about key elements of that model. For example, strategic therapy may reasonably appear on clinical exams for all of the mental health professions, given their scopes of practice and exam content areas. Should examinees base their answers on the Mental Research Institute group's version of strategic therapy, which emphasized paradoxical interventions and customizing therapy to each client?[32] Or should they answer based on the Milan group's version of strategic therapy, which emerged at roughly the same time, actively disavowed paradoxical interventions, and suggested using the same initial intervention with every client?[33] Examinees have no way of knowing. They may have significant, deep knowledge of a model, and their answers to questions about that model may be scored as incorrect anyway – simply because item writers drew from a different perspective on the model.

Similar issues can occur related to "correct" treatment for specific clinical issues. When a sample item for the ASWB Clinical Exam based its "correct" answer on the Kubler-Ross stage model of grief, an AI-powered chatbot flagged this as inconsistent with current standards of evidence.[34] Examinees and boards have nothing more to go on than developers' assurances that the knowledge assessed in their exams is current and appropriately evidence-based.

A better examination process would more transparently identify the anchor points for *all* the knowledge it ostensibly assesses. This would allow examinees to adequately prepare for their exams without relying on guesswork, leading exam scores to reflect the examinee's knowledge more accurately. It also would allow boards, examinees, and others to assess whether those anchor points are current and appropriate. Examinees should never be left to guess which version of an ethics code, treatment model, or set of diagnostic guidelines is considered the "correct" one for testing purposes. Developers should provide examinees with specific, comprehensive lists of the source material used in exam development. Without such clarity, what is measured by a clinical exam is not knowledge of the profession, it is *agreement with the anchoring of the item writers.*

Validity

Construct validity: Woven through many of the *Standards* is the importance of construct validity – the notion that a test should assess the actual construct that is intended to be assessed. As noted above, construct clarity – a prerequisite for construct validity – is a significant and foundational weakness for clinical exams. The *Standards* demand that developers go beyond simply *clarifying* the construct assessed, and *validate* their constructs and the resulting score interpretations.

The guidebooks that exam developers provide for examinees offer content outlines related to each clinical exam. However, it is not clear that the knowledge areas assessed in a clinical exam coalesce into a singular, cohesive construct of knowledge or competence that is best measured as a single variable. And there is at least some reason to suspect that they do not. Clinicians may reasonably be highly knowledgeable in some areas and lack basic knowledge in others. A clinician who is deeply lacking in knowledge about ethical practice or crisis intervention – fundamental public safety concerns – could still be deemed by the exam to be sufficiently competent to practice if they respond correctly to enough questions in other content areas, such as treatment planning.[17]

Furthermore, many content areas for all the exams considered here do not rationally connect with constructs

related to minimum competence or safety in practice. A clinician unsure of how to respond in a crisis, or who is unclear about ethical rules surrounding dual relationships, may indeed be unsafe or inadequately competent for independent practice. However, one who is not steeped in the specific language and interventions of a specific treatment model *that they do not use* would not appear to present a meaningful risk to the public,[17] or to be incompetent in the treatments that they do provide.

Impact of construct-irrelevant variance: For any measurement to be valid, it needs to measure the construct intended to be measured without significant interference from other factors or processes. Put more simply, licensure exams should fundamentally capture the *knowledge* they intend to capture, and not other factors like *test-taking skill*.

Test developers appear to turn a willful blind eye to the role of test-taking skill in exam results. The ASWB handbook tells examinees that "secrets and tricks don't really exist" (p. 29).[9] This is plainly inconsistent with decades of research showing that the use of test-taking strategies can significantly improve performance on multiple choice tests in general.[35-37] Research also suggests an important role of test-taking skill specifically on the ASWB Clinical Exam, as both clinicians-in-training[38] and an artificial intelligence engine[2] have been able to pass the exam *without even seeing exam questions,* choosing answers based only on cues and patterns in the available response options.

As part of its response to criticism over racial disparities in exam outcomes, ASWB itself has piloted an untested program teaching test-taking skills (in this case, a "mastery mindset;"[39] para. 1) to those who have failed an ASWB exam at least once.[40] ASWB cannot have it both ways: This pilot program is either an acknowledgement that test-taking skills *do* matter, and thus that their exams are subject to significant construct-irrelevant variance, or ASWB is providing failed examinees with a study aid that ASWB itself does not actually believe will help them.

