—HIGHER—
EDUCATION
—POLICY—

Higher Education Policy, **2019**, **32**, (513–536)

www.palgrave.com/journals

Original Article

# Measuring Postsecondary Achievement: Lessons from Large-Scale Assessments in the K-12 Sector

Daniel Koretz
Harvard Graduate School of Education, 415 Gutman Library, Cambridge, MA 02138, USA.
E-mail: daniel_koretz@gse.harvard.edu

Interest in using large-scale standardized assessments in the postsecondary sector has been growing rapidly in recent years. However, our experience is still limited, and there is a serious dearth of research investigating the characteristics and effects of testing in the postsecondary sector. We have far more extensive experience with large-scale testing in the K-12 sector, particularly in the USA. In this paper, I discuss a number of important issues that have arisen in K-12 testing and explore their implications for testing in the postsecondary sector. These include mistaking the part for the whole, overstating comparability, adding functions to extant tests without sufficient justification or validation, Campbell's Law, and unwarranted causal inference. All of these issues are relevant to assessment in the postsecondary sector, and some are more severe in that sector than in K-12 education. I end with recommendations for productive and appropriate uses of assessments in this sector.
*Higher Education Policy* (2019) **32,** 513–536. https://doi.org/10.1057/s41307-019-00142-4; published online 24 April 2019

**Keywords:** testing; assessment; high-stakes; large-scale testing; monitoring; accountability

---

Over the past few decades, interest in standardized measurement of the learning of postsecondary students has expanded rapidly in many nations. In this paper, I explore implications of the more extensive experience with K-12 testing for the use of assessments in the postsecondary sector. I draw substantially on experience in the USA because of the large US research literature, but the issues I describe are not specific to the US context. I discuss three broad issues:

- Matching assessments to inferences, that is, to the conclusions that are based on test scores;
- Campbell's Law — that is, the corruption of indicators induced by accountability — as it is manifested in educational assessment; and

- The problem of unwarranted causal inference, in particular, inferring the contribution of educational institutions to measured achievement.

Attending to these lessons from the K-12 sector can lead to more appropriate and effective uses of large-scale tests in the postsecondary sector.

I begin by presenting a framework for understanding the validity of conclusions based on test scores. I then provide evidence from the K-12 sector bearing on these issues. In final sections, I comment on the manifestation of these issues in the postsecondary sector and offer recommendations.

## Background

To start, it is necessary to clarify the limits of this discussion because the term "assessment" is used in many ways. In this paper, I address only direct measures of student achievement. Moreover, I limit the discussion to large-scale, external, summative assessments — that is, assessments that are developed outside of the educational institutions in which they are administered and that are designed to evaluate students' skill and knowledge. As shorthand, I will refer to assessments of this sort as direct assessments of student learning.[1]

There is a long history of assessments of other types in the higher education sector, including measurements of student inputs, resources, other institutional characteristics, and outputs (e.g., Astin and Antonio, 2012; Secolsky and Denison, 2012). Measurement of outputs, however, has largely focused on variables other than direct measures of student learning, e.g., degree completion rates (e.g., Moore *et al.*, 2014; Williams, 2014).

Nonetheless, concern about the performance of postsecondary students and efforts to measure it are not new. For example, Shavelson (2010) dates the first use of standardized postsecondary achievement tests in the USA to the first third of the last century. Placement tests have long been administered to incoming students in many institutions, both junior colleges and some senior colleges (e.g., the City University of New York). A number of postsecondary achievement tests have been developed and marketed in the USA — for example, the Collegiate Assessment of Academic Proficiency (CAAP) developed by ACT (but retired and not replaced in 2018; Allen, 2018), the HEIghten Outcomes Assessment Suite distributed by the Educational Testing Service, and the Collegiate Learning Assessment, distributed by the Council for Aid to Education. Moreover, policymakers' interest in external measures of postsecondary achievement and efforts to develop new postsecondary assessments have grown rapidly in recent years (e.g., Coates and Mahat, 2014; Judd and Keith, 2012; Shavelson, 2010; Yamada, 2014). Examples include the report of the U.S. Department of Education's *A Test of Leadership: Charting the Future of Higher Education* (U.S. Department of Education, 2006) and the OECD's

Assessment of Higher Education Learning Outcomes (AHELO) pilot study (Tremblay *et al.*, 2012).

Yet for all that, the use of standardized, external assessments remains limited in the postsecondary sector, as compared with the elementary and secondary (K-12) sector. In the USA, for example, the only external standardized tests taken by a large proportion of postsecondary students are admissions tests, which of course measure the learning of students before they enter a postsecondary institution. External testing of postsecondary achievement has been limited in large measure to research efforts, pilot programs, and testing programs adopted by a modest number of institutions (Shavelson, 2010). In contrast, large-scale external achievement tests have long history in the K-12 sector. Standardized achievement testing has been widespread in the US K-12 system for over half a century, and tests used both for selection into educational programs or schools and to certify completion of secondary education have long been in place in many other nations. During this time, the field has accrued extensive experience with a wide variety of assessments that have been used to serve diverse functions. Moreover, large-scale K-12 assessments have been the focus of a great deal of research investigating both psychometric characteristics of the assessments and the effects of their use on educational practice. Many of the issues raised by this experience and research in the K-12 sector apply to assessment in the postsecondary sector as well.

## A Validity Framework for Postsecondary Assessments

To understand the lessons from K-12 assessments, it is necessary to begin with a formal conception of validity. The framework I use here follows Koretz and Hamilton (2006). It extends the standard discussions by Messick (1989) and especially Kane (2006) to further clarify the effects of test *use* on validity — a critical consideration given the widespread interest in using assessment for purposes of monitoring and accountability in the postsecondary sector.

The term validity is used inconsistently in the measurement field. Much of the field uses the term to include both the justification for the inference based on a score and the effects of testing (e.g., Messick, 1989). I and others have argued that using this single term to refer to both the justification for an inference and impact is counterproductive (e.g., Cizek, 2016; Koretz, 2016). The justification for an inference and the effects of testing are largely independent; for example, a test may support a given inference well even if using the test has negative effects. Moreover, different evidence is required to evaluate impact and the justification for the inference (Cizek, 2016; Kane, 2016; Koretz, 2016). Finally, using "validity" to include impact is inconsistent with conventional English usage and therefore confuses non-technical audiences (e.g., Koretz, 2008). Therefore, although I will

also discuss the effects of testing in this paper, I will use the term validity only to refer to the justification for the inference based on scores.

