NBER WORKING PAPER SERIES

THE CONSEQUENCES OF USING ONE ASSESSMENT SYSTEM TO PURSUE TWO OBJECTIVES

Derek Neal

Working Paper 19214
http://www.nber.org/papers/w19214

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2013

I thank the Searle Freedom Trust for research support. I also thank Lindy and Michael Keiser for research support through a gift to the University of Chicago's Committee on Education. I thank Michael Greenstone, Diane Whitmore Schanzenbach and Robert S. Gibbons for useful comments. I thank Robert D. Gibbons and David Thissen for their insights on psychometrics and assessment development. I thank Ian Fillmore, Sarah A. G. Komisarow and Richard Olson for excellent research assistance. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The Consequences of Using One Assessment System To Pursue Two Objectives
Derek Neal
NBER Working Paper No. 19214
July 2013, Revised December 2013
JEL No. I20,I21

ABSTRACT

Education officials often use one assessment system both to create measures of student achievement and to create performance metrics for educators. However, modern standardized testing systems are not designed to produce performance metrics for teachers or principals. They are designed to produce reliable measures of individual student achievement in a low-stakes testing environment. The design features that promote reliable measurement provide opportunities for teachers to profitably coach students on test taking skills, and educators typically exploit these opportunities whenever modern assessments are used in high-stakes settings as vehicles for gathering information about their performance. Because these coaching responses often contaminate measures of both student achievement and educator performance, it is likely possible to acquire more accurate measures of both student achievement and education performance by developing separate assessment systems that are designed specifically for each measurement task.

Derek Neal
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and The Committee on Education
and also NBER
d-neal@uchicago.edu

Introduction

Assessment-based accountability systems in education typically pursue two distinct objectives using data gathered from a single assessment system. The first objective is to provide reliable information about student achievement and how levels of achievement vary among student populations defined by time and geography. The second objective is to induce educators to teach well by attaching consequences to assessment-based measures of their performance. The text of the No Child Left Behind Act of 2001 (NCLB) makes clear that these two objectives should be key goals in each state's accountability system, but these twin objectives were also present in systems that pre-date NCLB. Further, these dual goals are also implicit in the work of two consortia of states that are now developing new assessment systems designed to measure student achievement relative to the Common Core State Standards, a collection of curriculum and achievement standards that forty-five states have now adopted.[1]

The empirical literature on the effects of assessment-based accountability systems in education raises concerns about this approach and provides evidence that existing accountability systems have not done an effective job of pursuing these two objectives simultaneously. In a recent chapter in the Handbook of Economics of Education, Neal (2012), I review the empirical literature on high-stakes assessments systems used to create performance metrics for educators and note that many studies find the same pattern. Following the introduction of high-stakes assessment-based accountability or performance pay systems, student scores typically rise on the high-stakes assessments used to create educator performance metrics, but scores on other low-stakes assessments often show no improvement or rates of improvement that are much slower than those observed on high-stakes assessments of the same subject.

[1]The SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC) are the two groups assessment systems for the Common Core State Standards using funds awarded as part of the Obama Administration's Race to the Top initiative.

This common pattern suggests that educators usually respond to the introduction of assessment-based incentive systems by taking actions that raise the measured achievement of their students on exams generated by a particular assessment system, but these same actions may not create commensurate improvements in student subject mastery.[2] Thus, high-stakes assessment systems may be failing to meet both of their objectives. They may be inducing teachers to teach in ways that are not best for their students, and these same actions may be inflating test scores and contaminating public information about prevailing levels of student achievement at various points in time.[3]

There are large literatures in sociology, economics, and management that document unwanted consequences of incentive systems built around objective but poorly designed performance metrics.[4] In a seminal paper, Holmstrom and Milgrom (1991) highlight the central role of hidden action in these scenarios by demonstrating how incentive systems employed in government or the private sector should be expected to produce poor results in settings where employees can easily take unobserved actions that improve their measured performance without generating commensurate improvement in their true performance.

My thesis is that features of modern testing systems that promote reliable scoring when tests are given under low-stakes directly create opportunities for educators to take actions that are difficult to monitor and that improve student scores without generating commensurate improvements in student learning. These features of modern assessments are not problematic when tests are given in low-stakes settings where educators have no interest in student test scores per se but are interested in test scores only as sources of feedback about their students' skills. However, when modern tests become components of high-stakes accountability systems, teachers find that they do have a per se interest in

[2]The pattern described here is not definitive proof that student test score gains on high-stakes assessments do not reflect real gains in subject mastery since the parallel low-stakes assessments are never identical in terms of subject content. See Koretz (2002) for more on this point. However, many of the studies in this literature present results that are difficult to explain in a credible way without some story that explains how high-stakes assessments scores can rise quickly without commensurate improvements in true levels of math or reading achievement.

