NEPC
NATIONAL EDUCATION POLICY CENTER

Review of *District Awards for Teacher Excellence Program: Final Report*

Reviewed By

Donald B. Gratz

Curry College

March 2010

Summary of Review

Performance pay plans for teachers are a hot topic across the nation, so the country's largest program is bound to attract attention. But caution is advised. This is a state-sponsored program evaluation of Texas' most recent incentive pay program rather than research, and it suffers from two major constraints: data not collected and questions not asked. First, although billed as "final," data on teacher retention and teacher perceptions include only one award cycle, and student achievement data cover just two program years—not enough data to draw solid conclusions. In addition, little is known about how districts actually implemented D.A.T.E., or about such factors as school culture, leadership and class size. Second, the report primarily explores narrow descriptive questions about the Texas program—who participated and why, what design factors they chose to implement, and so forth. It does not consider broader questions such as whether incentive pay produces positive results, not just higher test scores, and under what conditions. The report simply answers the generally descriptive questions it poses, and its authors properly caution against drawing unwarranted conclusions beyond those questions.

Kevin Welner
Editor

William Mathis
Managing Director

Erik Gunn
Managing Editor

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: 303-735-5290
Fax: 303-492-7090

Email: NEPC@colorado.edu
http://nepc.colorado.edu

Publishing Director: Alex Molnar

This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at http://greatlakescenter.org.

Review of *District Awards for Teacher Excellence Program: Final Report*

Donald B. Gratz, Curry College

I. Introduction

District Awards for Teacher Excellence Program: Final Report[1] is a program evaluation of the latest in a long series of teacher incentive programs in Texas. Conducted by the National Center for Performance Incentives (NCPI) at Vanderbilt, the study presents findings describing the "experiences and outcomes for Cycle 1 districts participating in the first two years of the program" (p. viii).

The state-funded program provides grants to districts for the implementation of locally designed incentive pay plans; the grants fund merit awards made to teachers on the basis of scores on the state achievement test plus other factors determined by individual districts. School districts have flexibility in the specifics of how they implement the plan.

The program distributes $150 million to $197 million annually to 203 participating districts. The report, at nearly 500 pages, is correspondingly large. It was released in November, 2010 (partway through the project's third year) and is based on data from Years 1 and 2.

II. Findings and Conclusions of the Report

The report presents 63 "key findings," the bulk using descriptive rather than inferential statistics. Given the number of districts involved, the amount of data is massive. More than 100,000 surveys were distributed to educators each spring, for example, and the results were subjected to a series of multiple regression analyses using a large number of predictors. Similarly, evaluators analyzed individual student achievement data on the Texas Assessment of Knowledge and Skill (TAKS) from 1,773 schools. Because "raw scale scores from TAKS were not expressed on the same developmental scale from one year to the next or from one grade to the next," the authors constructed a "standardized test score gain" for each student (p. 228).

For this and many other elaborate analytical techniques, an extensive appendix is provided. At the same time, it is important to recognize that the available data come from a very short time period—either one or two years. The data set is immensely broad without being deep. Statistical significance is easily reached with such a large number, but meaningful effect sizes—what is

generally called “practical significance” —are a separate question and may not be reached here. Findings inferred from an aggregated number of these analyses are often expressed in the report in the form of broad generalizations. Some of those key findings are summarized below.

District Participation

The 203 participating districts in this voluntary program amount to 16% of all Texas districts. But the districts tend to be poor and urban and serve nearly half of all Texas students. Each district chose either to include all schools (“district-wide”) or to select particular underperforming schools (“select-schools”) to be included. Within some parameters—at least 60% of funding must go to teacher bonuses, and awards to teachers must be based in part on scores from the TAKS—districts have wide latitude in program design and implementation. The authors also note the issue of selection bias. Because districts “chose to participate in D.A.T.E., and to design their own incentive pay plans, … if schools that ended up participating in D.A.T.E. differed systematically from non-D.A.T.E. schools,” then findings of a difference in student performance between the two sets of schools may be due to non-D.A.T.E. factors (p. 83).

