Item Writing and Review for Bias and Sensitivity and Differential Item Functioning (DIF)

Including processes for EC, ESL, VI reviews

Defined

Item creation for the North Carolina Testing Program has an established history of inclusion of consideration for bias and sensitivity, and this has been considered as an integrated part of the development process prior to field testing. Vetting steps that specifically involve the EC/ESL/VI Specialists look for content that may present a bias or insensitivity issue such as contexts that might elicit an emotional response and inhibit students' ability to respond and contexts that students may be unfamiliar with for cultural or socio-economic reasons.

Participant Requirements

Teachers in North Carolina are the principal target population, but participants can be augmented with retired teachers and or those holding undergraduate degrees in the content area. The number of item writers and reviewers required during any item development period is determined by the need and the time allotted. All item writers and reviewers must be trained for bias and sensitivity.

Training Requirements

Item writers and reviewers must be trained on the standards and content being measured. All item writers and reviewers are subjected to extensive training on proper item design and they are also trained to consider bias and sensitivity of item content. Additionally, since the vetting process includes specific steps for EC, ESL, and VI check, training is required for these reviewers. Depending on the event and the experience of the group that is being asked to write and review, training may be best applied in a face-to-face session. However, the majority of training is designed to be delivered in self-directed online training modules.

Process and Timeline

Item writing can begin any time a change in standards has been initiated for any content that is required to be measured with a standardized test administration. See the flowcharts in the appendices for the process of writing and review that items must go through in order to be considered candidates for inclusion on either stand-alone field tests or as embedded experimental items on operational tests. Quantities and type of items per targeted standard and the time frame set by leadership of when operational tests are to exist helps determine the timeline for when items must be ready and how many item writers and reviewers are needed.

DIF Review

Defined

Per step 14 in the official SBE approved Test Development Process Flow Chart (http://www.ncpublicschools.org/docs/accountability/latestflowchart.pdf) bias reviews occur after items have been field tested and have data that supports further inspection of the items for bias or insensitivity. This is processed in steps within the online test development system (TDS) that are titled DIF Review.

The methodology used for the North Carolina Testing Program to identify items that show differential item functioning (DIF, sometimes called "statistical bias", is a concept that is different from the non-technical notion of "bias") is the Mantel-Haensel Delta-DIF method.

Calculating Statistical Bias using Mantel-Haensel Delta-DIF Method

Since the method depends on sample size, there is no single number or range of numbers that identifies an item as having moderate or more significant levels of DIF. Rather, the statistical methodology takes the sample size into account and determines whether an item should be rated as A, B, or C, according to whether it displays no significant DIF (A level), significant but still low level of DIF (B level), or more pronounced DIF (C level). A minimum number of 300 per subgroup is necessary in order to produce DIF values that are stable and do not exaggerate the counts of DIF in the B and C levels.

The current operational strategy is to reduce or eliminate the need for DIF Review by choosing not to use any item that has any significant degree of differential item functioning (C level DIF). In the rare case where an item is needed to fill test form design parameters and no A level DIF item exists, then an item in B (first choice) or C (last resort) DIF is put through an additional bias review process that content specialists coordinate.

The current subgroup analyses conducted are: Male/Female, White/Black, White/Hispanic, Urban/Rural, EDS/non-EDS.

This is the same system that the National Assessment of Educational Progress uses. For each analysis of DIF, there is a focal group and a reference group. For example in the male-female analysis, the focal group is females and the reference group is males. A plus (+) or minus (-) sign is used to indicate the direction of DIF. For example, if an item has a B- rating for the male-female analysis that means that the item slightly disfavors (minus sign) females (or slightly favors males). There may be many reasons for a B rating, and such a rating is by no means regarded as a reason to forbid the item to be on a test.

Below are some relevant links that describe the DIF methodology and related topics. The last link shows that NAEP sometimes does use items that have been flagged as having certain levels of DIF (click the individual links for the tests in the various NAEP content areas), provided that those items receive approval following the bias panel review and the subsequent content review. Ultimately, in NAEP's process, the final decision of whether to use an item is made by human beings based on all available info. It is not an automated decision produced purely by computer analyses.

- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_proced.aspx
- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_categ.aspx

- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_avoidviolat_results.aspx

Participant Requirements

DIF Review participants collectively must model the dimensions that are subject to the DIF parameters which match the Bias Review Panel participants. Since the volume of items that typically get flagged for non-A level values in the analysis that need to go through DIF Review is very small, the number of participants can likewise be a minimum set of five or six.

Training Requirements

DIF Review participants are required to go through the same training provided to the item writers and reviews and the Bias Review panel participants.

Review Process and Timeline

Tests are administered both fall and spring and the DIF analyses is done after the spring administration on combined data (fall and spring).

