A Perspective on Racial Bias within the Association of Social Work Boards (ASWB) Exams[1]

DOI: 10.55521/10-022-206

Stephen M. Marson, Ph.D., ACSW
smarson3@outlook.com
University of North Carolina at Pembroke
https://orcid.org/0000-0003-4948-5777

Guo Wei, Ph.D.
guo.wei@uncp.edu
University of North Carolina at Pembroke
https://orcid.org/0000-0001-9988-0498

Craig Taylor, BS
ctalkobt@ctalkobt.net
Computer Consultant
https://orcid.org/0009-0006-1535-6319

International Journal of Social Work Values and Ethics • Volume 22 (2025), Issue 2,
Copyright 2025 by IFSW

This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of IFSW.

Abstract

We present three issues for assessing minimum competency for clinical social workers as related to possible racial bias: 1) the inherent trade-offs between outcome alternatives; 2) the tension between subjective and objective measures of competence; and 3) the common misinterpretation of p-values as magical proofs. These issues are synthesized by examining the central question, "Which alternative is best?" Our analysis demonstrates that resolving this question ultimately

[1] Grace Gates, LCSW provided a critical review for this report.

depends on a value-based judgment, framed as, “How much are test-takers willing to pay for a given outcome?” We conclude by exploring advancements in artificial intelligence (AI). AI can provide a framework for navigating these complex trade-offs and informing critical decisions.

Key Terms:
ASWB, ACSW, Racial Bias, Assessing Competence, Finances

Introduction

A substantial body of literature examines potential racial bias in minimum competency examinations administered by the Association of Social Work Boards (ASWB). Among these articles include Albright and Thyer (2010), Curran and Joo (2025), DeCarlo and Bean (2024), DeCarlo (2022), Kim (2023), Kim and Joo (2025), Marson (2022), Marson, Kersting and DeAngelis (2011), Marson and DeAngelis (2010), Marson, DeAngelis and Mittal (2010), Rigaud (2024), Torres, Maguire and Kogan (2024), Victor, Kubiak, Angell and Perron (2023) and Zajicek-Farber (2024). Our contribution is to frame this debate through the inescapable statistical trade-off between Type I and Type II errors in competency assessment and to analyze the implications of this trade-off for both objective (e.g., ASWB) and subjective (e.g., ACSW) pathways.

Outcomes for Assessing Competence in Clinical Social Work

	PASS ASSESSMENT	FAIL ASSESSMENT
COMPETENT PRACTITIONER	A Knows the material (pass)	B Knows the material (fails)
DEFICIENT PRACTITIONER	C Doesn’t know (pass)	D Doesn’t know (fails)

Figure 1: Outcomes for Assessing Clinical Social Work Competence

Regardless of the method for assessing minimum competence for clinical social work, the decision model yields four possible outcomes, as illustrated in Figure 1. The horizontal axis represents the assessment result (pass or fail), while the vertical axis distinguishes between two types of candidates. The “competent practitioner” represents a social worker with the necessary knowledge, skills, and professional ethics. The “deficient practitioner” represents a social worker who lacks the required competencies or adherence to ethical standards.

The outcome in Cell A represents the correct classification of competent social workers deemed to have minimum competency. Cell B represents a *false negative* error, where social workers with the required knowledge and professional ethics are incorrectly assessed as lacking minimum competency. This error often correlates with below-average test-taking skills and can propel the social worker into an emotional crisis (Groshong & Roberson, 2023; Graybow, 2015; Lewis, 2023). Furthermore, this outcome deprives mental health consumers of access to qualified practitioners.

Cell C represents a *false positive* error, where *in*competent social workers are incorrectly assessed as possessing minimum competency. These individuals are often highly skilled test-takers, which allows them to pass the assessment despite their deficiencies. This failure in measurement poses a significant risk, as it permits people who should not be psychotherapists to practice, potentially causing emotional damage to clients. Lastly, Cell D represents a *true negativ*e, correctly identifying social workers who lack the basic competencies. This outcome demonstrates the assessment's validity in screening out incompetent candidates.

There are several key points that require further discussion. The first concerns the type of error inherent to any assessment method. Cell B represents a *false negative*, a serious error where competent professionals are denied the opportunity to practice, often due to factors like poor test-taking skills. On the other hand, Cell C represents a *false positive*, where incompetent practitioners are permitted to practice because of strong test-taking ability. This fundamental tradeoff between these two errors is central to the assessment debate.