[3] Glewwe, Ilias, and Kremer (2010), Jacob (2005), Klein et al (2000), Koretz and Barron (1998), Koretz (2002), and Vigdor (2009) all present results that show divergence between student assessment results on two assessments of the same subject matter in settings where one assessment became a high-stakes assessment for educators and the other assessment continued to involve relatively low-stakes. Neal (2012) provides a detailed discussion of this literature.

[4]See Campbell (1979), Kerr (1995), and Rothstein (2009) for many examples.

student test scores, and teachers respond by exploiting the opportunities for manipulation that are built into modern testing systems.

After explaining why modern assessments always invite coaching when they are administered in high-stakes settings, I discuss an alternative paradigm that requires two assessment systems. The first is a standard, modern testing system that tests students under low-stakes and provides reliable measures of student achievement. The second is a high-stakes system that is designed for the purpose of creating performance measures for educators. Because this second system would seek to remove coaching opportunities by avoiding repeated items and minimizing the predictability of item formats, it would likely not create the information necessary to produce scaled measures of student achievement that meet modern psychometric standards for reliability. However, such measures are not needed if the only purpose of the assessment system is to gather data on educator performance. Policy makers can form useful performance metrics for educators if they are simply able to reliably rank students at any test date according to their performance on a particular assessment form.

For the foreseeable future, public actors will continue to demand information concerning variation in levels of student achievement among various student populations as well as consequential measures of the relative performance of various school districts, schools, and teachers. To date, many attempts to satisfy both of these demands using a single assessment system have produced less than optimal results, and there are reasons to believe that no single system can ever simultaneously gather both types of information well. Thus, the obvious agenda for future research is the exploration of ways to develop two distinct assessment systems that gather these two types of information separately.

1. Screening versus Incentives

Modern assessment systems are used in many high-stakes settings that do not involve the creation of performance metrics for educators. These systems are employed as one component of various high-stakes screening systems, and in these contexts, there is considerable evidence that test takers prepare for these exams in ways that likely improve test-taking strategies without fostering commensurate improvements in subject

mastery. As an example, consider the Graduate Record Exam (GRE). Graduate schools use GRE scores to assess the aptitudes of applicants for graduate study, and many GRE takers spend significant time and money on books and courses that provide advice and practice on test-taking strategies that enhance GRE performance. These expenditures of time and money likely build few lasting skills, but if one is willing to assume that students who possess higher true levels of aptitude and subject mastery find it no harder to master test-taking strategies than other students, it is easy to write down economic models of the GRE market such that, ex post, GRE scores provide a reliable ranking of applicants with respect to the aptitudes that these scores are supposed to measure and also serve as useful predictors of graduate school performance. Further, since gathering information about the suitability of particular students for graduate study in different universities is costly, the GRE system as a whole may be a fairly efficient system for providing graduate schools with screening information about their applicants even when one accounts for the resources that students devote to test-preparation activities. Applicants pay the costs of preparing for and taking the exam once, but they are then able to simultaneously send a set of signals concerning their skills and aptitudes to all schools that they may wish to attend.

In contrast, the stated purpose of assessment-based accountability systems is not screening but on-going incentive provision. A key reason to hold educators responsible for the measured performance of their students is to induce educators to teach well in all classes at all times. In this context, if one concludes that the use of modern assessments in accountability systems is inducing teachers to consistently allocate class time to test-preparation activities that build few lasting skills, then one must also conclude that the decision by policy makers to measure educator performance using data from these systems has directly undermined a stated purpose of accountability systems.[5]

[5] The effort distortions induced by assessment-based accountability are not one-time costs. Any system that induces teachers to adopt teaching methods that raise test scores but degrade the true quality of instruction imposes an on-going cost on students, and students bear these costs throughout all grades and classes where their teachers are subject to assessment-based accountability.

2. Assessment Development

Here, I discuss features of modern assessment systems that make coaching students on test-taking strategies possible, profitable and likely when assessments are given under high stakes. I then briefly discuss alternatives to the current paradigm.

2.1 Standards and Prior Use

Consider the following hypothetical question: How do the math achievement levels of fifth grade students in Chicago in 2012 compare to those of fifth grade students in Chicago in 2011? It is hard to imagine how the results from any assessment system could provide an answer to this question unless the system found a way to link the assessments given in these two years.

To see this point, assume that a testing agency simply created an entirely new assessment for 2011. By "entirely new," I mean that no other students at other points in time or in other school systems have ever seen these questions on any exam. Now, also assume that the same testing agency created another entirely new assessment for 2012. Given this scenario, the results from these two assessments could never provide any information about whether or not achievement levels were higher or lower in 2012 compared to 2011. As long as the two assessments contain no common items and no items that have been previously administered to other populations of students, one can always explain any observed difference in the distributions of scores for 2011 and 2012 by asserting that the two populations of students shared a common achievement distribution but the two assessments differed in difficulty or discrimination in particular ways or by asserting that the two assessments were comparable but that the underlying distributions of student achievement levels differed in specific ways across years.

