THE GOVERNOR JOHN ENGLER
CENTER FOR CHARTER SCHOOLS
CENTRAL MICHIGAN UNIVERSITY

Differential Item Functioning

Context

The Governor John Engler Center for Charter Schools at Central Michigan University employs a holistic and nuanced methodology to evaluate the performance of all the charter public schools that we authorize. In making recommendations to the University Board of Trustees regarding the status of authorized schools, the Center endeavors to answer the following three questions:

1. Is the academic program successful?
2. Is the organization viable?
3. Is the Academy demonstrating good faith in following the terms of its Contract and all applicable law?

In pursuit of answering these questions, the primary components of the Center's evaluation framework include:

- Measures of student achievement and growth resulting from standardized assessments (NWEA, MSTEP, PSAT and SAT).
- Evidence collected from formal and informal site visits.
- Financial soundness, measured in part by identified key performance indicators.
- Evidence of Legal compliance and effective board governance.

The Center has developed and refined its evaluation methodology, importantly to include multiple measures that are complimentary, in order to account for the shortcomings of any one measure or metric. The Center believes that while all the components of its evaluation framework are valuable, there is often no single indicator alone that allows the Center to answer the three broad questions outlined above. Reauthorization and school closure decisions, which represent the most prominent application of this framework, are often the result of the analysis of the totality of evidence collected in these important areas of school performance and operations.

As the unit within the Center responsible for the appropriate use, analysis and distribution of standardized assessment data, it is the responsibility of the Research, Evaluation and Data Analysis team to address the questions that have arisen about the steps standardized assessment developers take to ensure that their tools are accurate and reliable across racial and ethnic subgroups.

Differential Item Functioning (DIF)

"A statistical characteristic of an item that shows the extent to which the item might be measuring different abilities for members of separate subgroups."

DIF is an important part of measurement theory and standardized assessment. DIF helps to ensure that standardized assessments measure the ability of a student independent of differences in racial and ethnic background.

NWEA

The following sections are taken from the NWEA MAP and MAP Growth Technical Manual and highlight NWEA's practices relative to DIF.

> "The fundamental assumption underlying item response theory (IRT) is that the probability of a correct response to a test item is a function of the item's difficulty and the person's ability. This function is expected to remain invariant to other person characteristics that are unrelated to ability such as gender, ethnic group membership, family wealth, eye color, or shoe size." Pg. 38

This basic assumption and necessary condition to use IRT is the reason that assessment developers like NWEA measure DIF. The objective in measuring DIF is to identify it and then minimize it. The presence of DIF has a negative impact on the validity of an assessment, meaning it may not accurately measure what it intends to measure. Assessment developers go to great lengths to eliminate it. They employ this strategy as part of their efforts focused on equity but also because it must be eliminated in order to meet the necessary conditions to use IRT. NWEA explains this necessary condition below:

> "Therefore, if two test takers with the same ability respond to the same test item, they are assumed to have an equal probability of answering the item correctly. To test this assumption, responses to test items by test takers sharing a particular aspect of a person characteristic (e.g., ethnic background, gender) are commonly compared to responses to the same test items by other test takers who share a different aspect of the same characteristic. Typically, the group representing test takers in a specific demographic group is referred to as the *focal* group. The group made up of test takers from outside this group is referred to as the *reference* group. When IRT's fundamental assumption does not hold, that is, test takers with the same ability in different groups of interest are shown to have different probabilities of correctly answering an item, the item is said to be functioning differently for the two groups. The accepted term for this effect is *differential item functioning* (DIF)." Pg. 38-39

NWEA uses a statistical method to sample items from test events, collect student response data, compare across demographic groups and classify and treat them differently based on the following 3 categories:

- A – The item shows very low levels of DIF
    - Remains on the assessment
- B – The item shows moderate levels of DIF
    - Undergoes a qualitative review from a panel of experts
    - The panel determines whether they will:
        - Revise the item and retest for DIF
        - Remove the item from the assessment
- C – The item shows high levels of DIF (bad)
    - Automatically removed from the assessment

Overall, the percentage of items that exhibit DIF between any subgroups range from 5.1% in Language Usage to 7.3% in mathematics. Given current test lengths, this equates to a potential of 2 items on an average assessment. NWEA is constantly monitoring and adjusting their item bank to minimize the number of questions exhibiting DIF.

