ISASP Fairness Review 2018

Documentation

Introduction

A primary goal of the *Iowa Statewide Assessment of Student Progress* (ISASP) is to ensure all students have the opportunity to demonstrate their levels of achievement with respect to the content of the assessment. The goal is to provide assessment results that are truly reflective of the achievement of all students. Ensuring test fairness is a fundamental part of validity and is an important feature built into each step of the test development process, starting with test design and integrated throughout item writing, item review, test administration, and scoring. Per the Every Student Succeeds Act, the ISASP was developed using principles of universal design for learning. This document will outline the steps taken to ensure the ISASP is accessible to all students and fair across student groups in its design, development, and analysis.

Bias and Sensitivity

According to the *Standards for Educational and Psychological Testing* (Standards), "Bias in tests and testing refers to construct-irrelevant [i.e., invalid] components that result in systematically lower or higher scores for identifiable groups of examinees" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 76; AERA, APA, & NCME, 2014, pp. 51–54). "Sensitivity" is used to refer to an awareness of the need to avoid bias in assessment. Reviews of tests for bias and sensitivity are reviews to help ensure that the test items and stimuli are fair for various groups of test takers (AERA, APA, & NCME, 2014, p. 64). The goal of fairness in assessment can be approached by ensuring that test materials are as free as possible of unnecessary barriers to the success of a diverse group of students. Iowa Testing Programs uses Fairness guidelines to help ensure that the assessments are fair for all groups of test takers, despite differences in characteristics including, but not limited to, disability status, ethnic group, gender, regional

background, native language, race, religion, and socioeconomic status. Unnecessary barriers can be reduced by following some fundamental rules (Educational Testing Service, 2016):

- Do not measure irrelevant knowledge or skills (i.e., construct-irrelevant content)
- Do not anger, offend, upset, or otherwise distract test takers
- Treat all groups of people with appropriate respect in test materials

Fairness

For review purposes, the term "fairness" can be defined as the extent to which test scores are valid for different groups of test takers. For example, if a test item contains complex language that acts as a more significant barrier for students who are not native speakers of English than for students who are native English speakers, then the item would be unfair. However, if items are more difficult for some groups of students than for other groups of students, the items may not necessarily be unfair. Fairness does not require that all groups have the same average scores. Fairness requires any existing differences in scores to be valid. An item would be unfair if the source of the difficulty were not a valid aspect of the item. For example, an item would be unfair if members of a group of test takers were offended or upset by an aspect of the item, but the item could be considered fair if the difference in difficulty reflected real and relevant differences in the groups' levels of mastery of the Iowa Core. A committee evaluated item fairness through a fairness review before the items were administered. After the items were field tested, fairness was further examined through differential item functioning.

Description of Fairness Review

Although the items are reviewed for fairness throughout the test development process (during item writing, item review, and item selection), a specific review addressing only issues related to fairness occurred in the fall of 2018. The Fairness review for all ISASP tests (all grades, all

subjects) took place remotely over the course of two weeks in November 2018. A committee of six reviewers were recruited based on ethnic, racial, and gender diversity, as well as diversity of the student population with which they have experience teaching. (One reviewer dropped out of the review process and was unable to complete the assignment). Reviewers received training on fairness guidelines, including handouts and a PowerPoint presentation that they could access during their reviews.

The trainings asked reviewers to consider the guiding principles of fairness, including to avoid construct-irrelevant variance and to allow all students the same opportunity to show what they know, as they considered each item. Specifically, in an effort to make items accessible to all groups of students, reviewers were asked to consider whether items contained the following:

1. **Unnecessarily difficult language**
   It is best practice to keep testing language simple and direct. The test should use accessible language. While the use of accessible language is particularly important for test takers who have limited English skills, it is beneficial for all test takers when linguistic competence is construct irrelevant.

2. **Unfamiliar language/vocabulary**
   The test should use language that is common. Items should avoid words or phrases that are associated with a particular social class.

