EPI BRIEFING PAPER

ECONOMIC POLICY INSTITUTE • APRIL 20, 2010 • BRIEFING PAPER #263

LET'S DO THE NUMBERS

Department of Education's "Race to the Top" Program Offers Only a Muddled Path to the Finish Line

BY WILLIAM PETERSON AND RICHARD ROTHSTEIN

Introduction

The American Recovery and Reinvestment Act of 2009 (the "stimulus" bill) provided $4.35 billion to the Department of Education for "Race to the Top" (RTT), a program in which states could apply for funds to implement education reform. Secretary of Education Arne Duncan established a competition to determine which states would receive the funds, and 40 states (plus the District of Columbia) entered. Of these, 16 were named as finalists, and in late March 2010, two states, Delaware and Tennessee, were announced as winners of the first round. The awards were substantial: Delaware got $100 million (or about $800 per pupil), and Tennessee got $500 million (or about $500 per pupil). In each case, the award represents about 7% of the total expenditures in these states for elementary and secondary education.

To compete for RTT funds, governors were faced with quickly organizing existing resources to underwrite extensive grant-writing efforts. Some invested significant political energy and leadership to persuade school districts and teacher unions to endorse the applications, while others had to press legislatures to change state laws on charter schools and teacher evaluation. When the winners were announced, some governors expressed concern and disappointment at what one called an "inscrutable process," leaving them to wonder whether it would be worth participating in future rounds of the competition.[1]

Delaware and Tennessee won because they got the most points (454.6 and 444.2, respectively) out of a total of 500 points available. Five outside panelists[2] reviewed each state's application, including interviews with delegations from the finalist states, and awarded points for states' compliance with policies promoted by Secretary Duncan, such as participating in a national consortium to develop common standards in reading and math (maximum of 20 points) or using data to improve instruction (maximum of 18 points).

Because the awards were based on precise numerical scores, the process was presented as objective and scientific.

TABLE OF CONTENTS

www.epi.org

However, further examination suggests that the selection of Delaware and Tennessee was subjective and arbitrary, more a matter of bias or chance than a result of these states' superior compliance with reform policies.[3]

At a time of widespread fiscal crises in the states, when receipt of Race to the Top awards can determine whether class sizes will be increased and teachers laid-off, such capricious decision-making is unfortunate. The Department of Education can use its distribution of funding as a "carrot" to stimulate states to improve their education policies, but when state budgets are as stressed as they are today, every state should get a fair share of federal funding, excepting only those that refuse to make good faith efforts to implement research-based improvements in elementary and secondary education.

The Obama administration intends RTT to be the model for a new approach to the distribution of federal elementary and secondary education aid. Whatever its merit in flush times, the substitution of competition for uniform funding has no place in this time of state fiscal crisis. The actual experience of RTT—in which the selection of particular states to receive competitive grants can't reasonably be justified—is further reason to abandon this approach for the future.

Dangers of metrics

Quantitative metrics are a popular management tool. Such metrics can be used to describe objective performance, such as total school lunches served per day, or subjective factors, such as an evaluator's judgment of a teacher's skill in teaching math. When managers use metrics to evaluate overall performance, they must assign weights, or relative importance, to the various metrics. For example, a school's overall rating could be determined by a combination of a rating for lunches served (weighted as 25% in importance) and a rating for the math teacher's skill (weighted as 75% in importance).

Subjective judgment is required both for assigning weights to metrics, and for making judgments regarding performance on most individual metrics. In the latter case, dangers of subjectivity can be reduced by providing evaluators with detailed checklists (sometimes called "rubrics") describing the components of performance (e.g., in the case of the math teacher, assigning so many points for demonstrating understanding of the lesson, assigning a certain number of points for calling on children from different parts of the room, etc.), and by training evaluators by asking them to observe identical lessons and comparing the ratings to ensure "inter-rater reliability."

If such precautions are not taken, or are insufficiently taken, then evaluations based on metrics can appear objective even though they in fact reflect only bias or chance. The RTT 500-point system suffers from several such deficiencies.

RTT weights

One source of false precision in the use of metrics for evaluation stems from the arbitrary assignment of weights to various indicators in a system. Some index systems make weights more credible by basing them on a survey (of opinion leaders, public officials, or the general public), asking respondents for their judgments regarding the relative importance of a list of factors, and then averaging the weights that respondents chose.

