

Letter Report to the U.S. Department of Education on the Race to the Top Fund

Board on Testing and Assessment; National Research
Council

ISBN: 0-309-14579-1, 19 pages, 8 1/2 x 11, (2009)

This free PDF was downloaded from:

<http://www.nap.edu/catalog/12780.html>

Visit the [National Academies Press](#) online, the authoritative source for all books from the [National Academy of Sciences](#), the [National Academy of Engineering](#), the [Institute of Medicine](#), and the [National Research Council](#):

- Download hundreds of free books in PDF
- Read thousands of books online, free
- Sign up to be notified when new books are published
- Purchase printed books
- Purchase PDFs
- Explore with our innovative research tools

Thank you for downloading this free PDF. If you have comments, questions or just want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, [visit us online](#), or send an email to comments@nap.edu.

This free book plus thousands more books are available at <http://www.nap.edu>.

Copyright © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

Division of Behavioral and Social Sciences and Education
Board on Testing and Assessment

500 Fifth Street, NW
Washington, DC 20001
Phone: 202 334 2353
Fax: 202 334 1294
E-mail: botal@nas.edu
www.nationalacademies.org

October 5, 2009

The Honorable Arne Duncan
Secretary of Education
U.S. Department of Education
400 Maryland Avenue, SW, Room 3W329
Washington, DC 20202

Dear Mr. Secretary:

This letter offers comments concerning the Department's Proposed Regulations on the Race to the Top (RTT) fund of the American Recovery and Reinvestment Act of 2009 (74 Fed. Reg. 37804, proposed July 29, 2009) from the Board on Testing and Assessment of the National Research Council. (See Attachment A for a list of members.) The comments reflect a consensus of the Board.

Under National Academies procedures, any letter report must be reviewed by an independent group of experts before it can be publicly released, which made it impossible to complete the letter within the public comment period of the *Federal Register* notice.¹ However, we hope that the Department will still find these comments helpful in revising the RTT plans.

The Board on Testing and Assessment stands ready to assist the federal government, Congress, and the states in addressing issues concerning the use of evidence to improve educational opportunities for the nation's young people.

Sincerely yours,



Edward H. Haertel, *Chair*
Board on Testing and Assessment

cc: Carmel Martin, Assistant Secretary for Planning, Evaluation, and Policy Development
Thelma Melendez, Assistant Secretary for Elementary and Secondary Education
Joanne Weiss, Senior Advisor to the Secretary and Director, Race to the Top
Marshall S. Smith, Senior Counselor to the Secretary, Director, International Affairs
John Q. Easton, Director, Institute of Education Sciences

¹ The reviewers for the report are acknowledged in Attachment B.

COMMENTS ON THE DEPARTMENT OF EDUCATION'S PROPOSAL ON THE RACE TO THE TOP FUND

This letter offers comments concerning the Department's Proposed Regulations on the Race to the Top (RTT) fund of the American Recovery and Reinvestment Act (ARRA) of 2009 (74 Fed. Reg. 37804, proposed July 29, 2009) from the Board on Testing and Assessment (BOTA) of the National Research Council. (See Attachment A for a list of members.) The comments reflect a consensus of the Board.

The Board held an open session at its meeting on July 30, 2009, to discuss the importance of evaluating RTT spending. This meeting had been planned before the Department's posting in the *Federal Register*. In deciding to hold this session, BOTA hoped to ensure that the unusual opportunity offered by RTT to invest in educational innovations be fully exploited through evaluation of the supported programs. Although RTT represents a substantial amount of money in absolute terms, it is small in comparison with the total spending on education in the United States. It is only through careful evaluations from RTT-supported innovations that the investment in RTT is likely to be leveraged to support improvements that can affect the entire educational system. Without the learning that results from careful evaluation, any benefits of this one-time spending on innovation are likely to end when the funding ends.

Tests often play an important role in evaluating educational innovations, but an evaluation requires much more than tests alone. A rigorous evaluation plan typically involves implementation and outcome data that need to be collected throughout the course of a project. In addition, a rigorous evaluation plan may affect the way that participants are selected to be included in the project.

Originally, BOTA's July 30 meeting was planned to include several members of the Department, to allow a discussion of possible approaches for requiring and conducting evaluations of RTT innovations. However, as a result of the *Federal Register* filing on the preceding day, most of the Department's representatives were unable to attend BOTA's meeting. In addition, the availability of the Department's full proposal caused the Board to consider and discuss a number of other testing-related elements in the proposal. As a result, the Board concluded that it would be important and appropriate to communicate directly with the Department in the form of a letter, both to underline the importance of requiring evaluations for RTT-supported innovations and to address several other concerns about the use of tests suggested in the *Federal Register* notice.

The comments we offer in this letter are based on more than 15 years of work by BOTA on a wide range of topics concerning the use of testing and assessment. We draw from that body of work and from our collective knowledge about accepted practices in educational measurement to offer our comments here. (See Attachment C for a selected list of BOTA reports.)

