# Information Economics and Education Policy

Derek Neal

World Bank's Learning Assessment Symposium

November 7, 2013

# Two Uses of Government Statistics

- Monitoring Social Conditions
  - provide information as a public good
  - e.g. Vital Statistics, Climate Measures, Demographic Measures, etc
- Regulation / Accountability / Incentives
  - hold public agencies or government contractors "accountable" for their performance
    - hopefully, induce good performance
  - e.g. clearance rates for reported crimes, EPA "Performance Measures", AYP measures under NCLB

# Campbell's Law

*"I come to the following pessimistic laws (at least for the U.S. scene): 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." Donald Campbell, (1976)*

# The Initial Questions

- Are Campbell's observations really Laws?
- If there are exceptions, why and how?

# Multi-tasking

- Campbell was an empiricist. Holmstrom and Milgrom (1991) provide the theory that economists use to understand Campbell's observation.
- Designers of incentive schemes must confront problems that arise from
  - hidden actions
  - hidden information

# Multi-tasking

- Every study that claims to find "gaming" or "corruption" of an incentive system can be understood as an application of H&M.
  - Campbell's Law (1979) in Sociology
  - Kerr's (1995) "On the Folly of Rewarding A, While Hoping for B" in Management Science

# Multi-tasking

- However, this "Jeopardy answer" understanding of H&M is too shallow
- The model teaches three lessons
  - scale
  - alignment
  - are "unwanted" hidden actions substitutes or additions to the efforts that best promote the mission of the organization?

# Education Policy

- Modern Assessments systems are designed to create "scaled" scores – BUT
  - It is difficult to map these scales into "dollars"
  - The design features that promote reliable scales under "low stakes" invite coaching (hidden action) when tests are used for accountability as well as assessment
  - Correct scaling is difficult to verify, so systems that rely on scales invite corruption (hidden action, hidden information)
  - Because instruction time is roughly fixed in schools, coaching replaces teaching

# Design Problem

- How do we compare 5th grade math achievement in 2012 with 5th grade math achievement in 2011?
  - Need prior information about at least some of the items on the 2012 assessment (relative to the 2011 assessment)
    - repeat items
    - pre-test items from a test bank
  - KEY is that at least some of the 2012 items must be given to some test takes before 2012
- Creates opportunity for coaching and test-prep, and these opportunities are problematic in high stakes settings.

# IRT Continued

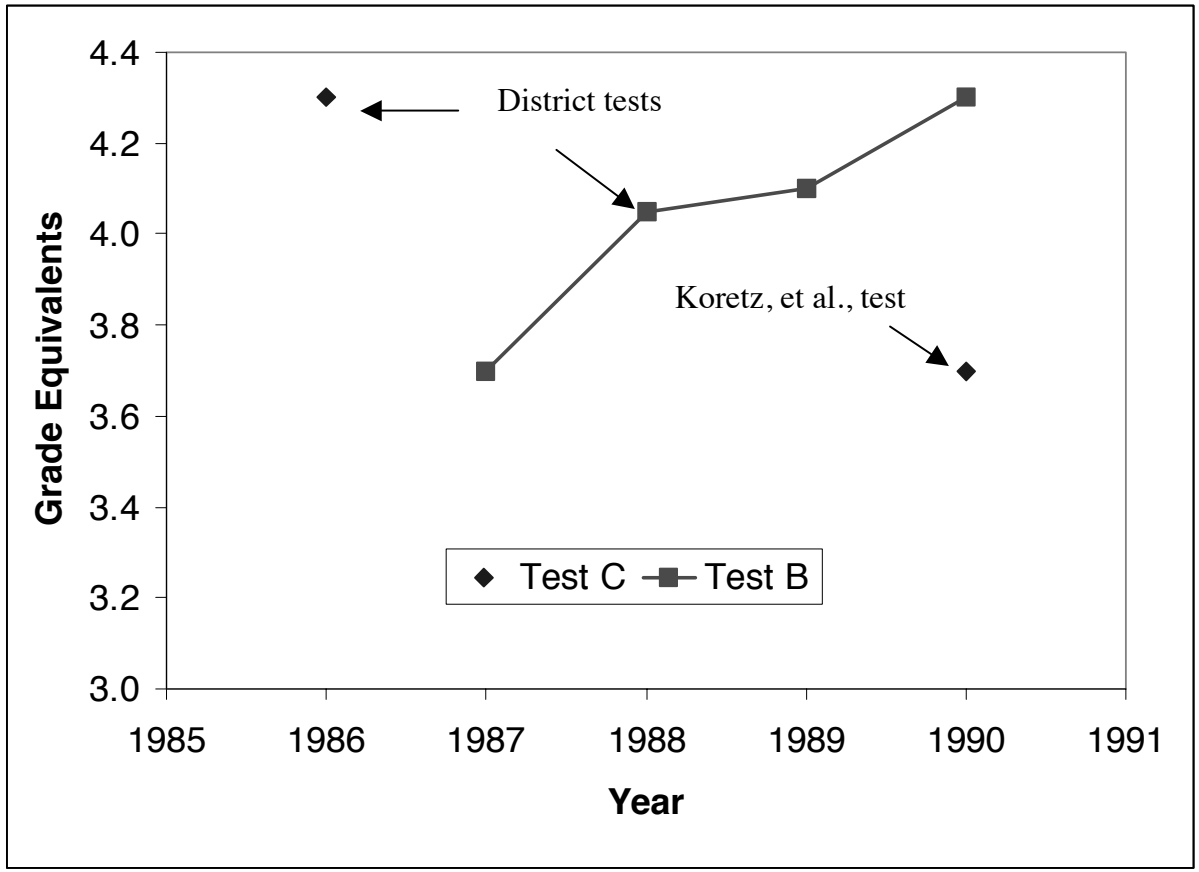
Bay-Borelli et al (2010) on the work of two consortia of states to develop new national assessments as part of the Race to the Top initiative.