In the case of RTT, Secretary Duncan and his staff chose provisional weights and then revised them after reviewing suggestions submitted by members of the public as part of a formal regulatory comment period. Several of the revisions made in this fashion made sense, but other well-founded suggestions were ignored.[4] These arbitrary weights have enormous consequence.

The RTT 500-point system

The RTT 500-point system, shown in **Table 1**, has six major categories, seven general categories, and various subcategories. The primary weighted metrics consist of the 30 categories whose points are shown in italics. The first column in Table 1 is a list of the various categories selected by the Department of Education. We raise many questions below concerning the particular categories chosen, but this listing and its subjective evaluation are a reasonable first step by the Department to describe how it believes states should proceed to improve their educational programs. However, by assigning numbers to this process, the Department implies it has a testable theory or empirical data to back up its quantitative method. By making RTT a competitive system, the Department then locks itself into accepting the numerical scores as

TABLE 1

Metric weighting for Race to the Top competition

	Possible points		Weight
A. State Success Factors	**125**		**25%**
(A)(1) Articulating State's education reform agenda and LEA's participation in it	**65**		13
(i) Articulating comprehensive, coherent reform agenda		*5*	1
(ii) Securing LEA commitment		*45*	9
(iii) Translating LEA participation into statewide impact		*15*	3
(A)(2) Building strong statewide capacity to implement, scale up, and sustain proposed plans	**30**		6
(i) Ensuring the capacity to implement		*20*	4
(ii) Using broad stakeholder support		*10*	2
(A)(3) Demonstrating significant progress in raising achievement and closing gaps	**30**		6
(i) Making progress in each reform area		*5*	1
(ii) Improving student outcomes		*25*	5
B. Standards and Assessments	**70**		**14**
(B)(1) Developing and adopting common standards	**40**		8
(i) Participating in consortium developing high-quality standards		*20*	4
(ii) Adopting standards		*20*	4
(B)(2) Developing and implementing common, high-quality assessments	***10***		2
(B)(3) Supporting the transition to enhanced standards and high-quality assessments	***20***		4
C. Data Systems to Support Instruction	**47**		**9**
(C)(1) Fully implementing a statewide longitudinal data system	***24***		5
(C)(2) Accessing and using state data	***5***		1
(C)(3) Using data to improve instruction	***18***		4
D. Great Teachers and Leaders	**138**		**28**
(D)(1) Providing high-quality pathways for aspiring teachers and principals	***21***		4
(D)(2) Improving teacher and principal effectiveness based on performance	**58**		12
(i) Measuring student growth		*5*	1
(ii) Developing evaluation systems		*15*	3
(iii) Conducting annual evaluations		*10*	2
(iv) Using evaluations to inform key decisions		*28*	6
(D)(3) Ensuring equitable distribution of effective teachers and principals	**25**		5
(i) Ensuring equitable distribution in high-poverty or high-minority schools		*15*	3
(ii) Ensuring equitable distribution in hard-to-staff subjects and specialty areas		*10*	2
(D)(4) Improving the effectiveness of teacher and principal preparation programs	***14***		3
(D)(5) Providing effective support to teachers and principals	***20***		4
E. Turning Around the Lowest-Achieving Schools	**50**		**10**
(E)(1) Intervening in the lowest-achieving schools and LEAs	***10***		2
(E)(2) Turning around the lowest-achieving schools	**40**		8
(i) Identifying the persistently lowest-achieving schools		*5*	1
(ii) Turning around the persistently lowest-achieving schools		*35*	7
F. General	**55**		**11**
(F)(1) Making education funding a priority	*10*		2
(F)(2) Ensuring successful conditions for high-performing charter schools and other innovative schools	*40*		8
(F)(3) Demonstrating other significant reform conditions	*5*		1
Competitive Preference Priority 2: Emphasis on STEM (Science, Technology, Engineering, Mathematics)	***15***		***3***
Total	**500**		**100%**

SOURCE: U.S. Dept. of Ed., 34 CFR Subtitle B, Chap II, RTT Fund, Final Rule, Federal Register 74 (221) Nov. 18, 2009. http://edocket.access.gpo.gov/2009/pdf/E9-27426.pdf

the specific criteria for selecting winners. The necessary subjective judgments required both for category selection and weight assignment makes a fair competition practically impossible, even if the competition is undertaken with great care.