Statistically, there is a seesaw relationship. As cell B shrinks, cell C volume increases. As cell C shrinks, the volume in cell B increases (as illustrated in Figure 2):

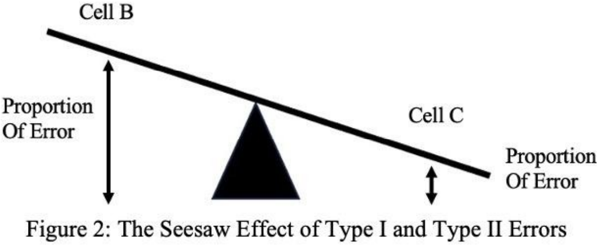

Figure 2: The Seesaw Effect of Type I and Type II Errors

Thus, the sad fact is as we vigorously control the elimination of error in one cell, the error in the opposite cell increases. We are in a position that requires us to decide what type of error is most important to control.

Controlling for error B, would increase practitioners who were not competent as psychotherapists. Controlling for error C, would prohibit good practitioners from practicing psychotherapy.

Regardless of what methodology we use to assess competence; errors will be made. This *cannot* be avoided. The question must become which type of error is the least problematic: error B or error C. Thus, the fundamental dilemma is that as we vigorously control one type of error, the other error increases. We are therefore faced with a critical decision: which error is less acceptable? Prioritizing the reduction of *false negatives* (Error B) would inadvertently allow more incompetent practitioners to pass. Conversely, prioritizing the reduction of *false positives* (Error C) would wrongly exclude more competent practitioners from the profession. The reality is that any methodology for assessing competence will produce errors; this inherent fallibility is unavoidable. The essential question, therefore, is not if errors will occur, but which type poses a greater risk to the public and the profession: falsely excluding a competent practitioner or falsely licensing an incompetent one?

The central challenge we must confront is racial bias, which is statistically manifested in the *false negatives* within Cell B. To combat this bias, we must deliberately reduce these errors, but this intervention has a direct, known consequence: an increase in the *false positives* within Cell C. This trade-off means that a testing protocol designed for greater equity would inevitably permit more incompetent psychotherapists to be licensed. This framing leads to the fundamental policy question: *Which outcome is worse?* Should we prioritize equity by accepting more incompetent practitioners into the profession to ensure we license all competent ones, including those with poor test-wiseness? Or should we prioritize consumer protection by denying licenses to some competent, poor test-takers in order to more rigorously exclude those who are incompetent?

ACSW Versus ASWB (Subjective Versus Objective)

The inescapable trade-off of the “seesaw effect” presents a core dilemma: which error is more acceptable? Some states are now pursuing the Academy of Certified Social Workers (ACSW) model as an alternative for assessing minimum competence. The ACSW relies on a portfolio-based evaluation conducted by professionals familiar with the candidate's work. However, this subjective approach raises several critical concerns:

- The candidate's selection of evaluators introduces a significant conflict of interest, as individuals are unlikely to choose assessors who might provide a negative review.
- The ACSW protocol lacks systematic, statistical guardrails to inhibit against racial bias, unlike the ASWB exam, which employs objective methods for this purpose. The ACSW model is therefore vulnerable to subjective discrimination.
- As modeled in Figure 1, the ACSW's subjectivity virtually eliminates Cell B errors but does so by significantly increasing Cell C errors, allowing more incompetent practitioners to be certified.

Ultimately, the choice between objective and subjective assessment methods is not about finding a perfect system but determining which is less problematic. The driving force behind establishing objective methodologies in the 1970s was the mandate to “protect the public,” a mantra that dominated legislative testimony and committee meetings. This foundational priority forces a difficult evaluation of whether the pursuit of equity through subjective measures unjustly compromises public safety.

Let us delve into a concern related to the employment of statistics for assessing competence.

An Issue with Statistical Analysis

In the context of testing for racial bias, p-values serve as a diagnostic tool to assess whether observed disparities are statistically significant or likely due to random chance. For example, if a study examines whether an algorithm favors one racial group over another, a small p-value would indicate that the observed disparity is unlikely to have occurred under the assumption of no bias (the null hypothesis). However, a crucial limitation is that p-values alone cannot quantify the magnitude or real-world impact of such bias. Moreover, sole reliance on p-values may overlook systemic inequities embedded in the data or the design of the test itself. Thus, while p-values can signal potential issues, they must be interpreted alongside the effect of sizes, contextual understanding, and ethical considerations to draw valid conclusions.