We begin our comments with a general discussion about the role of testing and assessment in educational reform. This discussion underlines some important aspects of testing and assessment that are widely understood but not always considered in the design of educational policies that use tests. We then discuss six testing-related topics that are reflected in the Department's proposal in the *Federal Register*:

1. the need for states to include evaluations in their RTT program activities;
2. the use of NAEP to evaluate RTT outcomes;
3. the use of student growth data to evaluate teachers and principals;
4. the use of data to improve instruction;

Comments on Race to the Top Proposal

5. challenges in developing common assessments; and
6. challenges in creating longitudinal data systems.

We focus our comments on these six issues because we have concerns in each case about how they will be addressed and the resulting implications for the way that tests and assessments will be used in implementing education policy. Throughout, the Departments' specific proposed language is in italics, with the page numbers taken from the notice in the *Federal Register*.

RELIANCE ON TESTING AND ASSESSMENT IN EDUCATIONAL REFORM

The proposed regulations rely heavily on the use of testing and assessment both to drive and to measure education reform. BOTA strongly supports the Department's desire to incorporate objective measures of student performance to inform the process of educational reform. Standardized assessments have much to offer in this regard. BOTA has noted in *Lessons Learned* (p. 5):

In many situations, standardized tests provide the most objective way to compare the performance of a large group of examinees across places and times.... Similarly, in K-12 education, statewide standardized tests are useful for comparing student achievement across classrooms, schools or districts, or across different times. When combined with other sources of information, these comparisons can help educators and policy makers decide how to target resources.

However, the *Lessons Learned* report (p. 5) also cautions that "A test score is an estimate rather than an exact measure of what a person knows and can do." The items on any test are a sample from some larger universe of knowledge and skills, and scores for individual students are affected by the particular questions included. A student may have done better or worse on a different sample of questions. In addition, guessing, motivation, momentary distractions, and other factors also introduce uncertainty into individual scores. When scores are averaged at the classroom, school, district, or state level, some of these sources of measurement error (e.g., guessing or momentary distractions) may average out, but other sources of error become much more salient. Average scores for groups of students are affected by exclusion and accommodation policies (e.g., for students with disabilities or English learners), retest policies for absentees, the timing of testing over the course of the school year, and by performance incentives that influence test takers' effort and motivation. Pupil grade retention policies may influence year-to-year changes in average scores for grade-level cohorts.

Moreover, test results are affected by the degree to which curriculum and instruction are aligned with the knowledge and skills measured by the test. Educators understandably try to align curriculum and instruction with knowledge and skills measured on high-stakes tests, and they similarly focus curriculum and instruction on the kinds of tasks and formats used by the test itself. For these reasons, as research has conclusively demonstrated, gains on high-stakes tests are typically larger than corresponding gains on concurrently administered "audit" tests, and sometimes they are much larger.² Improvements on the necessarily limited content of a high-stakes test may be offset by losses on other, equally valuable content that happens to be left untested.

² For references and further discussion of score inflation, see Koretz (2008, Chapter 10).

Comments on Race to the Top Proposal

These issues do not mean that test scores are unimportant or generally invalid. They do mean that test use should comport with relevant professional standards. A list of relevant professional and technical standards applicable to the use of educational assessments is included in Attachment D. Three core issues reflected in these documents are (1) the need to develop multiple measures of key outcomes, ideally using multiple assessment formats; (2) the need to validate these assessments for specific uses; and (3) the need to consider the populations involved and any associated validity and fairness issues. These documents also point to several technical issues (e.g., the reliability or precision of the measures, scaling and equating issues) and operational issues (e.g., data collection and processing) that can undermine the interpretability and usefulness of the test results if not handled appropriately. The documents provide guidelines on how to avoid these kinds of problems. In addition, a thorough account of the steps needed to improve the gathering and use of data by state assessment systems can be found in the BOTA report, *Systems for State Science Assessment* (2006).

We encourage the Department to pursue vigorously the use of multiple indicators of what students know and can do. A single test should not be relied on as the sole indicator of program effectiveness. This caveat applies as well to other targets of measurement, such as teacher quality and effectiveness and school progress in closing achievement gaps. Development of an appropriate system of multiple indicators involves thinking about the objectives of the system and the nature of the different information that different indicators can provide. Such a system should be constructed from a careful consideration of the complementary information that is provided by different measures.³

SPECIFIC AREAS OF CONCERN

Need for States to Include Evaluations in their RTT Program Activities

The Department specifically requests comment on whether states should be required to conduct evaluations of their support programs and indicates that the final notice for RTT applications will specify the nature of the evaluations that will be required:

The State and its participating LEAs must use funds under this program to participate in a national evaluation of the program, if the Department chooses to conduct one. In addition, the Department is seeking comment on whether a State should, instead of or in addition to a national evaluation, be required to conduct its own evaluation of its program activities using funds under this program. The Department will announce in the notice inviting applications the evaluation approach(es) that will be required. (Section II.D.a, p. 37808)

We strongly affirm the importance of independent evaluation of all RTT initiatives. These evaluations should be conducted at all levels: national, state, and local. Such evaluations will help foster a culture of continuous improvement in the nation's educational systems. As part of the evaluation, all consequences—both positive and negative, and intended and unintended—should be carefully monitored and weighed.

