Construct-Irrelevant Variance in High-Stakes Testing

Thomas M. Haladyna, Arizona State University West
Steven M. Downing, University of Illinois at Chicago

There are many threats to validity in high-stakes achievement testing. One major threat is construct-irrelevant variance (CIV). This article defines CIV in the context of the contemporary, unitary view of validity and presents logical arguments, hypotheses, and documentation for a variety of CIV sources that commonly threaten interpretations of test scores. A more thorough study of CIV is recommended.

Keywords: construct-irrelevant variance, high-stakes testing, validity

Currently, achievement testing can be characterized as driven by content standards that affect the planning and delivery of instruction and the design of student assessments. The alignment of content standards, instruction, and the tests used to evaluate student learning are commonly held paradigms in education. Most states either have in place or are developing comprehensive assessment systems that have these features.

In many states, test scores can have several high-stakes uses. For instance, students must pass tests to graduate from high school or to be promoted to the next grade. Schools are evaluated based on test scores and annual progress. Low-performing schools may be subject to intervention. Teachers and school leaders may be evaluated based on student test performance, and their employment and pay may be affected by this evaluation.

This article addresses a serious problem in high-stakes testing: construct-irrelevant variance (CIV). Part 1 reviews the contemporary view of validity, with its emphasis on construct validity. In Part 2, CIV is defined. In Part 3, a taxonomy is presented that organizes sources of CIV. Evidence is presented of CIV's extensiveness in high-stakes testing. Some of these sources of CIV have received more research attention than others.

Part 1: Validity

The most fundamental step in validation is defining the construct. Cronbach and Meehl (1955) called this construct formulation. Two kinds of achievement constructs seem represented in state and national content standards.

The first kind of achievement construct can be envisioned as a large domain of knowledge and skills, sometimes called declarative and procedural knowledge. Any achievement test is intended to be a representative sample from that domain (Haertel & Calfee, 1988). Although test specifications guide the design of these tests, the sample of this domain is usually small. Each student's test score is intended to show status in this domain. This type of achievement construct can be viewed as traditional. Tests of domains of knowledge and skills are consistent with the "criterion-referenced" movement of the 1970s and 1980s. Multiple-choice (MC) formats work well to measure this kind of achievement construct.

A second kind of achievement construct focuses on a cognitive ability, such as reading, writing, or mathematical problem solving. We can conceive of this ability as consisting of a domain of complex tasks. The theoretical rationale for any cognitive ability comes from cognitive psychology. Other terms used to signify a cognitive ability include fluid abilities (Lohman, 1993), developing abilities (Messick, 1984), and learned abilities (Sternberg, 1998). Each cognitive ability involves contextualized mental models, schemas, or frames, and complex performance that may have multiple correct pathways that depend on knowledge and skills. These abilities are slow growing. They are difficult to teach and learn. The item format with greatest fidelity to measure this kind of construct is performance. Messick (1984) referred to this measurement approach as construct-referenced because the performance test focuses on the ability itself and not on the domain of knowledge and skills that support it. Mislevy (1996) traced the history of achievement testing and forecasted that cognitive psychology will lead us in a direction that merges learning theory and test theory. For our purposes both kinds of achievement constructs are susceptible to CIV, but as we will see in different ways.

refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). Validation is an investigatory process by which we (a) create a plausible argument regarding a desired interpretation or use of test scores, (b) collect and organize validity evidence bearing on this argument, and (c) evaluate the argument and the evidence concerning the validity of the interpretation. Kane (1992) described the process for establishing a plausible argument and criteria we might use to evaluate this argument. Kane (2002) also described some nuances of descriptive and policy-based interpretations in high-stakes student achievement testing. He argued that current validation does not go far enough to justify the full array of interpretations of test scores entailed by proposed uses and theory support the interpretations. Kane (1992) described the investigative process by which we (a) create a plausible argument or test score use, Cronbach (1984) refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). Validation is an investigatory process by which we (a) create a plausible argument regarding a desired interpretation or use of test scores, (b) collect and organize validity evidence bearing on this argument, and (c) evaluate the argument and the evidence concerning the validity of the interpretation. Kane (1992) described the process for establishing a plausible argument and criteria we might use to evaluate this argument. Kane (2002) also described some nuances of descriptive and policy-based interpretations in high-stakes student achievement testing. He argued that current validation does not go far enough to justify the full array of interpretations and uses.

As the stakes for testing increase, the need for validity evidence also increases (Linn, 2002). The quest for validity evidence can be very complex. This evidence will likely consist of documentation of procedures in test development, administration, scoring and reporting, and empirical studies (Downing & Haladyna, 1996; Haladyna, 2002a). The Standards (AERA, APA, & NCME, 1999) present categories of validity evidence that include content, cognitive processes, internal structure of item responses or ratings of performance, reliability, relations of test scores to other variables, and consequences. Essays by Messick (1995a, 1995b) also provide suggestions for types of validity evidence and their importance.

Although we may view validation as a process for strengthening an argument about the validity of a particular interpretation or test score use, Cronbach (1988) noted that validation should also include the testing of alternate hypotheses concerning the validity of an interpretation. Crooks, Kane, and Cohen (1996) provided a comprehensive model for the study of threats to validity. They identified eight linked inferences and argued that a weakness in any link weakens the entire chain. As we acquire and evaluate validity evidence, we may conclude that some evidence is weak or negative. By eliminating or reducing these threats to validity, we can increase our confidence that a desired test score interpretation or use is more valid.

