PURPOSEFULLY CREATING SPACE FOR DIFFICULT CONVERSATIONS IN MEDICAL EDUCATION ASSESSMENT:

THE INAUGURAL EQUITY IN MEASUREMENT AND ASSESSMENT CONFERENCE (EMAC)

APRIL 25–26, 2023

NBME®

PURPOSEFULLY CREATING SPACE FOR DIFFICULT CONVERSATIONS IN MEDICAL EDUCATION ASSESSMENT:

THE INAUGURAL EQUITY IN MEASUREMENT AND ASSESSMENT CONFERENCE (EMAC)

April 25–26, 2023

Co-sponsored by National Board of Medical Examiners (NBME), American Board of Medical Specialties (ABMS), American Board of Internal Medicine (ABIM), & the Association of American Medical Colleges (AAMC)

The following NBME staff co-authored this publication:

Jerusha J. Henderek, PhD, Ye Tong, PhD, Jonathan D. Rubright, PhD, Christopher A. Feddock, MD, Lucia McGeehan, PhD, Michael A. Barone, MD, Linette P. Ross, PhD, & Amanda L. Clauser, EdD

With special thanks to the reviewers, who provided input, including:

Dana Dunleavy, PhD (AAMC), Reena Karani, MD, MHPE (The Icahn School of Medicine at Mount Sinai), Greg Ogrinc, MD (ABMS), & Jennifer Randall, PhD (University of Michigan & Center for Measurement Justice)

TABLE OF CONTENTS

INTRODUCTION

Researchers have demonstrated that a diverse, equitable, and inclusive physician workforce can lead to better health outcomes and patient experience for all patients, with particular benefit for traditionally underserved populations.[1-3] However, limited opportunities exist for medical education, measurement, and diversity, equity, and inclusion (DEI) professionals to learn from each other to expand their own perspectives and to collaborate on research or practice related to this imperative. Further collaborations could foster improvement in how these varied professional groups intersect to either facilitate or inhibit students' pathways into and through medical training. Lacking forums for such collaboration, it is extremely challenging to design more equitable assessment systems to admit students who reflect the diversity of our nation, support inclusive learning and evaluation, and support talented individuals throughout their medical education journey.

Recognizing this need to bring together diverse professionals who share common goals of dismantling structural racism and building equitable systems to enhance provider diversity and practice, leaders of NBME, American Board of Medical Specialties (ABMS), American Board of Internal Medicine (ABIM), and the Association of American Medical Colleges (AAMC) planned and launched an in-person conference to create opportunities for sharing, learning, and challenging conversing among critical stakeholders involved in the physician pathway. The EMAC planning committee included representatives across the four co-sponsoring organizations; the speakers and sessions from the EMAC agenda, as well as planning committee members, are listed in Appendix A. The intent of the conference was to:

- Create a collaboration forum for professionals in the three areas: measurement, medical education, and DEI,
- Continue to highlight the importance and necessity of equity and fairness along the medical education pathway, with an emphasis on assessments within medical education, and
- Most importantly, collaborate on concrete and practical solutions to embed DEI principles throughout assessment design and delivery in medical training.

FRAMING THE CONVERSATION

The inaugural invitational Equity in Measurement and Assessment Conference (EMAC) was held in Conshohocken, Pennsylvania on April 25-26, 2023. One hundred thirty-five participants were invited to present and converse with thought leaders at the intersection of the medical education, assessment, and DEI communities. The EMAC planning committee collaborated to invite individuals that represent different backgrounds: race/ethnicity, gender, years of experience, geography, size of the organizations, medical specialty, measurement experience, etc. There were four plenary sessions organized to thoughtfully frame the conversation and create shared learning opportunities for all attendees, which will be discussed in this paper. Full recordings of the plenaries are available on NBME's website: https://www.nbme.org/research/collaborations/conference-learnings.

Opening Plenary

An opening plenary, facilitated by Ye Tong, PhD, from NBME oriented the attendees to the purpose and content of the two-day meeting. In her introduction, Dr. Ye Tong discussed the importance for such a forum and relayed the fact that we all needed to lean into the difficult conversations to move equity forward. Leaders from each sponsoring organization served as panelists for the opening plenary and included:

- NBME President and CEO Peter Katsufrakis, MD, MBA
- ABMS President and CEO Richard Hawkins, MD
- ABIM President and CEO Richard Baron, MD
- AAMC Senior Director Stephen Fitzpatrick

These leaders set the stage with three shared messages:

- The conference is not intended to be a beginning or an end, but instead part of a journey that will necessarily provoke disagreement and discomfort,
- Organizations within the House of Medicine must commit to equity practices in medical education assessment, and
- It is necessary to identify, create, share and evaluate concrete methodologies and practices to continue to make progress toward equity in measurement and assessment.

These organizational leaders also shared work each organization is conducting in the DEI space in support of changes both at the assessment (e.g., content, volunteer recruitment) and organizational levels (e.g., hiring, staff education).

- Dr. Peter Katsufrakis noted that change is slow, especially in the culture of medicine. He pointed out a few positive changes that he has witnessed throughout his career (e.g., adoption of a broader understanding of sex and race, beyond biological variants). He shared the ongoing work of the United States Medical Licensing Examination® (USMLE®) Patient Characteristics Advisory Panel in better representing the patient population in test questions and NBME's Competency-Based Assessment Unit in advocating for approaches for holistic review. NBME is also initiating a collaboration with the Center for Measurement Justice (CMJ) to support CMJ dissertation fellows focusing on equity issues in measurement and supporting National Medical Fellowship (NMF) scholarships for medical students.