Program Components

District factors studied include the choice of approach (select-school or district-wide), whether all teachers can receive awards, and how non-designated funds are used (larger teacher awards, principal awards, professional development, hard to staff positions). School factors include who receives awards, award size, and the unit of accountability (individual, team, group, or hybrid). These factors are presented in a descriptive fashion and are also used to compare student outcomes and teacher awards (as exemplified below).

Teacher Factors

Based on Year 1 data only, some teacher factors appeared to matter. For example:

- Teachers with a bachelor’s degree were 12-17% more likely to receive an award than those with no degree, but master’s and doctoral degrees did not increase this likelihood (p. 80);
- Teachers new to a school were 12% less likely to earn an award than those who had been placed there in earlier years (p. 80); and
- Teachers with 20 years of experience were 2-4% less likely to receive awards than those with 5 years (p. 78).

Student Outcomes

The report presents “descriptive differences” between average student passing rates on TAKS reading and math tests for D.A.T.E. and non-D.A.T.E. schools. The authors looked at constructed individual student gain analyses based on TAKS scores, which they averaged by

school in order to compare D.A.T.E. and non-D.A.T.E schools. They could not "provide links of teachers to students," the authors point out, "so it is not possible to identify the most successful teachers or to identify the impact of specific teachers on student performance" (p. 83). However, the report's summary of findings from multiple grades and multiple subjects shows a "generally positive" result, and some contradictions and perplexities:

- TAKS passing rates for reading and math were lower at D.A.T.E. schools than non-D.A.T.E. schools, but the relationship between D.A.T.E. participation and average student achievement gains was "positive, statistically significant, but small in magnitude" (p. 84). For example, the difference in TAKS math passing rates between D.A.T.E. and non-D.A.T.E schools for grade seven went from more than 9% in 2005-06 to about 4% in 2009-10—with D.A.T.E. schools below the non-D.A.T.E. schools but moving closer.
- That is, "…D.A.T.E. schools exhibited negative gain scores, but their scores became less negative (closer to zero) over time" (p. 91).
- The unit of accountability (individual or group) "was related to student achievement in both reading and math, but not in a consistent direction" (p. 84).
- "Increasing the maximum award by $1,000 was associated with an increase in TAKS math scores of approximately one scale score point (p. 155), but this did not hold true for reading. One scale score point is a very small change.

Educator Opinions

Each district designed its own incentive plan within parameters, as described above. Teachers in D.A.T.E. schools tend to think their D.A.T.E. plans are fair, the goals worthy, and the recipients of awards deserving. Significantly, they do not believe the incentive plans contribute much to school improvement, but teachers who are positive about their plans tend to be positive about other aspects of their schools (p.125). Teachers in schools with group incentives express greater satisfaction with their schools and the incentive program than those with individual awards, and perceive a more satisfied and collegial workplace (p.154). Yet, individual awards produced reports of greater motivation and greater competition, plus higher test score gains and incentive payments (p.155). These findings are intriguing, but explanations are not suggested.

Teacher Turnover

Teacher turnover for 2008-09 declined in D.A.T.E. districts more than predicted based on previous years, but for district-wide plans (-1.3%) the change was "fully attributable to in-district turnover." In select-school plans (-2.2%), the reduction occurred regardless of whether the schools were D.A.T.E. schools or not. This result, according to evaluators, "raises the possibility that some other policies" in select school districts may have been the cause (p.108-09). Teachers who expected to receive awards tended to receive them, and those who received awards, particularly larger ones, were more likely to stay put.

Differences between High- and Low-Performing Schools

The value-added approach to measuring student gains was used to compare expected and actual performance on TAKS for D.A.T.E. and non-D.A.T.E schools, and to compare program design differences among D.A.T.E. schools.

High-performing schools were more likely to offer awards to principals (60%) than low-performing (15%), to offer larger awards to teachers, and to use a blended approach to teacher awards (group and individual elements). There were "profound differences" in school productivity (TAKS reading and math scores) between high- and low-performing D.A.T.E. schools, but few differences between schools in either district-wide or select-school districts. The authors caution that "This simple descriptive analysis does not establish a causal link" between specific D.A.T.E. design features and increased school effectiveness (p. 98).