SAT Suite of Assessments

Standards 3.1-3.5 of the American Educational Research Association (AERA), American Psychological Association (APA), and National Council of Measurement in Education (NCME) *Standards for Educational and Psychological Testing* requires test publishers to "minimize barriers to valid score interpretations for the widest possible range of individuals and relevant subgroups" (AERA/APA/NCME, 2014, p.63). To this end, the College Board, developer of the SAT suite of assessments, believes in providing all students with a fair opportunity to demonstrate their standing on the assessments they develop. For the College Board, fairness is seen as the equitable treatment of all test takers in the test administration and the equal measurement quality across subgroups and demographic populations.

The College Board takes item fairness seriously. Fairness issues are attended to at every stage of the test development process, from test design to delivery of score reports. The College Board believes that letting other factors influence the questions and test forms could cause the performances of test takers to be affected in ways that are unrelated to what they are testing (e.g. math ability). Consequently, this may result in scores that wouldn't provide accurate measure of student achievement. The College Board also believes that test takers demonstrating the same level of achievement in a content area should have similar chances of answering each question correctly regardless of gender or race/ethnicity.

The College Board conducts DIF on items, where two groups are compared on individual item performance to determine if one group is systematically performing differently than the other group. DIF analysis is done during the pretesting phase of test development and all items that the College Board use are tested for DIF. The presence of DIF indicates that an item functions differently for individuals in one subgroup from the way it functions for those of another subgroup who are at the same achievement level. In addition to DIF calculations, expert panel reviews are also used to ensure that individual items are not performing differently for different subgroups. Fairness reviewers are provided guidelines highlighting the following eight main themes (SAT Suite Assessments Technical Manual, p.23):

a) Topics to avoid
b) Portrayal
c) Stereotyping
d) Group identification
e) Language
f) Ethnocentrism
g) Regionalism
h) Testing context and its stressful nature

The Mantel Haenszel (MH) procedure (Dorans and Holland, 1993) is used for DIF analyses with the SAT Suite of Assessments. The MH D-DIF statistic ranges from negative infinity to infinity, with a value of 0 indicating no DIF. Both the MH D-DIF statistic and a significance test are used to evaluate the absence or presence of DIF, resulting in the following three categories:

- Negligible DIF – between 1.0 and -1.0 or are not statistically different from zero at .05 significance level.

- Moderate DIF – between 1.0 and 1.5 or and -1.0 and -1.5 or if they are greater than 1.5 and -1.5 and not statistically different from the absolute value of 1.0 at the .05 significance level.
- Sizable DIF – exceed 1.5 and -1.5 and are statistically different from the absolute value of 1.0 at the .05 significance level.

All items having sizable values of DIF undergo further review to determine whether some aspect of what the items are measuring is particularly related to subgroup membership and irrelevant to the dimension being tested (SAT Suite Assessments Technical Manual, p.45). When an item is identified as exhibiting sizable DIF, it is either revised and re-pretested or eliminated. Items exhibiting moderate DIF may be selected for a final form if items with negligible DIF are insufficient to meet particular specifications.

M-STEP

M-STEP addresses DIF at the beginning of their item development process by following the *Smarter Balanced 2017-2018 Technical Report*. The process by which the M-STEP addresses DIF is to train the item writers and a separate team of item reviewers to identify and avoid including bias into the items. This committee is called the Bias and Sensitivity Review Committee (BSC); items are also passed through a committee of Michigan educators called the Content Advisory Committee (CAC). The items that pass the qualitative review of both committees then proceed to field testing. The Michigan Department of Education (MDE) Office of Educational Assessment and Accountability (OEAA) uses two methods to field test items for DIF:

1. Randomly embed new items into operational computer adaptive tests.
2. If a new content area or test is being constructed then OEAA conducts separate field tests outside of the operational test.

After the data is collected, MDE performs the industry standard statistical analysis to identify DIF and review the items once more to insure they have minimized DIF. After an item passes all these steps it is considered "Ready for Operational" and included on the operational assessment as part of a student's score.

Conclusion

The standardized assessments the Center uses for evaluation purposes all employ strategies to minimize DIF. Because no solution is perfect and no standardized assessment can account for every contextual factor that effects student performance and measurement, the Center employs a multi-faceted approach to school evaluation. The holistic approach the Center uses helps to establish a body of evidence that is used to make decisions about school viability and performance.

References

MAP Growth Technical Report, 2019

Michigan Department of Education Technical Report, Michigan Student Test of Educational Progress (M-STEP), 2019

SAT Suite of Assessments Technical Manual (2017)

Smarter Balanced Assessment Consortium: 2017-18 Summative Technical Report

Technical Manual for Measures of Academic Progress MAP and Measures of Academic Progress for Primary Grades (MPG), 2011