Competency-Based Education Research Journal, 2026
https://journals.ku.edu/cberj/

A Data-Driven Evaluation of Assessment Equity in Competency-Based Education

Sean P. Gyll
Heather Hayes

Western Governors University

Achievement gaps in competency-based education can reflect systemic barriers and raise questions about the fairness of summative assessments. This study reports evidence from Western Governors University's Assessment & Learning (Equity Framework), which combines psychometric screening with expert review to monitor assessment fairness in courses showing attainment gaps. Using operational data from approximately 30,000 first-attempt records across 32 summative assessments in 30 courses (January 1, 2023-January 1, 2024), we evaluated differential item functioning (DIF) between historically underrepresented racial/ethnic groups (focal group) and White students (reference group) with item response theory residual DIF and iterative purification. Five assessments contained items that met a statistical flag criterion ($p \leq .01$), representing 2%-10% of the items in those assessments (20 items total). Subject-matter experts reviewed all flagged items for potential sources of construct-irrelevant variance (e.g., cultural references, unclear language, accessibility barriers); none were judged to contain content bias. These findings suggest that, in the sampled courses, assessment content is unlikely to be a primary driver of observed attainment gaps and highlight the importance of examining additional contributors such as opportunity-to-learn, learning-resource quality, readiness supports, and instructional interactions. DIF results are interpreted as one component of an ongoing equity monitoring system rather than as a comprehensive evaluation of equity initiatives.

Keywords: competency-based education (CBE); assessment fairness; differential item functioning (DIF); performance assessment; equity in higher education

Correspondence concerning this article should be addressed to Sean P. Gyll, Ph.D., Western Governors University, Salt Lake City, UT, USA. Email: s.gyll@wgu.edu

Western Governors University (WGU) was founded on an ideal, "some might have called it a fantasy - a vision of an institution that could value learning, not time; that could deliver knowledge where people are, as opposed to people having to go where the knowledge is" (King, 2017, p. 1). To achieve this vision, WGU crafted a bold and inspiring vision statement that drives change across the university to meet the needs of employers and, most importantly, students. The statement reflects its mission, cultural beliefs, and key results. It is designed to help all individuals better understand WGU's vision and align themselves with it, aiming to make the university the most inclusive in the world.

In a competency-based university like WGU, the quality of summative assessments is essential (Popham, 2017). Students' performance on these assessments must reflect their competency in the course's skills, enabling them to progress in their degree program. WGU follows a rigorous process for developing assessments and systematically monitors their quality over time. Suppose there are signs of poor quality, such as bias or unfairness. In that case, WGU will make the necessary adjustments, as these issues can negatively impact the validity and utility of the assessment outcomes.

Fairness in Assessment

Fairness in assessment refers to the absence of bias in the content and scoring processes, and is a multidimensional concept encompassing both procedural and substantive aspects (Kane, 2013; Messick, 1989). The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014) conceptualize fairness as the absence of bias, equitable treatment during testing, equality of outcomes, and the opportunity to learn. An assessment is considered fair when all students, regardless of background or identity, are afforded equal opportunity to demonstrate their competence through their performance.

Barriers to fairness may include aspects of the assessment that could unfairly advantage one group over another. These barriers can include discriminatory content, irrelevant constructs, slang, obscure vocabulary, or a lack of clarity about what is expected of students to complete a task successfully. Suppose an assessment contains items that result in biased outcomes for a particular group. In that case, competent students from that group may not pass, leading to performance gaps.

Bias in Assessment

Bias in assessment refers to systematic error that differentially affects individuals' performance based on their group membership (e.g., race, gender, socioeconomic status, language background). This systematic error introduces construct-irrelevant variance, measuring factors unrelated to the intended construct, which undermines the validity of inferences drawn from assessment scores. Bias can manifest at multiple levels:

- ***Content bias*** occurs when test items contain cultural references, language, or contexts that advantage certain groups while disadvantaging others.
- ***Methodological bias*** emerges from the procedures used in test development, administration, or scoring that systematically favor particular groups.
- ***Interpretation bias*** appears when similar scores lead to different consequences across groups due to flawed interpretations of what those scores signify.

Purpose

Observed differences in course completion and assessment pass rates are best interpreted as signals of structural and systemic barriers rather than as learner deficits. Within this context, WGU's Assessment & Learning Equity Framework (Equity Framework) treats assessment fairness as a necessary (but not sufficient) condition for equity: if summative assessments contain construct-irrelevant barriers for particular groups, attainment gaps may be amplified, and remediation efforts misdirected.

The purpose of this study was (a) to report the results of a differential item functioning (DIF) screening process applied to summative assessments in courses identified as having attainment gaps between historically underrepresented racial/ethnic groups and White students, and (b) to describe how DIF evidence is used within the Equity Framework to guide continuous improvement in assessment, curriculum, and learner support. Table 1 provides context for the attainment gap groups in the sampled courses and assessments.

Accordingly, DIF analyses were used to evaluate measurement fairness for course competencies rather than to assess Diversity, Equity, and Inclusion (DEI) competencies directly. Specifically, this study addressed the following three research questions:

(1) Across the sampled summative assessments, how frequently are items statistically flagged for DIF using IRT residual DIF with iterative purification?
(2) For statistically flagged items, how often do subject-matter experts identify plausible sources of content bias or construct-irrelevant variance?
(3) How are DIF findings integrated into a broader decision process for investigating and addressing attainment gaps (e.g., learning-resource review, readiness supports, and instructional interventions)?