III. The Report's Rationale for Its Findings and Conclusions

"The report's objective," say its authors, "is to inform policymakers and practitioners as they consider how to move forward, how to design and implement incentive pay and compensation reform for educators, and the implications of those policy choices" (p. 5).

The timeline of the study may have been dictated by the need to produce information for Texas' next funding cycle. Since the state's previous plan is described as achieving "dismal results,"[2] the authors' conclusion was probably helpful. Based on educator surveys, the "generally positive" test score gains on TAKS, and teacher turnover rates 1-2% lower than predicted, they conclude that: "…more often than not, participants in the D.A.T.E. program had a positive experience, student achievement gains and teacher turnover moved in a generally desirable direction, and teacher attitudes were favorable towards D.A.T.E."(p.xiii). Many of the findings are descriptive rather than analytic (such as reasons for district participation), but may be of interest to Texas policy-makers.

As a program evaluation, the report addresses questions of interest to its client rather than typical research questions. Still, in serving this purpose and meeting this deadline, the report may not meet its larger objective of informing policy-makers outside of Texas, and perhaps not even in Texas. The unanswered questions described below are questions that policy-makers *should* be asking,

IV. The Report's Use of Research Literature

NCPI is a major research center on performance pay, and it is not surprising that the authors reference historical and assessment literature—some of which they generated. However, they do not discuss other recent research that has produced different results. For example, the positive responses from teachers in Texas regarding the program and the value of bonuses conflicts with NCPI's own well-designed study in Nashville.[3] Similarly, a much publicized Mathematica study in Chicago of the Teacher Advancement Program found no significant results either in teacher

retention or student achievement.[4] Both the Nashville and Chicago studies used experimental or quasi-experimental designs, allowing for a richer analysis.

The on-going "ProComp" reform in Denver, arguably one of the most prominent experiments, is also not addressed. Denver experienced similarly modest test score gains during its Pay for Performance Pilot, for example, and teachers who participated in the pilot supported it. In Denver, however, we know why. Teachers liked the program and were happy to receive awards, but scorned the motivational value of these awards, which they found insulting. Instead, they believed that the realignment of school and district activities in support of teaching and learning made the difference.[5] This is significant, as D.A.T.E., like many such programs, is based on a theory of teacher motivation rather than district support.[6] That is, if D.A.T.E. teachers held views similar to Denver teachers—they liked the plan but did not believe that bonuses were effective change incentives—it would undercut the primary premise of incentive pay.

The evaluators steer clear of literature questioning the undergirding program foundations, such as whether high-stakes use of standardized tests is wise policy.[7] This is a particularly relevant issue given Texas' history of standardized testing, which has sometimes been detrimental to poor, urban children, who were drilled hard but taught little.[8] Though these kinds of references may not be common in program evaluations, the vital questions they pose require examination before programs such as D.A.T.E. are implemented.

V. Review of the Report's Methods

The report addresses 34 research questions of interest to policy makers in Texas. As noted, its data set is very broad. It contains significant gaps, however, and covers a time span that provides an inadequate base for such major policy decisions. Primary methodologies include the following:

- Program design factors based on district proposals are used to understand program results, including the differences between high- and low-performing schools. These factors, which include the size of the awards, the unit of accountability, and whether principals can receive awards, were used to compare teacher results and school productivity. Though most districts probably implemented what they proposed, the actual degree and fidelity of implementation is unknown.
- Teacher and administrator opinions are drawn from more than 100,000 annual spring surveys from Years 1 and 2. The analysis of these surveys is described in an elaborate technical appendix which includes reliability and correlational analyses of clusters from personnel surveys, means tables for survey item clusters, tables for hierarchical linear modeling, a description of how control schools were selected, copies of the instruments, and so forth (p. 124). Still, the results include only one award cycle.
- A value-added approach is used to compare expected and actual student performance between D.A.T.E. and non-D.A.T.E schools, as measured by TAKS pass rates from 2006-07 through 2009-10 (two years under D.A.T.E.). The evaluation includes a regression analysis "that allows evaluators to condition on many background characteristics" of students that may impact achievement (p.89). Student factors include the percentage of

white, LEP and gifted and talented students. Teacher factors include years of experience and average salary, and school factors include teacher-student ratio and type (traditional, charter, etc.) (p.231). As noted above, a value-added score is "constructed for each student based on previous years' scores" (p.228).