3. **Regionalisms**
   Test language should not require knowledge of words, phrases, or concepts more likely to be known in some regions of the United States than in others, unless it is important for valid measurement. It is best practice to use words and phrases that are understood across regions.

4. **Sports/jargon**
   Items should not contain specialized language used by particular groups that are difficult for others to understand. Test language should avoid technical terms relating to finance, politics, certain professions, cultures, or regions. Items should not require specialized knowledge of a sport to answer.

5. **Emotional topics**
   Test content that is unnecessarily controversial, offensive, or upsetting should be avoided when possible. It is best practice to avoid topics that may evoke feelings of discomfort, fear, sadness, or anxiety in test takers.

6. **Stereotypes**
   Test content should be respectful of all people in all different groups of the population. Stereotypes attempt to classify or group people based on a single aspect, such as age, race, ethnicity, religion, income level, geographic region, or gender. Some stereotypes are blatant and easy to eliminate, while others are less obvious and require careful reading of the material and attentiveness to cultural sensitivity. Fairness and sensitivity are not limited to the groups mentioned above. It is important to avoid biased language and stereotypes for any group.

The reviews took place within Pearson's ABBI item banking system. This allowed for maximized test and item security, as the items were reviewed with the use of an assigned username and password. Reviewers only had access to the specific items intended for a fairness review, they were unable to print the items, and their comments on and ratings of the items were securely recorded within the ABBI platform. As committee members completed their reviews, their access to the ABBI platform was removed.

Each reviewer had ten days to complete the review. The reviewers were instructed to use the following categories for ratings:

- Approved: The item is approved as is, with no changes.
- Approved with edits: The item has a small issue but can be approved following edits to fix the issue.
- Rejected: The item has inherent flaws that cannot be fixed. The item should be removed from the item pool.

Any item that received a "rejected" rating would be included in a conference call with participants to discuss the issues with the items. However, no items were flagged as rejected, so this call was not necessary.

The following tables show the item ratings. The majority of reviewers' ratings for each item was used to determine the final category for the item.

Table 1
Reading Fairness Ratings

Reading					
	Accepted	**Accepted with Edits**	**Rejected**	**Total Reviewed**	**Percent Accepted**
Grade 3	28	0	0	28	100%
Grade 4	29	0	0	29	100%
Grade 5	30	0	0	30	100%
Grade 6	31	0	0	31	100%
Grade 7	32	0	0	32	100%
Grade 8	32	0	0	32	100%
Grade 9	28	0	0	28	100%
Grade 10	28	0	0	28	100%
Grade 11	28	0	0	28	100%
Totals	**266**	**0**	**0**	**266**	**100%**

Table 2
Mathematics Fairness Ratings

Math					
	Accepted	**Accepted with Edits**	**Rejected**	**Total Reviewed**	**Percent Accepted**
Grade 3	32	3	0	35	91.4%
Grade 4	35	2	0	37	94.6%
Grade 5	39	1	0	40	97.5%
Grade 6	41	1	0	42	97.6%
Grade 7	45	0	0	45	100.0%
Grade 8	44	3	0	47	93.6%
Grade 9	35	0	0	35	100.0%
Grade 10	34	1	0	35	97.1%
Grade 11	35	0	0	35	100.0%
Totals	**340**	**11**	**0**	**351**	**96.9%**

Table 3
Language Fairness Ratings

Language (Includes Writing Prompt)					
	Accepted	**Accepted with Edits**	**Rejected**	**Total Reviewed**	**Percent Accepted**
Grade 3	22	3	0	25	88.0%
Grade 4	26	0	0	26	100.0%
Grade 5	27	0	0	27	100.0%
Grade 6	25	3	0	28	89.3%
Grade 7	28	1	0	29	96.6%
Grade 8	29	0	0	29	96.6%
Grade 9	29	1	0	30	96.7%
Grade 10	30	0	0	30	100.0%
Grade 11	30	0	0	30	100.0%
Totals	**246**	**8**	**0**	**254**	**96.9%**