At BOTA's July 30 meeting to discuss the Department's proposed regulations, there was strong support for rigorous evaluation requirements to help ensure both the success of the states'

³ For a thoughtful description of the creation of such a system of multiple indicators, see Chester (2005).

Comments on Race to the Top Proposal

RTT policy and program implementations and to increase scientific understanding of effective school reform. Strong evaluation requirements can also help to establish a culture of evidence-based reform at the state and local levels, which will pay dividends long after the short-term RTT stimulus funding has ended.

As part of its discussion, BOTA considered the example of welfare reform in 1990s, which was informed by a history of state experimentation, in which states received waivers from federal requirements under the condition that they conduct evaluations of those experiments (Harvey, Camasso, and Jagannathan, 2000). The evaluation requirements from the federal government led to the scientific study of the impact of crucial features of welfare policy that were tried out in these state experiments. The RTT fund offers an analogous opportunity in education: a careful choice of evaluation requirements can result in substantial opportunities for research and learning about the educational innovations that are supported.

BOTA's discussion on evaluation focused on six key principles: (1) theory of action, (2) appropriateness of design, (3) documentation of implementation, (4) formative and summative components, (5) independent evaluators, and (6) valid outcome measures. These principles are commonly accepted as best practice in evaluation.⁴ We briefly discuss each of these requirements below.

Theory of Action The design of each evaluation will benefit from a clear description by the applicant of the mechanism by which the reform initiative is expected to improve student learning outcomes. This “theory of action” will typically take the form of a connected set of propositions—a logical chain of reasoning that explains how the reform expenditures will lead to improved schooling practices. A theory of action “connects the dots,” explaining in commonsense terms which program features are expected to produce which results in order to reach the final desired outcome. A theory of action gives shape to any evaluation plan. The evaluation is framed as an investigation of each link in the chain, focusing on those links for which there may be only limited prior evidence. If the evaluation demonstrates that the reform initiative failed to meet its objectives, the investigation of the individual links will facilitate the redesign of the reform initiative to make it effective.

For example, if the theory of action for an accountability program based on tests assumes that performance will improve because educators will use test results to refine their instruction in areas in which students perform poorly, then an evaluation of that accountability program should look at the ways that the tests are being used to influence teaching. If the test results are used to refocus the curriculum towards areas in which students perform poorly, that would confirm the result expected by the theory of action; in contrast, if the test results lead educators to shift their attention from teaching to test preparation, then that would not confirm the result expected by the theory of action and suggest the possibility of unintended consequences.

Appropriateness of Design There is no one best design for program or policy evaluation. Whatever design is chosen must be appropriate to the program or policy being evaluated. A strong design will typically involve the collection of both quantitative and qualitative information. The choice of a design should be guided by the theory of action of the intervention or activity, by best practices in the field, and by a clear sense of the purpose of the evaluation: that is, what questions are to be answered and how they are to be prioritized. An evaluation

⁴ For further information, see *The Program Evaluation Standards* (Joint Committee on Standards for Educational Evaluation, 1994) listed in Attachment D.

Comments on Race to the Top Proposal

should be designed before the program is implemented so there is an opportunity to collect baseline data and for the evaluation plan to influence crucial decisions about the selection of participants for the program and the types of implementation and outcome information to be collected.

Documentation of Implementation The evaluation must include documentation of the extent to which the policy or program was in fact implemented as intended. Failure to implement a program adequately can be one explanation for a program's failure to produce the desired effect. Any significant problems encountered along the way should be documented so that future efforts can profit by lessons learned from those problems. Implementation evaluation can also help in spotting potential negative, unintended effects of reforms.

Formative and Summative Components Ideally, evaluation plans will provide for both short-term, "formative" feedback for program improvement, and longer-term, "summative" information to judge program impact. The formative component of an evaluation can be relatively informal, featuring quick turnaround, so that the local educators or state officials in charge of the program can adjust implementation and fine-tune policies as needed. Formative evaluation identifies ways to adjust or improve program implementation in real time. A summative evaluation is more formal, longer term, and typically requires more advance planning. This kind of evaluation does not influence the implementation of the policy or program; rather, it documents the extent of intended and unintended outcomes, using a rigorous research design and statistical methods.

For RTT initiatives, both formative and summative evaluation components should be included. Both components will benefit from documentation about implementation. When possible, summative evaluations should be designed in ways that provide generalizable knowledge to guide future education reform.

Independent Evaluators Evaluations should employ independent and well-qualified external evaluators. The field of evaluation is well developed, and much is known about sound evaluation practice that should be incorporated in an evaluation. Outside (external) evaluations will be more credible than evaluations conducted in-house by the states or LEAs themselves.