Criterion validity: As Callahan and colleagues[15] pointed out, validity of an exam is not limited to validity of exam *content*. The AERA standards require assessment of the validity of *score interpretations* – that is, how scores are actually used for decision-making in practice.

In general, tests can either be *norm-referenced* or *criterion-referenced*. In a norm-referenced test, an individual's score is compared against the performance of other individuals. The SAT would be an example of a norm-referenced test, where scores are based on individual performance in comparison to a cohort of examinees. Clinical exams in mental health care are criterion-referenced, meaning that score interpretations are ostensibly based not on comparison with other examinees, but with a specific criterion – in this case, the knowledge level that exam developers, through their development committees, have deemed necessary for minimally competent practice of the specific profession.

While the *clarity* of the construct being assessed by these exams is questionable as noted previously, it is evident that licensing boards believe an exam score means *something* more than just a number. Boards, at the encouragement of test developers, interpret clinical exam scores as meaningful gauges of readiness for independent practice.

AERA standard 5.5 demands that when scores are used for criterion-based interpretation, as they are in clinical exams, "the rationale for recommended score interpretation should be explained clearly" (p. 103).[3] In other words, developers and users have a responsibility to explain to examinees the basis for concluding that their exam performance makes them unfit for practice. For an exam with a passing score cutoff of 70%, why is a score of 69% considered unsafe for independent practice, while a score of 70% is safe? Rather than receiving such an explanation, examinees typically receive a one-page score report with a handful of subscale scores.

The *Standards* further suggest that "Serious efforts should be made whenever possible to obtain independent evidence concerning the soundness of such score interpretations" (p. 103).[3] No developer appears to have made any meaningful efforts in this direction, despite decades of opportunity to do so.[17] ASPPB has offered two defenses of its failure to seek evidence to validate score interpretations. The first, that such evidence is unobtainable because those who fail their exams do not get licensed,[22] is easily dismissible: Those who do not pass their exams are typically allowed to remain in practice, they simply must remain under supervision. Furthermore, thousands of practitioners across the mental health professions today are practicing with licenses obtained through grandparenting rather than through exams. Grandparented licensees, or those who failed an exam but remain in practice, could be compared with similar populations who passed an exam to see whether one group is more prone than the other to licensure complaints, disciplinary actions, civil judgments of liability, or other potential markers of safety in practice.

These potential comparison points relate to ASPPB's other explanation for why it has not sought evidence that its score interpretations are valid. ASPPB has taken the position that *no* criterion against which their exam could be presently compared would be as reliable or valid as the exam itself, and that therefore, no attempt at establishing criterion validity would hold value.[41] This is a surprising position for ASPPB to take, given APA's specific apology for how standardized testing has historically disadvantaged individuals on the basis of their race and ethnicity.[42] The

notion that the best and most fair way possible to assess one's knowledge is through a standardized, multiple-choice exam, with a cut score set by a committee, has proven itself time and again to not only be incorrect, but a point of view that upholds structural racism.

Improving validity: Exam developers should engage in meaningful post hoc analysis of whether their processes for developing exams and setting cut scores result in valid and appropriate decisions about whom to qualify for licensure. Rather than treating longstanding criticism of exam validity as a nuisance, developers must proactively establish that their exams are both valid and equitable. Exam users (licensing boards) should seek evidence to determine the full range of impacts of current examination regimens on the safety and competence of licensees, including impacts on the availability of an adequate and diverse range of licensed mental health providers.

Fairness

Closely linked to questions of exam validity are questions of fairness. "Fairness to all individuals in the intended population of test takers is an overriding, foundational concern" (p. 49).[3]

The data in ASWB's 2022 report revealed that ASWB's exams possess a high degree of disparity in outcomes based on the race, ethnicity, and age of the examinee. The ASWB report joins a significant sequence of studies of Psychology's EPPP,[10-14] and emerging scholarship on both the California and national MFT clinical exams,[16] in revealing that clinical exams across mental health professions produce disparate outcomes on the basis of race and ethnicity. These disparities are strikingly similar across exams, heavily favoring white examinees over other racial and ethnic groups, particularly Black examinees.