Several writers have proposed ways of organizing and describing threats to validity (e.g., Crooks, et al. 1996; Messick, 1984, 1989). At least five major threats to validity stand out: construct underrepresentation arising form poorly conceptualized or inadequately operationalized constructs, faulty logic of the causal inference regarding test scores, negative consequences of test score interpretations and uses, lack of reproducibility of test scores, and CIV. Although all of these threats deserve attention in validation research, this article concentrates on CIV.

Part 2: Construct-Irrelevant Variance

The major point here is that educational achievement tests, at best, reflect not only the psychological constructs of knowledge and skills that are intended to be measured, but invariably a number of contaminants. These adulterating influences include a variety of other psychological and situational factors that technically constitute either construct-irrelevant test difficulty or construct-irrelevant contamination in score interpretation. (Messick, 1984, p. 216)

CIV is error variance that arises from systematic error. A good way to think about systematic error is to compare it with random error. If we were to write the linear model representing what we know about random and systematic errors, the model would be:

$$y = t + e_r + e_s$$

where $y$ is the observed test score for any student, $t$ is the true score, $e_r$ is random error, and $e_s$ is systematic error due to CIV. Lord and Novick (1968, pp. 43–44) developed this idea and presented systematic error as a redefined true score that is essentially biased.

Random error is the difference between any observed and corresponding true score for each examinee. Both classical and generalizability theory (Brennan, 2001) present methods to study random error. Random error can be large or small, positive or negative. We never know the size of random error for any examinee. Reliability is the ratio of true-score variance and observed-score variance for a set of test scores. Random error and the observed scores are random variables, whereas the true score is a constant (Lord & Novick, 1968, p. 35). Random error is uncorrelated with true and observed scores. The expected value of random error across a set of test scores is zero.

Systematic error is not random, but group- or person specific. Construct-irrelevant easiness refers to a contaminating influence on test scores that tends to systematically increase test scores for a specific examinee or a group of examinees; construct-irrelevant difficulty does the opposite. It systematically decreases test scores for a specific examinee or a group of examinees. Lord and Novick (1968, p. 43) discussed systematic error as an undesirable change in true score. The change is caused by a variable that is unrelated to the construct being measured. Thus, the change in test score is construct irrelevant. Although random error is variable from examinee to examinee, systematic error is not. It is predictable. It also manifests in two types.

The first type of systematic error is constant error for all members of a particular group. This kind of error is characterized by members of a specific group having systematic over- or underestimation of their true scores. A good example of this type of constant error is rater severity in a performance test of a cognitive ability. Two raters are consistently harsh. Student papers evaluated by these two raters will likely result in systematic error that lowers their test scores. The group being scored by these two harsh raters gets lower scores than they deserve. Test form difficulty is another example of group-specific CIV. Those taking the more difficult test form will have underestimated scores unless test score equating is carried out.

The second type of systematic error is the over- or underestimation of individual examinee scores due to a CIV source that affects examinees differentially. Messick (1989) cited reading ability on subject-matter tests as an example of this kind of CIV. If performance on a test of a construct that is not reading comprehension is strongly dependent on one’s level of reading comprehension, then reading comprehension is construct-irrelevant, because the definition of achievement does not include reading comprehension. For instance, two students having equal science achievement may differ in their test performance because one is a better reader than the other. It is reading comprehension that differentiates these two students, not science achievement. By increasing the reading comprehension...
Random and Systematic Error

On a test, test anxiety, and fatigue. The Standards (AERA, APA, & NCME, 1999) also give some examples of potential sources of CIV, such as reading comprehension, item formats, anxiety, and test administration conditions. Both Messick (1989) and Lord and Novick (1968) offer examples. This article extends and expands these discussions of sources of CIV.

Contrasts and Comparisons of Random and Systematic Error

Some contrasts and comparisons between random and systematic error may further clarify this definition of CIV. While random error is uncorrelated to true and observed scores, systematic error is correlated to both true and observed scores. This is true because both individuals and groups are either affected or unaffected by CIV. The unaffected individuals or groups have observed scores that are closer to true scores. Although the expected value of random error is zero, if CIV is present, then the expected value of systematic error is a non-zero value. Systematic error is quantifiable. The larger this error variance, the more serious the threat to validity. Lord and Novick (1968) contended that if the estimation of systematic error is impossible, then test scores may not be validly interpreted and, thus, be of no use to the test sponsor.

Estimating CIV and Adjusting Test Scores to Eliminate CIV

The estimation of systematic error can be accomplished by using the general linear model. The test score serves as the dependent variable, and a CIV source serves as the independent variable. The percentage of accounted variance provides a measure of effect size for the source of CIV. With individual-specific error, the product-moment correlation provides this index. With group-specific error, \( R^2 \) provides this index. Analysis of variance also provides a basis for computing the effect size as well as the appropriate statistical test of significance.

There are many computer programs for the adjustment and elimination of some sources of CIV arising from MC or performance tests. These include ConQuest (Adams, Wu, & Wilson, 1998), Facets (Linacre & Wright, 2004), Parscale (Muraki & Bock, 2003), and RUMM (Andrich, Lyne, Sheridan, & Luo, 1997). Generalizability theory (Brennan, 2001) offers a useful way to study the precision of test scores and sources of random error, but it also provides a basis for studying CIV as well. However, not all sources of CIV are so easily identifiable, and adjustment to reduce or eliminate its influence will be a considerable challenge to measurement specialists.