Valid Outcome Measures Evaluations must use student outcome measures that are valid for the intended use. Outcomes must reflect intended program or policy goals. An otherwise sound evaluation may be undermined by reliance on a test that is poorly aligned with the intervention. An outcome measure should not be chosen simply because it is inexpensive or readily available. Even a test's alignment to content standards, as typically implemented, may be an insufficient criterion for the choice of an outcome measure. Alignment of a test to a state's academic content standards typically means that the items on the test can each be matched to one or more specific content standards, in accordance with the test specification. But the standards are often much broader and more complex than any of the corresponding items. Taken together, the items representing a standard may be a pale reflection of the intent of that standard. Moreover, some standards are typically omitted altogether from the test specification. For these reasons, "alignment," as generally implemented, is not sufficient to ensure that a test is appropriate as an outcome measure for an evaluation.

Use of NAEP to Evaluate RTT Outcomes

Although BOTA is a strong advocate for the importance of NAEP in measuring U.S. educational progress, NAEP cannot play a primary role in evaluating RTT initiatives, a role that might be mistakenly inferred from the language in the Department's proposal.

We propose using the NAEP to monitor overall increases in student achievement and decreases in the achievement gap over the course of this grant because the NAEP provides a way to report consistently across Race to the Top grantees as well as within a State over time. . . . (Section I, footnote 1, p. 37805)

It is necessary to be clear about the distinction between the requirements of an evaluation and the kind of "monitoring" that NAEP can provide. For the purposes of evaluating RTT initiatives, there are at least four critical limitations with regard to inferences that can be drawn from NAEP.

1. NAEP is intended to survey the knowledge of students across the nation with respect to a broad range of content and skills: it was not designed to be aligned to a specific curriculum. Because states differ in the extent to which their standards and curricula overlap with the standards assessed by NAEP, it is unlikely that NAEP will be able to fully reflect improvements taking place at the state level.⁵
2. Although NAEP can provide reliable information for states and certain large school districts, it cannot, as presently designed, fully reflect the effects of any interventions targeted at the local level or on a small portion of a state's students, such as are likely to be supported with RTT initiatives.
3. States are likely to undertake multiple initiatives under RTT, and NAEP results, even at the state level, cannot disaggregate the contributions of different RTT initiatives to state educational progress.
4. The specific grade levels included in NAEP (grades 4, 8, and 12) may not align with the targeted populations for some RTT interventions.

Consequently, NAEP will be of limited value in judging the success or failure of individual initiatives under RTT, even at the state level. The availability of NAEP does not in any way obviate the need to plan rigorous evaluations—at national, state, and local levels—that are appropriately designed to assess the implementation and outcomes of RTT initiatives.

These concerns do not diminish the importance of NAEP as an independent measure of educational progress at the state and national levels. NAEP has served and should continue to serve an essential "audit" function for evaluating states' reform efforts, using objective measurements that are comparable across states and over time. In order that it continue to serve that function, it is critically important that high-stakes decisions *not* be attached directly to NAEP results. What makes NAEP so valuable is precisely the fact that there are *no* incentives attached to it that might lead educators to "teach to the test." Consequently, NAEP is relatively immune to the pressures of score inflation that can distort any high-stakes measure. NAEP can

⁵ Continued progress toward common standards across the states, as expected by the Department's proposal, could lead to a shift in NAEP standards to align more closely with those emerging state standards. However, that will not have occurred during the period of RTT initiatives.

only be useful for the monitoring function envisioned in the regulations to the extent that it is not used for any high-stakes decision.

Use of Student Growth Data to Evaluate Teachers and Principals

We applaud the Department's desire to move educational data collection systems forward by supporting and encouraging the construction of data systems that can link students and their teachers.

We propose that to be eligible under this program, a State must not have any legal, statutory, or regulatory barriers to linking student achievement or student growth data to teachers for the purpose of teacher and principal evaluation. (Section II.A., p. 37806)

Such data systems are essential for conducting research related to the full range of potential approaches for evaluating educators and for developing pilot programs for evaluation approaches that will one day become operational. However, BOTA has significant concerns that the Department's proposal places too much emphasis on measures of growth in student achievement (1) that have not yet been adequately studied for the purposes of evaluating teachers and principals and (2) that face substantial practical barriers to being successfully deployed in an operational personnel system that is fair, reliable, and valid.

Differentiating teacher and principal effectiveness based on performance...The extent to which the State, in collaboration with its participating LEAs, has a high-quality plan and ambitious yet achievable annual targets to (a) Determine an approach to measuring student growth (as defined in this notice); (b) employ rigorous, transparent, and equitable processes for differentiating the effectiveness of teachers and principals using multiple rating categories that take into account data on student growth (as defined in this notice) as a significant factor; (c) provide to each teacher and principal his or her own data and rating; and (d) use this information when making decisions. (Section III.C.(C)(2), p. 37809)

The most prominent way to measure student growth involves "value added" approaches that use results from earlier tests, as well as other information, to adjust new test scores for pre-existing differences across students. The objective of these statistical techniques is to produce a measure of the "value added" to a student's achievement by a teacher or a school in a given year.