It is reasonable to assume good faith on the part of developers. These outcome disparities, and the resulting racial and ethnic disparities in licensee populations, are perhaps best categorized as unintended consequences of exam use. Helpfully, the *Standards* offer a clear and specific path forward for such events. "When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct" (standard 1.25, p. 30).[3] No examples were located of exam developers even acknowledging a failure of the test itself as a *possibility*, much less investigating it as the *Standards* require. Instead, developers have sought to shift blame for outcome disparities to graduate programs or other upstream factors.[43,44] While there indeed are disparities elsewhere in the professional pipeline,[45] this does not absolve exam developers of their responsibility to investigate whether their exams are failing to capture their intended constructs. Disparity through the professional pipeline and structural bias in exams are not mutually exclusive. Indeed, disparities throughout the pipeline would seem to make it *more* likely, not less, that those who make it through the existing pipeline and ultimately are involved in exam development might produce an exam that is biased.

Test users – the licensing boards utilizing these exams – also have specific responsibilities for ensuring fairness in testing. For example, most of these exams are offered only in English. Some licensing boards provide additional exam time for examinees whose native language is not English when taking English-language exams. However, the *Standards* demand that test users go farther (standard 9.11, emphasis added): "When circumstances require that a test be administered in the same language to all examinees in a linguistically diverse population, *the test user* should investigate the validity of the score interpretations for test takers with limited proficiency in the language of the test" (p. 145).[3] There does not appear to be any evidence to suggest that licensing boards are engaging in such investigations as the *Standards* require.

Considering that fairness is a foundational concern in testing, the response to this concern must be comprehensive. Exam developers should re-examine every facet of exam development, structure, and utilization for evidence of bias. While multiple-choice exams are convenient and efficient, this structure for performing and assessing professional knowledge bears little resemblance to actual clinical practice, where clinicians can ask follow-up questions, utilize resources, and consult with colleagues, often with minimal time constraint. There is ample evidence suggesting that the current exam structure contributes a unique source of bias to the licensing process. Licensing boards should take a more skeptical position toward developers, and engage in their own independent evaluation of whether current testing processes and outcomes are fair and equitable.

Statistical Analysis

Exam items should face careful scrutiny at each stage of development and usage. This scrutiny comes, in part, through statistical analysis of item performance. The *Standards* do not dictate specific analytic methodologies, as preferred methodologies are continually advancing. However, the *Standards* do identify appropriate *levels* of statistical analysis that exam developers should engage in. The *Standards* also specify appropriate responses when individual items or full exams show meaningful weaknesses.

Differential item functioning (DIF): At this level of analysis, individual exam items are tested to determine whether they perform differently for different groups

of examinees. For example, an individual item that Asian examinees answer correctly at a much higher rate than Black examinees should draw scrutiny. Such items, if allowed through, can result in an exam that is biased against specific groups for reasons unrelated to their underlying knowledge – the very definition of construct-irrelevant variance. This can undermine the overall validity of an exam.

ASWB has reported engaging in DIF analysis.[46,47] ASPPB has reported that individual items on the EPPP are specifically tested for bias,[48] suggesting DIF analysis even if not using the specific term. NBCC and AMFTRB both say that their exams are carefully validated,[7,49] though they do not specifically address whether they engage in DIF analysis.

The California MFT Clinical Exam, which has been in use in its current form since 2016, has not been subjected to DIF analysis *at all, ever*.[50] OPES, the exam developer, cited state law prohibiting the mandatory gathering of demographic data during the licensing process[51] as the reason for this lack of DIF analysis. However, existing law does not prohibit the gathering of such data on a voluntary basis.[52] OPES said only that it would "explore" such analysis in the future (p. 8).[53] California MFTs have been subjected to high-stakes testing for many years without even the minimal safeguard against exam bias that comes with testing for DIF.

Differential test functioning (DTF): Separate from DIF analysis is the process of analysis for differential *test* functioning. At this level of analysis, a full exam is reviewed in its totality to assess whether the full exam advantages specific demographic groups. Meaningful DTF may occur even in the absence of statistically significant DIF, if small amounts of DIF on individual items tend to "lean" in the same direction, favoring the same group of examinees.[54,55] This reality is acknowledged in the *Standards*, which specifically note that DIF and DTF each may occur in the absence of the other.[3]

AMFTRB, ASPPB, and NBCC do not appear to directly discuss their use of DTF analysis. As noted above, OPES does not gather the data necessary for either DIF or DTF analysis of the California MFT Clinical Exam, so neither has ever been done.