Questions on categories and initial weights

A review of Table 1 raises several general questions. What was involved in the decision to use a scale of 500 total points rather than, say, 10, or 100, or 1,000? Is there scientific support for the "State Success Factors" being 90.6% as important as the "Great Teachers and Leaders" factor? Should the "Great Teachers" maximum points be 140, or maybe 163, instead of 138? And are there missing factors, such as "Developing Techniques to Promote Creativity," or others?

In addition to these general questions, there are many specific ones. For example, most people would consider the following factors, included in Table 1, reasonably important:

- (A)(3)(ii), "Improving student outcomes," with a weight of 5% (i.e., 25 out of 500 points);
- (C)(3), "Using data to improve instruction," with a weight of 4%;
- (D)(2)(iv), "Using evaluations [of principals and teachers] to inform key decisions," with a weight of 6%; and
- (D)(3)(i), "Ensuring equitable distribution [of principals and teachers] in high-poverty or high-minority schools," with a weight of 3%.

Can we really be certain that these weights are appropriate? Even accepting the overall framework, it wouldn't be unreasonable to consider increasing each by a mere 3%, revising weights to:

- 8% (from 5%) for "Improving student outcomes;"
- 7% (from 4%) for "Using data to improve instruction;"
- 9% (from 6%) for "Using evaluations to inform key decisions;" and
- 6% (from 3%) for "Ensuring equitable distribution."

If the weights were increased in this manner (and the weights of the other 25 indicators reduced by roughly half a point,[5] so the total would remain 100%), Tennessee would no longer have won the competition—Georgia would have won instead. The official result can't be justified by a claim that Tennessee is more reform-minded than Georgia.

The selection of indicators themselves also seems arbitrary. Not everyone, of course, will agree with all of the indicators Secretary Duncan chose to include, but one can accept his policy preferences and still wonder about the selection. In March, Secretary Duncan presented to Congress his recommendations for re-authorization of the Elementary and Secondary Education Act (ESEA). These recommendations (entitled the "Blueprint") include proposals for additional competitions by which states can win added funds. Yet several innovative practices that states must follow to win proposed ESEA competitions were not given points in the RTT competition. For example, RTT awards 10 points for "developing and implementing common, high-quality assessments," referring to assessments that are aligned with the common standards in reading and math being developed by the National Governors Association (NGA), Council of Chief State School Officers (CCSSO), and a number of states. The Blueprint, however, also proposes competitive grants to develop "high-quality assessments in…science, history, or foreign languages; [and] high school course assessments in academic and career and technical subjects." But the RTT rubric awards no points for development of such assessments.

As another example, the Blueprint proposes a grants competition for states to develop "innovative programs [to] build the knowledge base about promising practices, and scale up effective practices to improve instruction for English Learners."

A state that proposed, in its RTT application, to develop high-quality assessments in subject areas other than reading and math, or that proposed to develop more effective practices for non-English speaking students, would have received essentially no points for such initiatives. Were such indicators included in the RTT list, perhaps other states could have outscored Delaware and Tennessee.

To repeat, these examples (and there are several others) were selected from the Department's own priorities, as described in its Blueprint. It cannot be argued that there is a rational basis for awarding RTT prize money for pursuing some of the Department's reform priorities, and not others, or that states that pursue policies in the RTT indicator list are more reform-oriented than states that pursue reform policies that, for some apparently arbitrary reason, were not included in this list. If state policy makers want to follow Secretary Duncan's agenda, should they pursue policies needed to win RTT grants (and with effort proportional to RTT weights), or policies needed to win prospective competitive ESEA grants? The two sets of policy priorities overlap, but are not identical.

With the RTT rubric, perhaps the Department of Education intended to encourage reform policies not specifically listed by awarding points for metric (F)(3) "demonstrating other significant reform conditions." Yet the system allows only five out of 500 points for such initiatives. This weight of a mere 1% for the judgment of governors and state officials about how to improve education is inconsistent with the administration's pledge to inject flexibility into federal education policy—when he released his proposals for re-authorization of ESEA, Secretary Duncan said, "We're offering support, incentives, and national leadership, but not at the expense of local control."[6]

Yet consider the case of Pennsylvania, which received the full five points for the "other significant reform" metric. Reviewers praised the state for "align[ing] early childhood education standards, curriculum, instruction and assessment to research on how young children learn, allowing more students to get a head start on learning before entering the elementary grades," and for "invest[ing] in programs to expose elementary school students to hands-on science." Both of these initiatives are supported by extensive research and are consistent with federal policy. Indeed, in the recent Congressional health care reconciliation bill, the Obama administration attempted to re-direct funds to early childhood challenge grants that would have supported just the kinds of reforms for which Pennsylvania was praised, and the federal government is in the process of redesigning the National Assessment of Educational Progress so that its science assessment includes more hands-on items.[7]