We can call the construct about which one is making an inference the *target of inference*. In the traditional language of achievement testing, the target is an inference or conclusion about a *domain* of achievement, such as "first year calculus" or "mastery of mathematics over the first 11 years of schooling." Because these domains are usually large and only limited time can be devoted to testing, most of the domain remains untested. For example, for years, Massachusetts has required students to pass two tests, one in mathematics and one in English language arts, in order to receive a high school diploma. These tests are first administered at the end of tenth grade. The portion of the math test that determines whether a student has mastered enough mathematics to deserve a diploma comprises only 42 test items (Massachusetts Department of Elementary and Secondary Education, 2018). Therefore, we are compelled to draw an inference about mastery of the domain on the basis of the severely restricted sample of student performance elicited by the test. In the current terminology in the measurement field, we must *extrapolate* from the limited sample of tested performance to the much larger domain from which it has been sampled.

Unfortunately, the tested sample is not only much smaller than the domain; it is also usually not fully representative of it (Kane, 2006; Koretz, 2008). One reason is that some parts of many domains are difficult or even impossible to assess with externally imposed, standardized assessments. For example, in a large-scale test administered in many different jurisdictions, it is difficult to assess students' ability to solve complex problems using a mix of familiar and novel information because students in different locations may have markedly different background knowledge. For that reason, some students may solve a problem that appears to require complex reasoning using simple recall (see, e.g., Hamilton *et al.*, 1997). However, even the portions of the domains that are practical to assess in large-scale standardized assessments are typically far too large to be tested exhaustively. Moreover, even once substantive material has been selected for measurement, the test authors must select from possible ways of presenting the material, response demands for examinees, and scoring procedures (e.g., Holcombe *et al.*, 2013), and this sampling is not random.

The sampling used to create tests often differs systematically among tests of the same domain. Some of these differences reflect intended uses. For example, a secondary school mathematics test used for college admissions is likely to give greater weight to advanced content than a test used to evaluate the performance of the entire school population, and it is likely to be designed to provide greatest precision at a higher point in the performance distribution. Some other differences in the selection of material for testing reflect judgments about the relative importance for the target of inference. However, many of the systematic choices do not have a substantive rationale of this sort (e.g., Holcombe *et al.*, 2013).

Under low-stakes conditions — that is, when neither educators nor students feel substantial pressure to raise scores as an end in itself — the incompleteness of tests has two main consequences. The first is simple measurement error: The performance of a student or a group of students will vary depending on the sample of content included in the test. The second is the risk of *construct underrepresentation*, that is, the exclusion of parts of the domain that are important for the intended inferences. The impact of construct underrepresentation may be modest under some circumstances, but it can be large. For example, one US state evaluated teachers in grades 3 through 8 based on students' gains on a basic skills test. In more advanced eighth-grade mathematics classes, however, students spend little if any time on basic skills; instead, they are studying material such as algebra. Because much of what they study is not measured, the estimated effectiveness of their teachers is downwardly biased. As I will discuss below, the problem of construct underrepresentation becomes much more severe when test scores have consequences.

The reverse of construct underrepresentation, generally called *construct-irrelevant variance* (i.e., variation in student performance that is unrelated to the construct the test is intended to measure), arises when a test measures something that is not included in the target domain. This may arise if the tested sample includes irrelevant content — for example, if the curriculum for a mathematics course includes no trigonometry but the external test does. However, construct-irrelevant variance can arise for other reasons as well, and it can pose a fundamental threat to accurate measurement of student performance when tests have consequences.

To extend this conventional framework to high-stakes tests — that is, to tests on which individuals feel pressure to raise scores — it is necessary to distinguish between *substantive* and *non-substantive* attributes of tests. Both the target and the test comprise sets of *performance elements*. This deliberately general term refers to all of the aspects of performance that affect scores on the test or inferences based on it. *Substantive* performance elements are directly related to the inference. For example, one of the substantive performance elements in a test administered to high school students in Massachusetts is finding the hypotenuse of a right triangle using the Pythagorean theorem. *Non-substantive* elements are not directly relevant to the inference but can nonetheless affect performance on the test. An example of a non-substantive performance element is a choice of item format that is unrelated to the intended inference but that does affect students' performance — for example, the decision to use a multiple-choice item rather than a constructed response item.

These performance elements vary in their impact on scores and their importance for the inference. We can call this importance *test weights* and *inference weights*.[2] It is critically important that test weights differ from inference weights. This can take the form of construct underrepresentation: Some elements that are important for the inference may be omitted from the test entirely or may be given less weight

than their importance for the target inference warrants. This is almost inevitable because the test is typically so much smaller than the target. Similarly, some elements may receive relatively more weight in the test than their importance to the inference warrants. Some elements may be given appreciable weight on the test even though they are entirely irrelevant to the inference. The performance elements that have too much or too little weight on the test may be either substantive or non-substantive. This is important because test preparation often focuses on non-substantive elements, that is, on elements that are not related to the inference at all. I will provide examples below.

Validity is then the extent to which an inference about the target — the weighted composite of all performance elements that are important for the inference — is justified by performance on the different and smaller weighted composite of tested elements. In other words, extrapolation from the test to the target inference hinges on alignment of the test and inference weights (Kane, 2006; Koretz and Hamilton, 2006). This becomes a critically important issue when tests have consequences.

## Matching Assessments to Inferences

It might seem obvious that the design of a test must be matched to the inference it is intended to support, but in practice, this is not always done in the K-12 sector, and problems of mismatch have already arisen in the postsecondary sector. The problems have been of three main types:

- Mistaking the part for the whole;
- Overstating comparability; and
- Encouraging "function creep."

### Mistaking the part for the whole

Users of scores from large-scale assessments have often ignored the systematic incompleteness of tests. That is, they have mistaken the smaller sampled part (the test) for the larger whole (the target).

An extreme form of this error has been in the use of test scores as measures of institutional quality and for accountability. This has often entailed ignoring the many aspects of quality that cannot be measured by standardized assessments of student achievement. Measurement experts have warned about the limited scope of achievement tests for well over half a century (e.g., Lindquist, 1951), and using scores from a single test without additional information to evaluate educational programs and institutions is a violation of accepted professional standards (e.g., American Educational Research Association, American Psychological Association,

and National Council on Measurement in Education, 2014, p. 213), but it has become common nonetheless.

Although this mistake has arisen in the postsecondary sector as well, a second, less extreme form of mistaking the part for the whole may be more relevant for postsecondary assessments: mistaking a test for a complete measure *of the testable portion of student achievement.* This occurs when one test is treated as a "gold standard," and the unavoidable incompleteness of tests and variations in results across tests are ignored. Relying on a single test for the test-based portion of a description or evaluation is risky because differences in the sampling used to construct different tests will often result in their yielding different results. This is true even when a test is not used to compare institutions or jurisdiction and when it is administered under low-stakes conditions — that is, when there is no pressure to raise scores on a particular test — but it can be an even more severe problem when scores are used for comparative purposes or with high stakes.