In order to place student results from 2012 and 2011 on the same scale, testing agencies need prior information about at least some of the items on the 2012 assessment. Typically, test developers acquire this information in one of two ways. They either repeat a portion of the items on a given year's assessment on the assessment for the

following year, or they pre-test items during experimental administrations and calibrate the difficulty of items that form a test bank. Given either approach, assessment developers can use their prior knowledge of the difficulty of specific items to isolate year-to-year performance variation that is due to changes in the underlying distribution of student achievement holding item difficulty constant, and this leverage allows researchers to place assessment results from different years on a common scale.

The key point here is that one cannot place the results of the 2012 assessment on the 2011 scale without some prior information about the difficulty of at least some of the items on the assessment, and one cannot acquire the needed prior information without administering these items to students. Thus, if the 2012 assessment is a high-stakes exam for educators, educators face a clear incentive to acquire copies of assessments given in previous years or in any experimental administrations used for calibration and then use these copies to drill their students on the answers to these specific items. The strength of this incentive varies with the details of the testing system in question, but in many modern testing systems, the assessment form for a given year often contains twenty percent or more of the items on the previous year's assessment.[6]

This practice creates strong incentives for educators to obtain copies of each year's assessment form or forms and then have next year's students memorize the answers to the questions found on these exams. Further, in their book on equating procedures, Kolen and Brennan (2010) argue that when two distinct exam forms are part of the same assessment system and test developers plan to use a set of common items to link these exams and equate scores on the two forms, test developers should make sure that these common items are representative of the entire test form. They write that the common-item set should be a "mini-version of the total test form."[7] This practice makes common-items more useful for equating, but it also makes prior assessments more useful as coaching material because the items repeated this year from last year's assessment will be representative of the items on the entire assessment.

---

[6] See Hambleton, Swaminathan, and Rogers (1991) p. 135.

[7] See Kolen and Brennan (2010), p 19.

2.2 Item Properties and Guidelines for Item Development

Further, high-stakes assessment systems create strong incentives for educators to obtain copies of prior exams even if only a tiny portion of the items on this year's exam came from prior assessments. Modern assessment systems employ screening and pre-testing procedures that impose certain types of uniformity on test items. While this uniformity promotes reliable measurement when students are tested under low-stakes, it also implies that educators benefit from being able to coach their students on test taking techniques that help them do well when questions are asked in a particular manner or format.

The dominant paradigm in modern test development is Item Response Theory (IRT).[8] IRT models express the likelihood that an individual test taker gives a particular response to a particular question as a function of both latent traits of the test taker and latent traits of the question, and IRT models assume that these traits are invariant in the following sense: the characteristics of an item are not influenced by the traits of the test taker, and the traits of the test-taker are not altered by the difficulty of the items she is asked to answer.

Thus, as an example, the IRT paradigm requires that if Suzy is at a higher latent achievement level than Joe, then the probability that Suzy gets any item on a given exam correct is always greater than the corresponding probability for Joe regardless of the particular item in question. Further, if Betty is less likely to answer question A correctly than question B, then Jim is also less likely to answer question A correctly than question B, and a similar result must hold for all possible pairs of students and questions.

Test developers use diagnostic tools to determine whether or not specific items used in experimental sessions held prior to the public introduction of an assessment series or in experimental sections of a particular assessment form exhibit these invariance properties, and items that do exhibit these properties are more likely to be used in future assessments as questions that count toward students' scores. Note that if item-developers come up with a format or style for asking questions that results in desirable invariance properties, they face a strong incentive to develop more items that follow this format

[8] For more on IRT models, see Hambleton, Swaminathan, and Rogers (1991).

because they know such items are more likely to be retained after they are given experimentally. Thus, the incentives that test developers face within the IRT framework create incentives for educators to not only coach students on the answers to prior assessments but also to create new questions that mimic the styles and formats that are most common on prior assessments and drill students on test-taking strategies that are suited to these formats.

Other steps in the modern assessment development process also foster predictability in item styles and formats, and this point is easily illustrated if one examines the literature surrounding recent efforts to develop two new assessment systems that may be widely used in future years for both accountability and assessment purposes. In September of 2010, the Obama Administration awarded $330 million for the development of new assessments that will be used in state accountability systems. Two consortia of states received funding to create new assessments that will be available for the 2014-2015 academic year. The states in these two consortia are coordinating on both the development of new assessments and the development of curriculum standards that the assessments will address.

In December 2010, Bay-Borelli et al (2010) produced a report for a leading educational testing firm that discusses the process of developing tests that reliably assess student performance relative to the curriculum standards that these consortia are developing. As expected, their report discusses the need for pre-testing items to gather information that will "inform writing efforts so that the statistical characteristics of the resulting items are consistent with (expectations)."[9] But, Bay-Borelli et al (2010) also provide other details concerning the need to regulate the item development process. They write, "close alignment between the content of the items developed and the standards is best supported by the establishment of clear and specific item development guidelines, which are also called item development specifications. These guidelines are used to clarify the intent of the curriculum standards for both item writers and item reviewers."[10]

At one level, the Bay-Borelli et al (2010) recommendations make perfect sense. If one is trying to design a series of assessments that can be scored in a manner such that

[9] See Bay-Borelli (2010), page 25.