We noted above that by adding a few points to some perhaps more important indicators, Georgia, not Tennessee, would have won the competition. What if, as well, the judgment of governors and other state policy makers were given greater respect, by giving a weight of 25%, not a mere 1%, to "demonstrating other significant reform conditions," such as those in Pennsylvania's plan? Because the criterion would still have required that the reform be "significant," reviewers would not have been required to award points for state policies that were not well-designed and research based. But if, with such a precaution, the weights had been modified in this way, Georgia would no longer have beaten out Tennessee. Neither Georgia nor Tennessee would have won—Pennsylvania would have been declared the winner.

Pennsylvania, in short, has now been told by the Department of Education that if it wants to compete successfully in the next round of RTT competition, then the state should downplay its focus on early childhood and science education, and put its efforts instead into categories that get more points but which, in fact, have a much weaker research base.

RTT panel judgments

Pseudo-scientific use of metrics can also imply false precision when evaluators are expected to make judgments that are too cognitively complex. This can happen if a scale has too many divisions.

Consider, for example, commonplace controversies about the accuracy of academic grading systems. Whenever we ask faculty to make subjective evaluations of students, we must always allow for bias and chance. When we assign numbers to such subjective evaluations, we are really just dividing the results into arbitrary groupings, in which the closer a judgment gets to a dividing line, the less accurate it will be. Intuitively we know that the fewer the divisions, the less chance for error. If we only have one division, we will always be 100% correct, but as we increase the number of divisions, the probability of error increases. For this reason, professional educators have long debated the relative advantages of pass/fail or

letter grading systems. (The letter systems are actually number systems, where each letter stands for a number that can be used to calculate a grade-point average.) In a pass/fail system, grades have a higher probability of error only if they are close to the passing line. In an A-B-C-D-F system, grades have a higher probability of error if they are close to any one of four cut points. For institutions where pluses and minuses are used, there may be as many as 13 cut points.[8] The more cut points there are, the more students there are whose grades will be inaccurate. Some professors, understanding that grades close to a cut point are really indistinguishable, attempt to account for this by giving students "the benefit of the doubt" and assigning a higher grade to students whom they believe deserve the next lower one. This, of course, introduces an upward bias to the entire grading system.

Similarly, many of us have taken surveys in which we are asked if, with regard to some product or process, we were very satisfied, satisfied, dissatisfied, or very dissatisfied. Although the dividing line between each of these categories is arbitrary, most people can make such a rational (i.e., justifiable) distinction. Asking us, however, to distinguish between very dissatisfied, and very, very dissatisfied, or very, very, very dissatisfied, would result in less reliable, more arbitrary responses. In short, where subjective ratings are involved, the number of rating categories should be small because of the natural limit of a typical human judge's capability to make rational distinctions between divisions.[9]

RTT judges (called "reviewers"), however, were asked to rate performance on scales that were impossibly large. Table 1 shows that of the 30 primary weighted metrics, 12 have scales of 20 points or more, requiring reviewers to make highly questionable judgments. Another 12 have scales of 10 to 19 points, also calling for highly challenging cognitive decisions. The Department should have made allowances in its RTT 500-point system for significant errors in judging the categories, but publishing such margins of error would have made it plain that the winning states won only by chance. When these judging errors are combined with expected errors in the design of the metrics, it is surprising that Department asserted that the final state scores were correct to one decimal place.

Consider two closely related indicators, (A)(1)(ii), "securing LEA [school district][10] commitment [to the state's education reform agenda]," and (A)(1)(iii), "translating LEA participation into statewide impact." Averaging the judges' awards on the first of these, Tennessee got 44 and Florida got 36 out of a possible 45 points. Averaging the judges' awards on the second, Tennessee got 14 and Florida got nine out of a possible 15 points.

However, 45 points is too large a scale to permit reviewers reasonably to make such fine distinctions. Can a reviewer—especially a non-professional reviewer with minimal training, conducting a one-time exercise—imagine 45 distinct degrees of effort to secure school districts' commitment? Can a reviewer imagine 15 distinct degrees of effort in translating school district participation into "statewide impact," whatever that means?