Table 4
Science Fairness Ratings

Science					
	Accepted	**Accepted with Edits**	**Rejected**	**Total Reviewed**	**Percent Accepted**
Grade 5	32	0	0	32	100%
Grade 8	32	0	0	32	100%
Grade 10	40	0	0	40	100%
Totals	**104**	**0**	**0**	**104**	**100%**

Table 5
Total Operational Pool Acceptance Rates

Acceptance Rate of Total Pool Review			
	Accepted without Edits	**Reviewed**	**Percent Accepted**
Reading	266	266	100.0%
Math	340	351	96.9%
Language	246	254	96.9%
Science	104	104	100.0%
Totals	**956**	**975**	**98.1%**

Summary of Fairness Review

The items were overwhelmingly approved without edits. No item was rejected. The group suggested edits for 1.9% of total items. The comments for these items were recorded, consolidated, and passed on to the content team for further review. In total, the committee accepted 98.1% of items as is, with no further edits.

Differential Item Functioning (DIF)

While the fairness review helped to ensure fairness in the content of the items, it is still possible for these items to function differently for different groups of students. DIF analyses identify items that function differently for two groups of examinees with the same total test score. In many cases, one group will be more likely to answer an item correctly on average than another group. These differences might be due to differing levels of knowledge and skills between the groups. For example, if members of one group tend to take more advanced classes or attend higher-performing schools than members of another group, then the performance of the two groups might differ on some items. DIF analyses take these group differences into account and help identify items that might unfairly favor one group over another. The items that are identified as potentially unfair by DIF are then presented for additional review.

The DIF analysis was conducted on the final edition of ISASP. Specific item-level comparisons of performance were made for groups of females and males, African Americans and Whites, Hispanics and Whites, SES-eligible and non-SES-eligible students, students with IEP and without IEP, and English Language Learners and non-English Language Learners.

The sampling approach for DIF analysis, which was developed by Coffman and Hoover, is described in Witt, Ankenmann, and Dunbar (1996). For each subtest and level, samples of students from comparison groups were matched by school building. Specifically, the building-matched sample for each grade level was formed by including, for each school, all students in whichever

group constituted the minority for that school and an equal number of randomly selected majority students from the same school. This method of sampling attempts to control for response differences between focal and reference groups related to the influence of school curriculum and environment.

The statistical analyses of items for DIF were based on variants of the Mantel-Haenszel procedure (Dorans & Holland, 1993). The DIF statistic *MH D-DIF* expresses the difference between the focal and reference groups after conditioning on the total test score; this difference is reported on the delta scale. Table 6 describes the DIF categories. The results of the DIF analyses are presented in the *ISASP Technical Manual 2018-2019*.

Table 6: DIF Classification Categories

DIF Category	**Description**
A *(negligible)*	The absolute value of the MH D-DIF is not significantly different from zero or is less than 1.0.
B *(slight to moderate)*	The absolute value of the MH D-DIF is significantly different from zero but not from 1.0 and is at least 1.0; OR the absolute value of the MH D-DIF is significantly different from 1.0 but is less than 1.5.
C *(moderate to large)*	The absolute value of the MH D-DIF is significantly different from 1.0 and is at least 1.5.

Conclusion

Fairness is a critical consideration during the test development process of the ISASP. The ISASP is designed to accurately and fairly assess students' knowledge and skills with respect to the content areas. The procedures described in this report reflect the Iowa Testing Programs' commitment to ensuring that the ISASP is accessible to all students and fair across student groups in its design, development, and analysis.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

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). Hillsdale, NJ: Lawrence Erlbaum.

Educational Testing Service (2016). *ETS guidelines for fair tests and communications.* Princeton, NJ: Author.

Witt, E. A., Ankenmann, R. D., & Dunbar, S. B. (1996, April). The sensitivity of the Mantel-Haenzsel statistic to variations in sampling procedure in DIF analysis. Paper presented at the annual meeting of the National Council on Measurement in Education, New York City.