Therefore, in the pursuit of fairness and accountability, p-values must be viewed as merely one piece of the evidentiary puzzle. They highlight inconsistency with the null hypothesis but do not measure the probability of a hypothesis being true or false. To effectively mitigate racial bias, we must integrate statistical significance with qualitative insights, transparency in methodology, and a commitment to addressing structural inequities. This comprehensive approach is essential to promote more equitable systems and move beyond the pitfalls of over relying on simplistic statistical thresholds.

Recommendations

The challenge highlighted in our analysis is not solvable by statistics alone, requiring a multi-faceted approach that may include:

- *Enhanced Exam Design:* Developing exams with demonstrably less bias and higher predictive validity for clinical skill and reliability.
- *Multiple Measures of Competence:* Incorporating portfolios, supervised evaluations with robust anti-bias safeguards and client outcomes alongside alternatives measurement models.
- *Transparent Acknowledgment of Trade-offs:* Openly discussing the inherent Type I/Type II error trade-off and the societal values that guide our choice of which error to prioritize.

While these strategies could enhance assessment quality, the primary obstacle is cost. In the 1990's, developing a competency exam was estimated to be $900 per item with 150 items on each exam (Marson, 2005). Today, the cost would be significantly higher. A potentially more cost-manageable pathway for integrating multiple measures lies in artificial intelligence (AI). For instance, AI could be employed as an objective tool to efficiently analyze test-takers' social histories, portfolios, and supervision protocols.

The use of AI carries its own risk of amplifying or introducing biases, potentially altering how discrimination manifests in the evaluation process. Therefore, rigorous analysis *and* continuous validation are essential throughout the AI's design and deployment. To mitigate the risks of both false positives and false negatives, a hybrid model is recommended, where a human reviewer evaluates all AI-generated pass/fail recommendations.

Artificial intelligence offers promising avenues for enhancing the evaluation processes within ASWB exams. AI could provide a more holistic assessment by objectively analyzing test-takers' social histories, portfolios, and supervision protocols. It could also enable the cost-effective evaluation of complex clinical skills, such as interviewing and diagnostics, and help identify and mitigate biases in exam questions and performance data. While initial development costs are significant, these tools may become more accessible over time, potentially improving equity in competency assessment.

However, most AI-driven proposals, while meritorious, face limitations that challenge their viability as comprehensive solutions. The substantial financial investment and practical implementation hurdles are particularly prohibitive within a realistic 5-10 year timeframe. Common obstacles include:

- Prohibitive infrastructure overhauls with protracted development cycles.
- Dependence on emerging technologies that lack the required maturity and cost-effectiveness.
- Questions of scalability and the ability to meet future demands.

A pragmatic approach is needed to identify solutions that are not only effective but also economically feasible and readily deployable. The following analysis examines some prominent proposals through this critical lens.

- **Multiple Measures (Human-Evaluated):** Incorporating human-evaluated portfolios, supervised hours, and client outcomes promises a more holistic view but faces prohibitive cost and scalability barriers for a national exam. Costs scale linearly with candidate volume, creating exorbitant expenses. Securing the expertise for fair, consistent, and bias-mitigated evaluation of thousands of candidates would be financially unsustainable, especially with the need for multiple evaluators to ensure reliability. Critically, as the ACSW model demonstrates, human evaluation—even with safeguards—can introduce subjective bias, potentially increasing Cell C errors (*false positives*) and failing to equitably resolve racial disparities.
- **AI for Dynamic Exam Customization:** The development of an AI for adaptive testing requires substantial, ongoing investment in data, algorithm development, and validation. A primary risk is that the adaptive logic could institutionalize new biases, for instance, by penalizing particular cultural communication styles or reasoning patterns. Maintaining the system's validity would demand continuous auditing to prevent

"drift," where the AI's definition of competence diverges from professional standards.

- **AI for Continuous Learning and Remediation:** This approach faces significant integration challenges with existing professional frameworks. It would necessitate a robust and potentially intrusive data infrastructure, raising major ethical concerns regarding practitioner privacy and surveillance. Furthermore, the AI's recommendations could perpetuate existing inequities if trained on biased data, stifling innovative practices. The cost of developing and maintaining personalized, AI-driven content remains prohibitively high.