Related Ideas in the Measurement Literature

Messick (1989) devoted only several paragraphs to CIV in his influential essay on validity. Lord and Novick (1968) recognized this problem and incorporated it into their discussion of true scores. The Standards (AERA, APA, & NCME, 1999, p. 10) identifies only six standards bearing on CIV, and provides a brief discussion. The six standards and their points of discussion are:

- Standard 7.1—subgroup differences
- Standard 7.2—the validity of inferences when CIV is present
- Standard 7.3—differential item functioning, one source of CIV
- Standard 7.10—the importance of studying systematic error among groups where no differences are believed to exist
- Standard 12.19—the interpretation of test results when CIV may be present
- Standard 13.18—concerns of CIV in computerized testing.

These standards are not very specific as to the many sources of CIV identified in this article. Considering the seriousness of the CIV threat to validity for high-stakes test score interpretations and uses, it is important to expand our understanding of CIV. Chapter 7 of the Standards (AERA, APA, & NCME, 1999) addresses fairness in testing. As pointed out in that chapter, there is no common meaning or technical definition for fairness. Fairness has four connotations: (a) all tests should be free from bias, (b) all examinees are treated equitably, (c) the test scores should be equal for groups thought to be equal, and (d) students being tested have equal and fair opportunities to learn before being tested and before we make decisions based on these test scores.

Cole and Moss (1989) described a duality with the term bias. The first connotation of bias involves social justice and equal treatment of students, such as is embodied in the idea of opportunity to learn. The second connotation is a more technical issue that relates to the test as a cause of differences between groups believed to be equal in achievement. This second meaning seems to resemble what we think of as CIV. In fact, the Standards (AERA, APA, & NCME, 1999, p. 76) stated, “The term bias in testing refers to constructs that result in systematically higher or lower scores for identifiable groups of examinees.” Neither the Standards nor Cole and Moss gave us detailed, specific information about categories of CIV.

Part 3: A Proposed Taxonomy for Studying CIV and Some Evidence

We begin this section by providing a simple taxonomy for classifying variables that produce CIV and then provide logical arguments, hypotheses, and empirical evidence for each source. The extensiveness of CIV weakens the validity of interpretations and uses of test scores in high-stakes accountability systems. Our purpose is to present some documentation of sources of CIV as a serious threat to validity and call for needed research or point out where research is being conducted.

The documentation has many origins. While we rely on scholarly reports as a primary source, secondary sources include newspaper articles, and other non-scholarly periodicals. One of the most productive secondary sources is the World Wide Web. Another useful secondary source is an in-house research report, such as those from the extensive collection of the Educational Testing Service. States and testing organizations often provide such information in their archives and on Web pages. We believe the mix of evidence for each source of CIV lends credibility to the claims made about the extensiveness of CIV in high-stakes achievement testing. Occasionally, we suggest viable research hypotheses to explore a particular source of CIV and indicate what research needs to be done.

Table 1 presents the taxonomy in four broad areas. In the three columns to the right, we name the type of achievement...
Table 1. A Taxonomy for the Study of Systematic Errors Associated with CIV

<table>
<thead>
<tr>
<th>Category</th>
<th>Instances</th>
<th>Construct</th>
<th>Need for Research¹</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniformity and types of test</td>
<td>1. Whether or not students get test preparation</td>
<td>Both</td>
<td>Adequate</td>
<td>Group</td>
</tr>
<tr>
<td>preparation</td>
<td>2. The extensiveness of test preparation</td>
<td>Both</td>
<td>Adequate</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>3. Unethical test preparation</td>
<td>Domain</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td>Test development, administration,</td>
<td>Test Development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and scoring</td>
<td>1. Item quality</td>
<td>Domain</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>2. Test item format</td>
<td>Both</td>
<td>Adequate</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>3. Differential item functioning</td>
<td>Both</td>
<td>Abundant</td>
<td>Group</td>
</tr>
<tr>
<td>Test Administration</td>
<td>1. Location of test site</td>
<td>Both</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>2. Altering the administration</td>
<td>Both</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>3. Participation and exclusion</td>
<td>Both</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>4. Computer-based testing</td>
<td>Domain</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>5. Calculators in testing</td>
<td>Domain</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td>Test Scoring</td>
<td>1. Scoring errors</td>
<td>Domain</td>
<td>Inadequate</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>2. Sanitizing answer sheet</td>
<td>Domain</td>
<td>Inadequate</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>3. Test form comparability</td>
<td>Domain</td>
<td>Inadequate</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>4. Rater severity and prompt choice</td>
<td>Ability</td>
<td>Adequate</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>5. Accuracy of passing scores</td>
<td>Both</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td>Students</td>
<td>1. The influence of verbal abilities on test performance</td>
<td>Both</td>
<td>Needed</td>
<td>Ind.</td>
</tr>
<tr>
<td></td>
<td>2. Test anxiety, motivation, and fatigue</td>
<td>Both</td>
<td>Needed</td>
<td>Ind.</td>
</tr>
<tr>
<td></td>
<td>3. Accommodations for special student populations</td>
<td>Both</td>
<td>Needed</td>
<td>Ind.</td>
</tr>
<tr>
<td>Cheating</td>
<td>1. Institutional</td>
<td>Both</td>
<td>Needed</td>
<td>Group</td>
</tr>
<tr>
<td></td>
<td>2. Individual</td>
<td>Both</td>
<td>Needed</td>
<td>Group</td>
</tr>
</tbody>
</table>

¹Rating Scale: Abundant research exists; research on this topic is adequate; more research is needed. Ind. = individual.

construct associated with the source of CIV (either domain, ability, or both), give a subjective appraisal of the adequacy of research (abundant, adequate, or needed), and identify the type of CIV error (individual or group specific).