Dr. Peter Katsufrakis, NBME

Dr. Richard Hawkins, ABMS

- Dr. Richard Hawkins agreed it was essential to "get this [work] right." ABMS is implementing a number of changes to their recruiting approaches to support workforce diversity and has provided educational opportunities for their board on social determinants of health, implicit bias, and anti-racism. He shared a 'must-do' list that ABMS used to guide their ten strategic imperatives; on that list was the creation of a DEI taskforce to prioritize and resource community needs and increased item writing on health disparities across boards. ABMS is considering providing structural support in diversity and equity for smaller specialty boards that may not have the resources to support a staff position themselves.
- Dr. Richard Baron described the value of DEI permeating all activities at ABIM, including thinking deeply about how assessments advance or hinder equity in the United States. He asked participants to continuously remember the *why* of DEI work and the critical nature of staying connected with the diverse communities, practice settings, and experiences of diplomates as well as patients. ABIM's current DEI efforts include increased development of health equity content and explicit inclusion of this content across all ABIM examinations to signal the value of equity to clinical practice.

- Stephen Fitzpatrick discussed the throughline connecting equity in assessments to the diversity of the physician workforce to health equity, sharing the AAMC's vision of a future in which everyone receives adequate health care. He reminded the audience that the organizations represented need to successfully navigate mistrust while working to ensure that we not only assess the necessary competencies but also appropriately signal our organizational values. He challenged the audience with the Howard Zinn quote, "You can't be neutral on a moving train," asserting that we must be cognizant of our desired destination, who is driving the train, how to reach passengers who are seemingly disinterested in the significance of this work, and what signal we are sending others by our actions.

Keynote: First, Undo Harm

In the inaugural EMAC keynote, Jennifer Randall, PhD, a professor at the University of Michigan and the founding President of CMJ, crafted and delivered a speech titled "First, Undo Harm: Disrupting Racist Logics and Their Consequences via Medical Assessment." Dr. Jennifer Randall set the context within which the current system operates—one in which historically marginalized populations not only experience racism as providers in the educational and care delivery system, but also as patients, exemplified by experiencing radically different health outcomes when seeking care from that system. Instead of relentlessly documenting those experiences and outcomes, she instead focused on opportunities for the assessment system to begin the process of undoing existing harm. She acknowledges that the assessment community cannot "undo all of the harm. Indeed, injustice is a wall that extends farther than our eyes can see; but you can begin to chip away at your part of that wall."

The current harm in medical education assessment can be viewed through theories of justice which force us to identify and acknowledge that medical assessments happen at a point in time well beyond when individuals first experience racism within educational and assessment systems. She outlined five main areas where we can begin to chip at this wall of injustice:

Dr. Jennifer Randall, University of Michigan & Center for Measurement Justice

- **Do not burden minoritized individuals.** In racist systems, minoritized groups are often asked to carry the burden of assessing fairness. Organizations should invest in professional development opportunities to raise the level of awareness—and responsibility—of all involved individuals.
- **Take/share responsibility.** We must acknowledge our part within the current racist system and work collectively to make positive change..
- **Actively disrupt conventional racist stereotypes.** Test development practices have previously suggested removing bias by excluding individual characteristics in items or leveraging statistical or keyword review of items. However, these methods are inadequate. Individuals tend to learn and study the content present on tests; therefore, we have a responsibility to include antiracist materials on examinations—including vignettes that intentionally disrupt racist stereotypes and content related to health equity.
- **Reveal oppressive sociopolitical injustices.** Validity arguments for test use that rest solely on merit are inadequate and misleading; we must recognize widespread, ongoing systems of oppression that seek to create the illusion of merit-based differences.
- **Don't make excuses**. We must evaluate current policies: when considering policy changes, it is imperative to evaluate who is advantaged by the current system or who would be with any system change. When given a choice, choose the option that benefits those at the margins.

Panel: Equity in Medical Education and Assessment

The second day of the conference began with a panel discussion facilitated by Linda Gadsby, Esq, from NBME. The distinguished panelists—Michellene Davis from National Medical Fellowships, Inc., Reena Karani, MD, MHPE, from The Icahn School of Medicine at Mount Sinai, and Robert Roswell, MD, from Zucker School of Medicine at Hofstra/Northwell—each shared moving stories of their personal experiences with inequity in medical education and assessment. The panelists spoke to how the system of medical education is built and structured around serving white patients; for instance, instruction still focuses on symptoms that are applicable to white patients (e.g., flushing) or use of pulse oximetry monitors known to not function as well on darker skin.

Michellene Davis, National Medical Fellowships, Inc.

The panelists encouraged attendees to become "co-conspirators" in the efforts to bring more equity to this space. In addition, they shared hopefulness for the potential promise of using assessment in different and improved ways to drive equity; Dr. Robert Roswell mentioned using patient characteristics research as one example of a way to take a scalpel approach (as opposed to a sledgehammer) to this issue. They also expressed optimism in the creation of spaces like EMAC to help organizations make positive changes and demonstrate active interest in course-correcting.

Closing Plenary & Overall Summary

The closing plenary of EMAC was a time of sharing feedback and key learnings among the attendees, during which Taniecea Mallery, PhD, from AAMC and Greg Ogrinc, MD, from ABMS challenged the audience to wrestle with what was heard at the conference.

Dr. Taniecea Mallery, AAMC

In the remainder of the white paper, we summarize three major themes from the conference gleaned from the plenaries and concurrent sessions: inequity in medical education, disrupting bias, and learner context considerations. Following those chapters, we provide a summary of actionable steps the community can take, including a discussion of future opportunities for collaboration.

INEQUITY IN MEDICAL EDUCATION

In the session, "Equity in Medical Schools," Nientara Anderson, MD, a psychiatry resident and medical education scholar from Yale School of Medicine, opened by describing the history of race in medical school assessment. Centuries of historical and political events have shaped American medicine and entrenched racism into the practice of medicine.[4] From the Enlightenment through the birth of "modern" medicine, European reason was held as a universal truth and superior to the reasoning of other ethnic groups. As an extension of this premise, scientists created biological races and designated racial hierarchies, maintaining the superiority of white persons. Physicians propagated these theories of racial superiority, theorizing physiological differences between races that had severe consequences for Blacks and other minoritized groups. For example, Black health care workers were assigned to tend to the sick during Philadelphia's 1793 yellow fever outbreak because of the belief that they were less susceptible to the disease. As a result, a disproportionately high number of deaths occurred in the Black population. Further, physicians suppressed minoritized groups from engaging in the practice of medicine, supporting laws criminalizing Black and Indigenous doctors. This culminated in the 1910 American Medical Association commissioning of the Flexner Report. Abraham Flexner recommended closing most Black medicine schools and training Black physicians in "hygiene" instead.