By specifying what DIF analyses can and cannot support, we aim to provide a practical model for institutions seeking to operationalize equity commitments through routine assessment monitoring, transparent decision rules, and iterative review.

Table 1
Attainment-Gap Context for Sampled Courses and Assessments

Metric	Reference group	Focal group	Gap (Ref - Focal)	Range across courses and assessments	Notes
Course completion rate (CCR)	74.07%	63.69%	10.38%	19.69-92.49%	Defined as course completion within the institutional monitoring window.
First-attempt assessment pass rate (PR)	79.11%	69.46%	9.65%	15.22-93.33%	Calculated on first-attempt outcomes for each assessment.
CPLT readiness score, *M* (*SD*)	49.93 (22.42)	47.24 (22.73)	2.69	0-100	Use the readiness measure(s) available for the sampled courses; specify scale.

Diversity vs. Equity vs. Inclusion

In higher education research and organizational scholarship, diversity is often used to describe the representation and meaningful engagement of people with different social identities and life experiences; equity emphasizes identifying and removing structural barriers that produce systematic disparities in opportunity and outcomes; and inclusion refers to conditions that support full participation and belonging while valuing individuals' uniqueness (Bensimon, 2007; Shore et al., 2011). At WGU, DEI is defined operationally as the range of individuals' unique attributes and experiences across ethnicities, ages, genders, and accessibilities. Specifically:

- ***Diversity*** is any dimension that distinguishes groups and individuals from one another. It reflects the mixture of differences and similarities worldwide and acknowledges the related tension as we strive to develop more inclusive and high-performing environments.
- ***Equity*** promotes fair and just procedures and practices throughout society and creates conditions where everyone can participate, prosper, and reach

their full potential. Equity recognizes that everyone starts from a different place and may require additional resources to reach their potential. It means fair treatment, access, opportunity, and advancement for all people, while striving to identify and eliminate barriers that prevent some from participating fully in society.

- ***Inclusion*** represents a state of feeling valued, respected, and supported. It is based on shared culture, organizational practices, and interpersonal relationships that support an institution's full utilization and diversity at all levels and functions. It ensures that everyone feels a sense of belonging. Simply put, inclusion is about making people feel welcomed and valued.

Based on these operational definitions, we use the term DEI to represent broad initiatives that leverage the potential benefits of DEI through inclusive instructional practices and institutional policies.

Identifying and Using Meaningful DEI Metrics

We now turn to what we believe is the most critical component of any change initiative and its corresponding strategies: identifying and utilizing meaningful DEI metrics. Most higher education institutions today recognize the importance of fully acknowledging the contributions of a diverse range of students to meet the demands of an increasingly diverse, global talent pool. They want to ensure their equity initiatives are planned and implemented so they can attract and leverage contributions from a broad student body. A set of key benefits in creating a diverse university helps institutions leverage their inclusivity to become more productive components of students' education, regardless of their background, socioeconomic status, or history.

WGU's Assessment & Learning (Equity Framework)

A common error when launching any change initiative is rushing into action without first understanding the real issues and opportunities. This can result in wasting valuable time and resources on ineffective interventions. Therefore, the first step towards a successful DEI initiative is to assess the current situation, identify areas for improvement, set realistic and measurable goals, and design effective interventions tailored to the university's unique needs (Bensimon, 2007; Ladson-Billings, 2006).

WGU has already been collecting quality outcome measures, such as attainment and graduation rates, that can effectively gauge progress on diversity. However, the key to success with this project lies in identifying and using meaningful equity metrics, evaluating the impact of the university's inclusion efforts on those metrics, and improving those that need improvement.

Using the Framework

When using the Equity Framework (see Appendix A), curriculum, content, and assessment developers are reminded that it is a dynamic tool that continues to evolve

as the university gains more knowledge about how to promote equity in learning and assessment materials. DEI competencies and relevant learning content are integrated into the Equity Framework during the design phase to ensure the success of all learners. These competencies encompass a range of topics, including self-awareness, privilege, stereotypes and bias, diversity, equity, and inclusion, worldviews, cross-cultural awareness, social perceptiveness, inclusive communication, and advocacy. By incorporating these competencies, learners gain a deeper understanding of DEI principles and their application throughout all course content.

Following the guidelines helps contribute to the formation of WGU courses and enhances quality through the Equity Framework's principles:

- Use the Equity Framework as a reference when creating learning content and consider how its guiding principles improve course elements.
- Consider how the course structure and student engagement align with the Equity Framework's learning development practices when designing content.
- During development, aim to create a learning experience that promotes student exposure and growth.
- Use the Equity Framework to determine the best approach to course materials that may impact the university's key results, specifically equity and attainment.

Implementation and Impact Measures

The Equity Framework identifies various Implementation and Impact Measures, each aligned to a Guiding Principle (see Table 2). The goal is to track outcomes over time and utilize data feedback loops to evaluate changes in learner knowledge, skills, attitudes, and behavior. The process is iterative and informs continuous improvement efforts to enhance the integration of equity principles across programs, courses, and assessments. By adopting this data-driven approach, WGU seeks to create a more equitable, inclusive, and socially conscious learning experience that empowers students to apply DEI principles to academic, professional, and civic contexts.

