Incentives and Test-Based Accountability in Education

ISBN
978-0-309-12814-8

150 pages
6 x 9
PAPERBACK (2011)

Michael Hout and Stuart W. Elliott, Editors; Committee on Incentives and Test-Based Accountability in Public Education; National Research Council

Summary

In recent years, there have been increasing efforts by the federal government and the states to devise systems that make students, teachers, principals, or whole school systems accountable for how much students learn. Large-scale tests are usually a key component of such systems. The No Child Left Behind (NCLB) Act of 2001 and the widespread use of high school exit exams in many states are two examples of a trend that has been going on for several decades.

The Committee on Incentives and Test-Based Accountability in Public Education was established by the National Research Council to review and synthesize research about how incentives affect behavior and to consider the implications of that research for educational accountability systems that attach incentives to test results. The committee focused on research about incentives in which an explicit consequence is attached to a measure of performance, starting first with basic research from the social and behavioral sciences and then turning to applied research in education.

BASIC RESEARCH ABOUT INCENTIVES

In reviewing basic research from the behavioral and social sciences about how incentives operate, the committee focused on theoretical research from economics and experimental research from psychology. Together, these two literatures show the way that subtle differences in the structure of incentives can be crucial in determining their effect. The

research review points to five key choices that should be considered in designing incentive systems:

1. *Who is targeted by the incentives*: In complex organizations, incentives can be designed for people in different positions who can affect outcomes in different ways.
2. *What performance measures are used*: The performance measures to which incentives are attached must be aligned with the desired outcomes for the incentives to have their desired effect.
3. *What consequences are used*: The size and structure of the consequences provided by the incentives will affect how the incentives operate and should be designed to be appropriate to the situation.
4. *What support is provided*: Without resources in support of organizational objectives, incentives can be discouraging to the very people they are intended to help, particularly if those people lack the capacity to reach the target that provides a reward or avoids a sanction.
5. *How incentives are framed and communicated*: To be effective incentives need to be framed and communicated in ways that reinforce people's commitment to the goal that incentives have been put in place to achieve, rather than in ways that erode that commitment.

The committee's research review also identified three issues related to evaluating the success of incentive systems:

1. *Nonincentivized performance measures for evaluation*: Incentives will often lead people to find ways to increase measured performance that do not also improve the desired outcomes. As a result, different performance measures—that are *not* being used in the incentives system—should be used when evaluating how the incentives are working.
2. *Changes in dispositions*: In addition to evaluating the changes in a set of defined objective outcomes, it is important to consider the way incentive systems affect people's dispositions to act when they are not being directly affected by the incentives.
3. *Weighing costs and benefits*: Incentive systems will typically generate a mix of costs and benefits that have to be weighed against each other to determine the net value of the system.

TESTS AS PERFORMANCE MEASURES

The tests that are typically used to measure performance in education fall short of providing a complete measure of desired educational

outcomes in many ways. This is important because the use of incentives for performance on tests is likely to reduce emphasis on the outcomes that are not measured by the test.

The academic tests used with test-based incentives obviously do not directly measure performance in untested subjects and grade levels or development of such characteristics as curiosity and persistence. However, those tests also fall short in measuring performance in the *tested* subjects and grades in important ways. Some aspects of performance in many tested subjects are difficult or even impossible to assess with current tests. And even for aspects of performance that can be tested, practical constraints on the length and cost of testing make it necessary to limit the content and types of questions. As a result, tests can measure only a subset of the content of a tested subject.

When incentives encourage teachers to focus narrowly on the material included on a particular test, scores on the tested portion of the content standards may increase while understanding of the untested portion of the content standards may stay the same or decrease. To the extent feasible, it is important to broaden the range of material included on tests to better reflect the full range of what students are expected to know and be able to do. And it is important to remember that the scores on the tests used with incentives may give an inflated picture of learning with respect to the full range of the content standards.

Incentives for educators are rarely attached directly to individual test scores; rather, they are usually attached to an indicator that combines and summarizes those scores in some way. Attaching consequences to different indicators created from the same test scores can produce dramatically different incentives. For example, an indicator constructed from average test scores or average test score gains will be sensitive to changes at all levels of achievement. In contrast, an indicator constructed from the percentage of students who meet a performance standard will be affected only by changes in the achievement of the students near the cut score defining the performance standard.