The most viable and impactful basis for enhancing competence assessment and reducing racial bias within a 5-10 year timeframe lies in the AI-powered evaluation of clinical vignettes. This approach strikes a superior balance of enhanced validity, scalability, and cost-effectiveness compared to other methods:

1. **Deeper Assessment of Clinical Judgment:** Vignettes require test-takers to demonstrate critical thinking and ethical reasoning in realistic scenarios, transcending the limitations of factual recall. AI evaluation of these responses can objectively analyze the nuance, coherence, and appropriateness of a candidate's clinical judgment, directly targeting a core weakness of multiple-choice exams.
2. **Superior Scalability and Long-Term Efficiency:** Although initial AI development requires significant investment, the long-term marginal cost per test-taker is low. A deployed system can instantly and consistently process thousands of responses, making it uniquely scalable for a national exam and directly mitigating the cost burden on candidates that plagues human-evaluated alternatives. This pathway presents the most pragmatic and defensible balance between the competing demands of validity, equity, and financial feasibility.
3. **Systematic Mitigation of Bias:** A core advantage of this approach is the potential for AI to be systematically calibrated for fairness. By training models on expert-curated, diverse datasets and evaluating responses against standardized, objective criteria, AI can deliver more consistent assessments that are less susceptible to the subjective biases of human reviewers. Crucially, unlike purely subjective models like the ACSW, the

AI's decision-making process can be continuously audited using statistical methods to detect and correct for emergent bias, creating a feedback loop for ongoing improvement.

4. **Mitigation of Test-Taking Disparities (Reducing Cell B Errors):** The vignette format inherently benefits competent practitioners who are poor test-takers by allowing them to demonstrate clinical reasoning in their own words, free from the constraints of multiple-choice questions. This directly targets the root of many Cell B (*false negative*) errors, helping to ensure that clinical competence, not test-wiseness, is the primary determinant of success.
5. **Technical Feasibility and Timeliness:** The existing trajectory of Natural Language Processing (NLP) confirms that robust AI evaluation of clinical vignettes is achievable within a 5-10 year horizon. The foundational technology is rapidly maturing beyond theoretical potential into practical application, making the assessment of complex written responses a realistic and imminent goal rather than a distant possibility.
6. **A Structured Defense Against Bias:** Critically, an AI-driven system offers a structured framework for bias mitigation that is inherently superior to purely subjective human evaluation. Bias can be proactively identified through transparent algorithms, continuously monitored via statistical auditing, and corrected through model retraining. A mandatory human-in-the-loop protocol, where a reviewer assesses all AI-generated pass/fail recommendations along with the AI's reasoning, serves as a final safeguard against both false positives and false negatives, ensuring accountability.

The adoption of AI-evaluated vignettes represents a strategically viable path forward. It directly tackles the profession's core challenges of racial bias and assessment validity, while avoiding the prohibitive costs and scalability limits of alternative solutions. This approach pragmatically leverages technological advancement to fulfill the dual mandate of protecting the public and ensuring equitable access to the profession.

Two Final Overarching Thoughts

One of the overarching themes within this paper can be summarized in one word *balance*. Every strategy for creating fair and equitable gatekeeping exams involves a series of tradeoffs. A central tenet of fairness is the realization that all strategies can have limitation and unintended consequences. While implementing guardrails to inhibit racism for all exams is *absolutely necessary*, the alternative is equally unacceptable. For example, supervisory evaluations for licensure face the same verification problems similar to Freudian theory: their validity, like that of Freudian concepts, has not been adequately established. In fact, current research suggests that reference letters cannot be trusted (Ibrahim, 2024; Reed, 2021). This inability to systematically verify their validity creates fertile soil for racism to persist.

Secondly, bivariate findings within social science research are *not* held in high esteem due to their significant limitations. Suggesting that the ASWB exams are racist, is an example of a bivariate analysis. In the current literature, alternative explanations that can lead to racial inequality *are not* controlled in any of the current research. However, a groundwork exists for establishing alternative explanations for the racial differences in outcome scores. For example, the passage rate for the clinical exam ranges from 37% to 100% among MSW graduate programs. This data raises questions about the wide variation in outcomes. In addition, this data suggests that educational competencies and standards vary at a rate that is *unacceptable*. Coupled with the well-established presence of institutional racism in social work education (Marson, 2022), a critical question emerges: “Do unequal academic standards exist within MSW academic programs across racial groups?” These alternative explanations have not yet been addressed but must be.