The term "value-added model" (VAM) has been applied to a range of approaches, varying in their data requirements and statistical complexity. Although the idea has intuitive appeal, a great deal is unknown about the potential and the limitations of alternative statistical models for evaluating teachers' value-added contributions to student learning. BOTA agrees with other experts who have urged the need for caution and for further research prior to any large-scale, high-stakes reliance on these approaches (e.g., Braun, 2005; McCaffrey and Lockwood, 2008; McCaffrey et al., 2003).

BOTA and the National Academy of Education conducted a joint workshop (November 13-14, 2008) to obtain expert judgments and assessments of issues related to the use of value-added methodologies in education. The workshop topics included both the potential utility of VAM approaches and the limitations of VAM techniques, particularly in making high-stakes

Comments on Race to the Top Proposal

individual, institutional, or district accountability decisions. (A report of the workshop is forthcoming; the workshop papers are available at www7.nationalacademies.org/bota.) The considerable majority of experts at the workshop cautioned that although VAM approaches seem promising, particularly as an additional way to evaluate teachers, there is little scientific consensus about the many technical issues that have been raised about the techniques and their use.

Prominent testing expert Robert Linn concluded in his workshop paper: “As with any effort to isolate causal effects from observational data when random assignment is not feasible, there are reasons to question the ability of value-added methods to achieve the goal of determining the value added by a particular teacher, school, or educational program” (Linn, 2008, p. 3). Teachers are not assigned randomly to schools, and students are not assigned randomly to teachers. Without a way to account for important unobservable differences across students, VAM techniques fail to control fully for those differences and are therefore unable to provide objective comparisons between teachers who work with different populations. As a result, value-added scores that are attributed to a teacher or principal may be affected by other factors, such as student motivation and parental support.

VAM also raises important technical issues about test scores that are not raised by other uses of those scores. In particular, the statistical procedures assume that a one-unit difference in a test score means the same amount of learning—and the same amount of teaching—for low-performing, average, and high-performing students. If this is not the case, then the value-added scores for teachers who work with different types of students will not be comparable. One common version of this problem occurs for students whose achievement levels are too high or too low to be measured by the available tests. For such students, the tests show “ceiling” or “floor” effects and cannot be used to provide a valid measure of growth. It is not possible to calculate valid value-added measures for teachers with students who have achievement levels that are too high or too low to be measured by the available tests.

In addition to these unresolved issues, there are a number of important practical difficulties in using value-added measures in an operational, high-stakes program to evaluate teachers and principals in a way that is fair, reliable, and valid. Those difficulties include the following:

1. Estimates of value added by a teacher can vary greatly from year to year, with many teachers moving between high and low performance categories in successive years (McCaffrey, Sass, and Lockwood, 2008).
2. Estimates of value added by a teacher may vary depending on the method used to calculate the value added, which may make it difficult to defend the choice of a particular method (e.g., Briggs, Weeks, and Wiley, 2008).
3. VAM cannot be used to evaluate educators for untested grades and subjects.
4. Most data bases used to support value-added analyses still face fundamental challenges related to their ability to correctly link students with teachers by subject.
5. Students often receive instruction from multiple teachers, making it difficult to attribute learning gains to a specific teacher, even if the data bases were to correctly record the contributions of all teachers.
6. There are considerable limitations to the transparency of VAM approaches for educators, parents and policy makers, among others, given the sophisticated statistical methods they employ.

Comments on Race to the Top Proposal

Many of these difficulties could be addressed in time—with further research and development of VAM statistical approaches, expansion of testing programs into more grades and subjects, improvement of data bases, and careful development of personnel evaluation systems that use multiple measures. However, it is unlikely that any state at this time could make a proposal for using VAM approaches in an operational program for teacher or principal evaluation that adequately addresses all of these concerns.

The use of test data for teacher and educator evaluation requires the same types of cautions that are stressed when test data are used to evaluate students: “Tests are one objective and efficient way to measure what people know and can do, and they can help make comparisons across large groups of people. However, test scores are not perfect measures: they should be considered with other sources of information when making important decisions about individuals” (*Lessons Learned*, p. 15). This caution is even more important when applied to complex statistics—like value-added analyses—derived from tests.

In sum, value-added methodologies should be used only after careful consideration of their appropriateness for the data that are available, and if used, should be subjected to rigorous evaluation. At present, the best use of VAM techniques is in closely studied pilot projects. Even in pilot projects, VAM estimates of teacher effectiveness should not be used as the sole or primary basis for making operational decisions because the extent to which the measures reflect the contribution of teachers themselves, rather than other factors, is not understood. Even in pilot projects, VAM estimates of teacher effectiveness should not be used to make operational decisions for teachers with students who have achievement levels that are too high or too low to be measured by the available tests because the estimates for such teachers will be essentially meaningless. Even in pilot projects, VAM estimates of teacher effectiveness that are based on data for a single class of students should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable.