[10] See Bay-Borelli (2010), page 19.

the scores generated in 2015, 2016, 2017 and beyond are consistently scaled measures of student achievement relative to a specific set of standards, one must make a commitment ex ante to develop items for each yearly assessment that possess similar psychometric properties and also probe the content of the standards in the similar ways. However, these consortia are also interested in creating assessments that will be used to gather information for accountability systems, and it takes little imagination to see how the ex ante commitments that Bay-Borelli (2010) recommend may provide strong incentives for school districts, consulting firms, or even individual teachers to gain all the information they can about these ex ante guidelines for item development in order to develop coaching and test-preparation strategies that boost student performance on items that follow these guidelines.

The psychometric reasoning behind all of the Bay-Borelli (2010) guidelines is sound[11], and in the best of all possible worlds, assessment developers could prevent educators or education consultants from learning anything directly or indirectly about the precise content of the test specifications and other guidelines that govern the development of modern assessment forms. However, basic economic reasoning and considerable empirical evidence suggests that, when stakes are high, assessment developers will find it practically impossible to keep items or item development guidelines secret. Interested parties typically find ways to build coaching strategies around at least indirect evidence (from prior assessment forms) concerning what the item development guidelines actually are. Thus, in the presence of high-stakes, the procedures that modern assessment developers employ to promote reliable assessment and proper equating inevitably produce the material that educators use to form coaching and test-preparation strategies that waste class time and, over time, inflate student scores relative to true subject mastery.

[11] Kolen and Brennan (2010) assert that proper ex post equating of the results from different exam forms is not possible without an ex ante commitment to systematic procedures that govern item and form development, and they give several examples of cases where equating procedures did not work well ex post because different assessment forms in a series were not developed and administered in a consistent manner. See Chapter 8.

3. More of the Same?

The SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness of College and Careers (PARCC) are the two consortia of states who are developing new assessment systems for the Common Core State Standards using funds awarded as part of the Obama Administration's Race to the Top initiative. Both SBAC and PARCC aim to reduce the prevalence of multiple-choice questions and to require students to convey their knowledge and understanding in more varied formats and settings. Both will include constructed-response items and performance events that attempt to measure how well students can analyze, use, and present information. Further, the SBAC plan employs computer-adaptive testing that presents more in-depth challenges for students who demonstrate competence on early portions of the exam.[12]

Given these new features, it may seem reasonable to hope that, even if SBAC and PARCC induce coaching, that the coaching induced by these schemes will involve activities that at least approximate good teaching. However, recent experience with an assessment system similar to those proposed by SBAC and PARCC suggests that this may not be the case.

The American Institute of Certified Public Accountants (AICPA) administers the Uniform CPA exam and awards the credential of "Certified Public Accountant" based on the results of this exam. The Uniform CPA exam contains four separate exams covering specific areas of accounting practice. The exams are Auditing and Attestation (AUD), Financial Accounting and Reporting (FAR), Regulation (REG) and Business Environment and Concepts (BEC). In 2004, the AICPA abandoned a paper and pencil form of the exam and adopted a computer-administered, IRT format for the exam, and this format contains many of the features proposed by SBAC and PARCC, e.g. constructed response items, simulation exercises, and computer-adaptive testing.

Between 2010 and 2011, the AICPA made several changes to the Uniform CPA exam. The AICPA made changes to the computer simulations and question formats on all four exams. Further, the AICPA added questions that cover aspects of International Financial Reporting Standards (IFRS) to three of the exams: AUD, FAR, and BEC.

[12] See http://www.k12.wa.us/smarter/ and http://www.parcconline.org/

However, the AICPA made no significant changes to the content specifications for the REG.

The first full year of testing under the new exam was 2005. Figure 1 plots pass annual rates on the REG exam for 2005-2012,[13] and the patterns in Figure 1 are striking. The pass rates on REG rose from just over 40 percent in 2005 to over 50 percent in 2010 while the content and format of the exam remained roughly constant. Then, when the AICPA adopted a new computer format for the REG exam in 2011, pass rates fell dramatically.[14] These patterns are interesting because (i) the noteworthy 2011 changes in the REG exam involved changes in format not content and (ii) the AICPA claims that the IRT methods used to score the exam make sure that the standard for passing remains constant over time.