Within this framework, assessment fairness is operationalized as one actionable component of equity monitoring: DIF screening is conducted under the 'Remove Barriers' principle to evaluate whether assessment items function similarly for students of equal proficiency across groups. DIF evidence is interpreted alongside other implementation and impact measures (e.g., learning-resource reviews, readiness indicators, and instructional interaction measures) to guide decisions about whether to revise assessments or to prioritize other course- and program-level interventions when attainment gaps persist.

Table 2

Equity Framework: Guiding Principles, Implementation, and Impact Measures

Guiding Principle	Implementation Measures	Impact Measures
Welcome Me: If I feel accepted and comfortable and can be my authentic self, I can focus more fully on learning and demonstrating competence.	Monitor content for intersectional perspectives, review learner profiles, and conduct peer reviews through ongoing research and engagement.	Persistence, satisfaction, surveys from faculty, learners, and alumni, and reenrollment rates.
Integrate Diverse Perspectives and Experiences: When I see others like me succeed, I am more likely to persist.	Monitor diversity in images and markers across case studies and other content, and conduct peer reviews of content against DEI standards and guidance.	Persistence; satisfaction; number of students who change programs; number of graduates in fields with goals to broaden diversity.
Make it Matter: I need to be able to trust that what I am learning is relevant to the workplace and matters to me.	Marketability index: map course content and assessments to needs analyses; map competencies and assessments to accreditation or professional standards.	Persistence, factored graduate impact, employer evaluation of graduate readiness, and graduates employed in the field.
Remove Barriers: I deserve an equal opportunity to pursue my education.	Review against UDL principles; correlate skills versus content types; review for DIF and other methods of assessment fairness.	Persistence of learners with specific needs; course completion rates for specific learners.
Personalize My Needs: Give me your best recommendations for pathways and content that align with my needs and goals but let me have the final say.	Review learner personas, analyze PA assessments, and track professional goals.	Persistence, attempts, and course completion rates.
Prioritize Value: My tuition and time are precious resources.	Reduce LR costs; apply budget management and program development efficiency measures; correlate skills versus content type; track per-student cost.	Persistence of low-income learners; number of learners citing cost as a reason for dropping.
Strengthen Opportunities for Impact: I will have to function in diverse environments and opportunities to promote equity-help me do both.	Track adoption of DEI competencies; map DEI competencies to accreditation and professional standards; produce marketable-skills reports tagged to DEI competencies.	Student success rates on DEI competencies, employer surveys tagged to DEI competencies, and student and alumni surveys.

Analyzing Bias in Student Data

Regarding equity and other inclusion initiatives, the conversation about metrics should happen early on and for a good reason. Any institution implementing such initiatives must have a system to measure their effectiveness. Implementing a thoughtful measurement plan provides structure and focus to DEI efforts, leading to more meaningful progress. A measurement plan clearly defines goals and determines the relevant metrics to track performance against them. As data are collected, problem areas become visible, and results can be benchmarked against goals. Measurement helps focus initiatives and resources where they are most needed. Rather than taking a scattershot approach, measuring change tracks how different initiatives and interventions impact their intended purpose. Over time, this enables adjustments to practices that improve results.

Differential Item Functioning (DIF)

WGU uses a multi-step process to evaluate assessment bias. When gaps in achievement or concerns about the fairness of an assessment are raised, the college's Lead Assessment Strategist (LAS) requests a statistical analysis from the psychometrics team, specifying the relevant groups. A psychometrician then conducts a DIF analysis to determine whether item bias exists. If the study reveals performance differences among groups for specific items, we flag those items for further review and analysis. At this point, we consider them to be potentially biased.

DIF is a psychometric concept that refers to how test items can perform differently for diverse groups while controlling for ability level. This property is essential in educational assessments to ensure fairness and validity across diverse populations. DIF occurs when individuals from different groups (e.g., gender, ethnicity, or cultural background) with equal ability levels have different probabilities of correctly answering a specific item. This suggests that the item may measure something other than the intended construct for at least one of the groups.

Statistical methods for detecting DIF include the Mantel-Haenszel procedure, item response theory (IRT) approaches, logistic regression, and various other techniques (Mantel & Haenszel, 1959; Dorans & Holland, 1993; Thissen et al., 1993; Raju, 1988; Sireci & Ríos, 2013; Osterlind & Everson, 2009). These methods help identify potentially biased items that may require revision or removal from assessments. DIF analysis has become a standard practice in high-stakes testing, ensuring that assessments provide valid and equitable measurements across diverse populations.

The method used in the current study is IRT-based residual DIF (RDIF), which identifies differences between observed dichotomous item responses and IRT-predicted probabilities of correct responses (Lim et al., 2022). More specifically, DIF is quantified as the difference in model misfit between groups. The advantages of this approach over other IRT methods include efficiency, simplicity, and straightforward

computations, as well as greater accuracy for smaller samples and a lower Type I error rate than other DIF methods (Lim et al., 2022).

If the results of a DIF analysis indicate that specific items exhibit performance differences across groups, we flag them for further review. Next, items with DIF are submitted to subject matter experts and assessment developers who conduct a content review to confirm bias. Reviewers use a checklist of potential sources of bias in item content to identify and address these issues. This checklist includes:

- Content interpretation (i.e., does the content appear to advantage or disadvantage any particular demographic group?).
- Lack of clarity or transparency of item stem or task requirements
- Obscure or overly technical language or vocabulary, inappropriate words or phrases (e.g., slang or regionalisms).
- Undue emotional distress (e.g., depictions of suffering, loss of life, loss of job), technology, and accessibility concerns.