Both sessions "Equity in Medical Schools" and "Bias in Medical Education and the Transition to Competency-Based Assessment" described how this history of bias persists in medical education today. Individuals who do not fit the white male prototype experience hostile learning environments and minority bias. Female, underrepresented in medicine (URiM), Asian, multiracial, and LGBTQ students bear a high burden of mistreatment compared to white male students.[5] In the Panel Discussion, both Drs. Reena Karani and Robert Roswell described their personal experiences with racial bias, specifically how hostile environments lead students to become less engaged yet more concerned over the hostility's impact on their career trajectories. In addition, URiM students and women tend to score lower on high-stakes multiple-choice question exams.[6] When schools assign grades, analysis has demonstrated that only a small proportion of the variance reflects student skill.[7] In fact, the majority of grade variance reflects error in measurement, due to assessors and other residual effects. Other studies have found multiple-choice exam performance to be the only consistent predictor of grades.[8] Thus, small differences in exam performance have substantial consequences as URiM students receive half the number of honors grades as majority students and are over three times less likely to be elected to the Alpha Omega Alpha (AOA) honor society.[9] For URiM students, these learning environment issues likely contribute to higher attrition rates.[10] For example, a URiM student with low family income raised in an under-resourced neighborhood is 3.7 times more likely to withdraw or be dismissed from school than students who were not URiM, not with low family income, and not raised in an under-resourced neighborhood..

Dr. Dowin Boatright, Grossman School of Medicine at NYU

Dowin Boatright, MD, from NYU's Grossman School of Medicine described how differences in academic measures are paralleled by the words used to describe learners.[11] White applicants were more likely to be described as "exceptional," "best," and "outstanding," while Black applicants were more likely to be described as "competent." Women were more likely to be described as "caring," "compassionate," and "empathic" compared to men. These descriptions are particularly important when program directors read the Medical Student Performance Evaluation (MSPE) for specific qualities. In fact, program directors describe comments about professionalism as particularly important in residency applications.[12] Yet, professionalism is poorly defined and is most understood within the context of dress, speech, work style, and timeliness.[13] This favors learners who reflect the norms of the evaluating faculty members, who are overwhelmingly white men.

Although holistic review has been lauded as a mechanism to increase diversity,[14] it cannot compensate for inconsistent practices that produce biased data for decision making. Christopher Feddock, MD, from NBME, described how MSPEs, key contributors to residency selection decisions, are poorly standardized. In fact, only 25% of schools provide complete information as suggested by AAMC recommendations.[15] Most schools use a keyword as an overall indicator of student performance. However, a total of 77 different keywords were used to describe students across schools, and only 55% provided a clear explanation of how those keywords are chosen. In fact, the meaning of any single keyword has high variability. For example, a student described as "very good" could reflect a student anywhere between the 0th and 71st percentiles, obscuring straightforward interpretation.

Overall, these practices have substantial repercussions for URiM and women in the selection process for residency and therefore, career trajectory and representation across medical specialties. Program directors prioritize USMLE scores, MSPEs, and clerkship grades when deciding whom to interview for their residency program, all of which have demonstrated differential performance that favors white men.

The differences in assessment persist in residency training, similarly favoring white men. Dr. Dowin Boatright described a study of milestone scores. Investigators examined 128 emergency medicine programs and 16,248 evaluations of their residents throughout training. At their first ACGME milestone reporting, most areas demonstrated no significant differences in the ratings provided across race, ethnicity, and sex. However, for each year of training, assessment ratings diverged with Asian and URiM residents receiving lower mean competency scores. Further, Asian and URiM women were rated lower than their male counterparts. For example, for the final end-of 4th year evaluation of medical knowledge, white women and Asian men were 0.25 lower, Asian women and URiM men were 0.5 lower, and URiM women were 0.75 lower than the white men mean competency score for ACGME milestones..

Inequity in Medical Education Recommendations

The inequity in medical education outcomes is a direct reflection of the inequity in the learning environment and assessment practices. Throughout the conference, several recommendations were consistently discussed across sessions, including themes from the conference keynote. Both Dr. Reena Karani in the plenary panel and Dr. Eric Holmboe in the "Bias in Medical Education and the Transition to Competency-Based Assessment" session discussed essential steps to confront the history of racism and achieve equitable outcomes for our trainees and ultimately our patients. These are the collective recommendations:

- Medicine must integrate the history of racism in medicine and medical education into the curriculum. Ignoring the complicated history only allows continued racist violence. American medical schools, institutions, and organizations must take accountability for their prior roles, and actively work to correct inequity stemming from those actions.
- Medical institutions must accelerate their transition to a competency-based education (CBE) framework. CBE focuses on outcomes or competencies, creates a developmental progression toward achieving competence, tailors learning experiences to meet student needs, includes instruction specifically designed to assist learners in achieving competence, and uses programmatic assessment to both provide learners with feedback and determine when competence has been attained.
    - Lisa Howley, PhD, from the AAMC explained the central role of formative assessment to support each learner's development. Formative assessment provides the feedback necessary for learners to improve prior to summative decisions and allows instruction to be tailored to match students' needs.
    - In addition, too many assessments rely on the global judgment of a learner's performance without direct observation of those skills. Inferring skills is insufficient to provide constructive feedback to learners. Plus, CBE requires criterion references to determine the student's developmental level and trajectory. The current system relies heavily on norm-referencing to rank students, which cannot contribute to learner growth.[16]