In fact, the following appears on the AICPA website:

> "In reviewing passing rates, it is important to remember that candidates are evaluated against an established standard of competence, and that the examination is scored and scaled so that scores are comparable across test forms and over time. The examination is not harder or easier to pass at different times. An increase in passing rates simply means that candidates are better prepared."[15]

Between 2005 and 2010, pass rates rose by almost 25%. Further, between 2010 and 2011, pass rates fell by more than 10% in one year. Since tens of thousands of

---

[13] The new exam was not given in the first quarter of 2004, and pass rates historically vary by quarter, with pass rates for first quarters being below the corresponding year-wide averages. The pass rate in the final three quarters of 2004 was almost identical to the 2005 annual pass rate and may have been slightly below if the exam was given in all four quarters of 2004.

[14] The pass rates for the other three components of the exam follow a similar pattern, but the patterns on other exams are more difficult to interpret because both the format and the item content of the other three exams changed substantially to reflect new international standards for accounting. The 2011 drop in annual pass rates is less than one percent for BEC and roughly five percent for AUD and FAR. For all three exams, the decline in pass rates is more pronounced when one compares pass rates from the first two quarters of 2011 to the pass rates from the first two quarters of 2010. The changes in content specifications for all exams were announced more than a year before the 2011 exams were administered.

[15] http://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/Passing\\Rates/Pages/default.aspx

persons take at least one section of the Uniform CPA exam in each quarter, these striking variations in pass rates cannot be attributed to sampling error.

It seems unlikely that swift improvements or collapses in the quality of college accounting classes could explain these patterns. So, given the AICPA's assertions concerning the consistency of the passing standard, one must ask what could be responsible for such dramatic changes in the preparation of candidates over such short intervals?

The CPA exam is a high-stakes exam for the examinees. Thus, examinees are willing to pay for advice concerning how to exploit the coaching opportunities that must be present given the exam's IRT format and the AICPA's desire for a fixed passing standard. In fact, a brief survey of websites for companies that offer CPA exam preparation classes reveals that these companies offer detailed advice and practice on test-taking strategies designed to maximize scores on exams that follow particular formats that are common on the Uniform CPA exam. Further, following the 2011 format change, one test-preparation service placed the following claim on their website,

> "(We used our team of) software experts to ensure that (our new test preparation) course's computer-based components mirrored exactly the new exam's functionality and replicated the exam-day experience for students."[16]

Thus, the most straightforward interpretation of the patterns in Figure 1 is that pass rates on REG rose as test-prep companies refined their test-taking strategies for the new exam, and scores fell as soon as the AICPA changed the exam format because the previous strategies were no longer optimal. In addition, the partial rebound in 2012 may be evidence that the test-preparation companies are now in the process of figuring out how to prepare candidates for the new format.

The Uniform CPA exam employs modern methods for assessment development and scoring. Thus, in a technical, psychometric sense, the AICPA's claims concerning a fixed standard for passing are true, but there are many ways to become "better prepared"

---

[16] http://www.becker.com/accounting/cpaexamreview/2011examchanges/\\nailed_the_exam.cfm?ff=true

for modern IRT exams, and it appears that some useful ways to prepare for the Uniform CPA exam involving practicing test-taking skills that are specific to the exam format used in a particular year. The fact that CPA candidates may waste considerable resources in their attempts to master test-taking techniques should be of little concern to the AICPA. If one assumes that the vast majority of examinees invest in learning the test-taking techniques taught in exam preparation classes, one must also believe that the AICPA is likely still awarding its credential to the examinees with the best command of accounting principles and practice. Further, the AICPA has no interest in minimizing the costs that new CPAs must pay to demonstrate their competence.

However, education policy makers have quite a different mission than the AICPA. If the exams being developed by the SBAC and PARCC consortia induce similar coaching activities that do not build real subject mastery, future accountability systems linked to these assessments will provide perverse incentives for teachers to allocate class time inefficiently.

The striking rise in REG pass rates between 2005 and 2010 suggests that, following the introduction of the SBAC and PARCC assessments, an entire industry of private firms may soon contact schools offering sophisticated test-prep services tailored to the sophisticated formats of these new exams. However, the sharp drop is REG passing rates between 2010 and 2011 raises concerns about the long-term value of the services these firms may offer.

4. A Different Approach

The current paradigm in education policy centers on defining achievement standards for students and then holding educators responsible for helping their students reach these standards. Thus, it is not surprising that the SABC and PARCC consortia are trying to develop single assessment systems that will simultaneously produce both reliable measures of student achievement and consequential measures of educator performance. However, given current testing technologies, there is no way to meet both of these objectives using a single assessment system. The very features of modern assessments systems that make reliable measurement relative to standards possible under

low-stakes testing create the opportunities for coaching behaviors that inevitably waste class time and inflate test scores when student exams become high-stakes assessments for educators.