If the reviewers identify a problem with the item, they indicate which issue(s) are present. However, if no source or explanation of bias can be determined, then this is noted in the review results.

If bias is confirmed during the review, a group of assessment experts (including the psychometrician) meets to discuss next steps. Next steps typically involve replacing the problematic item with a similar, high-performing one while determining how to improve it. If the item can be improved, it is modified and field-tested before being included in an updated version of the assessment. If no bias is confirmed during the review, the assessment remains unchanged. At this point, it is necessary to consider other potential causes of attainment gaps. A more in-depth overview of next steps beyond the evaluation of assessment quality is included in the discussion section of this paper.

Methods

This study used operational course and assessment data to evaluate the assessment-fairness component of the Equity Framework. We conducted a residual DIF (RDIF) analysis, followed by bias and sensitivity reviews, on summative assessments in 30 courses that were flagged by institutional equity monitoring as showing attainment gaps between historically underrepresented racial/ethnic groups (focal group) and White students (reference group) during the study window (January 1, 2023-January 1, 2024).

Course Selection and Attainment-Gap Identification

Courses were identified through routine equity monitoring dashboards that track course completion rates (CCR) and first-attempt pass rates (PR) by subgroup. For this study, a course was considered to show an attainment gap when the focal group's CCR or first-attempt pass PR was persistently lower than the reference group's, and

sample sizes were sufficient to support stable estimates for both outcome monitoring and DIF screening. Because traditional pre-entry achievement indicators (e.g., high school GPA, SAT/ACT) are not consistently available for adult learners in a competency-based institution, we contextualize course outcome gaps using available readiness indicators (e.g., Course-planning and Learning Tool (CPLT), Personal Learning Guide (PLG)), where applicable. Summary subgroup differences and content review results for the analyzed assessments are provided in Table 3.

Table 3
Summary of DIF Screening and Content Review Results

Outcome	Result
Assessments with ≥ 1 statistically flagged item	5 of 32
Total items flagged for expert review	20
Percent of items flagged within affected assessments	2%-10% (M = 3.5%)
Items confirmed as content-biased by expert review	0
Pattern of DIF across schools and assessment types	No consistent pattern observed
Primary interpretation for equity decision-making	Assessment content bias was not substantiated; prioritize investigation of other contributors to attainment gaps (e.g., opportunity-to-learn, learning resources, readiness supports, instructional interactions)

We estimated DIF using an iterative purification procedure, treating a provisional set of anchor items as DIF-free to link groups onto a common scale. Next, potential DIF items were flagged and removed from the anchor set, the model was re-estimated, and this process was repeated up to three iterations or until the flag set stabilized. Anchor items were chosen to be content-representative across domains and difficulty, showed good statistical fit and discrimination, and had no prior DIF indications; anchors were reviewed for content fairness before inclusion. Assumptions were checked (essential unidimensionality via parallel analysis, local independence, and item fit). Multiplicity was addressed through a stringent alpha level ($\alpha = .01$), and confirmation was achieved through content review. Analyses were conducted in R (package details available upon request).

Participants

Participants were approximately 30,000 students with first-attempt records on one of 32 summative assessments across 30 courses flagged for attainment gaps. The 30 courses span all schools within the university (Business, Information Technology, Health, Education, and General Education). Two of the 30 courses had two

summative assessments each, both of which were included in the analysis. For each assessment, only first-attempt data were retained, resulting in 300 to 1,000 students per assessment. Participant demographics (race/ethnicity, gender, and age) for the analytic sample are summarized in Table 4.

Table 4
Participant Demographics

Characteristic	n	%
Total analytic sample	31,504	100
Age (years), M (SD)	37.28 (9.73)	-
Age (years), range	16-78	-
Gender		
Female	22,108	70.2%
Male	7,648	24.3%
Nonbinary / another identity	N/A	N/A
Unknown / not reported	1,748	5.6%
Race/ethnicity (institutional categories)		
White (reference group)	18,517	58.8%
Black or African American	3,833	12.2%
Hispanic/Latino	4,208	13.4%
American Indian or Alaska Native	347	1.1%
Native Hawaiian or Other Pacific Islander	129	0.4%
Asian	941	3.0%
Two or more races	1,052	3.3%
Other	3	0.01%
Unknown / not reported (excluded from DIF grouping)	2,474	7.9%

Student demographic data were obtained from institutional records captured at enrollment. For DIF analyses, race/ethnicity was used to define comparison groups. The reference group was students who self-identified as White. The focal group aggregated students who self-identified with historically underrepresented racial/ethnic categories (e.g., Black or African American, Hispanic/Latino, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Asian, Two or more races, or Other. Students with missing or unknown race/ethnicity were excluded from focal-reference comparisons.

Although the Equity Framework supports examining multiple demographic comparisons when sample sizes allow (e.g., gender, first-generation status, English-language background, and intersections), this study focused on race/ethnicity comparisons because they were the primary equity-monitoring trigger for the sampled courses and because subgroup sample sizes for additional identities were not

consistently sufficient for stable DIF estimation across all assessments. Expanding DIF and other fairness evaluations to additional subgroups and intersections is an ongoing priority.

Independence considerations were addressed by analyzing each assessment separately and retaining only first-attempt records within each assessment. A student may appear in more than one course dataset across the broader project (e.g., if enrolled in multiple courses during the study window), but this does not affect within-assessment DIF estimation because models are fit independently for each assessment. Future work that pools data across courses would account for clustering at the student level (e.g., via robust standard errors or multilevel models).