Controlling for alternative explanations can be best implemented by a regression model (descriptive and variance decomposition). The regression model can illuminate the relationship between race and ASWB scores, *after accounting for* the other variables (like, GPA, GRE, age, school attended, year of graduation, etc.) in the model. Regression controls for other variables by holding them statistically constant. For describing complex patterns in data, regression analysis is both valid and powerful. The goal of such proposed research *is not* prediction but rather understanding how variables co-vary. In the end regression analysis will provide greater clarification on the validity of multiple choice items.

Email comments to: journal@ifsw.org

References

Albright, D. L., & Thyer, B. A. (2010). A test of the validity of the LCSW examination: Quis custodiet ipsos custodes? *Social Work Research, 34(4)*, 229-234.

Curran, L., Kim, J. J., & Joo, M. M. (2025). Promoting competence: Multistage conceptualization, empirical evidence, and current controversies on licensure to inform a future research agenda. *Journal of Social Work Education*, 1–16. https://doi.org/10.1080/10437797.2024.2420091

DeCarlo, M., & Bean, G. (2024). Assessing measurement invariance in ASWB exams: Regulatory research proposal to advance equity. *Journal of Evidence-Based Social Work*, 21(2), 214-235. https://doi.org/10.1080/26408066.2024.2308814

DeCarlo, M. P. (2022). Racial bias and ASWB exams: A failure of data equity. *Research on Social Work Practice, 32(3)*, 255-258. https://doi.org/10.1177/10497315211055986

Graybow, S. (2015). *Understanding failure: Social workers reflect on their licensing examination experience* (Order No. 3724587). Doctoral Dissertation, City University of New York.

Groshong, L. & Roberson, K. (2023). Editorial: The national clinical social work examination – ASWB. International *Journal of Social Work Values and Ethics, 20(1)*, 29-31. https://doi.org/10.55521/10-020-103

Ibrahim, H., Mohamad, M. K., Elhag, S. A., Al-Habbal, K., Harhara, T., Shehadeh, M., Alsoud, L. O., & Abdel-Razig, S. (2024). Components of effective letters of recommendation: A cross-sectional survey of academic faculty. *PLoS ONE, 19(1)*, e0296637. https://doi.org/10.1371/journal.pone.0296637

Kim, J. J., & Joo, M. M. (2025). Effect of race/ethnicity on the 2018–2022 Social work clinical license exam outcomes. *Research on Social Work Practice*. https://doi.org/10.1177/10497315251333373

Kim, J. J. (2023). The art and science of social work regulations: How values and data should guide regulatory practices. *Clinical Social Work Journal, 52(3)*, 287–299. https://doi.org/10.1007/s10615-023-00910-1

Lewis, A. (2023). *Impact of trauma and trauma-informed care*. Chapter 12: Preparing for the Masters ASWB Exam. https://umsystem.pressbooks.pub/aswbprep/chapter/trauma-informed-care-and-crisis-intervention/

Marson, S. (2022). Does racial bias exist in the ASWB social work exams? *International Journal of Social Work Values and Ethics, 19(2)*, 8-20. https://doi.org/10.55521/10-019-203

Marson, S. M. (2005). Social work. in Kemp-Leonard, K (ed). *The Encyclopedia of Social Measurement*. Academic Press.

Marson, S. M., Kersting, R, and DeAngelis, D. (2011). What do I do when I fail the social work exam? *The New Social Worker, 18(2)*, 16-18.

Marson, S. & DeAngelis D. (2010). 10 questions about the Association of Social Work Boards (ASWB) exams. *The New Social Worker* 17(3), 24-25.

Marson, S. M., DeAngelis, D. & Mittal, N. (2010). The Association of Social Work Boards' licensure examinations: Developing reliability and validity. *Research on Social Work Practice, 20(1)*, 87-99.

Reed, B. N. (2021). Rethinking letters of reference: Narrowing the gap between evidence and practice. *American Journal of Health-System Pharmacy, 78(23)*, 2175–2179.

Rigaud, J. (2024). Ethical challenges in social work licensing examinations: A call for integrity and strategies for success. *Social Work, 69(4)*, 395-402. https://doi.org/10.1093/sw/swae037

Torres, M. E., Maguire, S., & Kogan, J. (2024). "I was told to think like a middle-aged white woman": A survey on identity and the association of social work boards exam. *Social Work, 69(2)*, 185-196. https://doi.org/10.1093/sw/swae001

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

Zajicek-Farber, M. (2024). Social work licensing: then and now. *Clinical Social Work Journal*, 52(4), 368-381. https://doi.org/10.1007/s10615-024-00953-y