Nonetheless, policy makers can form useful performance metrics for educators without scaled measures of student achievement and without knowing anything about how students performed this year relative to previous cohorts of students. This claim likely sounds heretical to many in the education policy and testing communities. However, economists who work on personnel policy have long understood that it is possible to build useful incentive schemes without ever specifying absolute performance standards. Organizations can often provide effective incentives for their workers by having them compete in properly structured contests and then rewarding the winners of these contests with prizes in the form of raises, bonus pay, or promotions. Employers do not always need scaled performance measures to provide incentives. Ordinal performance measures alone may be enough.[17]

In recent work (Barlevy and Neal (2012)), Gadi Barlevy and I demonstrate that policy makers can, in fact, build useful performance metrics for educators if they are simply able to form groups of students who are comparable in terms of their baseline skills and then later rank all students according to their final performance on a given assessment.[18] Policy makers can effectively quantify educator performance using the ordinal content of assessment outcomes without ever measuring the distance between the achievement levels of any two particular students or placing assessment results from different years on a common scale.

Note that, if test developers do not need to equate the results of assessment forms used in different years to gather data for accountability purposes, they have greater freedom to develop and design assessments in ways that eliminate prior information about item contents or formats. Given this freedom, test developers may be able to design high-stakes assessments that induce teachers to teach in ways that build deep

[17] See Lazear and Rosen (1981) as well as Chapters 10 and 11 in Lazear and Gibbs (2008).

[18] The performance metric we propose is called the Percentile Performance Index (PPI). It is similar in construction to Student Growth Percentile measures (SGP) that are already being used in some states as accountability measures; see Betebenner (2009). Free PPI software is available at http://sites.google.com/site/dereknealresearch/home/pay-for-percentile-software.

subject mastery and also prepare their students to convey this mastery in response to questions that take many different forms and formats.

To see the potential value of such an assessment system, consider a recent proposal by Eric Hanushek. Hanushek (2009) proposes that policy makers improve assessment-based accountability systems by developing an assessment system such that "teaching to the test" is equivalent to teaching well. Hanushek's recommendation involves creating a large and diverse item bank and then generating yearly assessments by taking random draws from this item bank.

Implicit in this recommendation is the idea that test developers can make the item bank so large and so diverse that educators will not find it in their interests to have students memorize answers from past assessments or devote time to test-taking strategies that are only optimal given specific item formats. However, it is useful to consider two different scenarios for how this proposal might be implemented. In scenario A, test developers assemble the item bank that Hanushek envisions using standard procedures, i.e. they retain only items that adhere to strict development guidelines and exhibit appropriate psychometric properties in field testing. In scenario B, developers place few restrictions on the item development process, encourage item developers to experiment with many different items formats and styles, and they also prohibit pre-testing of items.

If education authorities implemented Hanushek's proposal under scenario A, the researchers involved in the effort would be able to place student results from each annual assessment on a common scale and make reliable inferences about student proficiency relative to a set of predetermined standards. However, shortly after the introduction of this assessment system, interested parties would be working diligently and effectively to reverse engineer the item development specifications used to develop the item bank, and it is reasonable to expect that widespread coaching tailored to these specifications would soon follow.

In contrast, if developers followed the procedures described in B, they would find it difficult if not impossible to reliably measure individual student achievement levels in each year relative to some fixed set of proficiency standards. The best researchers could hope for is that they would be able to assign meaningful percentile scores (ranks) to students within populations that take the particular exam forms generated for specific

years. Further, the plan B procedures still may not produce ordinal measures of student achievement levels that are as reliable those ordinal ranks implied by results from modern assessment systems.

However, as strange as it may sound to those in the educational testing community, this reduction in reliability is not necessarily a problem. Performance metrics for educators typically reflect the distribution of test score outcomes over many students, and it may be possible to create performance metrics for educators that are reliable enough for effective incentive provision without creating measures of student achievement that meet modern standards for reliability.[19]

Such an approach may seem bizarre to many testing experts, but it resembles the approach that many academics already employ on college campuses for their own classes. In our internet age, it seems that, once a college exam is given, a copy of the exam is soon posted on some student website or stored in some student-run test bank so that students who take the same class in future years can study the specific questions on the exam.

In this environment, professors have two choices. First, they can recycle old exams or give exams where questions differ in minor ways from those on previous exams. This strategy saves time and makes consistent grading easy ex post, but many students will respond ex ante by devoting too much time to studying answers to a small set of questions and may leave the class with a superficial understanding of the material. Second, professors can adopt a policy of asking brand new questions each year that vary greatly in format but always probe student mastery of the material. The second approach requires more effort and makes grading much more difficult. In fact, professors who adopt this approach typically resort to grading on a curve, i.e. assigning grades based on the ordinal ranks implied by the exam results. However, this second approach provides better incentives for students to study in ways that build knowledge and skills that last longer than the vacation breaks between semesters.

---

[19]Standard results on optimal incentive contracts show that, if educators are risk-neutral, a reduction in reliability does not hamper efficient incentive provision. On the other hand, if educators are risk-averse, they will demand to be compensated for assuming the extra risk created by any drop in reliability. However, as the number of students that any educator or group of educators teaches grows large, this effect may well become a second order concern.