Measures

Approximately half of the course's summative assessments were objective assessments (OAs), specifically multiple-choice or multiple-select item formats, and the other half were performance assessments (PAs). Using Bloom's taxonomy as a reference for the cognitive level of competencies measured (Anderson & Krathwohl, 2001), most OAs target lower-level competencies that indicate the student can remember, understand, and apply knowledge. PAs, on the other hand, target competencies at higher levels (e.g., applying, analyzing, evaluating, and creating). Thus, the cognitive complexity of tasks required to demonstrate competency is typically much higher for PAs than for OAs. Although most courses had only one summative assessment, which was either an OA or a PA, two courses had one of each.

Internal-consistency estimates (coefficient alpha) across the analyzed assessments ranged from .51 to .82 (M = .69; Table 5). Because performance assessments often involve heterogeneous tasks and rubric-based scoring by human evaluators, a primary source of score consistency or inconsistency is attributable to evaluation idiosyncrasies. Therefore, inter-rater reliability rather than alpha is the most appropriate reliability index for PAs. However, given the large-scale nature of PAs at WGU, there are insufficient evaluators to provide two or more evaluations for each PA task submission. Therefore, inter-rater reliability cannot be computed for our PA task scores. Table 6 summarizes recommended reliability evidence by assessment type (e.g., internal consistency and decision consistency for OAs; scoring-process and rater-consistency evidence for PAs).

Analysis

Data for each course assessment were submitted to a DIF analysis using the RDIF method. Part of the DIF procedure involved an iterative item purification process, in which any items with DIF were removed from the overall score used to match students from the two groups, until the overall score contained no DIF items. The purpose of this step is to improve the reliability and stability of overall scores as a matching criterion and ensure an accurate comparison of group performance on a given item (Hambleton, 2006). Any OA form or PA task in which at least one item had DIF in either direction (disadvantaging the focal group or the reference group) with a p-value of .01 or lower was submitted for a more in-depth content review.

Table 5
Study Sample and Assessment Characteristics

Characteristic	Value
Study window	January 1, 2023-January 1, 2024
Courses included (flagged for attainment gaps)	30
Assessments analyzed	32
Assessment attempts included	First attempt only (per assessment)
Approximate total student records	Approximately 30,000
Per-assessment sample size (range)	300-1,000
Focal group representation (range across assessments)	Approximately 10%-40%
Assessment types	Objective assessments (OAs) and performance assessments (PAs); approximately half of the assessments in each type
DIF method and flag criterion	IRT residual DIF with iterative purification; items flagged at $p \leq .01$ and forwarded to expert content review
Reported reliability (coefficient alpha)	Range = .51-.82; $M = .69$ (reported across assessments; interpretation varies by assessment type)

Results

The DIF analysis indicated that five of the 32 course assessments contained DIF. Three assessments (one OA in General Education, one OA and one PA in the School of Business) showed potential bias predominantly toward underrepresented students, and two (one PA in the School of Education, one OA in Health Professions) showed more DIF for underrepresented students. In these five assessments, 2% to 10% of the items demonstrated DIF (mean of 3.5%). There was no consistent pattern in which types of courses or assessments were more prone to DIF across groups. Moreover, none of the items with DIF in these five assessments were confirmed by experts as

Table 6
Recommended Reliability Evidence by Assessment Type

Assessment type	Primary evidence to report	Values for this study	Notes for interpretation
Objective assessments (OAs)	Internal consistency (e.g., alpha), item statistics, and (for cut scores) decision consistency/classification accuracy	Internal consistency: *M* = .69 *SD* = .07 Range = .51-.82	Internal consistency is most interpretable when items are intended to measure an essentially unidimensional construct; decision consistency is especially relevant in competency-based settings.
Performance assessments (PAs)	Rater agreement/consistency (e.g., ICC/kappa/percent agreement), rubric structure, and monitoring of scoring quality; consider generalizability evidence	Unavailable	Traditional alpha may be inappropriate for rubric-based or heterogeneous tasks; emphasize scoring-process evidence and consistency across evaluators.

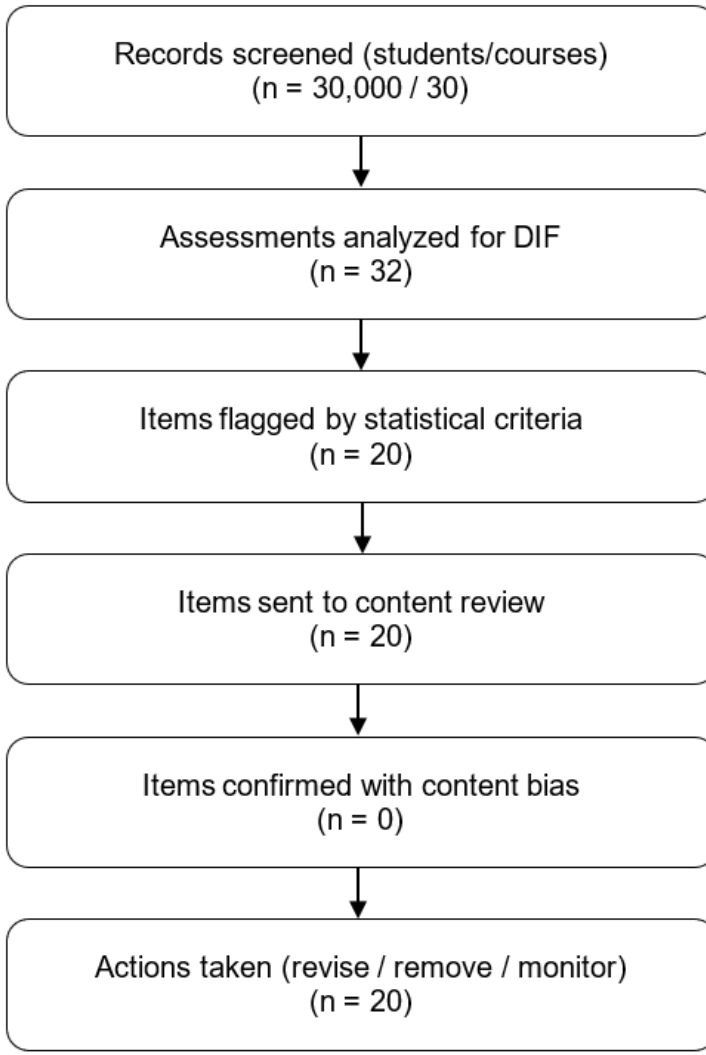
containing bias toward any particular group. The screening process for identifying and reviewing items with DIF is illustrated in Figure 1.