Use of Data to Improve Instruction

We applaud the Department’s support of “instructional improvement systems” that include the use of assessments, as well as student work, to help teachers focus instruction and evaluate the success of their instructional efforts. It is appropriate that the Department has included the use of such systems in its criteria for evaluating state proposals for the use of RTT funds.

The extent to which the State, in collaboration with its participating LEAs, has a high-quality plan to-- (i) Increase the use of instructional improvement systems (as defined in this notice) that provide teachers, principals, and administrators with the information they need to inform and improve their instructional practices, decision-making, and overall effectiveness. (Section III.B.(B)(3), p. 37809)

The choice of appropriate assessments for use in instructional improvement systems is critical. Because of the extensive focus on large-scale, high-stakes, summative tests, policy makers and educators sometimes mistakenly believe that such tests are appropriate to use to provide rapid feedback to guide instruction. This is not the case.

Tests that mimic the structure of large-scale, high-stakes, summative tests, which lightly sample broad domains of content taught over an extended period of time, are unlikely to provide the kind of fine-grained, diagnostic information that teachers need to guide their day-to-day instructional decisions. In addition, an attempt to use such tests to guide instruction encourages a

Comments on Race to the Top Proposal

narrow focus on the skills used in a particular test—“teaching to the test”—that can severely restrict instruction. Some topics and types of performance are more difficult to assess with large-scale, high-stakes, summative tests, including the kind of extended reasoning and problem-solving tasks that show that a student is able to apply concepts from a domain in a meaningful way. The use of high-stakes tests already leads to concerns about narrowing the curriculum towards the knowledge and skills that are easy to assess on such tests; it is critical that the choice of assessments for use in instructional improvement systems not reinforce the same kind of narrowing.

In applying its criteria to evaluate state proposals for RTT funds, BOTA urges the Department to clarify that assessments that simply reproduce the formats of large-scale, high-stakes, summative tests are not sufficient for instructional improvement systems. The multiple-choice format in particular lends itself more easily to measuring declarative knowledge than complex “higher-order” cognitive skills. Instructional improvement systems that rely heavily on such item formats may reinforce a tendency to narrow instruction to reflect little more than tested content and formats.

In addition, BOTA urges a great deal of caution about the nature of assessments that could meet the Department’s definition for inclusion in a “rapid-time” turnaround system.

Rapid-time, in reference to reporting and availability of school- and LEA-level data, means that data is available quickly enough to inform current lessons, instruction, and related supports; in most cases, this will be within 72 hours of an assessment or data gathering in classrooms, schools, and LEAs. (Section IV, p. 37811)

If the Department is referring to informal, classroom assessment methods that can be scored and interpreted by the teacher, a 72-hour turnaround is a reasonable goal. It is important to provide teachers with a better understanding about the types of assessment they can use informally in their classrooms.

However, if the Department is referring to more formal assessment methods, then a 72-hour turnaround is very difficult to attain. Even for multiple-choice items, it would be hard to provide a 72-hour turnaround. Assessment of complex reasoning and problem-solving skills typically demands assessment formats that require students to generate their own extended responses rather than selecting a word or phrase from a short list of options. Automated scoring of such “constructed responses” is an active field of psychometric research: to date, however, except for automated scoring of written essays, the “state of the art” extends only to small demonstration projects not large-scale applications. Likewise, the logistics of human scoring for large-scale assessments would make a 72-hour turnaround extremely costly. It is premature to specify such a required period of time within which assessment results must be returned to teachers. When seeking to provide quick turnaround of information, it is essential that the quality of that information not be sacrificed in the process. The first priority must be that the right information is produced and that the information meets professional standards for technical adequacy so that the information can accurately guide decision making.

Challenges in Developing Common Assessments

We applaud the Department's plan to invest in state consortia that will work toward developing high quality assessments. It is appropriate that the Department has included the commitment toward improving the quality of state assessment in its criteria for evaluating state proposals for the use of RTT funds.

Whether the State has demonstrated a commitment to improving the quality of its assessments by participating in a consortium of States that is working toward jointly developing and implementing common, high-quality assessments (as defined in this notice) aligned with the consortium's common set of K-12 standards (as defined in this notice) that are internationally benchmarked and that build toward college and career readiness by the time of high school graduation, and the extent to which this consortium includes a significant number of States. (Section III.A.(A)(2), p. 37808)

The opportunity of working toward common assessments of common standards raises the possibility of significant economies of scale in the development of high-quality assessments.

Although BOTA supports the value of a joint development effort toward common assessments, we want to stress that the requirements are quite high for producing common assessments that would truly allow comparisons in student achievement across states in the same way that NAEP currently does. BOTA's previous work on potential approaches to developing common standards and assessments (*Uncommon Measures*, 1999; *Embedding Questions*, 1999) concluded that this aspiration is very difficult to accomplish in practice. The fundamental problem relates to dissimilarities across states in their standards, instruction and curriculum, and uses of test scores, as well as the assessments themselves: these dissimilarities ultimately make assessment results incomparable.