5. Other Methods of Assessing Teacher Performance

If one starts with the premise that parents and taxpayers expect public schools to foster not only cognitive development in children but also social and emotional development, then it is obvious that any performance metrics derived from assessments of cognitive skills, even optimal ones, can never provide all the information required for a comprehensive system of performance measurement and accountability for educators. Thus, one can easily imagine the need for additional evaluation systems that involve direct observation of classroom practice, examination of lesson plans, and safety inspections.

Although there is considerable evidence from both the private sector and public schools that observational evaluations of worker performance are often compromised by favoritism or leniency, there is no logical reason that this has to be the case.[20] If systems require evaluators to produce complete performance rankings over all schools in a particular comparison sets, then within each comparison set, some school must be the best performer and another must be the worst, and all schools cannot be deemed "above average." A firm commitment to ranking the performance of schools relative to the performance of other schools instead of relative to some administratively-defined standard can mitigate and possible eliminate concerns about rating inflation or leniency while providing incentives for schools to foster important skills that may not be assessed by cognitive achievement tests.

Conclusion

An entire industry now offers classes, books, and on-line tutorials that help students prepare for high-stakes screening exams like the SAT, ACT, GRE, LSAT, etc., as well as professional exams like the Uniform CPA exam. The purpose of these exams is to screen potential entrants to educational programs and professions, and these systems may serve as fairly efficient screening mechanisms as long as the skills required to

[20] See Prendergast (1999), Neal (2012).

perform well on these exams are, in equilibrium, highly correlated with the skills that institutions wish to screen.

However, modern psychometrics does not address the question of how to create performance metrics that serve as incentive mechanisms for educators because psychometricians are not trying to design systems that direct the efforts of educators. Psychometricians are trying to build assessments systems that measure student achievement in a coherent way while taking as given the methods that educators employ to prepare students for exams. In contrast, the designers of accountability systems are, by definition, trying to influence how teachers teach. Thus, those who design accountability systems cannot ignore the coaching behaviors that are universal responses to high-stakes administrations of modern assessments.

Figure 1 above suggests that even the most sophisticated IRT exams create opportunities for coaching activities that inflate scores when these exams are administered under high stakes. The sharp drop in REG pass rates in 2011 hints that the dramatic rise in pass rates during 2005 to 2010 period did not result from a revolution in the quality of college accounting classes but steady improvement in the capacity of test-preparation services to both anticipate the item content of exams and devise test-taking drills that are specific to a particular format. It seems reasonable to expect similar responses to the new assessments that are now being developed by the SBAC and PARCC consortia.

References

Barlevy, Gadi and Neal, Derek, "Pay for Percentile," *American Economic Review*, 2012, 102(5), pp. 1805-1831.

Bay-Borelli, Michael; Rozunick, Christine; Way, Walter D., and Weisman, Eric, "Considerations for Developing Test Specifications For Common Core Assessments: Adopting Curriculum Standards – Only the First Step," A White Paper from Pearson, December 2010.

Betebenner, Damian W., "Norm- and Criterion-Referenced Student Growth," *Educational Measurement: Issues and Practice*, Winter 2009, 28(4), pp. 42-51.

Campbell, Donald T., "Assessing the Impact of Planned Social Change," *Evaluation and Program Planning*, 1979, 2, pp. 67-90.

Glewwe, Paul; Ilias, Nauman, and Kremer, Michael, "Teacher Incentives," *American Economics Journal: Applied Economics*, 2010, 2(3), pp. 205-227.

Hambleton, Ronald; Swaminathan, H., and Rogers, H. Jane, *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park, CA, 1991.

Hanushek, Eric, "Building on No Child Left Behind," *Science*, 2009, 326(5954), pp. 802-803.

Holmstrom, Bengt and Milgrom, Paul, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design," *Journal of Law, Economics and Organization*, 1991, 7, pp. 24-52.

Jacob, Brian, "Accountability Incentives and Behavior: The Impact of High Stakes Testing in the Chicago Public Schools," *Journal of Public Economics*, 2005, 89(5), pp. 761-796

Kerr, Steven, "On the Folly of Rewarding A While Hoping for B," *The Academy of Management Executive*, 1995, 9(1), pp. 7-14.

Klein, Stephen P.; Hamilton, Laura S.; McCaffrey, Daniel F., and Stecher, Brian M., "What Do Test Scores in Texas Tell Us?" RAND Corporation, 2000.

Kolen, Michael J., and Brennan, Robert J., *Test Equating, Scaling, and Linking: Methods and Practices*, Springer Science, 2010.

Koretz, Daniel M. and Barron, Sheila, *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS)*, RAND Corporation Monograph Reports, Santa Monica, CA, 1998.

Koretz, David M., "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources*, 2002, 37(4), pp. 752-777.

Lazear, Edward, and Rosen, Sherwin, "Rank Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 1981, 89(5), pp. 841-864.

Lazear, Edward, and Gibbs, Michael, *Personnel Economics in Practice*, John Wiley & Sons, Hoboken, NJ, 2008.