February through May:

- DIF reviews of DIF flagged items from the Fall

June through September:

- DIF reviews of DIF flagged items from the Spring

October through February:

- Spring base forms are assembled and embedded items are placed

DIF Review Questions

1. Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
   No
   Yes - Explain

2. Does the item contain any local references that are not a part of the statewide curriculum?
   No
   Yes - Explain

3. Does the item portray anyone in a stereotypical manner? (These could include activities, occupations, or emotions.)
   No
   Yes - Explain

4. Does the item contain any demeaning or offensive materials?
   No
   Yes - Explain

5. Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
   No
   Yes - Explain

6. Does the item assume that all students come from the same socioeconomic background?
(e.g., a suburban home with two-car garage)
   No
   Yes - Explain

7. Does the artwork adequately reflect the diversity of the student population?
   Yes
   N/A
   No - Explain

8. Is there any source of bias detected in this item?
   No
   Yes - Explain

Additional Comments:

Sample Bias and Sensitivity Training Materials

Instructions for Review

What is the purpose of this review?

After items are field tested, statistics are gathered on each item based on examinees' responses. Sometimes, the statistics indicate the possibility of Construct-Irrelevant Variance – "noise" in the item that prevents us from knowing something about the student's abilities and is measuring something else instead. Your part in this review is to judge whether the content of the item is in fact measuring something about the student other than his or her ability or knowledge in the content area that the question was intended to measure.

How were these items identified for review?

Through a statistical technique called "Differential Item Functioning" (DIF). After controlling for students' ability, are there differences in performance on the item between groups? If an item behaves differently statistically for one group of examinees than it does for another group of examinees, it is flagged for review.

The content of the items was not considered during the statistical analysis. So, these items were flagged for review because we need to determine if there is anything about these items that may be a source of bias.

What is bias?

TRUE Bias is when

- An item measures membership in a group more than it measures a content objective.
- An item contains information or ideas that are unique to the culture of one group AND this information or idea is not part of the course of study (North Carolina Essential Standards or North Carolina Common Core Standards).
- The item cannot be answered by a person who does not possess some certain background knowledge.

Sensitivity is another issue that could occur in an item. Sensitivity issues occur when

- An item contains information or ideas that some people will find objectionable or raise strong emotions AND this information or idea is not part of the course of study.
- Assumptions are made within the item that all examinees come from the same background.

Bias is NOT

- Just having a boy's name or a girl's name in the item
- Just mentioning a part of the state, country, or world
- Just mentioning an activity that is variably familiar to certain groups (e.g., vacations, using a bank)
- Just mentioning a "boy" activity (e.g., sports) or a "girl" activity (e.g., cooking) Think about: Jackee Joyner-Kersee or Babe Zaharias; Emeril or The Cajun Chef

DIF versus Bias

There is, then, a distinction between DIF and bias. DIF is a statistical technique whereas bias is a qualitative judgment. It is important to know the extent to which an item on a test performs differently for different students. DIF analyses examine the relationship between the score on an item and group membership, while controlling for ability, to determine if an item may be behaving differently for a particular group. While the presence or absence of true bias is a qualitative decision, based on the content of the item and the curriculum context within which it appears, DIF can be used to quantitatively identify items that should be subjected to further scrutiny.

Guidelines for Bias Review

All groups of society should be portrayed accurately and fairly without reference to stereotypes or traditional roles regarding gender, age, race, ethnicity, religion, physical ability, or geographic setting. Presentations of cultural or ethnic differences should neither explicitly nor implicitly rely on stereotypes nor make moral judgments. All group members should be portrayed as exhibiting a full range of emotions, occupations, activities, and roles across the range of community settings and socioeconomic classes. No one group should be characterized by any particular attribute or demographic characteristic.

The characterization of any group should not be at the expense of that group. Jargon, slang, and demeaning characterizations should not be used, and reference to ethnicity, marital status, or gender should only be made when it is relevant to the context. For example, gender neutral terms should be used whenever possible.

In writing items, an item-writer, in an attempt to make an item more interesting, may introduce some local example about which only local people have knowledge. This may (or may not) give an edge to local people and introduce an element of bias into the test. This does not mean, however, that no local references should be made if such local references are a part of the curriculum (in North Carolina history, for example). The test of bias is this: Is this reference to a cultural activity or geographic location something that is taught as part of the curriculum? If not, it should be examined carefully for potential bias.

Name of Reviewer: _____ Date: _____

When reviewing testing materials for bias, consider the following:

1. Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
2. Does the item contain any local references that are not a part of the statewide curriculum?
3. Does the item portray anyone in a stereotypical manner? (These could include activities, occupations, or emotions.)
4. Does the item contain any demeaning or offensive materials?
5. Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
6. Does the item assume that all students come from the same socioeconomic background? (e.g., a suburban home with two-car garage)
7. Does the artwork adequately reflect the diversity of the student population?
8. Other comments
9. No source of bias detected in the item