Given the broad outcomes that are the goals for education, the necessarily limited coverage of tests, and the ways that indicators constructed from tests focus on particular types of information, it is prudent to consider designing an incentive system that uses multiple performance measures. Incentive systems in other sectors have evolved toward using increasing numbers of performance measures on the basis of their experience with the limitations of particular performance measures. Over time, organizations look for a set of performance measures that better covers the full range of desired outcomes and also monitors behavior that would merely inflate the measures without improving outcomes.

INCENTIVE PROGRAMS REVIEWED

The committee's literature review focused on studies that allowed us to draw causal conclusions about the overall effects of test-based incentive programs. We looked specifically for information about outcomes *other* than the high-stakes tests that have incentives attached in order to avoid having our conclusions biased by the test score inflation that the incentives may have caused. We also attempted to contrast different incentive programs according to the key features identified by the basic research in economic theory (the first four features noted above): who is targeted by the incentives, what performance measures are used, what consequences are used, and what support is provided. The existing literature did not allow us to contrast incentive programs according to the way they frame and communicate incentives, the key feature identified by the basic research in psychology (the fifth feature noted above).

We focused on 15 test-based incentive programs, including the large-scale policies of NCLB, its predecessors, and state high school exit exams, as well as a number of experiments and programs carried out in both the United States and other countries. These various programs involved a number of different incentive designs and substantial numbers of schools, teachers, and students.

CONCLUSIONS

Conclusion 1: Test-based incentive programs, as designed and implemented in the programs that have been carefully studied, have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries. When evaluated using relevant low-stakes tests, which are less likely to be inflated by the incentives themselves, the overall effects on achievement tend to be small and are effectively zero for a number of programs. Even when evaluated using the tests attached to the incentives, a number of programs show only small effects. Programs in foreign countries that show larger effects are not clearly applicable in the U.S. context. School-level incentives like those of the No Child Left Behind Act produce some of the larger estimates of achievement effects, with effect sizes around 0.08 standard deviations, but the measured effects to date tend to be concentrated in elementary grade mathematics and the effects are small compared to the improvements the nation hopes to achieve.

Conclusion 2: The evidence we have reviewed suggests that high school exit exam programs, as currently implemented in

the United States, decrease the rate of high school graduation without increasing achievement. The best available estimate suggests a decrease of 2 percentage points when averaged over the population. In contrast, several experiments with providing incentives for graduation in the form of rewards, while keeping graduation standards constant, suggest that such incentives might be used to increase high school completion.

RECOMMENDATIONS FOR POLICY AND RESEARCH

The modest and variable benefits shown by test-based incentive programs to date suggest that such programs should be used with caution and that substantial further research is required to understand how they can be used successfully.

Recommendation 1: Despite using them for several decades, policy makers and educators do not yet know how to use test-based incentives to consistently generate positive effects on achievement and to improve education. Policy makers should support the development and evaluation of promising new models that use test-based incentives in more sophisticated ways as one aspect of a richer accountability and improvement process. However, the modest success of incentive programs to date means that all use of test-based incentives should be carefully studied to help determine which forms of incentives are successful in education and which are not. Continued experimentation with test-based incentives should not displace investment in the development of other aspects of the education system that are important complements to the incentives themselves and likely to be necessary for incentives to be effective in improving education.

Recommendation 2: Policy makers and researchers should design and evaluate new test-based incentive programs in ways that provide information about alternative approaches to incentives and accountability. This should include exploration of the effects of key features suggested by basic research, such as who is targeted for incentives; what performance measures are used; what consequences are attached to the performance measures and how frequently they are used; what additional support and options are provided to schools, teachers, and students in their efforts to improve; and how incentives are framed and communicated. Choices among the options for some or all of

these features are likely to be critical in determining which—if any—incentive programs are successful.

Recommendation 3: Research about the effects of incentive programs should fully document the structure of each program and should evaluate a broad range of outcomes. To avoid having their results determined by the score inflation that occurs in the high-stakes tests attached to the incentives, researchers should use low-stakes tests that do not mimic the high-stakes tests to evaluate how test-based incentives affect achievement. Other outcomes, such as later performance in education or work and dispositions related to education, are also important to study. To help explain why test-based incentives sometimes produce negative effects on achievement, researchers should collect data on changes in educational practice by the people who are affected by the incentives.