In Tennessee, every superintendent and school board president endorsed the state's application, but according to the notes of one reviewer, "the State expects some attrition of districts."[11] Is "some attrition" equivalent to one point on the 45-point scale, resulting in Tennessee's score on this metric of 44 rather than 45? Why not a three-point, or a five-point penalty for "some attrition"? In Florida, 89% of all school districts in the state "signed on with full endorsement to the RTT application." Why didn't Florida get 40 points (89% of 45), rather than 36 for this indicator?

If, in fact, Tennessee had received only 40 points for this indicator, and Florida had also received 40, Florida would have won the overall competition, not Tennessee.

The recent Winter Olympics skating competition provides an interesting point of comparison. For this competition, judges awarded points on a scale for the quality of skaters' performances. The Olympics skating rules, however, attempt to account for the arbitrariness of judgment by excluding "outliers"—the highest and lowest of judges' ratings—from the final average scores. An alternative way of reducing the influence of outliers would be to use the judges' median score, not their average.

The RTT process took no such precautions and thus, winning could be the result only of individual reviewers' occasional quirkiness. In the case of Florida's score for LEA commitment, the five reviewers respectively awarded

points of 38, 40, 35, 40, and 25. If the outlying score of the reviewer who awarded only 25 points had been discarded, Florida would have received 38 points, not 36 for this indicator.[12] Eliminating a few other outliers in other categories could easily have tipped the balance in Florida's favor. For example, on indicator (D)(2)(iv), "using evaluations to inform key decisions," Florida received 24 out of a possible 28 points. But the judges' awards on this indicator were 15, 28, 25, 24, and 28, respectively. The second and fifth reviewers awarded nearly twice as many points as the first. If the first reviewer's points had been excluded, Florida would have received not 24, but 26 points. There are several other such examples in the judging of Florida's RTT application. The judges in Tennessee, in contrast, were considerably more consistent. As a result, if outlying scores had been discarded, Florida, not Tennessee could have won the competition.[13]

When such judging errors are combined with expected errors in the design of the metrics themselves (described in the previous section), it is surprising that the Department of Education claimed that the process had sufficient precision to justify basing hundreds of millions of dollars in awards to some states and not to others.

Devil in the details

As an additional exercise, we examined the case of Massachusetts, which scored surprisingly low (13th of the 16 finalists) for a state with a reputation of having unusually high academic standards and achievement. The first thing we noted was that on nine of the 30 metrics, Massachusetts got higher scores than the winning state of Tennessee (and on another six it had an identical score). So we modeled what would have happened if, on these nine metrics, we increased the weights (as in the examples above) by 3 percentage points.[14] We found that Massachusetts' score was still behind the winners, so we examined further.

Massachusetts' problem, it turned out, centered on metric (B)(1)(ii) "adopting standards." The RTT guidelines required states to participate in the effort to develop common standards in reading and math. For this participation, Massachusetts, like Tennessee, was awarded the full 20 points allowed. But the guidelines also required states then to adopt these standards by next August. Massachusetts, as we noted, already has very high academic standards, so state policy makers might have had reason to wonder whether hasty adoption of these new common standards would improve or harm Massachusetts education. As a result, the state decided to permit a period of public comment between the time these new common standards are completed, and their formal adoption. For permitting this period of public comment, the state was deemed in violation of the competition rules, and the RTT reviewers docked Massachusetts a whopping 15 (out of 20 possible) points on this metric.

In sum, Massachusetts' willingness to permit the public to comment on its academic standards, combined with a few quirks in the weighting system, cost the state hundreds of millions of dollars.

Needless complexity

The Department's 500-point system is needlessly complex. Its implied precision makes the results seem less affected by human judgment than is the case. The Department could have accomplished an almost identical result with a much simpler system, for example, one utilizing only 70 points.

Table 2 shows the maximum points allowable for each of the seven general categories, and the actual scoring, by general category, for the top 10 finalist states. Numbers in the table have been rounded to whole integers.