Differences in results across tests are frequently modest, but they are occasionally large. In the US context, one of the most important recent examples of large differences among tests is trends in elementary school mathematics. We have three nationally representative survey-based assessments that provide estimates of that trend: the main sample of the National Assessment of Educational Progress (NAEP); the NAEP long-term trend assessment, which is kept unaltered for long periods of time; and the US sample of the Trends in International Mathematics and Science Study (TIMSS). The main NAEP is most often used in research and in public discussion of trends. Between 1990 and 2007, the mean fourth-grade mathematics score on the main NAEP increased dramatically, by more than 0.84 standard deviation. Leaving aside trends that have been biased by changes in selectivity or score inflation (discussed below), this is one of the most rapid large-scale changes in mean scores in more than half a century of US data from large-scale tests. However, the NAEP long-term assessment showed a gain that was only about half as large during that period (Figure 1). The total increase in fourth-grade mathematics in the US sample of the TIMSS assessment is more similar to that of the long-term trend NAEP than to the main NAEP. No one has offered a convincing explanation of these differences.

Faced with these discrepant trends, it is safe to conclude that the mathematics performance of US elementary school students has increased substantially, but it is risky to conclude that this increase is as large as it is in the main NAEP. Yet the latter is precisely the conclusion many have reached. Advocates of the test-based accountability policies that have dominated US K-12 education for the past two decades often point to the main NAEP mathematics trends as evidence that their policies have been successful, and the most credible research estimating the impact of those policies on elementary school mathematics achievement has relied on this test. Those estimates would be considerably less positive if there were based on either of the other two assessments (Koretz, 2017). However, the disparity in trends

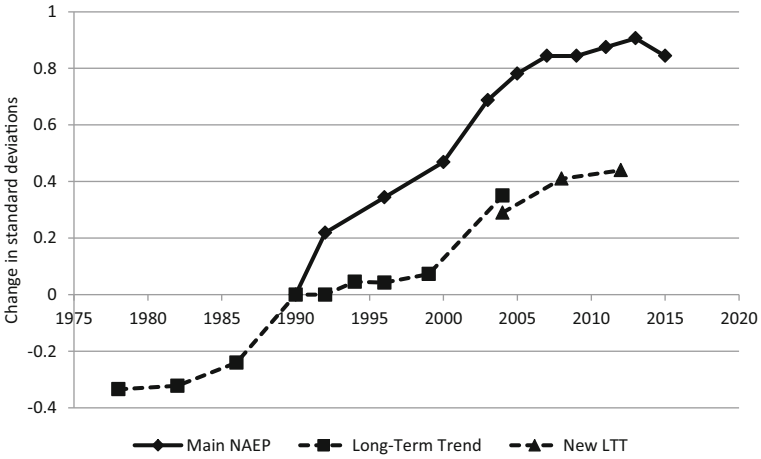


**Figure 1.** Trends in elementary school mathematics on the main NAEP (grade 4) and long-term trend NAEP (age 9), relative to 1990.

among these assessments is rarely even mentioned by researchers or education policymakers.

Differences among tests are also commonly ignored in the test-based accountability systems used in US education. It is now common to evaluate teachers based on their students' scores on a single test. Yet research has shown that value-added estimates can vary markedly from one test to another (e.g., Corcoran *et al.*, 2012; Lockwood *et al.*, 2007).

## Overstating comparability

The systematic incompleteness of tests can be an even more serious limitation when tests are used for comparative purposes — for example, to compare national K-12 systems or postsecondary institutions. Because of differences in curricula and educational goals, some tests will align more closely than others with the intended curriculum in a given jurisdiction or institution. The consequence is that score differences among these institutions or systems, and sometimes even their ranks, will vary depending on the choice of test.

This problem appears clearly in the results of large-scale international K-12 student achievement tests (ISATs), such as TIMSS and PISA (Programme for International Student Assessment). The correlations among country means on ISATs are sometimes high. For example, Klieme (2016) noted that the correlation between country means on the 2015 PISA and TIMSS mathematics assessments was 0.92. However, the correlations among country means have not always been that high. For example, the correlations between TIMSS grade 8 and PISA means

in mathematics have varied markedly and ranged from 0.78 to 0.83 in the first three iterations of PISA. For purposes of comparison, in a single population, *between-subject* aggregate correlations are in some instances higher than some of the *within-subject* correlations between TIMSS and PISA country means, despite the extremely high level of aggregation in the latter. For example, in the US national standardization sample of the Iowa Tests of Basic Skills, Form A, the correlation between school means in mathematics and reading in grade 8 was 0.86 (Hoover *et al.*, 2003, Table 8.4). Clearly, this high correlation does not imply that the reading and mathematics tests are substitutes that measure the same target of inference. It would not justify using the reading test to rank schools in "mathematics," treating the 26 percent of variance in school mathematics means not predicted by reading means as simple noise. By the same token, the correlations between TIMSS and PISA means do not justify treating either one as the "correct" measure of mathematics and ignoring the differences between the assessments merely as error.

Moreover, if one looks beyond simple correlations of country means, one finds numerous important differences in the results provided by different ISATs that reflect sampling decisions made in constructing the tests. For example, in general, TIMSS shows a large gap between East Asian countries, which are the highest scoring nations in mathematics, and European countries. Some of these differences are far smaller in PISA. Wu (2009) analyzed performance by item type and concluded that this difference between the two assessments reflects a systematic difference in the sampling of performance elements. Specifically, TIMSS includes many fewer items that present mathematics in a realistic, complex context. Another example is provided by content strands in both assessments. In TIMSS, for example, the mean performance of some countries varies markedly across the content strands included in the mathematics assessment. In the 2007 grade-8 mathematics assessment, the highest and lowest strand means of many countries were more than 50 scale points apart (Mullis *et al.*, 2008, Chapter 3). By way of comparison, the difference in grand means between Japan and the USA was 39 points (Mullis *et al.*, 2008, 34). This indicates that different weighting of the content strands in calculating the TIMSS composite score would substantially affect some grand mean scores and hence the relative positions of some countries.

Some other differences between ISATs are less apparent but are important nonetheless. Both TIMSS and PISA are reported in terms of an "international standard deviation," but I have argued elsewhere that this is not a useful statistic for calculating effect sizes (Koretz, 2008). The international standard deviation reflects the happenstance set of nations that choose to participate in a given assessment; it therefore does not display performance relative to a clear reference population. A logical alternative is to use within-country standard deviations, as these are readily interpretable and yield effect sizes that are comparable to others to which users may wish to contrast the findings, such as within-country differences between racial/

ethnic groups. If one calculates effect sizes in this manner, one finds that the results can be dramatically different across ISATs. Again using eighth-grade mathematics as an example, the mean difference between Korea and the USA was 1.06 US standard deviations in the 2003 TIMMS, but only 0.62 standard deviation in the 2003 PISA.