INCENTIVES AND TEST-BASED ACCOUNTABILITY IN EDUCATION

Committee on Incentives and Test-Based Accountability
in Public Education

Michael Hout and Stuart W. Elliott, *Editors*

Board on Testing and Assessment

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Awards B7990 and D08025 from the Carnegie Corporation of New York, and Awards 2006-7514 and 2007-1580 from the William and Flora Hewlett Foundation. Additional funding was also provided by the Presidents' Committee of The National Academies. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Carnegie Corporation of New York or the William and Flora Hewlett Foundation.

International Standard Book Number-13: 978-0-309-12814-8
International Standard Book Number-10: 0-309-12814-5

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu

Printed in the United States of America

Suggested citation: National Research Council. (2011). *Incentives and Test-Based Accountability in Education*. Committee on Incentives and Test-Based Accountability in Public Education, M. Hout and S.W. Elliott, *Editors*. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON INCENTIVES AND TEST-BASED ACCOUNTABILITY IN PUBLIC EDUCATION

Michael Hout (*Chair*), Department of Sociology, University of California, Berkeley
Dan Ariely, Fuqua School of Business, Center for Cognitive Neuroscience, and School of Medicine, Duke University
George P. Baker III, Harvard Business School
Henry Braun, Lynch School of Education, Boston College
Anthony S. Bryk, Carnegie Foundation for the Advancement of Teaching (*until 2008*)
Edward L. Deci, Department of Psychology, University of Rochester
Christopher F. Edley, Jr., School of Law, University of California, Berkeley
Geno Flores, California Department of Education
Carolyn J. Heinrich, LaFollette School of Public Affairs, University of Wisconsin–Madison
Paul Hill, School of Public Affairs, University of Washington
Thomas J. Kane, Graduate School of Education, Harvard University, and Bill & Melinda Gates Foundation, Seattle, Washington (*until February 2009*)
Daniel M. Koretz, Graduate School of Education, Harvard University
Kevin Lang, Department of Economics, Boston University
Susanna Loeb, School of Education, Stanford University
Michael Lovaglia, Department of Sociology, University of Iowa, Iowa City
Lorrie A. Shepard, School of Education, University of Colorado, Boulder
Brian Stecher, RAND Corporation, Santa Monica, California

Stuart W. Elliott, *Study Director*
Naomi Chudowsky, *Senior Program Officer* (until 2009)
Rose Neugroschel, *Research Assistant* (2009-2010)
Teresia Wilmore, *Senior Program Assistant* (until 2009)
Kelly Duncan, *Senior Program Assistant* (2009-2010)
Kelly Iverson, *Senior Program Assistant* (since 2010)

Preface

This project originated in the Board on Testing and Assessment (BOTA) in 2002 as the No Child Left Behind (NCLB) Act of 2001 was in its early stages of implementation. The initial discussions were sparked by the different perspectives on the use of test-based incentives by the board members, whose expertise included a wide range of disciplines. In particular, the board's interest in the topic was animated by the apparent tension between the economics and educational measurement literatures about the potential of test-based accountability to improve student achievement.

As a result of its early discussions, BOTA held workshops about the use of incentives in 2003 and 2005. These early discussions were funded, in part, by support for BOTA from the U.S. Department of Education and the U.S. National Science Foundation. After these workshops the board identified, defined, and sought support for the research synthesis the board concluded could be undertaken. With generous funding from the Carnegie Corporation of New York and the William and Flora Hewlett Foundation, the Committee on Incentives and Test-Based Accountability in Public Education was appointed in early 2007 to carry on the work that BOTA had started.

The charge called for the committee to examine research related to the use of incentives and to synthesize its implications for the use of test-based incentives in education. The committee held three meetings, as well as a workshop on multiple measures and NCLB that was supported by

additional funding from the Carnegie Corporation, the Hewlett Foundation, and the Presidents' Committee of The National Academies.

When work began on this topic 9 years ago, no one expected that the project would occupy most of a decade or that it would provide such an opportunity to survey a remarkable period of educational change. As the report notes in Chapter 1, the use of test-based incentives in education has been growing for several decades. However, it was in the first decade of the 21st century—which saw the enactment of NCLB, the maturation of the state movement for using high school exit exams, and the strong interest in using newly-available student test data to tie teacher pay to value-added analyses of their students' test results—that the use of test-based incentives truly took hold of the education policy world. At the same time, there has been a transformation in the rigor of the methods used to analyze educational data. The combination of policy experimentation and new research methods has produced the set of studies that are reviewed in this report. We note that few of these studies were available when BOTA started down this path in 2002.

