The Achievement Gap: Test Bias or Real Differences? A Test Publisher’s Perspective

Margaret A. Jorgensen, Ph.D.

This report was previously published at the NATD symposium of the American Educational Research Association Annual Conference by the National Council for Educational Measurement in San Diego, California on April 14th, 2004.

May 2004
The Achievement Gap: Test Bias or Real Differences?  
A Test Publisher’s Perspective

According to [U.S. Secretary of Education Rod] Paige, despite the years of effort since Brown to equalize education, the achievement gap persists as our most pressing social problem. “It is the civil rights issue of our time,” he asserted. . . . “It is devastating for a child to be provided no intellectual tools, and to be set adrift with no means of finding his way back,” said Paige. “When a child is left behind, it is not just a problem for that child; it is a problem for the rest of the nation.” (Feinberg, 2004)

Introduction

As noted by Schellenberg (2004) and Sireci (2004), the achievement gap in the United States is one of the most pressing concerns being addressed in education today. Professionals involved in the development of achievement tests and the interpretation of results have acknowledged the need to address the question: “are the gaps the result of true differences in achievement or are they the result of bias in the measurement instruments?” (Schellenberg, 2004, p. 2).

Long before the present era of accountability heralded by the No Child Left Behind Act of 2001 (NCLB), test publishers had made great strides in removing bias from their assessment products. Sensitivity reviews and statistical analysis, such as analysis of differential item functioning, have significantly reduced both overt and less obvious types of bias (Nitko, 2004; Sireci, 2004). Despite all the attention and resources directed at closing the achievement gap, a disparity in performance remains. Sireci (2004) correctly notes that blaming the achievement gap on standardized testing “is like blaming the thermometer for a fever (p. 3).” Yet, Schellenberg (2004) cautions that although “the most obvious biases are no longer evident” in published tests, “those of us who work in testing should not be
lulled into a false sense of calm (p. 4).” Indeed, test publishers remain vigilant in addressing concerns of bias to produce assessment products that are reliable, valid, and fair.

Recognizing these dual concerns, this paper discusses a leading test publisher’s approach to the minimization of bias in its assessment systems. To address the second concern, that the achievement gap remains, this paper presents an approach to item design that enables assessments to not only measure a student’s academic achievement but also to provide detailed information about what learning objectives a student has not yet reached. By both serving as a thermometer and also offering educational prescriptions, standardized assessments may emerge as a tool for narrowing the achievement gap.

**Identifying the Differences**

Test publishers extensively rely on both the *Standards for Educational and Psychological Testing* (American Education Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999) and the *Code of Fair Testing Practices in Education* (AERA, APA, and NCME, 1988) to produce high-quality assessment instruments. Both documents assert that a published test must be designed and empirically proved to be both valid (by measuring the qualities the test is intended to measure) and reliable (by producing consistent measurements over time). Test publishers have a professional and ethical responsibility to remove barriers of bias so that every student has an equal opportunity to demonstrate what they know and can do. For a published test to be fair and unbiased, it must measure a student’s achievement without being affected by extraneous factors such as the student’s gender, culture, ethnicity, geography, or socioeconomic status.

Deeply underpinning this discussion is the fundamental issue of the purpose for which a test is developed. The clear purpose of educational assessments is to measure the differences in student achievement. There is no legitimate reason to build and administer a test that confirms sameness. For example, consider a norm-referenced test that does not reveal differences among a group of students, or, similarly, a standards-based test that judges every student to be “proficient” in a subject area. To administer either of these assessments seems untenable, as it would waste both valuable resources and scarce instructional time. Only by understanding the differences between students can we improve instruction for each student.

For educators, there is less insight to be gained from measurements that identify sameness than there is from measurements that identify differences. There is also the intuitive recognition that if a measured variable indicates that we are all
exactly the same, the variable itself is probably too broadly defined. If tests are expected to detect differences in the measured variable (such as achievement in an academic subject area), how do test developers ensure that bias is minimized so that only achievement differences are measured? Potential for bias must be considered at all stages of the test development process: conceptualization, alignment of tests to standards, mapping content, item construction, test format, test structure, administration directions, scoring, and score reporting (Massad, 2003).

For a test publisher developing tests for a wide range of customers, there are accepted and well-documented procedures for satisfying this mandate. Leading test publishers have adopted the solutions to bias concerns that Schellenberg (2004) and Sireci (2004) mention, especially the application of the principles of universal design to address issues for minorities and students with disabilities. The principles of universal design include equitable use, flexibility of use, simplicity, and intuitiveness so that a product can be used by the broadest range of users possible (Case, 2003). An example of a test feature that follows these principles is untimed administration. When time limits are removed from a standardized test, students can demonstrate what they know and can do according to their individual needs without boosting the scores of non-disabled students who are no longer affected by time-related strains or errors (Brooks, Case, and Young, 2003).