I have used international comparisons because they provide a clear illustration, but this issue is not limited to them. For example, similar patterns have been noted in comparing districts and states within the USA. These are merely examples illustrating that it is risky to use a single test to compare aggregates that have different educational goals and intended curricula. When units differ in intended curricula, some of the measured differences among them are likely to be sensitive to decisions about the test weights used in constructing tests.

## "Function creep"

Many people falsely believe that a "good" test that serves one function well, once it is in place, can be used for other purposes as well, even if it was not designed to do so. However, validity is an attribute of a particular inference, not of a test. A given test may be a good match to one target of inference but a poor match to another. Some additional functions may be appropriate, but others may not be.

While this principle is axiomatic, it has been widely ignored in K-12 assessments. Elsewhere I have labeled this problem *function creep*: adding new functions to an extant assessment, without due consideration of the test's appropriateness for the inferences entailed by the new uses. There are many examples, but two particularly extreme instances both happen to involve the SAT, one of the two principal undergraduate college admissions tests in the USA. In the 1980s, the federal Department of Education published "wall charts" that used states' mean SAT scores as a purported indicator of educational quality. The SAT was not designed to measure mastery of K-12 curricula, but even more important, at that time, states differed dramatically in the proportion and characteristics of students who took the test. In some states, a substantial majority of high school seniors took the SAT, while in some others, only a small minority did so (e.g., 6 percent in Minnesota), primarily high-achieving students applying to elite eastern universities that required SAT scores. Even if the content of the SAT had been appropriate for this use in terms of content, it is absurd to compare the mean of a very small, elite group to the mean of the majority of students in other states. Another example is a current merit-pay program for teachers in Florida called the "Best and Brightest Scholarship Program." To receive the bonus pay, experienced teachers must meet two criteria: They must be rated as highly effective based on their actual teaching, and they must have scored above the 80th percentile on either the SAT or ACT at any point in their lives. That is, the state is using a test designed to predict performance in undergraduate education, not to predict the effectiveness

of teaching, to evaluate teachers. Moreover, they are using tests that are designed to predict *unobserved* performance to "evaluate" performance after a measure of actual performance is taken into account.[3]

While these two examples are more extreme than most, the problem of function creep is common. For example, the Iowa Tests of Basic Skills (ITBS), a developmental test battery that has been in use for well over half a century, is designed to provide diagnostic information, and its developers have stated explicitly that it is not intended for use as a high-stakes summative test. Despite this warning, some jurisdictions, e.g., the Chicago public school district, have used the ITBS as a high-stakes summative assessment. PISA — both the tests and the survey in which they are embedded — was designed to provide comparative descriptions of performance and is not suited to supporting causal inferences. That is, PISA is well designed to describe how countries differ in both student achievement and characteristics of national educational systems, but it is poorly designed to evaluate which characteristics of the system contributed to differences in measured achievement. PISA scores are frequently used to justify causal inferences nonetheless.

## Campbell's Law

More than four decades ago, Donald Campbell, one of the founders of the discipline of program evaluation, wrote what has since become known as *Campbell's Law*:

> The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell, 1976, 49).

Although Campbell was describing what he saw as a general problem, he noted specifically that it arises in educational testing:

> Achievement tests may well be valuable indicators of… achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways (Campbell, 1976, 51).

A similar warning was offered a quarter century earlier by E. F. Lindquist, who was one of the most important developers of standardized achievement tests in the history of testing:

> The widespread and continued use of a test will, in itself, tend to reduce the correlation between the test series and the criterion series [the later behavior, outside of the testing situation, that is our real concern] for the population involved. Because of the nature and potency of the rewards and penalties associated in actual practice with high and low achievement test scores of students, the behavior measured by a widely used test tends in itself to become the real objective of instruction, to the neglect of the (different) behavior with which the ultimate objective is concerned (Lindquist, 1951, 152–153).

As Campbell expected, these phenomena have been documented in a wide variety of disparate fields — to name just a few, healthcare quality control, the management of airline delays, and the industrial system in the Soviet Union. (For an overview of instances of Campbell's Law in many fields, see Rothstein, 2008.) The ubiquity of the phenomenon is the reason it is routinely labeled Campbell's "Law."

In educational testing, Campbell's Law takes the form of *score inflation*, increases in test scores that are greater than the increase in the target of inference — that is, greater than the inferred increase in learning — warrants. The logic of most studies of score inflation reflects the validity framework above. If an increase in the tested sample accurately signals a commensurate increase in mastery of the target, then similar increases should appear in other samples from that target, that is, on other tests designed to measure similar constructs. Therefore, these studies compare trends in scores on the test that has high stakes to trends on an *audit test*, which is a test that is designed to measure a similar target but that has low (or at least lower) stakes for educators and students and that is administered to the same students or to a randomly equivalent sample of them. In the USA, NAEP has often been used as the audit test because NAEP's target reflects a degree of national consensus, scores are readily available down to the level of states and large districts, and teachers have no direct incentive to prepare students for that particular test.

Research has shown that while score inflation is not inevitable, it is common and often very large. For example, a number of studies have found gains on high-stakes tests that are three to six times as large as gains on an audit test, and at least one study found very large gains on a high-stakes test with no gains whatever on an audit test (e.g., Ho, 2007; Jacob, 2005; Klein *et al.*, 2000; Koretz *et al.*, 1991; Koretz and Barron, 1998).

One well-known example of score inflation occurred after New York State introduced a new high-stakes testing program in 2006. In eighth-grade mathematics, the statewide mean score in mathematics increased by 0.57 standard deviation in the space of only three years. Those familiar with historical trend data knew that an increase of 0.57 standard deviation in three years is highly suspect, and their doubts were corroborated by the trend on the state's representative sample in the

**Table 1** Mean increase in eight-grade mathematics scores in New York State, on state's test and NAEP, in standard deviations

| | *Total* | *White* | *Black* |
|---|---|---|---|
| New York test (2006–2009) | 0.57 | 0.50 | 0.72 |
| NAEP, NY sample (2005–2009) | 0.08 | 0.11 | 0.08 |

NAEP assessment. During the corresponding period, the state's mean on the NAEP mathematics assessment increased by only 0.08 standard deviation — that is, approximately one-seventh as much (Table 1).[4]

The effects of Campbell's Law need not be uniform, and therefore, it can undermine comparisons among jurisdictions or institutions as well as overall estimates of performance. A small but growing body of research suggests that in the USA, a disturbing variation in Campbell's law is that inappropriate test preparation and the resulting score inflation tend to be more severe in schools serving disadvantaged populations. For example, in the New York case, the mean score of black students on the state test increased by 0.22 standard deviation more than that of whites, reducing the black-white gap by roughly one-fourth in the space of only three years. In contrast, the black-white gap remained essentially unchanged in the state's representative NAEP sample (Table 1). It is not inevitable that disadvantaged students will experience Campbell's Law in more severe form; rather, this finding appears to reflect particular characteristics of US schools and the features of US test-based accountability. However, this illustrates how severely Campbell's Law can undermine comparisons among groups.