Uniformity and Types of Test Preparation

As noted previously, AERA (2000) has provided a useful set of guidelines regarding high-stakes testing that includes advice about alignment of content and cognitive processes, instruction, and assessment. These guidelines also address opportunity to learn and the providing of remedial opportunities. After assuring that these guidelines have been met, we should also consider the issue of uniform and ethical test preparation. Most testing specialists recommend test preparation (e.g., Nitko, 2001, chapter 14). Two guiding principles in test preparation are (a) no test preparation should violate ethical standards of our profession and (b) increases in test scores should be correlated with a corresponding increase in student learning (Popham, 1991).

There are many aspects to test preparation, including (a) giving advice to parents, (b) instructing students based on the curriculum represented by the test, (c) providing examples of different test item formats, (d) motivating students to do their best, and (e) teaching testwise-ness—test-taking strategies that include efficient time use, error avoidance, informed guessing, and deductive reasoning.

Whether or not students received test preparation can be a source of CIV. If some students in a reportable unit of analysis, such as a school or school district, have received test preparation and another group of these students has not, does this difference in test preparation affect the validity of test score interpretations? Differences in performance might not be attributable to sound curriculum design and appropriate and effective instruction, but to the fact that some students received test preparation and others did not.

A second type of CIV associated with test preparation is its extensiveness. There should be some evidence that all students received uniform test preparation. For instance, Nolen, Haladyna, and Haas (1992) reported considerable variation in the amount of test preparation by teachers in one state. Lomax, West, Harmon, Viator, and Madaus (1995) provided evidence of excessive test preparation with educationally disadvantaged students.

A third type of CIV involves unethical types of test preparation. In an article on preparing for a performance test, Mehrens, Popham, and Ryan (1998) offered a set of guidelines that seems applicable to all high-stakes tests. Their first guideline has to do with criterion performance being task- or domain specific. Their second guideline is that if the criterion performance is domain specific, we should not teach to the ex-
tent that the inference from the test score to the domain of knowledge or skills or a cognitive ability is inaccurate. By focusing test preparation on a subset of the domain that happens to include items appearing on the test, the test scores would be higher than deserved. If, however, the achievement construct is an ability, such as writing, it would be misleading to the consumers of test score information to teach writing in only the one genre that is assessed in a high-stakes test. These examples are instances of construct-irrelevant easiness, and may also be related to construct under representation. Specifically, unethical test preparation might include (a) developing a curriculum based on test content instead of following the state’s content standards, (b) presenting items that are similar or identical in content and format to those on the test, or (c) using published instructional materials that focus on exactly what the specific test measures or instructional practices directly aimed at the specific content of the test, pejoratively called narrowing the curriculum. All these practices can increase test scores without materially affecting the broader domain that the test samples. Studies attest to different problems associated with test preparation (e.g., Herman & Golan, 1993; Mehrens & Kaminski, 1989; Nolen et al., 1992). Mehrens and Kaminski voiced the concern that if test preparation is unethical, then the public is led to believe that students achieved more than they really did achieve.

Although test preparation should be a standardized practice in all classrooms, the variations in test preparation and the use of unethical test preparation constitute sources of CIV. We need more research on the unethical types of test preparation and better documentation in all student achievement testing programs of the role of test preparation. All students should receive ethical test preparation and the extensiveness of this ethical test preparation should be uniform.

Test Development, Administration, and Test Scoring

Item quality. While the principles of writing effective MC test items are documented in textbooks and research (Haladyna, Downing, & Rodriguez, 2002), the writing of performance test items has little scientific or research basis. What effect, if any, does the use of poorly crafted items have on item and test characteristics? A study by Downing (2002a) showed that flawed MC test items were about 7 percentage points more difficult than non-flawed items measuring the same content. Lower achieving students had greater difficulty with these flawed items than did higher achieving students. Although this problem may be more common with teacher-made tests, it points to item quality as a potential CIV threat to validity. Most professionally developed, large-scale tests are systematically reviewed and edited to remove such CIV-inducing items, but locally developed tests may have a greater tendency to exhibit this source of CIV (Downing, 2002b).

Test item format. Although research suggests that item format may not introduce CIV for boys and girls, the threat is omnipresent that some formats tend to advantage some groups of students while other formats do not. For example, Beller and Gafni (2000) found a gender-by-format interaction in 2 different years of the International Assessment of Educational Progress that essentially reversed itself. Upon closer examination of this interaction, they found that difficult performance items were more difficult for girls than for boys. Thus, under some circumstances, performance differed as a function of both the item format and item difficulty. DeMars (1998) studied the consequences of performance under varying item format conditions by gender and ethnicity. Gender-by-format interactions were reported. Rodriguez (2002) reviewed research on format differences and found, typically, that format is not construct irrelevant. However, Martinez (1999) provided evidence that occasionally item format can be construct irrelevant. Test item format continues to be a fruitful area of study of CIV.