ASWB has publicly argued that DTF analysis on its exams is unnecessary, because as they describe it, DIF analysis is "more stringent" (para. 6).[46] They base this conclusion on the general conclusion in testing that "DIF does not typically favor one examinee group consistently" (para. 6).[46] However, the very data ASWB produced in 2022 suggests at least the meaningful *possibility* that small amounts of DIF on individual items may be moving generally in the same direction, creating disadvantages for Black and older examinees at the exam level. Failure to actually check for DTF, when available data raises the possibility of DTF, would appear to violate both the language and intent of the *Standards*.

Scoring error: With thousands of questions in use across these exams at any given time, some small level of scoring error is inevitable. Such errors, in and of themselves, are not necessarily a sign of weakness in testing. They do not necessarily indicate non-compliance, in letter or spirit, with the *Standards*. It is the *failure to identify and respond to* such errors that would represent a significant concern, and a potential compliance issue.

When scoring errors are suspected or identified, the AERA standards demand rescoring (standard 9.5).[3] The *Standards* do not specify *how* this rescoring should be done, and it reasonably may vary depending on the nature of the error. Problematic items may be re-keyed; multiple response options on an item may be counted as correct; an item may be removed from scoring, with a test-taker's score determined based on the percentage of remaining scored items answered correctly; or another process may be more appropriate, so long as it supports the intention of maximizing fairness and accuracy in test scoring and interpretation.

ASWB has repeatedly stated that they continue to monitor items for DIF after those items enter the pool of scored items, and that they remove scored items that show DIF.[47,56,57] (They report that "typically < 5% of items" are removed due to DIF.[58]) Yet they have also said that removed items never impact scores.[58] These two statements are contradictory. Regardless of the specific rescoring method, removal of scored items would necessarily impact the scores of those who had taken an exam with one or more problematic items on it.

Indeed, the only way for both of these statements to be truthful is if ASWB is failing to rescore exams that included scored items later removed due to DIF. This would represent a major deviation in both letter and spirit from the *Standards*, which demand rescoring and user notification in such instances. If scored items are in fact being removed for DIF and rescoring of impacted exams is *not* occurring, this issue *by itself* could be keeping hundreds or even thousands of ASWB examinees each year unfairly closed out of licensure.[59]

Language from NBCC, meanwhile, similarly suggests that scored NCMHCE items flagged for DIF or other deficiencies may not result in rescoring. NBCC's most recent description of its exam development process simply says that released (scored) items found not to meet appropriate statistical standards are "flagged to be revised or retired" (p. 10).[49]

Improving statistical analysis: Exam developers should use multiple forms of analysis to identify racial

and other forms of bias in the exam process. DIF and DTF analysis are both warranted, considering the available data. When scored items are removed from the test pool due to DIF or other concerns with item performance, developers should treat the items' inclusion in scored exams as a scoring error, and rescore impacted tests. They should be transparent with licensing boards about how often this occurs, and for what reasons.

Conclusion

Professions owe it to new professionals and the communities they serve to ensure that barriers to licensure are valid and useful. The bar for such conclusions should be particularly high when evidence exists of disparate exam outcomes on the basis of examinee demographics. Limiting the public's ability to access a diverse workforce of licensed mental health professionals, on the basis of an exam, should *only* be acceptable if that exam has been convincingly demonstrated to be valid, free from bias, and consistent with industry standards.

Exam developers have shown little interest to date in critical examination of their own tests. Instead, they hypothesize that outcome disparities result primarily or even exclusively from various "upstream factors," such as graduate education, clinical supervision, and the historical impacts of oppression and marginalization.[43,44] These factors may very well contribute to disparities in exam outcomes, and are worthy of investigation. However, they do not absolve developers of their responsibility to investigate *the exams themselves* as a potential source of disparity, particularly given the available data.