What makes score inflation possible is the fact that in most testing programs, the test weights of both substantive and non-substantive performance elements, including omissions, are to some degree consistent over time and predictable (e.g., Holcombe *et al.*, 2013). This predictability is in part deliberate. For example, considerable similarity in successive test forms is needed in order to link scores over time. However, much of the predictability is not needed for technical reasons. It arises because of the time and financial costs entailed in item development and piloting, convenience, and happenstance.

This predictability creates the opportunity for two different types of behavior that can inflate scores (Koretz and Hamilton, 2006). The first, *reallocation*, entails shifting instructional resources to better match the test weights in the particular tests used. Reallocation is not necessarily undesirable, and it does not necessarily induce score inflation. Inflation occurs when reallocation reduces resources allocated to elements that are *important for the inference based on scores* but that have small test weights or are omitted from the test altogether. Performance on these de-emphasized elements may stagnate or deteriorate, even as scores rise. The second method is *coaching*. This term is used in many different ways, but I follow

Koretz and Hamilton (2006) in using it to refer to focusing instruction or other test preparation on incidental aspects of a test that can affect scores but that are not important for the inference. These may be non-substantive elements or small details of content that are not important for the inference.

An example of coaching can be found in commercial test preparation materials sold to prepare students for a mathematics test that students had to pass in order to earn a high school diploma in Massachusetts. The authors of the materials noticed that the test often included an item about the Pythagorean theorem. In writing items about the Pythagorean theorem, test authors are constrained by the fact that few students know how to calculate square roots. Therefore, if an item has a solution that does not entail an obvious square root, even students who know the theorem are likely to answer incorrectly because they cannot compute the root, which would be misleading. The solution is to use simple squares, that is, triples such as 3:4:5 and 5:12:13. The appearance of these triples on the test is an entirely incidental performance element; the item is intended to measure whether students know the Pythagorean theorem, not whether they know that 3:4:5 is a Pythagorean triple. The test preparation book labeled these triples "common" and "popular" Pythagorean triples (they are common in test items, not in the real world) and informed students that these two triples or multiples of them will solve items they encounter on that test (Rubinstein, 2000). This coaching enabled students who do not know the theorem to answer the item correctly, thus inflating scores.

## The Problem of Causal Inference

Many of the most important — although often unwarranted — uses of large-scale K-12 testing entail causal inferences. For example, test scores are commonly used not only to describe students' performance, but also to make claims about the effectiveness of teachers, institutions, or even entire national systems.

Scores are often used to evaluate educators or institutions with no adjustment or control for other factors that influence student performance. For example, at various times, American schools have been evaluated on the basis of simple mean scores, percents above a cut score (most often, the "proficient" standard), or changes from cohort to cohort in one or the other of those two statistics. Some advocates of these policies simply ignore the powerful influences on test scores of other factors, such as family background, proficiency in the language of testing, and peer effects. Others recognize these influences but argue that educators should be able to compensate for them.

In recent years, however, some systems have attempted to remove the effects of other factors and isolate the impact of educational quality by using some form of "value-added model" (VAM). A diverse variety of models have been used for this purpose (see Castellano and Ho, 2013), but they share the principle of controlling in

some manner for students' earlier scores in an attempt to isolate the effects of education in the current time period.[5]

While the strengths and limitations of VAM remain the subject of intense debate, VAMs do not necessarily isolate the effects of current educational experience. This view was summarized in a recent position paper issued by the American Statistical Association, the primary professional association of statisticians in the USA, which included the following summary statements:

- VAMs typically measure correlation, not causation: Effects — positive or negative — attributed to a teacher may actually be caused by other factors that are not captured in the model.
- Under some conditions, VAM scores and rankings can change substantially when a different model or test is used, and a thorough analysis should be undertaken to evaluate the sensitivity of estimates to different models (American Statistical Association, 2014, 2).

The American Statistical Association statement noted class size, the inclusion of high-needs students, and the inclusion of students receiving supplementary tutoring as examples of potentially relevant influences on scores that are typically not included in VAMs and that may be confounded with estimates of educational effectiveness (American Statistical Association, 2014, 4).

McCaffrey *et al.* (2003) noted that VAM estimates can vary substantially from one test to another, and they explained that differences in test construction (i.e., differences in test weights), cross-sectional scaling methods, and in some instances, methods used to place different ages or grades on a single scale can contribute to this variation. With respect to test weights, they noted:

> Although most common scaling models treat the construct of interest as unidimensional, this is a simplification. In most cases, a test of any broad domain of achievement…will assess a variety of different dimensions of performance. The process of constructing the test requires decisions about the relative emphasis given to various aspects of performance. Moreover, the actual emphases inherent in a test may differ independently of the intent of designers because of a variety of factors….The ordering of means, such as district or state means in U.S. comparisons or country means in international comparisons, is sometimes sensitive to these differences in test construction….It is generally assumed that a primary reason for this sensitivity is variations in curricular alignment (McCaffrey *et al.*, 2003, 64).

All of these differences in VAM estimates — across statistical models, test weights, and scaling methods — bear on a concern noted above: overstating comparability. VAM does not free us from the risk that variations in the alignment

between curricula and tests will lead to different rankings of institutions or systems when different tests are used. Indeed, when the institutions or systems compared are very different, VAMs may actually exacerbate this problem. For example, consider a test that comprises content that is too basic for students in some institutions. Students in those institutions are likely to score very well on that test regardless and will thus rank highly in cross-sectional comparisons. However, they are likely to show relatively little *growth* on that test because the test lacks content that would measure their growth and will therefore rank poorly on VAM measures.

The problem of score inflation is also not reduced by the use of VAMs. Using the evaluation of teachers as an illustration, McCaffrey *et al.* (2003) explained that score inflation can undermine VAM estimates in two entirely different ways: biasing the estimates for individual teachers and exaggerating variations in effectiveness (the "teacher effect"):[6]

> Any appreciable variability in the extent of inflation could substantially bias the inferences based on VAM. Even a random distribution of inflation would upwardly bias the estimated variance of the teacher effect, and the rankings of many individual teachers would be rendered meaningless. To the extent that the distribution of score inflation is systematically related to important characteristics of teachers or their contexts, inferences about the characteristics of teachers are likely to be severely distorted (McCaffrey *et al.*, 2003, 100).