Differential item functioning. An item shows differential item functioning (DIF) if the probability of a correct response depends on group membership and the groups are assumed to be equal with respect to the achievement construct that the test measures. Standard 7.3 specifically provides for research on DIF and the taking of appropriate action to eliminate DIF. Methods for the study of DIF are numerous. Holland and Wainer (1998) provided a thorough treatment of this source of CIV for MC tests. Penfield and Lam (2000) reviewed methods of study of DIF with performance formats and made recommendations for new methodology for the study of DIF. Thus, DIF as a source of CIV applies to both kinds of achievement constructs, domain-based and abilities. Research on DIF seems abundant, and the mapping of test items for DIF seems to be standard practice in high-quality testing programs.

Test Administration

Location of the testing site. Adverse testing conditions may be a source of CIV. The location in which a test is given is not necessarily standardized. The classroom may be the most natural environment, but when students are displaced from their classrooms to take a large-scale test, are the results the same as those from comparable students taking the test in their own classroom? Is there an age-location interaction? Younger students may perform better in their own classroom, but as they mature, they might be relocated for testing without affecting their performance. Research studies are much needed to evaluate the effect of specific testing environments on test performance.

Altering the administration. We can extend the administration time, hoping that this extended time will help students who need more time. Scores might be higher than comparable groups where administration time was not extended. To what extent do such non-standardized time extensions occur? What is the effect of this alteration on test scores? Nolen et al. (1992) reported that 8% of elementary and 3% of secondary teachers altered the administration time. Wodkire, Harper, Schommer, and Brunelli (1989) concluded after their observation of 10 kindergarten classrooms that standardized tests were administered in such a non-standardized way as to render them incomparable. They also stated that non-standardized administration may result in confusion, anxiety, behavioral resistance, negative attitudes, and other problems that caused these researchers to wonder why we are bothered to test. Research on test administration, more frequent monitoring, and independent supervision of testing are needed to ensure that altered administration does not contribute CIV.

Participation and exclusion. The percentage of participation in a school, school district, or state can contribute CIV. By excluding a group of low-scoring students, we can raise average test scores? Nolen et al. (1992) reported that 8% of elementary and 3% of secondary teachers altered the administration time. Wodkire, Harper, Schommer, and Brunelli (1989) concluded after their observation of 10 kindergarten classrooms that standardized tests were administered in such a non-standardized way as to render them incomparable. They also stated that non-standardized administration may result in confusion, anxiety, behavioral resistance, negative attitudes, and other problems that caused these researchers to wonder why we are bothered to test. Research on test administration, more frequent monitoring, and independent supervision of testing are needed to ensure that altered administration does not contribute CIV.
scores and by so doing misrepresent the achievement of a class, school, or even a school district. Thus, differences in group performance may not be based on actual achievement differences but on who was sampled and excluded.

The problem is more serious when one considers the recent policy where schools are labeled as failing as called for in the new Elementary and Secondary Education Act—No Child Left Behind (http://www.ed.gov/nclb/landing.html). A report from the Nation’s Report Card (NCES, 2002) for NAEP Science shows participation rates by states range considerably from national averages. A National Assessment of Educational Progress (NAEP) report showed participation rates for students with disabilities can vary by state from 2.6% to 6.7%. Given that these students tend to be low scoring, greater fluctuations in participation can contribute sizably to CIV (Grissmer, Flanagan, Kawata, & Williamson, 2000). Large disparities in participation rates for students with disabilities have also been observed (Erickson, Ysseldyke, & Thurlow, 1996). They stated that such variability in participation rates may be due to the need for accountability and achieving high test scores. Erickson, et al. concluded:

Such variability prohibits valid comparisons between states, and prevents policy-relevant findings to be drawn about how students with disabilities are benefiting from their educational experiences.

Without a doubt, there is an urgent need to ensure through policies and procedures that standardization exists in test participation and exclusion. Variations in these rates directly contribute to CIV when comparisons are made within any unit of analysis. Policies that provide clear guidelines regarding participation and exclusion coupled with research and documentation of uniform practices would help alleviate this threat to valid interpretations of achievement scores for schools, school districts, and state.

Computer-based testing. We would not offer computer-based testing (CBT) to any student if we thought the results would be lower than those obtained by paper-and-pencil administration. There is increasing use of CBT, but less frequently do we see documented evidence of the equivalence of CBT and paper-and-pencil testing. Huff and Sireci (2001) raised several important CIV issues related to CBT, such as student proficiency in taking a computerized test, computer platform familiarity, user interface, speededness, and test anxiety. They also noted the potential of incorrect estimates of student scores due to problems with scoring algorithms.

Another potential problem with computerized adaptive testing is the heavy demand on mid-difficult items that provide maximum information. Since these items are the most frequently used, these items quickly become overexposed, which is another source of CIV. The potential threats of CIV in the CBT environment have only begun to be explored at this time. Indeed, Standard 12.19 of the Standards (AERA, APA, & NCME, 1999) provides specific warning about the dangers of CIV related to computerized testing. Besides research reports addressing these problems, technical reports on such testing programs offer an opportunity to document that CBT does not contribute CIV.

Calculators in testing. The role of calculators in testing has been an active research topic in item development and test design (Haladyna, 2004). The plausible hypothesis is that students who have calculators have an added advantage over those without calculators in mathematics tests and in other content that may require calculation. A recent report in The Nation’s Report Card (NCES, 2002) presented results from the 2000 NAEP showing an interaction of grade level with calculator usage. More frequent use of calculators was correlated with lower scores in grade four, but the opposite was true at grades eight and 12. Also, some item types seem more susceptible to better performance by using calculators. Thus, calculator usage seems associated with CIV and the type of item being offered. The use of calculators would seem to enhance testing of many types of achievement by providing a higher fidelity experience. At the same time, the use of calculators must not be permitted to increase CIV. Thus, research is constantly needed to address each new application involving calculators or any other technological innovation.