Illustrative Examples of Statistically Flagged Items

To protect item security, we provide paraphrased composite examples that illustrate why an item may be statistically flagged for DIF yet judged construct-relevant during structured content review:

- A scenario-based objective item describing a common workplace context. Although the focal group showed lower performance after matching on overall proficiency, reviewers determined the context was broadly familiar and directly tied to the targeted competency, with no clear cultural references or differential access assumptions.
- An objective item with dense syntax and technical vocabulary. Reviewers noted that the language could increase cognitive load, but concluded the terminology was construct-relevant and explicitly covered in the learning resources; the item was retained with a recommendation to monitor readability in future revisions.

Figure 1
DIF *Screening & Review Flow*



- A performance-assessment rubric criterion emphasizing professional communication conventions. Reviewers considered whether the criterion could introduce construct-irrelevant barriers but concluded it aligned with an explicitly taught course competency; reviewers recommended clarifying rubric descriptors to reduce ambiguity for all learners.

Discussion

This study examined whether summative assessments in courses flagged for attainment gaps showed evidence of DIF between historically underrepresented racial/ethnic groups (the focal group) and White students (the reference group). Using IRT residual DIF with iterative purification and a stringent flag criterion ($p \leq .01$), five of 32 assessments contained at least one statistically flagged item (20 items total; 2%-10% of items within affected assessments). Importantly, none of the flagged items were judged by subject-matter experts to contain content bias or other clear sources of construct-irrelevant barriers.

These results are best interpreted as evidence that, for the sampled courses and time window, assessment item content is unlikely to be a primary driver of observed attainment gaps. At the same time, DIF screening is not a definitive test of fairness. DIF statistics identify items with differential performance after matching on overall proficiency. However, differential performance can arise from multiple sources (e.g., curricular emphasis, differential opportunity to learn, random sampling error, or model misspecification). Consistent with recommendations that DIF be paired with expert review as part of routine test development and maintenance (Martinkova et al., 2017), we treat DIF as a screening signal that triggers structured content review and, when warranted, revision and re-evaluation.

More broadly, fairness is multidimensional. In addition to item-level content bias, fairness considerations include equitable treatment in administration and scoring, the opportunity to learn, and the consequences of score use (AERA, APA, & NCME, 2014). This study focused on item-level DIF and content review. It did not evaluate predictive bias or differential validity of score interpretations, which are important complements to DIF in high-stakes contexts (Sackett et al., 2008). Accordingly, the findings support a narrow conclusion about the fairness of assessment content within the evaluated assessments. They should not be interpreted as evidence that attainment gaps have been eliminated or that broader equity goals have been fully achieved.

Beyond Assessment Bias: Course-Level Interventions

When DIF flags are not confirmed as content bias, equity work should expand to other plausible contributors to gaps in course outcomes. Within the Equity Framework, the absence of confirmed content bias is used to avoid unnecessary assessment changes and to prioritize investigation of course experiences and supports that may differentially affect students' opportunity to learn and demonstrate competence. Examples of course-level follow-up questions include:

- Learning resources and course activities: Do examples, scenarios, images, and readings reflect diverse perspectives, avoid stereotypes, and remain accurate and current? Are materials accessible (e.g., readability, captions, assistive-technology compatibility) consistent with WGU principles?

- Readiness and opportunity to learn: Do readiness indicators (e.g., CPLT/PLG measures) suggest differential preparation for the competencies assessed? Are there targeted just-in-time supports that address prerequisite knowledge and test-preparation strategies?
- Instructional interactions and feedback: Are outreach, coaching, and feedback timely and actionable for students who struggle? Are there systematic differences in the types or quality of instructional interactions received across groups?
- Formative assessment and practice: Do students have sufficient low-stakes practice aligned to the summative assessment, with feedback that helps close gaps before the first attempt?

Several WGU initiatives described elsewhere in this manuscript (e.g., readiness dashboards, structured outreach, and quality-monitoring of instructor-student interactions) are intended to support these follow-up analyses and interventions. Future work should connect these supports to measurable changes in course completion and pass rates over time.