Over the course of this work, we have benefited from the generous contributions of many individuals. Three members of BOTA provided the key impetus in the initial development of the ideas and the definition of the current project: Chris Edley, Daniel Koretz, and Edward Lazear. The project would never have come together without their suggestions and encouragement. In addition, the suggestions of the staff of the project's funders—Barbara Gombach and Talia Milgrom-Elcott at the Carnegie Corporation of New York, and Marshall (Mike) S. Smith at the William and Flora Hewlett Foundation—helped define a balanced and workable study. We are grateful for their suggestions for shaping the project and for their patience as the work has unfolded.

In addition to the members of BOTA, a number of individuals made invited presentations at the initial 2003 and 2005 workshops that developed the project, and we thank them: Hilda Borko, University of Colorado; Edward Deci, University of Rochester; Eric Hanushek, Stanford University; Carolyn Heinrich, University of Wisconsin, Madison; Richard Ingersoll, University of Pennsylvania; Richard Koestner, McGill University; Michael Kramer, Harvard University; Victor Lavy, Hebrew University of Jerusalem; Harry O'Neil, University of Southern California; and Brian Stecher, RAND.

The committee's workshop on multiple measures in 2007 included a number of invited presentations that helped the committee explore the use of multiple measures and refine its thinking about their use, and we are grateful for this input: Robert Bernstein, California Department of Education; Kerri Briggs, U.S. Department of Education; Mitchell Chester, Ohio Department of Education; Daniel Fuller, Association for Supervision and Curriculum Development; Drew Gitomer, Educational Testing

Service; Kati Haycock, Education Trust; Jan Hoegh, Nebraska Department of Education; Lindsay Hunsicker, Office of Senator Enzi; Robert Linn, University of Colorado; Jill Morningstar, House Education and Labor Committee; Roberto Rodriguez, Office of Senator Kennedy; and William Taylor, Citizens' Commission on Civil Rights.

As we finalized the report's text, we received assistance from a number of the authors of studies cited to ensure that we were accurately describing their study conclusions. We thank the following researchers for their assistance: Eric Bettinger, Stanford University; Thomas D. Cook, Northwestern University; Roland Fryer, Harvard University; Steven M. Glazerman, Mathematica Policy Research; Brian A. Jacob, University of Michigan; Victor Lavy, Hebrew University of Jerusalem; Jaekyung Lee, State University of New York, Buffalo; Karthik Muralidharan, University of California, San Diego; Sean F. Reardon, Stanford University; John Robert Warren, University of Minnesota; and Manyee Wong, Northwestern University.

The committee's work was assisted by members of the National Research Council (NRC) staff. Naomi Chudowsky worked closely with the committee members to turn their discussions into initial draft text. Teresia Wilmore, Kelly Duncan, Rose Neugroschel, and Kelly Iverson provided administrative support and research assistance throughout the course of the project. The text was greatly improved by the expert editing of Chris McShane, Eugenia Grohman, and Yvonne Wise. Finally, a project of this duration experiences more than its share of institutional hurdles; we are deeply indebted to the efforts of several NRC staff: Michael Feuer, Patricia Morison, Connie Citro, and Robert Hauser for their help and encouragement throughout the project.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the NRC Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report: Eric Bettinger, School of Education, Stanford University; Martha Darling, consultant, Ann Arbor, MI; David P. Driscoll, consultant, Melrose, MA; Amanda M. Durik, Department of Psychology, Northern Illinois University; Edward Haertel, School of Education, Stanford University; Jane Hannaway, Education Policy Center, Urban Institute, Washington, DC; Joseph A. Martineau, Office of Educational Assessment and Accountabil-

ity, Michigan Department of Education; Lorraine McDonnell, Department of Political Science, University of California at Santa Barbara; Michael S. McPherson, Office of the President, Spencer Foundation, Chicago, IL; Barbara Reskin, Department of Sociology, University of Washington; and Lauress (Laurie) L. Wise, Human Resources Research Organization (HumRRO), Monterey, CA.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions and recommendations nor did they see the final draft of the report before its release. The review of this report was overseen by Charles E. Phelps, university professor and provost emeritus, University of Rochester and Richard J. Shavelson, School of Education, Stanford University. Appointed by the NRC, they were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report, however, rests entirely with the authoring committee and the institution.

Michael Hout, *Chair*
Stuart W. Elliott, *Study Director*
Committee on Incentives and Test-Based
Accountability in Public Education

Contents