Test Scoring

Scoring errors. Standard 11.10 reads, “Test users should be alert to the possibility of scoring errors; they should arrange for rescoring if individual scores or aggregated data suggest the need for it” (AERA, APA, & NCME, 1999, p. 115). Indeed, an epidemic of scoring errors has arisen throughout the United States. For example, in Minnesota, 47,000 students received incorrect scores, leading to serious negative consequences for these students and to subsequent lawsuits (Henriches & Steinberg, 2001). More than 20 states have been affected by scoring errors. In Arizona, 12,000 students received incorrect scores due to an error in the scoring key (Bowman, 2001). In Washington, 204,000 essays had to be rescoring. Scoring errors or delays also occurred in California, Florida, Georgia, Indiana, Mississippi, New York City, Nevada, North Carolina, South Carolina, Tennessee, and Wisconsin. In the Education Week on the Web Archives (2004), there are 8 listings for scoring error incidents. In high-stakes testing, especially where critical pass-fail decisions are made, we need stronger, more independent assurance of score accuracy and additional documentation of extra scrutiny in scoring.

Sanitizing answer sheets. “Cleaning up answer sheets” is a practice that is recommended. For instance the National Association of Test Directors (2004) provides specific examples of how answer sheets should be sanitized: “Erase all stray marks, darken light marks, and clean up incomplete erasures.” Volunteer parents may be asked to “clean up” answer sheets before scoring. For example, they might make incomplete erasures more thorough, since double-marked items are scanned as incorrect. That some schools and districts may sanitize answer sheets while other schools and school districts introduce potential CIV. The solution to this validity threat is to have all answer sheets sanitized as is recommended by nearly all test scoring services. This threat to validity is primed for studies that explore the frequency of sanitizing and its consequences on test scores.

Test form comparability. The equating of test forms is a standard practice in testing programs. There are many methods for adjusting test scores so that one test form is no more difficult or easy than any other test form. However, it is possible that errors can occur in equating studies. Although research on equating methods is active and important, we have few mechanisms and little documentation for ensuring that equating is
done accurately. Without adequate quality control and technical reports that thoroughly document these procedures, such errors can go undetected and introduce CIV.

A related problem is test-score drift. From the time of first administration of a new test, test scores tend to increase for comparable groups as a function of test age. It would be easy and presumptuous to conclude that these increases are due to improved learning. However, this growth should be validated by comparison to growth on another standardized achievement test, such as the NAEP. In a study by Linn, Grane, and Sanders (1990), comparisons of publishers’ standardized achievement test results with NAEP test results suggested that test scores may depend on non-achievement factors, such as item exposure. Test-score drift may be related to unethical test preparation previously discussed or to other methods used to increase scores for some examinees in a construct-irrelevant way. The problem of drift implies that we have continuing problems with the accuracy of equating.

Rater severity and prompt choice.
The threat of CIV when equating performance tests principally comes from rater effects, particularly rater severity, and the issue of prompt choice in writing assessments. Linn, Baker, and Dunbar (1991, p. 8) stated, “The training and calibration of raters is critical in this regard.” The practice of letting students choose writing prompts assumes that prompt choice does not contribute to CIV. Testing specialists have argued that prompt choice indeed introduces CIV (Linn, Betebenner, & Wheeler, 1998; Wainer & Thissen, 1994). Fortunately, we have an extensive literature on rater severity and a growing literature on the problem of prompt choice (Engelhard, 2002). However, this research literature has not yet affected practice of scoring performance tests and removing CIV due to rater severity, other rater effects, and prompt choice.

Accuracy of passing scores. The establishment of a passing score for a pass-fail decision or the setting of multiple points on a test score scale for setting benchmarks, (as in the NAEP) is usually a judgmental process involving subject-matter experts (SME). The threat of CIV is that the SME group, which recommends a passing score or set of benchmarks, may not be representative of the population of content experts from which a committee was convened.

Another dimension to this problem arises when a jurisdiction applies standards in non-uniform ways. States vary in the way they identify and label “failing schools.” While Michigan scores above the national average on the NAEP, it reported the most failing schools (40%), while states like Arkansas and Wyoming reported no failing schools. Thus, we have a clear indication that CIV is present in standard setting. Rothstein (September 8, 2002) reported that the different standards employed in states to designate schools as failing has resulted in some states engaging in more busying so that students can change schools.

These differences in passing standards and standard-setting outcomes cause us to question the extent to which the method of standard setting may contribute CIV. A confounding factor is the type of achievement test used in a state and its connectedness to instruction. Based on the disparities among states on how failing schools are identified and labeled, one is inclined to conclude that standards are indeed very arbitrary. Research is sorely needed into the validity of standards. Part of this validity research should examine the consistency of judges who set standards and potential bias introduced by panels of judges who may differ from the population of potential judges. Consistency of ratings is another issue within this need for research on standard setting.

Students
Compared with others sources of CIV, students potentially provide the most serious CIV threat to validity. In these instances, students comprise individual-specific CIV versus the group-specific CIV.