If states indeed adopt fully common standards and develop common assessments, these concerns would be reduced, but even seemingly small deviations from fully common standards and assessments will introduce important limitations in the degree of comparability of outcomes. For instance, to be fully comparable, the assessments would need to be administered under standardized conditions across the country. This means that time limits, the length of testing sessions, the instructions given to test takers, test security provisions, and the use of manipulatives and calculators would need to be the same everywhere. The test administration would need to be scheduled at an appropriate time during the school year such that all the students who will take the test have been exposed to and have an opportunity to learn the content covered on the test. The stakes attached to the assessment results would also need to be constant across the country to ensure that students are equally motivated to perform well. Including state-specific content on the assessments—that is, content that differs across states—introduces the issue of context effects. These state-specific items would need to be placed on the assessment in a location where they would not influence how students perform on the common items. See *Embedding Questions* (1999, pp. 20-36) for a fuller discussion.

In addition to the aspiration of creating common assessments, the Department's proposal also notes the objective of creating assessments that are "internationally benchmarked." There are several different ways this phrase might be interpreted. However, for assessment results that could be directly compared to their international counterparts, we note that the difficulties that arise in comparing test results from different states apply even more strongly for comparing test

Comments on Race to the Top Proposal

results from different countries. For making comparisons internationally, the problems with differing standards, assessments, instruction, curricula, and testing regimes are magnified. In addition, international test comparisons raise difficult problems related to language translation. Because of these challenges, the Department should think carefully about the kind of “international benchmarking” that it wants to encourage states to pursue.

The aspiration for higher quality in standards and assessments is one that BOTA members share, and over the coming months, BOTA will be conducting a series of workshops on improving state assessment systems, with particular attention to the opportunities offered by the current interest in moving toward common assessments and by the ARRA funding set aside for this purpose.

Challenges in Creating Longitudinal Data Systems

We applaud the Department’s desire to support statewide longitudinal data systems. It is appropriate that the Department has included the commitment toward development such a system in its plan for evaluating state proposals for the use of RTT funds.

The extent to which the State has a high-quality plan to ensure that data from the State's statewide longitudinal data system are accessible to, and used to inform and engage, as appropriate, key stakeholders (e.g., parents, students, teachers, principals, LEA leaders, community members, unions, researchers, and policymakers); that the data support decision-makers in the continuous improvement of instruction, operations, management, and resource allocation. (Section III,B,(B)(2), p. 37809)

The development and application of such longitudinal data systems are considerably more difficult than many people realize. State proposals for RTT funds should indicate a familiarity with the full complexity of the challenges and describe appropriate plans to begin to address them. The workshop on VAM approaches discussed above highlighted the challenges associated with creating data systems that produce understandable and useful information for all stakeholders, including the challenge of accurately connecting students to their teachers. An ongoing set of evaluation activities, as described above, should inform states’ efforts in this regard. Another upcoming joint report from BOTA and the National Academy of Education on calculating dropout and graduation rates will also offer advice on building longitudinal data systems and making use of the data to inform decision making. Issues such as dealing with transfer students and students who drop out, some of whom re-enroll and drop out repeatedly, pose challenges in building data systems that accurately code and track students’ status. We hope to address the qualities for best practices in state assessment systems in the workshop series we will be conducting over the coming year.

CONCLUSION

In closing, we return to the beginning of this letter, with the importance of rigorously evaluating the innovations supported by RTT funds. Careful evaluation of this spending should not be seen as optional; it is likely to be the only way that this substantial investment in educational innovation can have a lasting impact on the U.S. education system. BOTA urges the Department to carefully craft a set of requirements for rigorous evaluation of all initiatives support by RTT funds.

Comments on Race to the Top Proposal

Attachment A	Membership of the Board on Testing and Assessment
Attachment B	Reviewer Acknowledgments
Attachment C	Selected Reports of the Board on Testing and Assessment
Attachment D	Key Professional and Technical Standards Documents Relevant to the Use of Educational Assessments
Attachment E	References

Comments on Race to the Top Proposal

Attachment A: Membership of the Board on Testing and Assessment

Edward H. Haertel (*Chair*), School of Education, Stanford University
Lyle Bachman, Department of Applied Linguistics, University of California, Los Angeles
Stephen Dunbar, College of Education, University of Iowa
David J. Francis, Department of Psychology, University of Houston
Arthur Goldberger⁶, Department of Economics, Emeritus, University of Wisconsin–Madison
Michael Hout, Graduate Group in Sociology and Demography, University of California,
Berkeley
Michael Kane, Research and Development Division, Educational Testing Service
Kevin Lang, Department of Economics, Boston University
Michael Nettles, Policy Evaluation and Research Center, Educational Testing Service
Diana Pullin, Lynch School of Education, Boston College
Brian Stecher, Education Program, RAND
Mark R. Wilson, Graduate School of Education, University of California, Berkeley
Rebecca Zwick, Gevirtz Graduate School of Education, University of California, Santa Barbara

Stuart W. Elliott, Board Director
Judith Anderson Koenig, Senior Program Officer

⁶ Was unable to participate in BOTA's Committee to Respond to the Department of Education RTT Proposal.