Dr. Lisa Howley, AAMC

- Successful implementation of CBE is only possible through faculty education and professional development on bias in medical education and assessments to mitigate sources of bias, including from raters. The AAMC, ACGME, ACCME, and AACOM have produced the Clinician Educator Milestones to support the development and improvement of faculty teaching and assessment skills.[17]
- Even with faculty development, assessments should be centralized and monitored for bias. Institutions must accept accountability for the impact of their assessments on learners, continually striving for fair and unbiased assessment. Centralized reporting systems can improve learning environments for all learners and faculty.
- Institutions must seek to diversify their faculty across the spectrum of roles. Diverse faculty are necessary to serve as role models and mentors, and to lead programs that are inclusive of all learners. Their presence can enhance environmental inclusion. Likewise, institutions should seek to adjust the visual environment to ensure that portraits and other displays are representative of diverse learner and patient populations.
- Medical schools and residency programs must embrace and expand holistic review as a historically informed approach to admissions. Again, institutions must acknowledge their history and current practices that disadvantage specific demographic groups and seek those that exemplify the changes desired in the practice of medicine. Students, faculty, and residents all deserve processes that consider the whole individual and that do not disproportionately prioritize specific academic factors.

DISRUPTING BIAS IN ASSESSMENT CONTENT

Another major topic at the inaugural EMAC was how to reduce bias and stereotypes in our assessment content. Practical strategies were offered on a number of dimensions: subject matter expert (SME) composition and training, item writing guidelines, statistical techniques to detect bias, and important considerations when using technology for content creation.

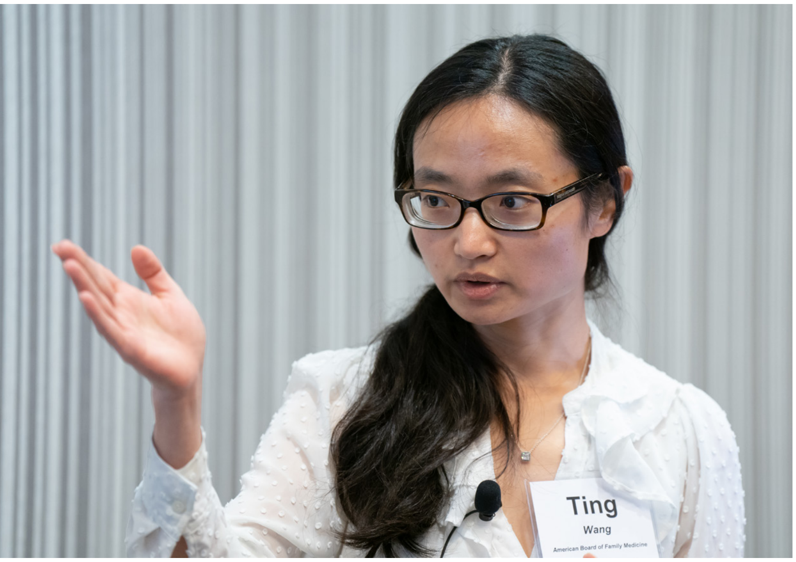
In the "Content Development Through the Equity Lens" session, Connie Murray, Kris DeRuchie, and Amy Morales from NBME discussed the importance of the SME committee composition and selection process, the training of SMEs to write equitable and unbiased items, and the challenges they face with content and psychometric reviews. The important takeaway is not to treat the process as a series of checklists, but rather to embed DEI principles in every step along the way. When we have diverse representation of SMEs on the committee, covering a wide range of race/ethnicity, gender, region, medical specialty, years of experience, size of medical school, practice type, and so on, such characteristics will be reflected in the assessment content created. While this representation matters, as the keynote speaker Dr. Jennifer Randall reminded attendees in her keynote, diverse SME composition is necessary *but not sufficient* to develop content through an equity lens.

Amy Morales, NBME

Conference attendees discussed the importance of including DEI staff or others with expertise on DEI issues in the content development and review process.

Careful training of SMEs to adhere to guidelines that embrace DEI practices will help develop content with a holistic approach from the beginning. This session also emphasized the importance of thoughtful use of patient characteristics in test items: avoiding bias, eliminating stereotypes, being mindful of the intent to include specific patient characteristics, and balancing how collectively patient populations should be represented on exams. The presenters also provided practical examples to make it more concrete for the audience, such as "expand reference to use partner in addition to spouse, husband and wife," "varied the sex of family caregivers, such as father brings child to well visit, son brings parent to ER," and so on. One of the goals for EMAC was to discuss concrete strategies and examples on how to promote equity in assessments, and this session contributed to that goal.

Psychometric analysis called Differential Item Functioning (DIF) is a statistical process to identify whether a certain item is performing differentially for different groups, such as male versus female, or white versus Black. EMAC presented a session devoted to DIF processes in a variety of board certification exam settings and was probably the most psychometrically-charged session at the conference. Jordan Yee Prendez, PhD, and Yang Zhao, PhD, from ABIM discussed their process in avoiding, detecting, and mitigating against bias in their test content. They shared procedures to detect bias, challenges in how to avoid bias, and offered some powerful potential strategies to eliminate bias from test content. When the testing volume is relatively low for certain exams, the desired sample size for statistical procedures is much harder to achieve. Seohee Park, PhD, and Dr. Yang Zhao from ABIM conducted a simulation study to address the small sample size issue for statistically flagging bias and differential performance. They recommended some empirical baseline guidance for conducting DIF analysis when the group size is small. They also reemphasized that statistical techniques are intended to help focus SMEs' attentions but should not be the sole criterion to determine whether items have differential functioning or not. In the same DIF session, Ting Wang, PhD, and Thomas O'Neill, PhD, from ABFM presented a score-based test of DIF approach to deal with the fact that some auxiliary grouping variables are categorial whereas others are ordinal or continuous. These presentations were more technical in nature, and helped illustrate various ways psychometrics can support reducing bias in test content.