## Manifestation of These Issues in Postsecondary Assessment

Many of these issues have already arisen in large-scale assessments in the postsecondary sector, and all are potentially important. For instance, the Collegiate Learning Assessment (CLA) provides an example of function creep. The Council for Aid to Education has suggested fully nine different uses for the CLA:

> CAE has pioneered the use of performance-based tasks in our Collegiate Learning Assessment to evaluate critical thinking skills of college students. CLA + measures critical thinking, problem solving, scientific and quantitative reasoning, writing, and the ability to critique and make arguments. Over 700 institutions…have used the Collegiate Learning Assessment to benchmark value-added growth in student learning [1] at their college or university compared to other institutions [2]….Student-level metrics provide guidance to students and data to faculty and administrators for making decisions about grading, scholarships, admission, or placement [3–6]. Institutions can use CLA + for additional admissions information for college applicants [7], to evaluate the strengths and weaknesses of entering students. Results for graduating seniors may be used as an independent corroboration of the rapid

> growth [of] competency-based approaches [8] among colleges. Graduating seniors can also use their scores to provide potential employers with evidence of their work readiness skills [9] (Council for Aid to Education, 2013; emphasis and enumeration added).

The OECD's trial comparative postsecondary assessment, AHELO, was a less extreme example, but it had at least three different intended purposes: institutional improvement, comparative monitoring, and accountability.

However, it would be a mistake to conclude that the postsecondary and K-12 sectors are similar with respect to these issues. In a number of ways, the issues confronting large-scale assessment are more difficult in the postsecondary sector.

One important difference between the sectors is the degree of curricular differentiation. In the K-12 sector, curricular differentiation increases as students progress through school. For example, in the USA, there is far more differentiation in the mathematics students study in secondary school than in the elementary grades. Differentiation is far greater yet in the postsecondary sector, and this affects test-based inferences about both individuals and institutions or systems.

One reason for this greater curricular differentiation is the sorting of students into disciplines. Obviously, whether students take any coursework in mathematics and, if they do, the amount and type of mathematics they study, depends on their chosen field of study. It is necessary for students in chemistry, physics, and engineering to take courses in single- and multivariable calculus and linear algebra. Students in psychology rarely need to study multivariable calculus, but they do need to study statistics. Students in the humanities may study none of these.

Moreover, similar course titles may obscure important curricular differences. For example, there are many varieties of introductory statistics courses. Some of this variation is happenstance, and some is unrelated to the target of inference. For example, faculty may impose different levels of demand even then their course goals are nominally the same. However, some of the variation is related to the intended targets of inference.

These differences will cloud comparisons of the performance of individual students and, even more, comparisons of aggregate levels of achievement in their institutions. It is difficult to draw conclusions about differences in proficiency in mathematics, for example, when institutions differ in the proportion of students studying mathematics and, for those who are, what mathematics they are studying and what curricular goals their faculty have for them.

In the postsecondary sector, there is yet another difficulty drawing comparisons at the aggregate level. Many students in postsecondary education allocate a substantial part of their coursework to specialty courses that enroll a very small percentage of students — courses such as "Topics in Late [Chinese] Imperial History" or "Kinetics of Condensed Phase Process." (These are random selections from Harvard's course catalog.) The consequence is that any given common

assessment — such as a test of writing or mathematics — will necessarily represent a smaller share of the output of postsecondary institutions than of K-12 schools, particularly elementary schools. Moreover, the results of a common assessment will often represent a larger share of student learning in some institutions than in others.

The issue of causal inference is also more problematic in the postsecondary sector because of student selection. Both selection by schools or systems and self-selection are present in many K-12 systems. For example, in Singapore, selection into secondary schools hinges in large part on students' performance on the Primary School Leaving Examination. In the Netherlands and some US systems, students (or their families) can self-select into K-12 schools. Nonetheless, in many countries, both institutional selection and self-selection are far more substantial in the postsecondary sector.

This greater selectivity undermines causal inference for a number of reasons. First, it makes it even more difficult to control for differences in the intake characteristics of individual students. Second, it can affect processes within institutions other than instruction. For example, students at highly selective postsecondary institutions are surrounded by highly able and academically highly motivated peers, many of whom are studying similar course material, and students may learn from their peers. Therefore, growth in test scores could reflect instruction, intake characteristics, or peer effects (among other things), and the available data generally would not allow one to separate their effects.

## Implications and Recommendations

As substantial as they are, the issues that have arisen in large-scale K-12 assessments are not a reason to forgo external assessment in the postsecondary sector. Such tests provide specialized information that is often unavailable from any other source. For example, it is only because of standardized tests that we know that in the USA, the achievement gap between low-SES and high-SES students has been growing even as racial and ethnic gaps have been gradually shrinking (Reardon, 2011). However, the issues described above signal the need to exercise restraint in the use of large-scale standardized assessments. I will close with a number of specific recommendations.

*Avoid function creep* It is costly in both time and effort to construct a good assessment, and it is therefore always tempting to use an extant one for multiple purposes. In some instances, this is not problematic, but it is often a serious mistake. Assessments should be designed to serve a specific use and to support specific inferences. Assessments should be applied to new uses only with care, as is clearly spelled out in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological

Association, and National Council on Measurement in Education, 2014, especially Chapter 1). When an assessment is applied to a new purpose, it is the obligation of the agency responsible for that use to warn users that the test is being used for a purpose other than then one for which it is designed and to obtain validity evidence supporting that new use (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, validity standards 1.3 and 1.4).

*Avoid using the same test for monitoring and accountability* This is a particularly important and burdensome instance of the previous recommendation. We know from experience in K-12 testing that educational tests are highly vulnerable to Campbell's Law when they are used for accountability. Therefore, it is usually not appropriate to use the same test for both accountability and monitoring because educators' responses to accountability will often bias scores and thereby undermine the assessment's value for honest monitoring. Moreover, when two different tests are used for monitoring and accountability, the more similar they are in terms of both substantive and non-substantive performance elements, the less credible the results from the monitoring assessments will be.

While separating the monitoring and accountability functions of testing is essential, it is unfortunately not a guarantee that Campbell's Law will not bias scores on the test intended for monitoring. Once a monitoring test is implemented and results are made public, administrators and policymakers may feel pressure to improve their scores. That is, it may gradually come to function as an accountability test even though that was not its original purpose. One example is the well-known "PISA shock," the reaction of policymakers in numerous countries to what they considered to be unacceptable or embarrassing rankings on the PISA assessment.[7] The responses to this pressure may include reforms that truly improve student learning, but they also may include the types of shortcuts that lead to Campbell's Law. For example, it was recently revealed that the department of education in one US state had developed plans to distribute NAEP-specific test preparation materials to schools selected to participate in that monitoring assessment.[8] It is therefore prudent to watch for possible corruption of scores even when a test is used for monitoring but not for explicit accountability.