Beyond the Course: Multi-stage Interventions

Because attainment gaps reflect cumulative barriers across the student journey, course-level review is necessary but may be insufficient. Multi-stage supports can begin at enrollment and continue through program completion. Examples include:

- Early-term readiness supports: Use first-term readiness indicators to route students to targeted help centers (e.g., writing, math, study strategies, time management) before high-stakes milestones (Tinto, 2012).
- Mentoring and belonging: Strengthen mentoring structures and peer support to reduce isolation and improve persistence, particularly during setbacks (Crisp & Cruz, 2009; Walton & Cohen, 2011).
- Program-level monitoring: Pair course-level assessment monitoring with program-level dashboards that track progress, reenrollment, and key checkpoints by subgroup, triggering proactive outreach when delays emerge.

Importantly, these recommendations extend beyond what was directly tested in this study; they represent a structured response to the finding that assessment content bias was not substantiated among statistically flagged items.

Challenges

Implementing data-driven equity monitoring in operational settings involves practical trade-offs. Subgroup sample sizes can limit power to detect DIF and constrain the evaluation of additional identities and intersections. Stringent statistical thresholds reduce false positives but may miss small effects. Performance assessments introduce added complexity because scoring depends on rubrics and rubric consistency, and traditional internal-consistency indices may be inappropriate. Final-

ly, equity monitoring requires sustained institutional capacity: clear governance for acting on results, protections for item security, and ongoing investment in inclusive assessment development so that post hoc analyses complement, rather than replace, equitable design practices.

Overall, pairing inclusive development practices with routine DIF screening and expert review provides a defensible, scalable approach to assessment fairness in competency-based education. Equally important, the process helps institutions avoid overattributing gaps to assessments when the evidence instead points to opportunity-to-learn factors that must be addressed through curriculum, supports, and instructional practice.

Limitations

It is worth noting that our research reflects a single, competency-based institution and an observational design using operational data; causal claims cannot be made. Although we used RDIF with iterative purification and expert content review, anchor misspecification, model assumptions (e.g., essential unidimensionality and local independence), and measurement error may still leave residual bias. Small cell sizes for specific student subgroups may reduce power to detect DIF and yield unstable estimates. Reliability varied across assessments (especially shorter forms), and DIF detection for performance assessments is constrained by rubric structure and rater variability. Finally, stringent multiple-comparison control ($\alpha = .01$) prioritizes specificity over sensitivity and may have resulted in false negatives for small effects; therefore, results should be interpreted with these trade-offs in mind. In addition, predictive bias and differential validity of score interpretations were not evaluated in this study and should be examined in future work (Sackett et al., 2008).

Final Thoughts and Next Steps

The impact of equity ripples beyond classrooms, touching lives, communities, and future generations. By fostering a culture where every voice is heard, every perspective is valued, and every story matters, we cultivate a fertile ground for empathy, understanding, and collaboration. Through this, we unlock the transformative power of education-one that shapes minds and hearts, paving the way for a brighter, more inclusive future.

Next Steps

Institutions with diverse cultures view equity as a strategy to:

- Building a foundation - Gaining a competitive advantage.
- Establishing priorities - Assessing the university's mission and goals.
- Developing leaders - Ensuring that organizational goals are achieved.
- Designing and measuring outcomes - Building systems and supports to ensure success.

- Refining and expanding - Continuing to evolve and expand in response to student and workforce demand.

Institutions of higher learning have an opportunity to recommit to addressing the inequality gap in our nation, which is essential to education, employers, and overall workforce goals. This commitment stems from the lack of educational opportunities for women, students of color, students from low-income families, and students with disabilities, which violates the belief in equity at the heart of the American promise (U.S. Department of Education, Office of Vocational and Adult Education, 2012). America can only lead the world in meeting workforce demand if we extend opportunities to everyone fairly and equitably. As we embark on this transformative journey, let us be the torchbearers of change, the champions of equity, and the architects of a higher education landscape that truly reflects our diverse world. With equity as our guide, we can create a tapestry of equitability that enriches individuals' lives and reshapes society's trajectory.

References

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

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman.

Bensimon, E. M. (2007). The underestimated significance of practitioner knowledge in the scholarship on student success. *The Review of Higher Education, 30*(4), 441–469.

Crisp, G., & Cruz, I. (2009). Mentoring college students: A critical review of the literature between 1990 and 2007. *Research in Higher Education, 50*(6), 525–545.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11 Suppl. 3), S182-S188.

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

King, H. T. (2017). *Reinventing higher education, changing lives: The story of Western Governors University.* Western Governors University.

Ladson-Billings, G. (2006). From the achievement gap to the educational debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*(7), 3–12.

Lim, H., Choe, E. M., & Han, K. C. T. (2022). A residual-based differential item functioning detection framework in item response theory. *Journal of Educational Measurement, 59*(4), 678-701. https://doi.org/10.1111/jedm.12364

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of the National Cancer Institute, 22*, 719–748.

Martinkova, P., Drabinova, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE-Life Sciences Education, 16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning.* Sage.

Popham, W. J. (2017). *Classroom assessment: What teachers need to know* (8th ed.). Pearson.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215–227.

Shore, L. M., Randel, A. E., Chung, B. G., Dean, M. A., Ehrhart, K. H., & Singh, G. (2011). Inclusion and diversity in work groups: A review and model for future research. *Journal of Management, 37*(4), 1262–1289. https://doi.org/10.1177/0149206310385943

Sireci, S. G., & Ríos, J. A. (2013). Decisions that make a difference in detecting DIF. *Educational Measurement: Issues and Practice, 32*(2), 35–43.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response theory. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum.