Dr. Ting Wang, ABFM

ABMS member boards also discussed their strategies on how to prevent, detect, and mitigate bias in their initial certification examinations. Dr. Greg Ogrinc from ABMS, Sarah Schnabel, PhD, from ABO, Andrew Dwyer, PhD, from ABP, and Andrew Jones, PhD, from ABS participated in the session. They differentiated relevant versus irrelevant constructs, gave concrete examples on "what not to do" in developing content, and emphasized providing standardized training to raters to guard against bias. They also provided examples of items that either contain or do not contain cultural bias, and suggested ways to mitigate against cultural bias. Human scoring was a major piece of a certification assessment for one of the specialty boards at the session. The presenter discussed in detail the types of training and guardrails being put in place to mitigate potential bias in the scoring process. By including real examples of (publicly released) test content and scoring methodology, the presenters combined theory and practice effectively for the audience.

With advances in technology, especially considering generative AI, NBME organized a session devoted to the use of technology in content development and discussed its impact on equity in assessments. In this session, Allison Kulesher, Kimberly Swygert, PhD, and Victoria Yaneva, PhD, from NBME, discussed both the cognitive modeling approach for item generation and the large language model (generative AI) approach for item generation. While technology could potentially scale item development exponentially, it is also susceptible to bias and stereotypes in ways that are different from humans. The training of the computer models determines what gets generated. The presenters focused on the challenging elements of such approaches and offered mitigation strategies. Their conclusion? Technology can greatly assist in content development and can safeguard against bias, with great intentional approaches and principled care. The audience appreciated the timeliness of the presentations and walked away with a healthy level of optimism regarding the use of technology to promote equity in assessments.

These sessions highlighted some ways organizations are trying to promote equity in their assessment offerings (e.g., eliminating stereotypes in their test content). It is clear that DEI principles need to be embedded within the entire process of assessment creation, delivery, scoring, and reporting; it cannot simply be a checklist and as an afterthought.

Additionally, and most importantly, given assessments have a place in curriculum and teaching, are there opportunities to use assessment to further promote equity and justice? The EMAC keynote speaker Dr. Jennifer Randall powerfully argued that assessment professionals can actively create content to not only avoid bias and stereotypes, but also to disrupt and dismantle racism and bias. She stated that, as an example, it is not enough to craft items that do not imply that Black patients are more difficult; it is of fundamental importance that we create items/scenarios where Black patients are exceedingly thoughtful and actively engage in their health and treatment plans. Without question, her talk provided further ideas and inspiration to all the organizations participating in the conference.

Disrupting Bias in Assessment Content Recommendations

- While it is important to reduce and eliminate bias in assessment content, assessment designers and developers should also consider moving one step further and develop content that actively disrupts conventional stereotypes about historically marginalized populations. Assessment has a place in curriculum and instruction; it has the potential to play a bigger role in us making progress in equity.
- While a checklist can be important to implement to protect assessments from potential bias, it should be noted that DEI principles should be embedded throughout the assessment development process.
- Even though having a diverse group of SMEs does not guarantee that assessments will be free of bias, it must be one of the very first steps. When considering diversity, besides race/ethnicity, we should also consider other demographic and cultural variables such as gender, region, medical specialty, years of experience, size of medical school, student population characteristics, practice type, and so on.
- Patient characteristics can be used as one way to combat bias in assessment by presenting patients in situations and scenarios that directly challenge these stereotypes.
- Psychometric analysis is an important tool to evaluate fairness of a given assessment. It is important that psychometric analyses (such as DIF) are routinely carried out for the assessments.
- While AI could potentially scale content development exponentially, it is also susceptible to bias and stereotypes in ways that are different from humans. As we move into the generative AI era, assessment developers need to pay close attention to how AI is being used in assessments and exercise principled care to ensure equity.

CONSIDERING LEARNER CONTEXT

In addition to disrupting bias in test content and inequity in medical education, an additional theme emerged from the EMAC sessions: being aware of and attentive to the learner context, including accessibility and an environment of trust (or lack thereof) in the assessment process.

Test Accommodations

Another important theme of the conference was the need to focus on appropriate test accommodations. In the "Equity in Measurement through Test Accommodations" session, Erin Convery and Lucia McGeehan, PhD, from NBME, Dara Greenberger from AAMC, and Maryellen Gusic, MD, from the Lewis Katz School of Medicine at Temple University focused on the role accommodations play in allowing for valid and reliable inferences about the performance of a diverse group of test takers.

Presenters discussed various ways in which testing agencies and medical schools ensure accommodations provide access and support equity and inclusion. Testing agencies provide equity in measurement through test accommodations by relying on an individualized review process for fair decision making; collaborating with colleagues on determining appropriate and reasonable accommodations for test content; advocating for universal design features; and promoting online resources such as guidelines, frequently asked questions, and practice materials. In the context of medical education, medical educators must consider the implementation of appropriate and reasonable accommodations to provide access to different learning environments, such as classroom-based versus clinical-based, and use caution so that the essential requirements of medical coursework are not altered because of an approved accommodation. As highlighted by the presenters across educational and testing spectrums—academic, high stakes, and licensure—test accommodation needs are relevant to content, location, and purpose of the examination or coursework. Presenters considered how accommodations support both learners and test takers and also raised the importance of including accommodated populations in research, item, and assessment development.

The key takeaways from this session focused on how future research and DEI partners can support equitable assessment. Test development staff should define clear test constructs to support disability service professionals as they evaluate the accommodation needs of a test-taker, proactively discuss new item types with disability service professionals to better understand appropriate and reasonable accommodations, and develop research studies to better understand the needs of accommodated test-takers. With regard to partnership, the discussion focused on including disability as part of the overall DEI conversation in testing organizations and medical schools and DEI partnerships with disability resource professionals and personnel to better understand and learn about the environment. This session was well received and contributed to the overall goal of the conference by highlighting the diverse needs of a population often overlooked—accommodated test takers and learners. We look forward to additional opportunities for growth and learning in this space.