Existing literature posits a number of other potential explanations for exam outcome disparities that are centered in the testing process itself. Exam developers may not recognize when test items reflect the biases of item writers; the National Education Association cites an example from school-based testing where students taking a science test were asked about decomposition of grass clippings after mowing the lawn. Many examinees, who may have understood the scientific process of decomposition, did not understand the concept of grass clippings.[60] They may have lived in apartment buildings with no lawns, in rural desert environments with no lawns, or in other places where a landscaper would mow the grass. Exam developers in mental health care may similarly write questions attuned to the practices and populations served by the item writers, failing to recognize where that context differs from the practices and populations served by examinees. While exam developers sometimes use committees to identify potentially biased content, such committees have been shown to be ineffective.[61]

Perhaps most compelling is the argument that the *overall structure* of these exams (primarily four-option multiple choice, with a single correct answer, often based on a very brief case vignette) is an inappropriate vehicle for performing and assessing "knowledge" relevant to professional clinical practice. This structure may significantly advantage test-taking skill and English language comprehension over actual knowledge, safety, or competence in the profession being assessed. In real practice, clinicians can ask follow-up questions, consult with colleagues, and utilize external resources, none of which are possible in a clinical exam.[17]

The lack of competing measures of mental health clinician competency, outside of the deeply problematic exams in current use, places licensing boards and policymakers in a complicated position. The AERA standards provide significant leeway for deviation from individual standards in light of the overall context of measurement[3] – and a lack of competing measures is surely part of that context.[62]

However, the clinical exams in current use for mental health licensure appear inconsistent with foundational concerns of the *Standards*. Even applying a reasonably low bar for adherence, these exams do not appear to reach a reasonable degree of accord with the letter or spirit of the *Standards*. Instead, these exams appear to suffer from major deficits related to construct clarity, validity, fairness, and statistical analysis. The selected areas highlighted here appear to be more than sufficient to support immediate action by licensing boards and other policymakers, to ensure that professional licensure in mental health care is fair and equitable.

As others have also suggested,[1,2,34] social work boards should suspend their utilization of the ASWB Clinical Exam. APA and its sister organizations, AAMFT and ACA, should follow the lead of NASW,[1] and withdraw support for clinical exams in their respective professions as well. As APA itself noted, the field of psychology has not done enough to put "an end to the misuse of testing and assessment practices (including standardized assessments) and interventions in education and the workplace developed by psychologists and others that perpetuated racial inequality" (para. 23).[42] Withdrawing support for the EPPP would be a worthwhile way for APA to show its commitment to correcting its own historical wrongdoing in this area, and advancing equity in the psychology profession.

Licensing boards across the mental health professions should suspend use of their clinical exams, in light of the exams' overlapping development processes, similar outcome disparity data, and appearance of similar weaknesses with foundational adherence to industry standards. Boards for all of these professions should work with appropriate policymakers to establish alternative pathways to licensure that do not require deeply flawed assessment instruments that limit the public's access to qualified mental health professionals.

At a minimum, boards must demand new and convincing evidence of construct clarity, validity, fairness, and proper statistical analysis before accepting developers' assertions that clinical exams in mental health care are in line with industry standards. A wealth of available evidence suggests otherwise.

Disclosure

The author provides exam preparation for one of the license exams discussed in this manuscript. To the extent that this manuscript calls for clinical exams to be suspended, the author is arguing here against his personal economic interest.

Financial Support

The author attests that there is no outside financial support for this work.

Acknowledgements

The author gratefully acknowledges the contributions of Matthew DeCarlo and Jen Hirsch, each of whom provided valuable feedback on an earlier version of this manuscript, and the #StopASWB group, which provided reference materials utilized in development of this manuscript.