*Avoid too broad an inference* It is all too tempting to conceptualize a test as measuring a broad domain, without reference to its limitations. This is a fundamental error and the source of many of the specific problems that have arisen in large-scale K-12 assessment. For example, neither PISA not TIMSS measures "mathematics." Both measure portions of that domain, and while the tested portions overlap considerably, they are not the same.

By the same token, avoid spurious precision. Different tests often provide different estimates. The results of TIMSS and PISA provide a clear example: It is safe to say that students in the USA perform considerably less well in mathematics than students in developed East Asian countries, but it is not safe to say precisely how large that gap is — unless one wants to narrow the inference to "in mathematics as measured by PISA" or "in mathematics as measured by TIMSS."

*Be wary of comparative uses of large-scale tests* The discussion above clarified that differences in the targets of inference can threaten comparative uses of large-scale tests and that these problems are typically more severe in the postsecondary sector than in K-12 education. Moreover, the more unlike the institutions or systems that are compared, the more serious this difficulty will become. To address this requires narrowing the inference to shared targets of inference and acknowledging explicitly both that this subset is incomplete and that it is more incomplete in some institutions than in others. In this case, it is particularly important to avoid spurious precision and to look for the robustness of results across different measures.

*If possible, use multiple measures* If multiple measures are available, it is usually best to use them, considering respects in which their findings are both similar and dissimilar.

*When using scores for instructional or institutional evaluation, combine them with other data* Because tests are incomplete, scores should be supplemented when feasible with relevant information of other types. Although this principle is often ignored, it has been axiomatic in the measurement field for many decades, and it is clearly stated in the *Standards*:

> In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, 213).

*Monitor the effects of testing* In the K-12 sector, the same test has often been used both as the tool to induce improvement and measure the resulting change. This was a fundamental error in that it ignored Campbell's Law. Not only were scores inflated; in the process of inflating scores, educational practice was undermined (Koretz, 2017). To ascertain the effects of large-scale testing requires examining data other than scores on those specific tests — for example, data from other tests and from more direct measures of educational practice.

In sum, the overarching lesson from testing in the K-12 sector is "less is more." Testing is a powerful tool that can provide important information we cannot obtain

in any other way, and it can have a powerful influence on educational institutions. However, ignoring the limitations of testing can lead to distorted conclusions and can undermine precisely the processes it is intended to improve.

## Notes

1 As Shavelson (2010) pointed out, there is a difference between assessing what students *know* and what they have *learned* during a period such as their time in postsecondary education. The former is a simple cross-sectional measure, while the latter evaluates change over time. For present purposes, however, that distinction is not necessary, and I will use *student learning* to refer to both.
2 The test weight of a performance element is the sensitivity of the test score to changes in performance on that element and can be represented formally as the partial derivative of the score with respect to performance on the element. The inference weight, however, is rarely clearly defined and may vary from one user of scores to another.
3 In the interest of full disclosure, I am the expert witness for the Florida Education Association, which has sued the Florida Department of Education and all Florida school districts to end this program.
4 NAEP is administered only every second year, so it was necessary to compare NAEP the NAEP increase over four years to the New York test increase over three years.
5 As Castellano and Ho (2013) point out, some of the commonly used models properly should not be called value-added models because they do not entail direct measures of growth. For example, one of the most common approaches, predicting current-year scores from prior scores and other variables in a regression model, is technically a "conditional status" model. However, these distinctions are not important for present purposes, so I will label all of the approaches that control for prior scores as value-added models.
6 The amount of variation in VAM estimates for schools and especially for teachers — the "teacher effect" — has played a large role in debates about K-12 education. This was in response to a widespread view that schools had relatively little impact and that inequities in educational performance stem primary from factors independent of schooling, such as family background. The more substantial the variation in VAM estimates, the more credible the argument that student learning can be increased substantially by improving the quality of teaching.
7 This term originated in Germany, where PISA shock resulted in major changes to the educational system (Waldow, 2009), but it occurred in many other countries as well (Breakspear, 2012).
8 Specifically, the department developed materials to encourage reallocation, showing the teachers in sampled schools which of the state standards would be tested by NAEP. The plan was abandoned after it was brought to the attention of the authority responsible for NAEP. This has not been documented in the press or the archival literature, but a video of the state's Superintendent discussing the plan with his board was posted for some time at http://www.alsde.edu/sites/boe/Pages/VideoLargeItem.aspx?ID=1522.

## References

Allen, J. (2018) Personal communication, May 18.
American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014) *Standard for educational and psychological testing (2014 Edition)*, Washington, DC: Authors.

American Statistical Association. (2014) *ASA statement on using value-added models for educational assessment,* Author. https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf. Accessed 11 Apr 2019.

Astin, A.W. and Antonio, A.L. (2012) *Assessment for excellence: the philosophy and practice of assessment and evaluation in higher education*, Lanham, MD: Rowman & Littlefield.

Breakspear, S. (2012) *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance,* Paris, OECD Publishing (*OECD Education Working Papers* No. 71), http://dx.doi.org/10.1787/5k9fdfqffr28-en.

Campbell, D.T. (1976) 'Assessing the Impact of Planned Social Change,' Occasional paper #8, in G.M. Lyons (ed.) *Social Research and Public Policies,* Hanover, NH: Dartmouth College.

Castellano, K.E. and Ho, A.D. (2013) *A practitioner's guide to growth models*, Washington, DC: Council of Chief State School Officers.

Cizek, G.J. (2016) 'Validating test score meaning and defending test score use: different aims, different methods', *Assessment in Education: Principles, Policy and Practice* 23(2): 212–225.

Coates, H. and Mahat, M. (2014) 'Advancing student learning outcomes' in H. Coates (ed.) *Higher education learning outcomes assessment: international perspectives*, Frankfurt am Main: Peter Lang, pp. 15–32.

Corcoran, S.P., Jennings, J.L. and Beveridge, A.A. (2012) *Teacher effectiveness on high- and low-stakes tests,* New York University, working paper. Retrieved from https://www.nyu.edu/projects/corcoran/papers/Corcoran_Jennings_Houston_Teacher_Effects.pdf. Accessed 11 Apr 2019.

Council for Aid to Education. (2013) *Performance assessment: CLA+ overview*, New York, Author. Retrieved from https://2014.accreditation.ncsu.edu/pages/3.5/3.5.1/CLA.pdf. Accessed 11 Apr.

Hamilton, L.S., Nussbaum, E.M. and Snow, R.E. (1997) 'Interview procedures for validating science assessments', *Applied Measurement in Education* 10(2): 181–200.

Ho, A.D. (2007) 'Discrepancies between score trends from NAEP and state tests: a scale-invariant perspective', *Educational Measurement: Issues and Practice* 26(4): 11–20.