Neal, Derek, "Providing Incentives for Educators," in *Handbook of Economics of Education*, Eric Hanushek, Steve Machin, and Ludger Woessmann, eds., 4 (Amsterdam: Elsevier), 2012.

Prendergast, Canice, “The Provision of Incentives in Firms,” *Journal of Economic Literature*, 37:1, 1999, pp. 7-63.

Rothstein, Richard, “Holding Accountability to Account, ” in *Performance Incentives: Their Growing Impact on American K-12 Education*, Matthew Springer, ed., (Washington DC: Brookings), 2009.

Vigdor, Jacob, “Teacher Salary Bonuses in North Carolina,” in *Performance Incentives: Their Growing Impact on American K-12 Education*, Matthew Springer, ed., (Washington DC: Brookings), 2009.

Figures

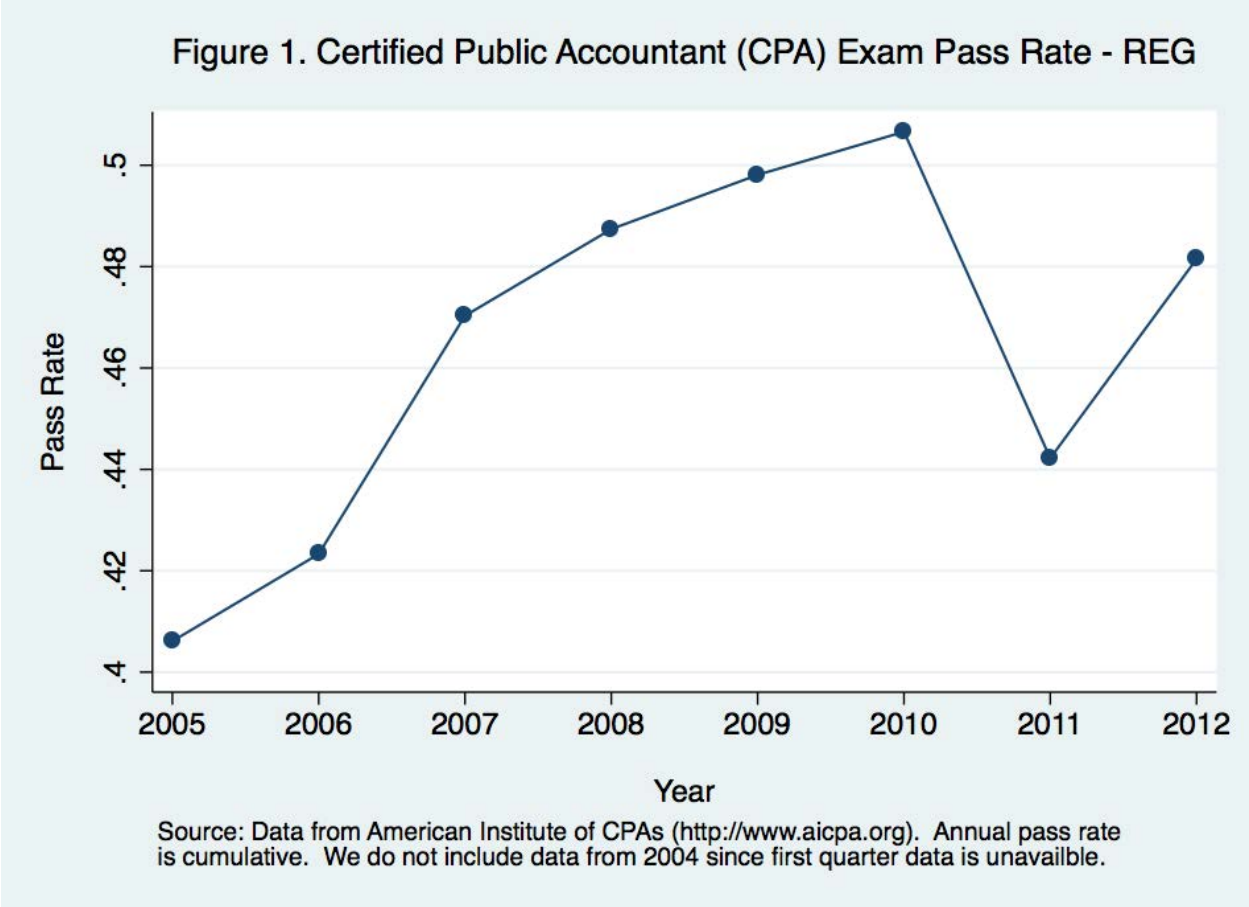
Figure 1. Certified Public Accountant (CPA) Exam Pass Rate - REG
Pass Rate
.4
.42
.44
.46
.48
.5
2005
2006
2007
2008
2009
2010
2011
2012
Year
Source: Data from American Institute of CPAs (http://www.aicpa.org). Annual pass rate is cumulative. We do not include data from 2004 since first quarter data is unavailble.