"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."

This makes perfect sense if you want a series of assessments that can be scored reliably and consistently in 2015, 2016, 2017, etc. ... BUT what about the problems created by this type of predictability?

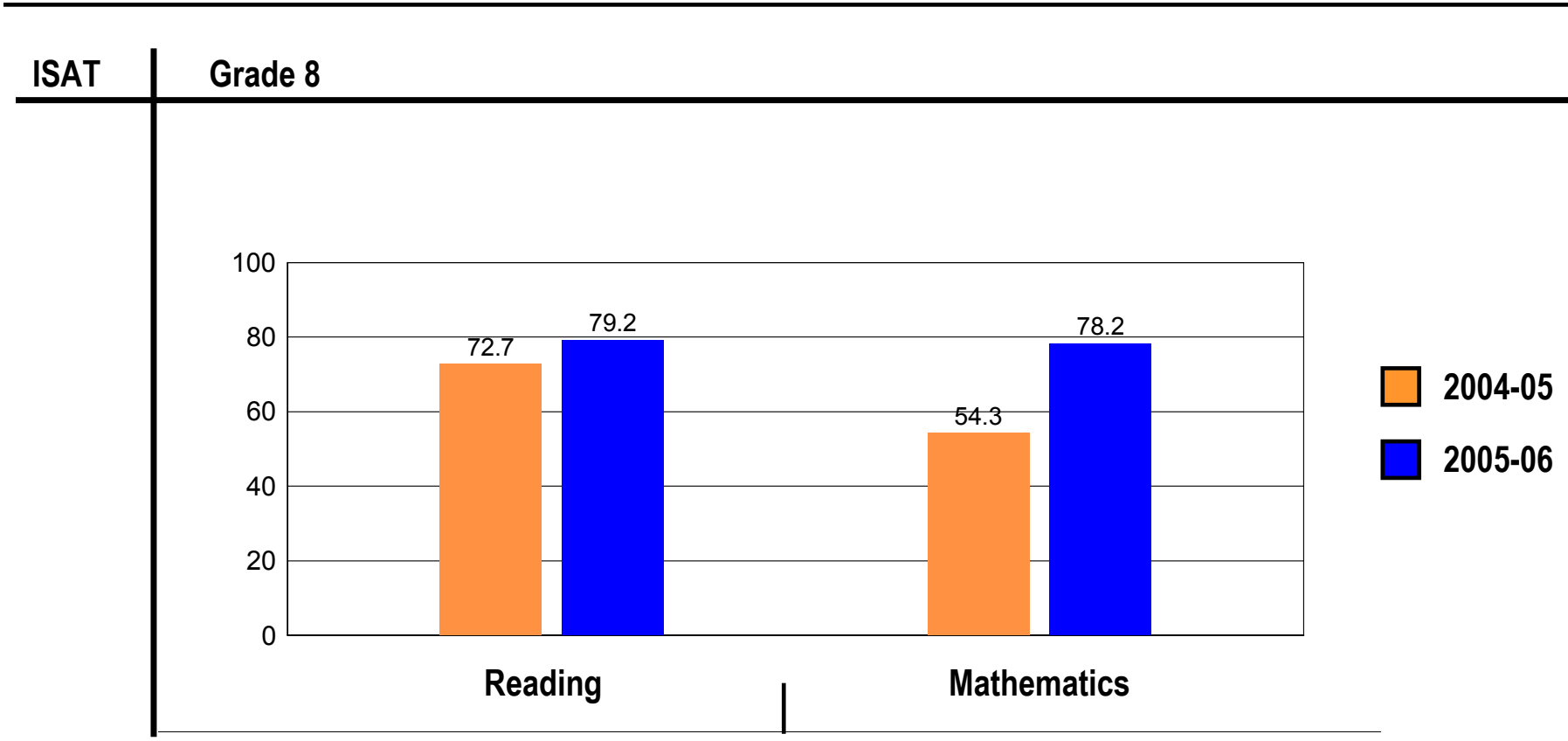
# Coaching vs Teaching

**Figure 2. Performance on Coached and Uncoached