Holcombe, R., Jennings, J. and Koretz, D. (2013) 'The roots of score inflation: an examination of opportunities in two states' Tests', in G. Sunderman (ed.) *Charting reform, achieving equity in a diverse nation*, Greenwich, CT: Information Age Publishing, pp. 163–189. http://dash.harvard.edu/handle/1/10880587. Accessed 11 Apr 2019.

Hoover, H.D., Dunbar, S.D., Frisbie, D.A., Oberley, K.R., Ordman, V.L., Naylor, R.J., Bray, G.B., Lewis, J.C., Qualls, A.L., Mengeling, M.A. and Shannon, G.P. (2003) *The Iowa tests: guide to research and development, Forms A and B*, Itasca, IL: Riverside Publishing.

Jacob, B.A. (2005) 'Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools', *Journal of Public Economics* 89(5–6): 761–796.

Judd, T. and Keith, B. (2102) 'Student learning outcomes at the program and institutional levels', in C. Secolsky and D.B. Denison (eds.) *Handbook on measurement, assessment, and evaluation in higher education*, New York: Routledge, pp. 31–46.

Kane, M.T. (2006) 'Validation', in R.L. Brennan (ed.) *Educational measurement* (4th ed.), Westport, CT: American Council on Education/Praeger, pp. 17–64.

Kane, M.T. (2016) 'Explicating validity', *Assessment in Education: Principles, Policy and Practice* 23(2): 198–211.

Klein, S.P., Hamilton, L.S., McCaffrey, D.F., and Stecher, B.M. (2000) *What do test scores in texas tell us?* Santa Monica, CA: RAND (Issue Paper IP-202).

Klieme, E. (2016) *TIMSSS 2015 and PISA 2015: How Are They Related At The Country Level*? DIPF Working paper, Frankfurt, Germany: Deutsches Institut für Internationale Pädagogische Forschung.

Koretz, D. (2008) *Measuring up: what educational testing really tells us,* Cambridge, MA: Harvard University Press.

Koretz, D. (2016) 'Making the term "validity" useful', *Assessment in Education: Principles, Policy, and Practice* 23(2): 290–292.

Koretz, D. (2017) *The testing charade: pretending to make schools better,* Chicago: University of Chicago Press.

Koretz, D., and Barron, S.I. (1998) *The validity of gains on the Kentucky Instructional Results Information System (KIRIS),* Santa Monica, CA: RAND (MR-1014-ED).

Koretz, D. and Hamilton, L.S. (2006) 'Testing for accountability in K-12', in R.L. Brennan (ed.) *Educational measurement* (4th ed.), Westport, CT: American Council on Education/Praeger, pp. 531–578.

Koretz, D., Linn, R.L., Dunbar, S.B. and Shepard, L.A. (1991) 'The effects of high-stakes testing: preliminary evidence about generalization across tests,' in R.L. Linn (chair), *The effects of high stakes testing*, symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April. http://dash.harvard.edu/handle/1/10880553. Accessed 11 Apr 2019.

Lindquist, E.F. (1951) 'Preliminary considerations in objective test construction', in E.F. Lindquist (ed.) *Educational measurement,* Washington, DC: American Council on Education, pp. 119–184.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B., Le, V. and Martinez, J.F. (2007) 'The sensitivity of value-added teacher effect estimates to different mathematics achievement measures', *Journal of Educational Measurement* 44(1): 47–67.

McCaffrey, D.F., Lockwood, J. R., Koretz, D.M. and Hamilton, L.S. (2003) *Evaluating value-added models for teacher accountability*, Santa Monica, CA: RAND (MG-158-EDU). Retrieved from http://www.rand.org/pubs/monographs/MG158.html. Accessed 11 Apr 2019.

Massachusetts Department of Elementary and Secondary Education. (2018) *2019 Next-generation MCAS test information for Grade 10 Mathematics,* Malden, MA, Author (Revised September 7.) Retrieved from http://www.doe.mass.edu/mcas/tdd/math.html?section=nextgen. Accessed 11 Apr 2019.

Messick, S. (1989) 'Validity', in R. Linn (ed.) *Educational measurement* (3rd ed.), Washington, DC: American Council on Education, pp. 13–100.

Mullis, I.V.S., Martin, M.O. and Foy, P. (2008) *TIMSS 2007 International Mathematics Report*, Newton, MA, TIMSS & PIRLS International Study Center, Boston College.

Moore, K., Coates, H. and Croucher, G. (2014) 'Understanding and improving higher education productivity', in E. Hazelkorn, H. Coates and A.C. McCormick (eds.) *Research handbook on quality, performance and accountability in higher education,* Cheltenham, UK: Edward Elgar, pp. 161–177.

Reardon, S.F. (2011) 'The widening academic achievement gap between the rich and the poor: new evidence and possible explanations', in R. Murnane and G. Duncan (eds.) *Whither opportunity? Rising inequality and the uncertain life chances of low-income children,* New York: Russell Sage Foundation, pp. 91–116.

Rothstein, R. (2008) *Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education*, Nashville: National Center on Performance Incentives, Vanderbilt Peabody College. Retrieved from http://www.epi.org/files/2014/holding-accountability-to-account.pdf. Accessed 11 Apr 2019.

Rubinstein, J. (2000) *Cracking the MCAS Grade 10 Math*, New York: Princeton Review Publishing.

Secolsky, C. and Denison, D.B. (2012) *Handbook on measurement, assessment, and evaluation in higher education*, New York: Routledge.

Shavelson, R.J. (2010) *Measuring college learning responsibly*, Stanford, CA: Stanford University Press.

Tremblay, K., Lalancette, D. and Roseveare, D. (2012) *AHELO feasibility study report, Volume 1*, Paris: OECD.

U.S. Department of Education. (2006) *A test of leadership: charting the future of U.S. higher education,* Washington, DC: Author.

Waldow, F. (2009) 'What PISA did and did not do: Germany after the 'PISA-Shock"', *European Educational Research Journal* 8: 476–483. Published online 1 January. http://dx.doi.org/10.2304/eerj.2009.8.3.476.

Williams, R. (2014) 'Comparing and benchmarking higher education systems', in E. Hazelkorn, H. Coates and A.C. McCormick (eds.) *Research handbook on quality, performance and accountability in higher education,* Cheltenham, UK: Edward Elgar, pp. 178–188.

Wu, M. (2009) *A Critical Comparison of the Contents of PISA and TIMSS Mathematics Assessments,* unpublished working paper, University of Melbourne. Retrieved from https://edsurveys.rti.org/PISA/documents/WuA_Critical_Comparison_of_the_Contents_of_PISA_and_TIMSS_psg_WU_06.1.pdf. Accessed 11 Apr 2019.

Yamada, R. (2014) 'Comparative analysis of learning outcomes assessment policy contexts', in H. Coates (ed.) *Higher education learning outcomes assessment: international perspectives*, Frankfurt am Main: Peter Lang, pp. 33–48.