Tinto, V. (2012). *Completing college: Rethinking institutional action.* University of Chicago Press.

U.S. Department of Education, Office of Vocational and Adult Education. (2012). *Investing in America's future: A blueprint for transforming career and technical education.* https://files.eric.ed.gov/fulltext/ED536889.pdf

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*(6023), 1447–1451.

Appendix A: The Assessment & Learning Equity Framework

DE&I Assessment and Learning Framework

GUIDING PRINCIPLES	LEARNING DEVELOPMENT PRACTICES	MEASURE OF IMPLEMENTATIONS	MEASURE OF IMPACT ON STUDENTS
Welcome Me *If I feel accepted and comfortable and can be my authentic self, I can focus more fully on learning and demonstrating competence.*	- Building SEL support and development into the course - Including community-building activities and tools - If content with problematic references or stereotypes must be used (such as in primary sources in history), addressing it directly with critical analysis in the explanations	- Monitor content for intersectionality (aspects of an identity that can intersect across multiple demographics) through tagging in the Digital Asset Management System (DAMS) - Iterative review of target learner profiles and clusters - Peer review by internal and external specialists through continuing research and engagement	- Persistence measures - Satisfaction ratings from students - Surveys of faculty, learners, and alumni - Reenrollment rates for learners from various backgrounds
Integrate Diverse Perspectives and Experiences *When I see that others like me have succeeded, I am more likely to persist. Seeing others who I perceive to be different helps me to grow.*	- Including images and voices in the course that demonstrate diversity and avoid stereotypes - Including career highlights that show real diversity in target occupations - Providing opportunities to overtly analyze matters related to diversity, bias, inclusion, etc. - Identifying and encompassing industry case studies that demonstrate workplace diversity standards and pedagogy	- Use DAMS tags to monitor diversity in images use as well as diversity markers in case studies, examples, and other content - Peer review of content against DE&I Standards and Guidance to prevent bias	- Persistence measures (drop rate, program change rates, grad rate) - Satisfaction ratings - Number of students who change programs (and to what) - Number of graduates in fields with goals to broaden diversity, such as nursing and IT
Make It Matter *I need to be able to trust that what I'm learning has workplace relevance, but it also needs to matter to me personally in ways to which I can relate.*	- Completing needs analysis of target learner audiences and target occupations - Utilizing the skills libraries and marketability index to ensure career relevance - Supporting and fostering connections between content and learners' identities and communities and to other cultures - Including case studies that contain various environments with diverse protagonists	- Marketability index - Mapping course content and assessments to needs analyses - Mapping of competencies and assessments to programmatic accreditation or professional standards	- Persistence measures - Factored Graduate Impact - Employer evaluation of graduate readiness across all courses - Graduates employed in field of study

WGU
PROGRAM DEVELOPMENT

Appendix A (cont.)

DE&I Assessment and Learning Framework

GUIDING PRINCIPLES	LEARNING DEVELOPMENT PRACTICES	MEASURE OF IMPLEMENTATIONS	MEASURE OF IMPACT ON STUDENTS
Remove Barriers *I deserve an equal opportunity to pursue my education.*	- Adhering to principles of Universal Design for Learning (UDL) - Providing varying access points to read, listen to, and view course material - Providing support for students to advance at an individual pace and raise the challenge level for themselves - Selecting tools that do not exacerbate the digital divide	- Review against UDL principles by internal and external reviewers - Correlation of tags for certain skills versus variety of content types in the DAMS - Review Differential Item Function and other methods of monitoring assessment fairness	- Persistence of learners with specific access needs - Course Completion Rates of specific learner personas
Personalize to My Needs *Give me your best recommendation about pathways and content for my needs and goals, but let me have the final say. Agency matters in learning, but if wrong choices are made, we should both learn from them.*	- Surfacing prerequisite skills and providing just-in-time awareness of any related foundational skills - Providing choice of examples that resonate with various learner personas - Designing constructed response assessments to allow learners to contextualize their work in ways important or familiar to them	- Review of when/where/how learner personas - Tagging and analyzing constructed response (PA) assessments that allow learners to contextualize and tracking learner behavior and success - Analytic tracking to assess the potential match/pathway for the learner relative to degree/ credential attainment to meet professional goals	- Persistence - Attempts - Course Completion Rates
Prioritize Value *My tuition and time are precious resources. Help me optimize the way I spend both.*	- Helping keep tuition low via efficiency - Providing content in multiple modalities (mobile, downloadable, text, video) to accommodate less-than-ideal study contexts - Leveraging scale to reduce per-student costs	- Reduction of learning resource costs over time - Budget management and program development efficiency measures - Correlation of tags for certain skills versus variety of content types in the DAMS - Per-student cost measures	- Persistence of low-income learners - Number of learners who cite costs as reason for dropping
Strengthen My Opportunities for Impact *I will have to function in diverse environments. I will have opportunities to promote equity. Help me learn to do both well.*	- Appropriately including in our programs DE&I competencies that prepare our graduates to work in diverse environments and promote equity - Creating high-quality learning and assessment practices to help our learners attain these competencies	- Track adoption of these DE&I competencies - Map these DE&I competencies to programmatic accreditation or professional standards - Marketable skills reports focused specifically on DE&I skills tagged to these competencies	- Student success rates on these competencies - Employer surveys regarding preparedness and application of these competencies - Student and alumni surveys

We welcome your questions and feedback.
Reach out to WGU's Program Development at programdevdei@wgu.edu. 2