References

1. National Association of Social Workers (2023, February 3). NASW opposes Association of Social Work Boards (ASWB) exams. Retrieved June 29, 2023 from https://www.socialworkers.org/News/News-Releases/ID/2611/NASW-Opposes-Association-of-Social-Work-Boards-ASWBExams
2. Victor BG, McNally K, Qi Z, et al. Construct-irrelevant variance on the ASWB Clinical Social Work Licensing Exam: A replication of prior validity concerns. Research on Social Work Practice [advance online publication]. 2023. https://doi.org/10.1177/10497315231188305
3. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. APA.
4. Association of State and Provincial Psychology Boards (2022 October 28). Information Regarding the EPPP (Part 2-Skills) Required 1/1/2026. Retrieved June 23, 2023 from https://www.asppb.net/news/621595/Information-regarding-the-EPPP-Part-2-Skills-Required-112026.htm
5. National Board for Certified Counselors (2023). NBCC Examinations. Retrieved June 23, 2023, from https://nbcc.org/exams
6. Association of Social Work Boards (2022a). 2022 ASWB Exam Pass Rate Analysis. Retrieved June 23, 2023 from https://www.aswb.org/wp-content/uploads/2022/07/2022-ASWB-Exam-Pass-Rate-Analysis.pdf
7. Association of Marital and Family Therapy Regulatory Boards (2023). Examination in Marital and Family Therapy. Retrieved June 23, 2023 from https://amftrb.org/exam-info
8. Board of Behavioral Sciences [California] (2023). Handbook for Future LMFTs: Licensed Marriage and Family Therapists. Retrieved June 23, 2023 from https://www.bbs.ca.gov/pdf/publications/lmft_handbook.pdf
9. Association of Social Work Boards (2022b). ASWB Examination Guidebook. Retrieved June 23, 2023 from https://www.aswb.org/wp-content/uploads/2023/01/ASWB-Examination-Guidebook.pdf
10. Bowman N, Ameen E. Exploring differences in pass rates on the Examination for Professional Practice in Psychology. Communique. 2018. https://www.apa.org/pi/oema/resources/communique/2018/06/pass-rates
11. Macura Z, Ameen EJ. Factors associated with passing the EPPP on first attempt: Findings from a mixed methods survey of recent test takers. Training and Education in Professional Psychology. 2021; 15(1): 23-32. https://doi.org/10.1037/tep0000316
12. Sharpless BA. Are demographic variables associated with performance on the Examination for Professional Practice in Psychology (EPPP)? The Journal of Psychology: Interdisciplinary and Applied. 2019; 153(2): 161-172. https://doi.org/10.1080/00223980.2018.1504739
13. Sharpless BA. Pass rates on the Examination for Professional Practice in Psychology (EPPP) according to demographic variables: A partial replication. Training and Education in Professional Psychology. 2021; 15(1): 18-22. https://doi.org/10.1037/tep0000301
14. Sharpless BA, Barber JP. Predictors of program performance on the Examination for Professional Practice in Psychology (EPPP). Professional Psychology: Research and Practice. 2013; 44: 208-217. http://dx.doi.org/10.1037/a0031689
15. Callahan JL, Bell DJ, Davila J, et al. The enhanced examination for professional practice in psychology: A viable approach? American Psychologist. 2020; 75(1): 52-65. https://doi.org/10.1037/amp0000586
16. Lyness K, Gehart D. (in draft). Evidence of bias in clinical exams for marriage and family therapists.
17. Caldwell BE, Rousmaniere T. (2022). Clinical Licensing Exams in Mental Health Care [white paper; October 2022 version]. Retrieved July 23, 2023 from https://www.psychotherapynotes.com/wp-content/uploads/2022/10/Clinical-Licensing-Exams-in-Mental-Health-Care-October-2022.pdf
18. Board of Psychology [California] (2018). Examination for Professional Practice in Psychology (EPPP) Part 2 Update. Retrieved July 29, 2023 from https://www.psychology.ca.gov/applicants/eppp_exam_info_2.shtml
19. Dardard S. (2018). ASPPB Uses Carrot & Stick for New EPPP2. The Psychology Times. Retrieved July 29, 2023 from https://thepsychologytimes.com/2018/12/18/asppb-uses-carrot-stick-for-new-eppp2/
20. Caldwell B. (2023). The controversy over racial bias in mental health clinical exams. Pollen [online magazine]. Retrieved April 28, 2023 at https://www.simplepractice.com/blog/racial-bias-mental-health-clinical-exams/
21. House Bill 2365 (Illinois) (2023). Retrieved July 10, 2023 from https://ilga.gov/legislation/billstatus.sp?DocNum=2365&GAID=17&GA=103&DocTypeID=HB&LegID=147441&SessionID=112
22. Association of State and Provincial Psychology Boards (2023). Frequently Asked Questions about the EPPP. Retrieved July 10, 2023 from https://cdn.ymaws.com/www.asppb.net/resource/resmgr/eppp_2/faq_revisedeppp_feb2023.pdf
23. Turner MD, Hunsley J, Rodolfa ER. Appropriate validation standards for licensure examinations: Comment on Callahan et al. (2020). American Psychologist. 2021; 76(1): 165-166. https://doi.org/10.1037/amp0000695
24. Mir TH. Adherence versus compliance. HCA Healthcare Journal of Medicine. 2023; 4(2): 219-220. https://doi.org/10.36518/2689-0216.1513
25. California Association of Marriage and Family Therapists (2019). CAMFT Code of Ethics. Retrieved July 23, 2023 from https://www.camft.org/Membership/About-Us/Association-Documents/Code-of-Ethics