Attachment B: Reviewer Acknowledgments

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 National Research Council's 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 study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report: Jonathan G. Dings, Office of Planning and Assessment, Boulder Valley School District; Mark Dynarski, Center for Improving Research Evidence, Mathematica Policy Research, Inc.; Margaret E. Goertz, Graduate School of Education, University of Pennsylvania; Jane Hannaway, Education Policy Center, The Urban Institute; Scott F. Marion, Office of the Associate Director, National Center for the Improvement of Educational Assessment; Lorraine McDonnell, Department of Political Science, University of California, Santa Barbara; and Laress (Laurie) L. Wise, Human Resources Research Organization (HumRRO), Monterey, CA.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by and Robert L. Linn, Professor Emeritus, Department of Education, University of Colorado, and Stephen E. Fienberg, Department of Statistics, Carnegie Mellon University. Appointed by the National Research Council, 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 rests entirely with the authoring committee and the institution.

Attachment C: Selected Reports of the Board on Testing and Assessment

NOTE: All of the reports listed here were or will be published by the National Academies Press.

- Assessing Accomplished Teaching: Advanced-Level Certification Programs* (2008)
Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment (2003)
Educating One and All: Students with Disabilities and Standards-Based Reform (1997)
Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests (1999)
Evaluation of the Voluntary National Tests: Year 1 and Year 2 (1999)
Getting Value Out of Value-Added: Report of a Workshop (forthcoming, 2009)
Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress (1999)
High Stakes: Testing for Tracking, Promotion, and Graduation (1999)
Incentives and Test-Based Accountability in Education (forthcoming, 2010)
Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessment (2004)
Knowing What Students Know: The Science and Design of Educational Assessment (2001)
Lessons Learned About Testing: Ten Years of Work at the National Research Council (no date)
Measuring High School Graduation and Dropout Rates: Next Steps for Research and Policy (forthcoming, 2010)
Measuring Literacy: Performance Levels for Adults (2005)
Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions (1999)
NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting (2001)
Systems for State Science Assessment (2005)
Testing, Teaching, and Learning: A Guide for States and School Districts (1999)
Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality (2001)
Uncommon Measures: Equivalence and Linkage Among Educational Tests (1999)
Understanding Dropouts: Statistics, Strategies, and High-Stakes Testing (2001)

Attachment D: Key Professional and Technical Standards Documents Relevant to the Use of Educational Assessments

- The Personnel Evaluation Standards: How to Assess Systems for Evaluating Educators* (1988). Joint Committee on Standards for American Educational Research Educational Evaluation. Thousand Oaks, CA: Corwin Press.
- Position Statement of the American Educational Research Association Concerning High-Stakes Testing in PreK-12* (2000). American Educational Research Association, Washington, DC
- The Program Evaluation Standards: How to Assess Systems for Evaluating Educators. 2nd Edition* (1994). Joint Committee on Standards for Educational Evaluation. Thousand Oaks, CA: Corwin Press.
- Standards for Educational Accountability Systems* (2002). Baker, E., Linn, R.L., Herman, J. and Koretz, D. CRESST Policy Brief 5. Available through the Educational Resources Information Center (ERIC).
- Standards for Educational and Psychological Testing* (1999). American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA/APA/NCME), Washington, DC.
- The Student Evaluation Standards: How to Improve Evaluations of Students* (2003). Joint Committee on Standards for Educational Evaluation. Thousand Oaks, CA: Corwin Press.

Attachment E: References

- Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.
- Briggs, D.C., Weeks, J.P., and Wiley, E. (2008). *The sensitivity of value-added modeling to the creation of a vertical score scale*. Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison, April 22-24.
- Chester, M. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice*, 24(4), 40-52.
- Harvey, C., Camasso, M.J., and Jagannathan, R. (2000). Evaluating welfare reform waivers under section 1115. *Journal of Economic Perspectives*, 14:4, 165-188.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.
- Linn, R.L. (2008). *Measurement issues associated with value-added models*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. Available: http://www7.nationalacademies.org/bota/1VAM_Workshop_Agenda.html [September 2009].
- McCaffrey, D., and Lockwood, J.R. (2008). *Value-added models: Analytic issues*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. Available: http://www7.nationalacademies.org/bota/1VAM_Workshop_Agenda.html [September 2009].
- McCaffrey, D., Lockwood, J.R., Koretz, D.M., and Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.
- McCaffrey, D., Sass, T.R., and Lockwood, J.R. (2008). *The intertemporal stability of teacher effect estimates*. Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison, April 22-24.