Trust in Assessment

Another vital aspect impacting equity in assessment is the trust extended to the assessment systems by the learners and the organizational contexts within which the assessments are part. Michael Barone, MD, MPH, and Linette Ross, PhD, from NBME and Javarro Russell, PhD, from AAMC highlighted the continuing importance of trust, and how changes to systems may either erode or increase that trust over time, in the session "Assessment 20 Years from Now: The Importance of Trust."

Dr. Linette Ross, NBME and Dr. Javarro Russell, AAMC

Trust, defined as an assured reliance on someone or something's character, ability, or truth,[18] is vital on both individual and organizational levels. Individual behaviors such as care, sincerity, reliability, and competence contribute to building trust; organizations, too, need to understand trust building during flourishing periods and trust recovery during crises. This includes understanding shared values, aligned interests, benevolence, competence, integrity, and effective communication.[19-21]

In health care and health professions education, trust is crucial for clinical practice and training the next generation of providers. Assessments play a significant role in education, licensure, and certification. However, several threats to trust exist at the individual, organizational, and systemic levels, often due to organizational and systemic challenges.

At the individual level, the documented experiences of health professions trainees demonstrate this lack of trust in assessment systems.[22] At the organizational level, medical school competency assessments have been shown to be measuring as much or more error than construct-relevant performance.[7] At the systems level, many organizations in the House of Medicine have been viewed as having conflicts of interest related to assessment.[23-24] To rebuild trust, assessment systems need to embrace transparency, acknowledge expertise beyond their own, and focus on long-term trust-building. At the assessment program level, achieving intrinsic, contextual, and instrumental equity is essential.[25]

Equity in assessments can be promoted by considering various criteria like validity, reproducibility, equivalence, and acceptability.[26] In medical education assessments, trust is built when assessments are fair, accurate, and aligned with their intended purposes, while emphasizing intrinsic, contextual, and instrumental equity in assessments can help in evaluating success and promoting trust. Trustworthy learning environments are crucial for the development of competent health professionals. An understanding of the abilities and level of trust a learner brings with them when interacting with assessment systems is vital for further progress in equitable assessment models. Finally, involving learners in the assessment development itself can be pivotal in building their trust in the system and advancing equity.

Considering Learner Context Recommendations

- Clear communication is essential for allowing for equitable accommodations.
  - Test constructs must be defined clearly to ensure accommodations can be structured appropriately.
  - New item types should be discussed with disability service professionals to ensure appropriate accommodations in advance of incorporation.
- Research should be prioritized to better understand appropriate and reasonable accommodations.
- Disability and disability resource professionals should be included as part of the overall DEI conversation in testing organizations and medical schools.
- Organizations should focus on building trust with their users to ultimately improve equity by:
  - Providing all learners with the opportunities and resources to learn and be assessed,
  - Aligning assessment use with appropriate standards of psychometric rigor,
  - Acknowledging that we are not the only experts,
  - Being transparent and communicating our work, and
  - Focusing on building trust over the long term, not in a "project by project" manner.
- Inviting learners to be co-creators in the process of developing assessments is a key component in establishing trust and ensuring fairness and equity.

CONCLUSION

Overall, the inaugural EMAC was very well received by attendees. EMAC strove to combine theory and practice: while highlighting the importance of equitable practice in assessments, the conference also aimed to focus on concrete steps we should take. The presentations were uniformly powerful, provocative, and insightful. Participants were willing to lean into uncomfortable conversations and were eager to translate words into actions.

Dr. Jennifer Randall issued a call to action in her plenary to harness the collective power of expertise. EMAC sought to bring together a diverse group of professionals from many areas of expertise, who, as Dr. Jennifer Randall reminded us, can each keep chipping away in our various places of work to ultimately do the greatest good. The conversations at EMAC, and following this inaugural event, were the start of this work.

This paper sought to elaborate on a few of these conference learnings by synthesizing discussions occurring in many of the sessions at the conference that we believe can result in actionable, positive change. In addition to the recommendations included in each chapter, a summary of the top 10 key conference learnings is provided in Appendix B.

Due to the important and inspirational directions generated from the first EMAC, the organizers are planning a second EMAC event occurring in the fall of 2024. We look forward to continued dialogue in the areas highlighted in this white paper and beyond: what we can do collectively to address inequity in medical education, avoid and disrupt bias, increase focus on equitable accommodations, and continue to build trust.

REFERENCES

1. Jetty A, Jabbarpour Y, Pollack J, et al. Patient-physician racial concordance associated with improved healthcare use and lower healthcare expenditures in minority populations. J Racial Eth Health Disparities. 2022;9(1):68–81. doi: 10.1007/s40615-020-00930-4

2. Ku L, Vichare A. The Association of Racial and Ethnic Concordance in Primary Care with Patient Satisfaction and Experience of Care. J Gen Intern Med. 2023 Feb;38(3):727-732. doi: 10.1007/s11606-022-07695-y. Epub 2022 Jun 10. PMID: 35688996; PMCID: PMC9186269.

3. Marrast LM, Zallman L, Woolhandler S, Bor DH, McCormick D. Minority physicians' role in the care of underserved patients: Diversifying the physician workforce may be key in addressing health disparities. JAMA Intern Med. 2014;174(2):289-291

4. Anderson N, Nguyen M, Marcotte K, Ramos M, Gruppen LD, Boatright D. The Long Shadow: A Historical Perspective on Racism in Medical Education. Acad Med. 2023 Apr 19. doi: 10.1097/ACM.0000000000005253. Epub ahead of print. PMID: 37071703.

5. Hill KA, Samuels EA, Gross CP, Desai MM, Sitkin Zelin N, Latimore D, Huot SJ, Cramer LD, Wong AH, Boatright D. Assessment of the Prevalence of Medical Student Mistreatment by Sex, Race/Ethnicity, and Sexual Orientation. JAMA Intern Med. 2020 May 1;180(5):653-665. doi: 10.1001/jamainternmed.2020.0030. PMID: 32091540; PMCID: PMC7042809.