26. American Association for Marriage and Family Therapy (2015). AAMFT Code of Ethics. AAMFT. Retrieved July 23, 2023 from https://www.aamft.org/Legal_Ethics/Code_of_Ethics.aspx

27. PearsonVUE (2021). California Board of Behavioral Sciences Examination Candidate Handbook, January 2021. Retrieved July 23, 2023 from https://home.pearsonvue.com/getattachment/8cb12fa9-e6da-4c06-8a83-361602243036/California%20Behavioral%20Sciences%20Candidate%20Handbook.aspx

28. S Sodergren [Executive Officer, California Board of Behavioral Sciences], private communication via email, March 2023.

29. National Board for Certified Counselors (2022). Candidate Handbook for National Certified Counselor Certification - National Clinical Mental Health Counseling Examination. Retrieved June 23, 2023 from https://www.nbcc.org/assets/exam/handbooks/ncmhce_applicant_handbook_for_national_certification.pdf

30. American Counseling Association (2014). 2014 ACA Code of Ethics. ACA. Retrieved July 23, 2023 from https://www.counseling.org/resources/aca-code-of-ethics.pdf

31. American Mental Health Counseling Association (2020). AMHCA Code of Ethics. AMHCA. Retrieved July 23, 2023 from https://www.amhca.org/viewdocument/2020-amhca-code-of-ethics

32. Haley J. (1987). Problem-Solving Therapy (2nd ed.). Jossey-Bass.

33. Selvini-Palazzoli M, Viaro M. The anorectic process in the family: A six-stage model as a guide for individual therapy. Family Process. 1988; 27(2): 129-148. https://doi.org/10.1111/j.1545-5300.1988.00129.x

34. Victor BG, Kubiak S, Angell B, et al. Time to move beyond the ASWB licensing exams: Can generative artificial intelligence offer a way forward for social work? Research on Social Work Practice. 2023; 33(5): 511-517. https://doi.org/10.1177/10497315231166125

35. Chittooran MM, Miles DD. (2001). Test-Taking Skills for Multiple-Choice Formats: Implications for School Psychologists. Paper presented at the Annual Conference of the National Association of School Psychologists (Washington, DC, April 17-21, 2001).

36. Dolly JP, Williams KS. Using test-taking strategies to maximize multiple-choice test scores. Educational and Psychological Measurement. 1986; 46(3): 619-625. https://doi.org/10.1177/0013164486463014

37. Heist BS, Gonzalo JD, Durning S, et al. Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: A mixed methods study. Journal of Graduate Medical Education. 2014; 6(4): 709-714. https://doi.org/10.4300/JGME-D-14-00176.1

38. Albright DL, Thyer BA. A test of the validity of the LCSW examination: Quis custodiet ipsos custodes? Social Work Research. 2010; 34(4), 229-234. https://doi.org/10.1093/swr/34.4.229

39. Association of Social Work Boards (2023). Association of Social Work Boards pilots free support program for test-takers [news post]. Retrieved June 23, 2023 from https://www.aswb.org/association-of-social-work-boards-pilots-free-support-program-for-test-takers/

40. Association of Social Work Boards (2023). The testing experience: How ASWB's examination practices support access and equity [webinar].

41. Turner MD, Rodolfa ER. (2019). Letter to the California Board of Psychology. Retrieved July 29, 2023 from https://www.psychology.ca.gov/applicants/bd_lttr_to_asppb_2019_01_29.pdf

42. American Psychological Association (2021). Apology to People of Color for APA's Role in Promoting, Perpetuating, and Failing to Challenge Racism, Racial Discrimination, and Human Hierarchy in U.S. (Resolution adopted by the APA Council of Representatives on October 29, 2021). Retrieved July 30, 2023 from https://www.apa.org/about/policy/racism-apology

43. Hardy-Chandler S. (2022). Beyond Data: A Call to Action. Retrieved July 10, 2023 from https://www.aswb.org/beyond-data-a-call-to-action/