Tests, Third-Grade Mathematics**




SOURCE: Adapted from Koretz, Linn, Dunbar, and Shepard (1991).

# Measurement & Manipulation



# Fighting the Law

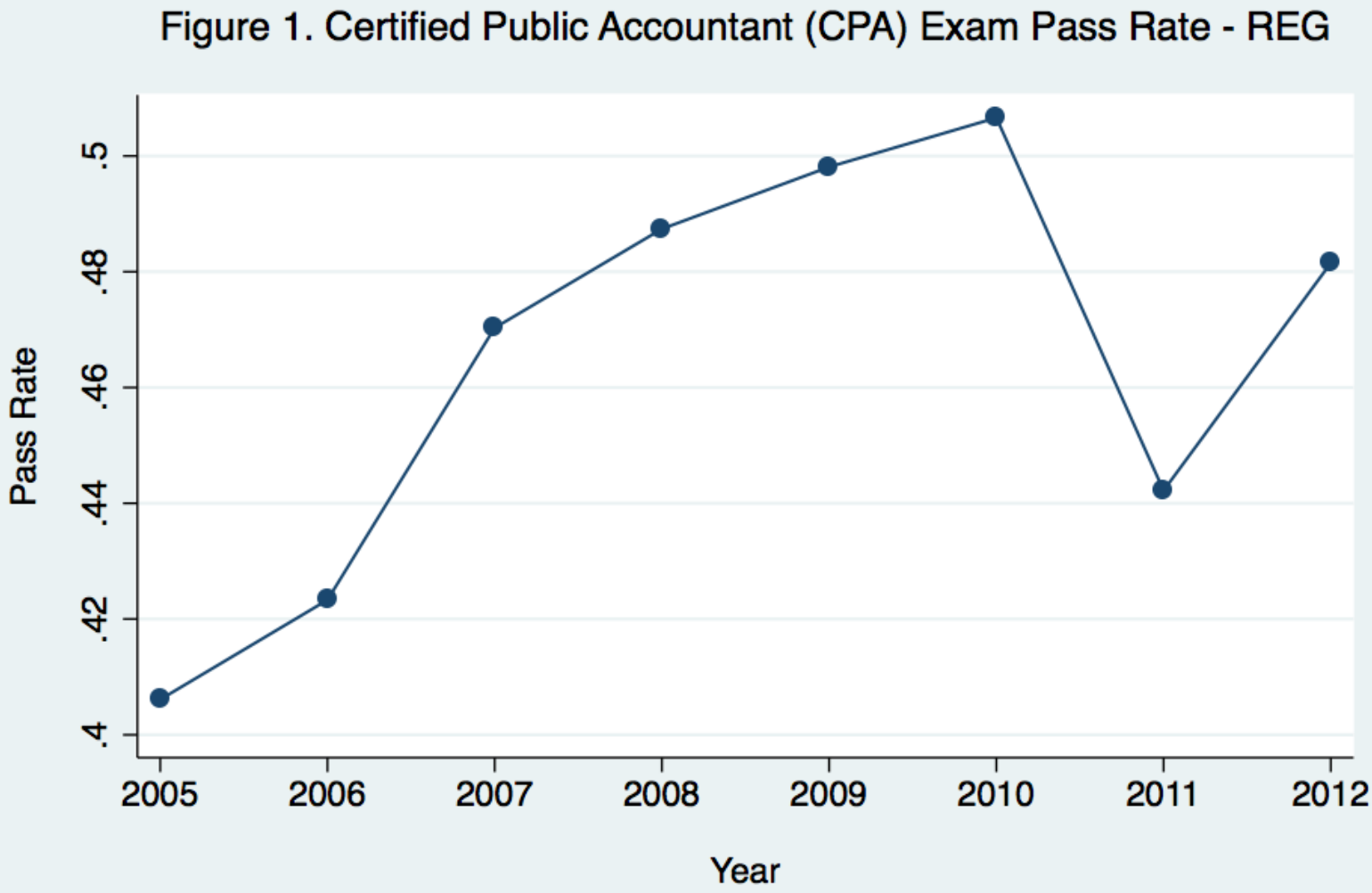
- External assessment administration is the first step in combating the third.
- Two Assessments for Two Purposes
  - More sophisticated versions of the current approach are not promising