The Influence of Verbal Abilities on Test Performance
Another important source of CIV may be the demand for verbal abilities needed in the measurement of another ability. By verbal abilities, we mean reading, writing, speaking, and listening. Ryan and Demark (2002) hypothesized four types of achievement constructs:

1. A construct, such as writing, where the highest fidelity measurement technique is a performance test.
2. A construct where verbal abilities are very important in the performance test. A good example of this would be an advanced placement test that requires reading a passage, exercising critical thinking, and writing ability.
3. A construct where verbal abilities are used but are not deemed crucial in the performance task. This construct might involve a domain of knowledge and skills with a low cognitive demand. Minimal reading and writing skills may be needed to perform adequately on a test of this kind of construct.
4. A construct where there is little demand for reading and writing abilities. This construct might involve computation, symbolic representation associated with chemistry or physics, or one of the performing arts.

To what extent do deficits in reading interfere with performance on an achievement test? Research involving students with limited English proficiency (LEP) by Abedi, Lord, Hofsstetter, and Baker (2000) showed that vocabulary has a powerful influence on test performance. They experimented with NAEP test items, using simplified vocabulary to improve students’ test performance by allowing students to use a glossary and permitting students extra testing time. Fitzgerald (1995) characterized students with LEP as slow readers, whose test performance is obviously impaired with time limits. Garcia (1991) showed that LEP students may have trouble with the familiarity of mainstream American topics, contributing to their lower performance. Thus, we have increasing evidence that reading comprehension plays a vital role in test performance. Students with LEP are probably most affected by deficits in reading comprehension. The measurement of achievement in other subject matter may be hopelessly contaminated by their deficiency in reading comprehension. However, this problem is not limited to LEP students. Non-LEP students can also have low reading comprehension. The treatment of this threat to validity should include a careful consideration of the definition of the construct and the role that verbal ability plays in this definition. For the measurement of some abilities, reading and writing at a high level are necessary, and for other abilities, the demand for these verbal abilities is less important.
With so many students being deficient in these verbal abilities, the threat of CIV in these challenging performance tests suggests that research on this problem is very much needed.

Test Anxiety, Motivation, and Fatigue

We know that test anxiety can increase test performance but more generally lowers test performance. In a meta-analysis of 562 studies, the pattern of student performance in relation to test anxiety is unmistakable and conclusive (Hembree, 1988). Test anxiety can be pernicious in three ways. First, it affects many test takers. Test anxiety is estimated to include about 25% of the general population (Hill & Wigfield, 1984). Second, test anxiety can be exacerbated or reduced by imposing certain conditions on the examinees. Hancock (2001) provided experimental evidence in a study with college students that an evaluative threat can increase test anxiety. Zohar (1998) also provided complex experimental evidence that disposition to anxiety and the high-stakes situation contribute to test anxiety. Third, test anxiety can have consequences. For example, Thornton (2001) reported that teachers in training in Great Britain have been so intimidated by teacher testing that they are dropping out of their teacher education programs and making alternative plans. As we can see, not only is test anxiety a powerful source of CIV, it also affects students and preservice teachers.

The motivational level of students may affect test score performance, no matter the achievement level of the student. The manifestation of low motivation may be non-compliance with the test-taking protocol. Students may seriously underperform, make random marks on the answer sheets, omit answers, or not finish the test. The frequency of omitted responses and items not reached are signals of low motivation and non-compliance. Paris, Lawton, Turner, and Roth (1991) found that younger students take large-scale tests more seriously than older students. Schools and school districts take very different approaches to motivating students to perform on these tests. Tactics include threats, parties, prizes, awards, and pep rallies. Whether the tactic is positive or negative, knowing the extensiveness of these practices and the degree of the influence of each of these motivational tactics is important. If the motivational strategies work, then test scores contain CIV, because not all students or schools receive uniform motivational stimulation from school leaders. What may be accounting for differences among schools or school districts might not be real learning, but more effective motivational techniques. Although these motivational techniques are desirable, these techniques should be uniformly applied to ensure that motivation does not become a source of CIV.

While there is no research to report about fatigue in testing, we hypothesize that young students may be more susceptible to fatigue in long testing situations than older students, and the conditions for test administration may interact with different types of students. The effects of fatigue are not well understood or studied, but should we be concerned with the energy level of students as they take long, high-stakes tests? Or is fatigue not a factor in test performance? A related area of concern is the extent to which students eat before testing and are allowed breaks and snacks during a long testing day.

Although we have a promising emerging science of person-fit analysis (Meijer & Sijtsma, 1995), we do not routinely study item response patterns of students to find out if students' response patterns suggest anxiety, poor motivation, or fatigue. Some students are plodders who work slowly and correctly but do not finish tests in the allotted time. Studies of examinee fit ought to be routine in large-scale, high-stakes assessments, and evidence supporting any of these student sources of CIV should validate the scores or cause us to look for reasons for underperformance other than inadequate learning. Another aspect of this problem is non-response, items omitted or not reached (Koretz, Lewis, Skewes-Cox, & Burstein, 1993). The frequency of omitted and not-reached items should signal potential problems with test anxiety, motivation, fatigue, or timing. Yet, there is surprisingly little research on threats to validity.

Unique Problems of Special Populations

Keeping in mind the admonitions of Messick (1984) that CIV can contaminate both interpretations of test scores and implications we make from knowledge of test scores, we confront the unique problems associated with four often co-mingled populations: students with disabilities, LEP students, students living in poverty, and students living in cultural isolation.