44. Marson S. Editorial: Does racial bias exist in the ASWB social work exams? International Journal of Social Work Values and Ethics. 2022; 19(2): 8-20. https://doi.org/10.55521/10-019-203 (fulltext https://jswve.org/wp-content/uploads/2022/08/10-019-203-IJSWVE-2022.pdf)

45. O'Connor ST. (2010). Why don't they get licensed? Investigating success in the California clinical social worker and marriage and family therapist licensing process (unpublished master's thesis, California State University, Sacramento). Available online at https://scholars.csus.edu/esploro/outputs/graduate/Why-dont-they-get-licensed/99257831127301671

46. Association of Social Work Boards (no date). DIF vs. DTF. Retrieved July 10, 2023 from https://www.aswb.org/exam/measuring-social-work-competence/measuring-competence-fairly/dif-vs-dtf/

47. Owens S. How does ASWB guard against bias on the social work licensing exams? The New Social Worker. 2021; 28(3): 21-24. https://www.socialworker.com/feature-articles/education--credentials/aswb-guard-against-bias-social-work-licensing-exams/

48. Santoro H. New psychology licensing exam expands. Monitor on Psychology. 2023; 54(3): 24-25. Retrieved June 29, 2023 from https://www.apa.org/monitor/2023/04/psychology-licensing-exam-expands

49. National Board for Certified Counselors (2019). Assessment Development: An Introduction to NBCC & CCE Assessment Processes. NBCC.

50. Montez T. (2022). Comments on licensing exams presented at the November 4, 2022 meeting of the California Board of Behavioral Sciences. Available online at https://www.youtube.com/watch?v=99iEOFDmJmc

51. California Government Code, section 8310.

52. Board of Behavioral Sciences [California] (2021). California State Board of Behavioral Sciences Bill Analysis: AB1236. Retrieved June 29, 2023 from https://www.bbs.ca.gov/pdf/agen_notice/2021/20210416_pa_item_xiv.pdf

53. Board of Behavioral Sciences [California] (2022). Minutes of the November 2022 Board Meeting. Retrieved July 10, 2023 from https://www.bbs.ca.gov/pdf/board_minutes/2022/20221103.pdf

54. Chalmers RP, Counsell A, Flora DB. It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. Educational and Psychological Measurement. 2016; 76(1): 114-140. https://doi.org/10.1177/0013164415584576

55. Halamová J, Kanovský M, Gilbert P, et al. Multiple group IRT measurement invariance analysis of the Forms of Self-Criticising/Attacking and Self-Reassuring Scale in thirteen international samples. Journal of Rational-Emotive & Cognitive-Behavior Therapy. 2019; 37: 411-444. https://doi.org/10.1007/s10942-019-00319-1

56. Hardy-Chandler S. (2023). Presentation to the Maryland Board of Social Work, January 13, 2023. Available online at https://youtu.be/WRk1wbNFm7Y

57. Harless L. (2020). Putting social work values to the test — ASWB's commitment to Diversity, Equity, and Inclusion. Social Work Today, 20(3). https://www.socialworktoday.com/archive/MJ20p24.shtml

58. Association of Social Work Boards (2023). The art and science of exam development: Exploring best practices for building reliable, valid, and fair exams [online webinar; relevant section begins at 52:16]. Archive video available at https://vimeo.com/803202910/7303b63b47

59. DeCarlo MP. (2023). ASWB hides data from state social work boards that could license thousands of excluded social workers [blog post]. Retrieved June 29, 2023 from https://opensocialwork.org/2023/04/29/aswb-hides-data-from-state-social-work-boards-that-could-license-thousands-of-excluded-social-workers/

60. Long C. (2023). Standardized testing is still failing students. NEA Today. Retrieved July 10, 2023 from https://nea.org/nea-today/all-news-articles/standardized-testing-still-failing-students

61. Engelhard G Jr., Hansche L, Rutledge KE. (1989). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, March 27-31, 1989.

62. Kansas News Service. (2022). A social work licensing exam that people of color fail more often is under scrutiny in Kansas. Kansas Public Radio. Retrieved July 10, 2023 from https://kansaspublicradio.org/local-news/2022-11-08/a-social-work-licensing-exam-that-people-of-color-fail-more-often-is-under-scrutiny-in-kansas