To address content bias during the development process, test publishers follow a set of procedures that frequently involve thorough, sequential reviews of all content. A thorough content-sensitivity review process entails having representatives of various geographic, educational, social, and ethnic groups participate in the repeated examination of content to ensure that no word, graphic, content presentation, or other attribute of the assessment is biased for or against any test taker. Table 1 presents the first in a series of checklists used to guide content reviews and shows an application of the principles of universal design:
Table 1. Item Bias and Sensitivity Issues

<table>
<thead>
<tr>
<th>Does the item provide access for the greatest number of test takers?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the item free from bias in the areas of:</td>
</tr>
<tr>
<td>Gender?</td>
</tr>
<tr>
<td>Race?</td>
</tr>
<tr>
<td>Religion?</td>
</tr>
<tr>
<td>Socioeconomic status?</td>
</tr>
<tr>
<td>Age?</td>
</tr>
<tr>
<td>Culture?</td>
</tr>
<tr>
<td>Is the item sensitive to:</td>
</tr>
<tr>
<td>Special-needs groups, such as physically disabled, visually impaired, and deaf and hard-of-hearing people?</td>
</tr>
<tr>
<td>Second-language learners?</td>
</tr>
<tr>
<td>Does the item avoid potentially offensive or disturbing information?</td>
</tr>
</tbody>
</table>

Other methods used to eliminate bias are similar to those currently presented by Sireci (2004). These methods require sufficient representative sample sizes, appropriate analytic tools, expertise in interpreting analyses, and strategies for making adjustments based on these analyses while sustaining the purpose of the assessment.

For proprietary assessments developed by test publishers, the procedures to minimize bias reflect the best practices articulated by experts and in the scientific literature. As customers are increasingly sophisticated and knowledgeable, the ultimate test of acceptability, of course, lies in their response to the final product. For custom-developed assessment products, the customer is frequently very involved in the subjective, often public review process that is the subject of open discourse. Upon signing confidentiality agreements with the test publisher and following procedures to maintain the integrity of the assessment, politicians, media representatives, teachers, and education agency staff participate in content review before and after the test is developed and administered.

**Understanding the Differences**

From the subjective review of content to the empirical review of item and test performance across groups, the processes for measuring achievement differences are established and accepted. However, these processes do not take us where we need to be in terms of understanding the differences. Can test publishers progress beyond their role in merely confirming the existence of the achievement gap and facilitate a deeper understanding of the root causes of these differences?
To address this possibility may require a shift in the traditional conception of the function and design of standardized assessment systems and individual items. What if items were constructed differently than they are for today’s high-stakes and norm-referenced tests? What if items were constructed so that the responses could be interpreted against a knowledge hierarchy? To explore this idea, let us focus on a specific type of item found in most standardized achievement tests: the multiple-choice item.

Multiple-choice items have a long history of helping educators and policymakers understand what students know and can do. However, the incorrect answer options, or *distractors*, have been far less important in the writing and editing of items than the correct answer option. Traditionally, distractors are designed to reflect typical student errors, with the reasoning behind the distractor referred to as its rationale. However, information about those errors and what they reveal about student cognition has generally not been collected and analyzed. If a central goal of educational assessment is to improve instruction, it makes sense to use assessment to understand student misconceptions. Distractors clearly present an opportunity to gather this kind of information about students.

For distractors to function well in this role, it is important for the types of errors that distractors can reflect to be clearly stated and distinguishable. Hence, it is useful to organize the types and levels of distractors that can appear in a multiple-choice item into *taxonomy of distractor rationales*. Using this taxonomy to develop an item’s distractors enables an item to reveal a pattern of misconception or weakness that a student has in the corresponding learning objective. An instructor can use this information to rapidly intervene with targeted instruction for the student.

Multiple-choice items are written to assess a particular learning standard associated with tasks ranging from simple to complex. Hence, the taxonomy of distractor rationales will reflect a hierarchy of four cognitive levels in the manner of similar cognitive models, such as Bloom’s taxonomy (Nitko, 2004). Very simple Cognitive Level 1 items that assess student recall may make little distinction between the incorrect answer options. However, for more complex items, options that reflect a complete range of the taxonomy of distractor rationales are possible.