6. Rubright JD, Jodoin M, Barone MA. Examining Demographics, Prior Academic Performance, and United States Medical Licensing Examination Scores. Acad Med. 2019 Mar;94(3):364-370. doi: 10.1097/ACM.0000000000002366. PMID: 30024473.

7. Zaidi NLB, Kreiter CD, Castaneda PR, Schiller JH, Yang J, Grum CM, Hammoud MM, Gruppen LD, Santen SA. Generalizability of Competency Assessment Scores Across and Within Clerkships: How Students, Assessors, and Clerkships Matter. Acad Med. 2018 Aug;93(8):1212-1217. doi: 10.1097/ACM.0000000000002262. PMID: 29697428.

8. Hauer KE, Lucey CR. Core Clerkship Grading: The Illusion of Objectivity. Acad Med. 2019 Apr;94(4):469-472. doi: 10.1097/ACM.0000000000002413. PMID: 30113359.

9. Teherani A, Hauer KE, Fernandez A, King TE Jr, Lucey C. How Small Differences in Assessed Clinical Performance Amplify to Large Differences in Grades and Awards: A Cascade With Serious Consequences for Students Underrepresented in Medicine. Acad Med. 2018 Sep;93(9):1286-1292. doi: 10.1097/ACM.0000000000002323. PMID: 29923892.

10. Nguyen M, Chaudhry SI, Desai MM, Chen C, Mason HRC, McDade WA, Fancher TL, Boatright D. Association of Sociodemographic Characteristics With US Medical Student Attrition. JAMA Intern Med. 2022 Sep 1;182(9):917-924. doi: 10.1001/jamainternmed.2022.2194. PMID: 35816334; PMCID: PMC9274446.

11. Ross DA, Boatright D, Nunez-Smith M, Jordan A, Chekroud A, Moore EZ. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. PLoS One. 2017 Aug 9;12(8):e0181659. doi: 10.1371/journal.pone.0181659. PMID: 28792940; PMCID: PMC5549898.

12. Bird JB, Friedman KA, Arayssi T, Olvet DM, Conigliaro RL, Brenner JM. Review of the Medical Student Performance Evaluation: analysis of the end-users' perspective across the specialties. Med Educ Online. 2021 Dec;26(1):1876315. doi: 10.1080/10872981.2021.1876315. PMID: 33606615; PMCID: PMC7899642.

13. Gray, A. The Bias of 'Professionalism' Standards. Stanford Social Innovation Review. 2019. https://doi.org/10.48558/TDWC-4756

14. Otugo O, Alvarez A, Brown I, Landry A. Bias in recruitment: A focus on virtual interviews and holistic review to advance diversity. AEM Educ Train. 2021 Sep 29;5(Suppl 1):S135-S139. doi: 10.1002/aet2.10661. PMID: 34616988; PMCID: PMC8480505.

15. Tisdale RL, Filsoof AR, Singhal S, Cáceres W, Nallamshetty S, Rogers AJ, Verghese AC, Harrington RA, Witteles RM. A Retrospective Analysis of Medical Student Performance Evaluations, 2014-2020: Recommend with Reservations. J Gen Intern Med. 2022 Jul;37(9):2217-2223. doi: 10.1007/s11606-022-07502-8. Epub 2022 Jun 16. PMID: 35710660; PMCID: PMC9296706.

16. Ryan MS, Lomis KD, Deiorio NM, Cutrer WB, Pusic MV, Caretta-Weyer HA. Competency-Based Medical Education in a Norm-Referenced World: A Root Cause Analysis of Challenges to the Competency-Based Paradigm in Medical School. Acad Med. 2023 Mar 24. doi: 10.1097/ACM.0000000000005220. Epub ahead of print. PMID: 36972129.

17. Accreditation Council for Graduate Medical Education (ACGME), Accreditation Council for Continuing Medical Education (ACCME), Association of American Medical Colleges (AAMC), American Association of Colleges of Osteopathic Medicine (AACOM). The Clinician Educator Milestone Project. ©2022. Accessed at: https://www.acgme.org/globalassets/pdfs/milestones/standalone/2022/clinicianeducatormilestones.pdf

18. "trust." Merriam-Webster.com. 2023. https://www.merriam-webster.com/dictionary/trust#dictionary-entry-1 Accessed 6/23/2023

19. Cruess SR, Cruess RL. Professionalism and Medicine's Social Contract with Society. Virtual Mentor. 2004 Apr 1;6(4)

20. Feltman C. The Thin Book of Trust: An Essential Primer For Building Trust at Work. Thin Little Book Publishing, July 2021. ISBN: 0988953862

21. Hurley, RF, Gillespie, N, Ferrin, DL, Dietz, G. Designing Trustworthy Organizations. MIT Sloan Management Review, 2013 June, 54 (4), 74-82

22. Marte D. Can a Woman of Color Trust Medical Education? Acad Med. 2019 Jul;94(7):928-930

23. Carmody JB, Rajasekaran SK. On Step 1 Mania, USMLE Score Reporting, and Financial Conflict of Interest at the National Board of Medical Examiners. Acad Med. 2020 Sep;95(9):1332-1337

24. Gesundheit N. A Crisis of Trust Between U.S. Medical Education and the National Board of Medical Examiners. Acad Med. 2020 Sep;95(9):1300-1304

25. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical Education's Wicked Problem: Achieving Equity in Assessment for Medical Learners. Acad Med. 2020Dec;95(12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments):S98-S108.

26. Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V., et al. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 33:206–214.