Students with Disabilities or LEP

Federal guidelines and the new Standards (AERA, APA, & NCME, 1999) give considerable attention to the necessity of altering the administration conditions or the test itself to eliminate a disability as a source of CIV. As discussed previously, reading comprehension may be a serious source of CIV. With LEP students, this type of CIV is likely to occur. Chapters 9 and 10 of the Standards provide considerable discussion and offer many standards bearing on what is needed to eliminate CIV when testing students with disabilities and students with LEP. Policies of excluding these students from assessments vary not only within classrooms and schools, but also across school districts and states. Federal law requires that students with disabilities be included in assessments, but the law does not explain which accommodations are acceptable or specify the criteria for accommodation. If such accommodations are carried out uniformly in all school districts and states, then differences in performance will not be due to this source of CIV. Until we have full participation and more uniformity in the way accommodations are offered, comparisons of performance of students with disabilities and LEP among units of analysis such as classrooms, schools, and states cannot be considered reasonable or validly interpreted.

Students Living in Cultural Isolation

The measurement of achievement of students living in culturally homogeneous, isolated communities can be affected in many ways. For instance, students living on Native American reservations have a variety of characteristics that work against effective test performance (Haladyna, 2002b). These students, too, need accommodations in testing and, in some circumstances, alternative assessments. The same case might be made for racial or ethnic communities that live in isolation from the rest of society. While test scores may traditionally be low for these groups, the lack or failure of accommodations and modifications in assessments might account for some of this low performance. The threat of CIV for these populations is similar to that of students with disabil-
and with LEP. Research on the motivation and preparedness for achievement tests for these special population students can reveal much about how they learn and why their performance is lower than expected. Such studies can benefit from methods where students are actually interviewed after the test or “think-aloud” during a test-taking session.

Cheating

Any deception committed to misrepresent a group’s or a student’s level of achievement is cheating. Institutional cheating is a deception used to misrepresent student achievement for a class, school, school district, or state. Some examples of institutional cheating include teachers getting an advance copy of the test and planning some lessons based on specific items or providing students “practice” with actual, secure test items (Popham, 2003); reading answers to students during the test or helping students select correct answers; giving hints to correct answers in the classroom during the test or changing wrong answers to right answers after the test is given. Unintentionally excluding low-achieving students from a test may raise the average score for a unit of analysis, but intentionally excluding such students can also be considered cheating. This problem was a topic of discussion over a decade ago (Cannell, 1989). One of the most publicized cheating scandals occurred in the New Orleans public schools (Meitrodt & Nabonne, 1997). A variety of score-boosting techniques were used that included exclusion of low-scoring students, unethical test preparation, early distribution of tests and subsequent teaching of the test, and unusual test administration arrangements. We have witnessed outbreaks of cheating in Austin, Texas; Chicago, Illinois; New York City; Reston, Virginia; and Rhode Island (Hoff, 2000).

Cizek (1989, pp. 62–69) provided extensive evidence of institutional cheating in schools and school districts. The trustworthiness of student achievement data should be questioned, particularly when large gains are observed for a school or school district, as in the New Orleans scandal. Like other sources, institutional cheating in a school or school district level introduces CIV that is difficult to detect and document.

Individual cheating is also a very serious source of CIV. We have many examples. The Educational Testing Service (ETS) conducted an investigation of student cheating on one of its tests that led to the arrest of 61 students who were accused of fraudulent test taking (Li, 2002). This type of cheating involved hiring professional test takers as substitutes. There have been recent incidents of electronic devices used to copy test items or pirate test items. In China, 64 students had their scores invalidated for copying from other sources in a performance test for the Graduate Record Examination (Taipei Times, 2002). Two students from Columbia University were arrested for using high-tech transmitters and walkie-talkies to cheat on the Graduate Record Examination (Chronicle of Higher Education, 2002). This small sample of incidents shows that CIV from individual student cheating is a worldwide problem.

The volume by Cizek (1999) is a milestone in the study and documentation of cheating as a source of CIV. Nonetheless, more research is needed into the motivation for institutional cheating and its extensiveness particularly within high-stakes accountability testing programs.

Conclusion

Every interpretation or use of an achievement test score in a high-stakes environment is vulnerable to many validity threats, such as inadequate construct definition, construct under representation, illogical reasoning regarding the causes of student learning, negative consequences of test score uses, and low reliability of test scores. CIV is one of these threats to validity. We have shown that CIV has many sources, and evidence has been presented of its extensiveness. We have argued that CIV needs increased attention, especially in high-stakes testing programs.

Researchers should systematically address understudied sources of CIV. Identification and assessment of the seriousness of each source in high-stakes testing programs is the first step. This research should build a robust literature that provides a clear picture of the seriousness of this threat to validity.

Documentation is needed to assure the public that each specific threat to validity is not serious for the particular score interpretations proposed. In instances where this documentation shows that a threat is serious, we need to take appropriate action to study the CIV threat, and then eliminate or reduce it. A primary type of documentation is the annual technical report. This report should contain references to validity evidence and its relationship to the appropriate standards (AERA, APA, & NCME, 1999) and detail specific CIV threats and their management.

Sponsors of state and local high-stakes achievement testing programs and test companies that develop these tests should work together to consider the seriousness of all threats to validity, including CIV. A shared responsibility and honest examination of these CIV threats are likely to improve the effectiveness and increase the validity evidence for these testing programs.

References


Spring 2004

26