Table 2 presents a hierarchy of the four levels of cognition and the corresponding taxonomy of distractor rationales for reading items.
Table 2. Taxonomy of distractor rationales for reading items

<table>
<thead>
<tr>
<th>Cognitive Level</th>
<th>Student Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>Makes errors that reflect focus on decoding and retrieving facts or details that are not necessarily related to the text or item. Student invokes prior knowledge related to the general topic of the passage, but response is not text-based. These errors indicate that the student is grabbing bits and pieces of the text as he or she understands them, but the pieces are unrelated to the information required to answer the question being asked.</td>
</tr>
<tr>
<td>Level 2</td>
<td>Makes errors that reflect initial understanding of facts or details in the text, but inability to relate them to each other or apply them to draw even a weak conclusion or inference.</td>
</tr>
<tr>
<td>Level 3</td>
<td>Makes errors that reflect analysis and interpretation, but conclusions or inferences arrived at are secondary or weaker than ones required for correct response.</td>
</tr>
<tr>
<td>Level 4</td>
<td>Correct response.</td>
</tr>
</tbody>
</table>
Examples of items designed to incorporate this taxonomy of distractor rationales follow:

**Objective 1: Vocabulary**

An item designed with distractor rationales that represent each level of the taxonomy:

**Read this sentence from the story “Frogs and Toads.”**

Both frogs and toads have a tail at first that **disappears** when they get older.

**Which word has the same meaning as **disappears** as it is used in this sentence?**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><strong>disagrees</strong> [Cognitive Level 1: look-alike word]</td>
</tr>
<tr>
<td>B</td>
<td><strong>vanishes</strong> [Cognitive Level 4: correct answer]</td>
</tr>
<tr>
<td>C</td>
<td><strong>can be seen</strong> [Cognitive Level 2: antonym]</td>
</tr>
<tr>
<td>D</td>
<td><strong>becomes small</strong> [Cognitive Level 3: related to the meaning, but not precise]</td>
</tr>
</tbody>
</table>

The same item designed with distractors serving their traditional function:

**Read this sentence from the story “Frogs and Toads.”**

Both frogs and toads have a tail at first that **disappears** when they get older.

**Which word has the same meaning as **disappears** as it is used in this sentence?**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>turns green</td>
</tr>
<tr>
<td>B</td>
<td><strong>vanishes</strong>*</td>
</tr>
<tr>
<td>C</td>
<td>jumps</td>
</tr>
<tr>
<td>D</td>
<td>breathes</td>
</tr>
</tbody>
</table>

All distractors are related to frogs, not to the meaning of the word “disappears.”
Objective 2: Main Idea

An item designed with distractor rationales that represent each level of the taxonomy:

What is the main idea of the story “Frogs and Toads”?

A  Frogs and toads share many differences and similarities.
    [Cognitive Level 4: correct answer]
B  Frogs and toads are cute.
    [Cognitive Level 1: prior knowledge, not text-based]
C  Toads have shorter legs than frogs have.
    [Cognitive Level 2: text-based detail unrelated to main idea]
D  Frogs are different than toads.
    [Cognitive Level 3: only part of main idea]

The same item designed with distractors serving their traditional function:

What is the main idea of the story “Frogs and Toads”?

A  Frogs and toads share many differences and similarities.*
    [Cognitive Level 3: They are all related to the main idea, but are not the best answer.]
B  Frogs live closer to water than toads.
C  Frogs and toads are like cousins.
D  Frogs are different than toads.
Objective 3: Identifying Conflict

An item designed with distractor rationales that represent each level of the taxonomy:

What is the hare’s main problem in “The Tortoise and the Hare”?

A  He does not like the tortoise.  
   [Cognitive Level 1: based on title]
B  He wants to run faster than the owl.  
   [Cognitive Level 2: incorrect character]
C  He loses the race even though he is fast.  
   [Cognitive Level 3: summary]
D  He is sure he will win so he stops trying.  
   [Cognitive Level 4: correct answer]

The same item designed with distractors serving their traditional function:

What is the hare’s main problem in “The Tortoise and the Hare”?

A  He is lazier than the hare.  
B  He falls asleep during the race.  
C  He loses the race even though he is fast.  
D  He is sure he will win so he stops trying.*

All distractors are essentially cognitive level 3: They are all related to the main problem, but are not the best answer.

Conclusion

By building assessments that provide increasingly precise information about why students choose the incorrect answer option, test publishers make a direct contribution to the improvement of classroom instruction. Once the taxonomy of distractor rationales has identified the nature of a student’s misconceptions, the results can be used to give feedback to the student and teacher as well as to
recommend corresponding learning activities. This application of the taxonomy of distractor rationales is supported by recent scientific studies in which education researchers demonstrated that assessments can positively influence student achievement when they provide students with corrective, timely feedback that references a specific level of skill or knowledge in a subject area (Marzano, Pickering, and Pollock, 2001).

If measured differences in achievement are a function of learning, and if our society hopes to eliminate these differences, it seems necessary to probe deeply into learning. Validity is defined too narrowly if it focuses solely on measuring the right thing in the right way, if it focuses only on the correct answer. Now, validity must also address how a student arrives at a wrong answer and ways that information can be used to assist with learning. Our energy must shift from separating students into groups who have learned and those who have not. It must move to a profound exploration of each student’s individual learning path. Reaching this goal will allow us to fully understand the real differences in student learning that are the root cause of the nation’s achievement gap.

References