APPENDIX A: EMAC SESSION LIST

The following sessions were organized by the EMAC Planning Committee:

Ye Tong, PhD,[1] Linda Gadsby, Esq.,[1] Greg Ogrinc, MD,[2] Katherine Torres-Hertz, MOL,[2] Rebecca Lipner, PhD,[3] Dana Dunleavy, PhD,[4] and Taniecea Mallery, PhD[4]

[1]NBME, [2]ABMS, [3]ABIM, [4]AAMC

Plenary Sessions

- Plenary: Equity in Measurement and Assessment
  - Richard Baron, MD, ABIM
  - Stephen Fitzpatrick, AAMC
  - Richard E. Hawkins, MD, ABMS
  - Peter J. Katsufrakis, MD, NBME
  - Ye Tong, PhD, NBME
- Keynote: First, Undo Harm: Disrupting Racist Logics and Their Consequences via Medical Assessment
  - Jennifer Randall, PhD, University of Michigan and Center for Measurement Justice
  - Introduction by Rebecca Lipner, PhD, ABIM
- Panel: Equity in Medical Education and Assessment
  - Michellene Davis, JD, National Medical Fellowships, Inc
  - Reena Karani, MD, Mount Sinai
  - Robert Roswell, MD, Zucker School of Medicine at Hofstra/Northwell
  - Facilitation by Linda Gadsby, JD, NBME
- Plenary: Challenge and Feedback
  - Taniecea Mallery, PhD, AAMC
  - Greg Ogrinc, MD, MS, ABMS

Concurrent sessions

- Equity in Medical Schools
  - Nientara Anderson, MD, Yale School of Medicine
  - Christopher Feddock, MD, NBME
  - Lisa Howley, PhD, AAMC
- Assessment 20 Years from Now: The Importance of Trust
  - Michael Barone, MD, NBME
  - Linette Ross, PhD, NBME
  - Javarro Russell, PhD, AAMC
- How to Assess Item Bias in Standardized Assessments
  - Seohee Park, PhD, ABIM
  - Jordan Prendez, PhD, ABIM
  - Yang Zhao, PhD, ABIM
  - Ting Wang, PhD, ABFM
  - Thomas O’Neill, ABFM
- Facilitated Conversation with Dr. Jennifer Randall
  - Jennifer Randall, PhD, University of Michigan and Center for Measurement Justice
  - Facilitation by Dana Dunleavy, PhD, AAMC
- Development Through the Equity Lens
  - Kris DeRuchie, NBME
  - Amy Morales, NBME
  - Connie Murray, NBME
- The Inclusion of Patient Characteristics in Test Items: Current Practices and Future Directions
  - Pamela Kaliski, PhD, ABIM
  - Chunyan Liu, PhD, NBME
  - Lorna Lynn, MD, ABIM

- Using Assessments in Context
  - Michael Bastedo, PhD, University of Michigan
  - Leila Harrison, PhD, Washington State University Elson S. Floyd College of Medicine
  - Sunny Nakae, PhD, California University of Science and Medicine
  - Cindy Searcy, AAMC
- Bias in Medical Education and the Transition to Competency-Based Assessment
  - Dowin Boatright, MD, NYU Department of Emergency Medicine
  - Eric Holmboe, MD, Accreditation Council for Graduate Medical Education
- How Can Assessments Contribute to Health Equity?
  - Rebecca Fraser, PhD, AAMC
  - Ann Harman, PhD, ABIM
  - Kamilah Weems, MS, AAMC
- Equity in Measurement through Test Accommodations
  - Erin Convery, NBME
  - Dara Greenberger, AAMC
  - Maryellen E. Gusic, MD, Lewis Katz School of Medicine at Temple University
  - Lucia McGeehan, PhD, NBME
- Preventing, Detecting, and Mitigating Bias in ABMS Member Board Initial Certification Examinations
  - Andrew Dwyer, PhD, ABIM
  - Andrew Jones, PhD, American Board of Surgery
  - Greg Ogrinc, MD, ABMS
  - Sarah Schnabel, PhD, American Board of Ophthalmology
- Use of Technology in Item Development and the Impact on Equity in Assessment
  - Allison Kulesher, NBME
  - Kimberly Swygert, PhD, NBME
  - Victoria Yaneva, PhD, NBME

APPENDIX B: TOP 10 TAKEAWAYS

The inaugural Equity in Measurement and Assessment Conference (EMAC) brought together measurement, assessment, medical education, and diversity, equity, and inclusion (DEi) experts to answer a central question within medical education assessment—how can we ensure that everyone, regardless of their background, has an equitable opportunity to demonstrate their knowledge and skills?

View some of our top takeaways from the conference below, and visit reassessthefuture.org to learn more about how NBME is working to improve fairness in testing.

1 Experts from all backgrounds, including measurement and assessment, medical education, and DEI, should collaborate and continue to bring justice to medical education assessments, and we are all responsible for the solutions.

2 Leaning into difficult and uncomfortable conversations about bias and inequity in assessment is needed to enact change.

3 Assessments need to do more than simply avoid perpetuating negative bias—they must actively disrupt conventional stereotypes about historically marginalized populations.

4 Patient characteristics can be used as one way to combat bias in assessment by presenting patients in situations and scenarios that directly challenge these stereotypes.

5 Questions that don’t specifically mention patient characteristics can still be biased, as authors often write through the lens of their own personal experiences.

6 Evaluating questions for bias isn’t enough if the right people aren’t in the room, which makes diverse representation on test development and review committees essential.

7 Although large language models, such as the one used by ChatGPT, show promise in advancing the test development process, they must be specifically designed and trained on appropriate databases to avoid bias.

8 Automated item generation has the potential to reduce bias in assessment by producing a high volume of questions with different variations of patient characteristics, but careful review of items is necessary.

9 Formative assessments are a critical component of supporting equitable practices, since they enable educators to provide tailored support for learners from different backgrounds.

10 Inviting learners to be co-creators in the process of developing assessments is a key component in establishing trust and ensuring fairness and equity.

Source: https://www.reassessthefuture.org/nbme-in-action/fairness-in-testing/emac-top-10-learnings/