Table 3 recalculates these data by eliminating the complex weighting scheme, and instead gives each major category the identical weight of 10, for a total maximum score of 70. It then applies the reviewers' actual ratings to these simple weights (i.e., it uses the same relative scores as Table 2) and then rounds the decimals.[15]

There is only a slight difference between the results of Tables 2 and 3. Florida moves up ahead of Georgia to a tie for second place (before rounding, it would have moved up to third), and Ohio moves up ahead of Rhode Island to eighth place. This massive shift of weights, which appear so precise in RTT, makes almost no difference. The inaccuracy and subjective nature of the inputs makes the ordering of states fuzzy. In fact, state policy makers who looked at Table 3 might well conclude that there was very little difference between states' scores, and certainly not

TABLE 2

Race to the Top points awarded to top 10 finalists

	Maximum	Del.	Tenn.	Ga.	Fla.	Ill.	S.C.	Penn.	R.I.	Ky.	Ohio
State Success Factors	**125**	119	112	103	100	93	100	107	99	114	101
Standards and Assessments	**70**	69	68	66	69	69	68	65	66	68	69
Data Systems to Support Instruction	**47**	47	44	41	41	39	41	36	32	43	39
Great Teachers and Leaders	**138**	119	114	111	109	110	114	106	121	111	103
Turning Around the Lowest-Achieving Schools	**50**	43	48	47	44	49	44	45	45	45	43
General	**55**	42	43	50	54	49	41	45	41	22	49
Emphasis on STEM	**15**	15	15	15	15	15	15	15	15	15	15
Totals	**500**	**455**	**444**	**434**	**431**	**424**	**423**	**420**	**419**	**419**	**419**

SOURCE: U.S. Department of Education, detail chart of the Phase 1 scores for each State. Scores by Criterion. http://www2.ed.gov/programs/racetothetop/phase1-applications/phase1-scores-detail.xls.

sufficient difference to justify extremely consequential decisions about federal funding.

It would seem that the hassle of having developed a complex metric represented only misplaced effort and expense. The ordering of states in Table 3 is no more or less plausible than the ordering of states in Table 2, for what we have are just descriptive evaluations pretending to be numbers. The only apparent reason for 500 total points as opposed to, say, 70 total points is to provide sufficient artificial variability in scores to make the differences between nearly identical states seem plausible.

Conclusions and recommendations

In short, the Race to the Top 500-point rating system presents a patina of scientific objectivity, but in truth masks a subjective and somewhat random process.

This competition was a trial run for Secretary Duncan of a policy approach he hopes to make permanent. The Obama administration has proposed that formula-driven Title I funding[16] be frozen at its present level, without future adjustment for inflation, and that increases in federal education spending be devoted entirely to a new

TABLE 3

A simplified weighting scheme for Race to the Top

	Maximum	Del.	Tenn.	Ga.	Fla.	Ill.	S.C.	Penn.	R.I.	Ky.	Ohio
State Success Factors	**10**	10	9	8	8	7	8	9	8	9	8
Standards and Assessments	**10**	10	10	9	10	10	10	9	9	10	10
Data Systems to Support Instruction	**10**	10	9	9	9	8	9	8	7	9	8
Great Teachers and Leaders	**10**	9	8	8	8	8	8	8	9	8	7
Turning Around the Lowest-Achieving Schools	**10**	9	10	9	9	10	9	9	9	9	9
General	**10**	8	8	9	10	9	7	8	7	4	9
Emphasis on STEM	**10**	10	10	10	10	10	10	10	10	10	10
Totals	**70**	**66**	**64**	**62**	**64**	**62**	**61**	**61**	**59**	**59**	**61**

SOURCE: Authors' calculations from data in Table 2.

collection of competitive grants, some of which have similar requirements to RTT, and some of which, as indicated above, attempt to create incentives for initiatives not included in RTT. Because such a reduction in real Title I funding would further exacerbate state fiscal crises, and because this trial run of a competitive system has proven to have little credibility, the administration should rethink its approach to federal education aid and its relationship to school improvement.

Yet for now, the Department of Education proposes to go through an identical process for judging a second round of applications by July. States that lost in the March competition have been invited to re-apply, and several are doing so, again investing time and expense to re-do their applications. Experts in these states are likely to spend many hours studying the review process employed in March, so they can recommend small changes in their states' applications to exploit the quirks of the Department's rating system. Such gaming is unlikely to reflect an actual improvement in the education policies of applicant states.

We recommend instead that the Department abandon this complexity, and move to a simpler "pass/fail" system for the next round of the competition. Even a pass/fail system will have errors, as states that are close to whatever standard the Department employs will either arbitrarily receive awards or be denied. So the benefit of the doubt should be given to borderline states: any states that undertake reasonable efforts to improve their elementary and secondary education systems should receive awards. Only those patently contemptuous of the reform process should be denied. Such a system would sacrifice little in national efforts to enhance the performance of American schools, and would spare states the pain of engaging in unreasonable competition where bias and chance play more of a role than educational improvement.