# Tests Worth Teaching To?

SMARTER Balanced Assessment Consortium (SBAC) & Partnership for Assessment of Readiness of College and Careers (PARCC)

- Reduce reliance on multiple choice
  - e.g. PARCC promises math "problems worth doing."
- Require students to convey knowledge in multiple formats
- Constructed response items and performance events
- SBAC – computer adaptive testing

# What Should We Expect?



# Catch 22

- The desire for validity and reliable scales pushes modern methods toward predictable tests that capture a single dimension of achievement (invariance)
- Predictability and the requirements of invariance imply that simply seeking to "master" subject matter is never an optimal strategy for test takers
- Instead, coaching and test prep become optimal responses
- AND, over time, improvements in coaching techniques undermine the validity and reliability that modern systems are supposed to deliver.
- In the end, society gets wasted class time, distorted student effort, and contaminated measures of secular trends in achievement

# The Law

- In the context of education policy, Campbell's Law appears to be just that.
- Attempts to make one assessment system serve two functions create two failures
- Such systems "*distort and corrupt the educational processes they are intended to monitor.*"

# Ranks Are Enough for Incentives

- Teachers cannot coach if the item formats and specific items are not predictable or one dimensional
- There is no scope for scale manipulation if there is no scale
- Pay For Percentile is an assessment based-incentive scheme that
  - (i) employs new assessments and formats each period
  - (ii) employs only the ordinal content of assessment results
- The question is whether or not ordinal information is enough is a multi-output setting

# Barlevy & Neal (2012)

- Place each classroom in a league based on student characteristics.
- Place each student in a league based on his/her characteristics (including past scores)
- At end of year, give each student a percentile score that reflects rank within league
- Form weighted averages of these percentile scores to get Percentile Performance Indices (PPI)
- Pay Bonuses proportional to PPI
- Elicits efficient effort on ALL tasks that contribute to the education of ALL students

# Comparison Sets Are Key

- PFP amounts to competition among teachers within leagues defined by classroom type.
- Because competition involves all students in each classroom, teachers internalize the consequences of instructional spillovers.
- Relative Performance for a fixed prize pool implies that the scheme cannot be manipulated into a change in base pay.

# One Task At a Time

- Policy makers will do a better job of designing properly aligned performance metrics if they abandon the goal of scaling.
- Policy makers will do a better job of measuring student progress if they use a distinct measurement system that has no impact on the distribution of rewards and sanctions among teachers and principals.
- HOWEVER, an open question remains. Can we develop tests that induce "pursuit of mastery" while providing ranks that are "reliable enough" for use in incentive schemes?

# Big Picture

- The social goal of assessment-based accountability is to improve the allocation of teacher effort in classrooms, i.e. better teaching and more of it.
- Accountability programs will never induce the intended improvements unless the metrics they employ are designed to be properly aligned with this goal.
- Educators must believe that teaching well is their best strategy for improving their accountability measures.
- We do not yet know how to design such measures.
- We do know that "just use the tests we have" does not work.