*—**William Peterson** (bpeterson1931@yahoo.com) is a retired marine engineer with over 35 years experience in the management and maintenance of large commercial tankers and Navy ships, a lifelong interest in education, and in the use and misuse of numbers—especially by managers.*

*—**Richard Rothstein** (riroth@epi.org) is a research associate of the Economic Policy Institute.*

Endnotes

1. Sam Dillon 2010. "States Skeptical About 'Race to Top' School Aid Contest." *The New York Times*, April 5. http://www.nytimes.com/2010/04/05/education/05top.html. Governor Ritter of Colorado, who made the "inscrutable process" comment, subsequently announced that his state would, after all, participate in the next round. Todd Engdahl. 2010. "Ritter: Colorado in for Second 'Race to Top'." *Education News Colorado*, April 7. http://www.statebillnews.com/2010/04/ritter-colorado-in-for-second-race-to-top/
2. A total of 49 panelists worked on evaluating RTT applications. They were selected from 1,500 applicants and paid $5,000 each for work that spanned a two-month period.
3. When we use the term "bias" here, we do not imply that the Department of Education or its reviewers deliberately skewed the results. We refer only to the inevitable unconscious factors that influence subjective judgments.
4. For example, in response to observations by researchers that no statistically valid methods exist to use student test scores for teacher evaluation, the final RTT regulations award only five out of 500 points for policies that use scores in this fashion—although with so little weight ultimately assigned to this metric, it would have made more sense to follow the recommendations of researchers and simply eliminate it.
5. For this estimate, the weight of each of the other indicators was reduced by 0.48%, or 12/25.
6. U.S. Department of Education. 2010. "Press Release: Obama Administration's Education Reform Plan Emphasizes Flexibility, Resources and Accountability for Results." March 15. http://www2.ed.gov/news/pressreleases/2010/03/03152010.html
7. To the administration's disappointment, its supporters in Congress were forced to eliminate early childhood grant money from the reconciliation bill because of a Congressional Budget Office ruling that reduced estimated savings from federalizing the college loan program. The earlier, higher estimate of savings had been designated, in part, to fund early childhood services.
8. Assuming that there is no "F+" or "F-" grade available.
9. This conclusion is consistent with discussions in the survey research literature. See, for example, Nora Cate Schaeffer and Stanley Presser (2003) "The Science of Asking Questions." *Annual Review of Sociology 29:* 65-88: "The choice of the number of categories represents a compromise between the increasing discrimination potentially available with more categories and the limited capacity of respondents to make finer distinctions reliably and in similar ways. Based largely on psychophysical studies, the standard advice has been to use five to nine categories…"
10. "LEA" is the Department's abbreviation for "Local Education Agency," commonly known as a school district.
11. The Department has made the results, scoring tables and notes of all reviewers publicly available at: http://www2.ed.gov/news/pressreleases/2010/03/03292010.html and at http://www2.ed.gov/programs/racetothetop/phase1-applications/index.html
12. If a high score (one of the 40s) and low score were both discarded, Florida would have received 38 points as well, after rounding.
13. We have not taken the time to go through the full scoring tables for each state, discarding high and low ratings, and then re-calculating the winners. We invite an enterprising reader to do so.
14. And, as in the previous examples, the added 27 points were then evenly subtracted from the remaining 21 metrics. The indicators on which Massachusetts scored higher than Tennessee were: (A)(3)(ii) Improving student outcomes; (B)(2) Developing and implementing common, high-quality assessments; (B)(3) Supporting the transition to enhanced standards and high-quality assessments; (C)(3) Using data to improve instruction; (D)(1) Providing high-quality pathways for aspiring teachers and principals; (D)(5) Providing effective support to teachers and principals; (E)(2)(ii) Turning around the persistently lowest-achieving schools; (F)(1) Making education funding a priority; and (F)(3) demonstrating other significant reform conditions.
15. For example, in Table 2 from the actual RTT competition, Delaware was awarded 119 of a possible 125 points for "State Success Factors," or 95.2%. Table 3 awards Delaware 95.2% of a possible 10 points, rounded to the nearest integer, or 10 points.
16. "Title I" (of the Elementary and Secondary Education Act) funds are presently distributed to states in approximate proportion to the number of low-income students in those states and to the level of existing state education spending. No competitions are required for states to qualify for such funds.