- Teacher turnover data for each district (one year) is compared to projected turnover based on six previous years for both D.A.T.E. and non-D.A.T.E. schools. The analysis attempts to separate out types of turnover (leaving the school, district, or profession), and to control for non-D.A.T.E. factors in determining whether receiving an award and the size of the award influences teacher retention.

VI. Review of the Validity of the Findings and Conclusions

The report is generally thorough and professional, using extensive statistical techniques where possible, but suffers from two significant limitations: data not available and questions not asked. These limitations weaken the analysis, and prevent the reader from discovering meaning in its findings. The authors regularly insert caveats as to how much can be inferred from their results, though they do not mention the limitation of conducting a final study in the second year of a three-year project. For example:

Student Performance

TAKS passing rates in the two years assessed increased slightly at D.A.T.E. schools, meaning that they lost less ground than in previous years to non-D.A.T.E. schools. But the effects were small, as noted above, and the timeline was short. While it may be that scores "moved in a generally desirable direction," two years is too short a period to confirm a trend and the impact of the program is unclear. The validity of TAKS as a single measure of achievement is not raised.

Differences Between High- and Low-Performing Districts

Large differences in school "productivity" (as measured by TAKS reading and math scores) were found between high- and low-performing D.A.T.E. schools, suggesting "profound differences" in implementing D.A.T.E. (p.99). High-performing districts proposed higher awards for teachers, rewards for principals, and hybrid approaches, among other factors. These are important factors, but other implementation and school differences may exist that are not explored. Given past instances of drilling low-income students on test-taking rather than learning,[9] we should draw conclusions with care. It's possible to believe that $3,000-6,000 bonuses may change teacher behavior, but the report can't determine whether higher test scores are due to better teaching, more drill, professional development, or something else.

Teacher Turnover

Teacher turnover is based on one year of program data. Turnover diminished more in D.A.T.E. districts than statewide, as described above, but this decline "was fully attributable" to in-district

transfers for district-wide districts. The decline in turnover in select-school districts occurred whether the school was a D.A.T.E. school or not, suggesting that "some other policies may have changed" in these districts—a non-D.A.T.E factor.

At the same time, individual teachers who received large awards were more likely to stay put than those who did not. The authors proclaim that the turnover rate "surged" among teachers not receiving awards, and "fell sharply" among teachers who did, terms that will doubtless be repeated in the press (p.109). Many unknown factors cloud these conclusions, however, and it is not clear that such dramatic descriptions are appropriate for the small magnitude of the changes.

Educator Opinions

Surveys were distributed twice, covering one award cycle. They produced many interesting findings but results may change after the first year.

In one example, teachers who expected awards most often report receiving them. They also report greater use of professional practices and development—a potentially important finding. Some districts plans proposed funding professional development, but we don't know whether the districts or schools providing the professional development are those referenced by teachers, as occurred in Denver, or whether teachers initiated their own professional development. This is unfortunate. The latter practice could indicate the successful use of bonuses to motivate teachers to improve—a significant result, quite different from other studies.[10]

Teachers who were positive about the program were also positive about their schools generally, to cite another example. But it's possible that school satisfaction leads to program satisfaction, rather than the reverse. Indeed, evidence from Denver suggests that teachers rarely think in their daily work about possible bonuses in June, but are affected by the school culture and climate every day.[11] It could be that teachers simply like their schools, with or without D.A.T.E, a different conclusion from the report. In sum, despite the report's significant breadth, it lacks the data to explore its findings in greater depth.

VII. Usefulness of the Report for Guidance of Policy and Practice

The study presents many intriguing findings about teacher attitudes, the potential impact of group versus individual awards, and other program factors. But the evaluation's design limitations undercut its broader use. We learn that teachers support the program, that it is associated with a slight improvement in test scores and a decline in teacher turnover, and that schools where teachers and principals received larger awards posted slightly higher test scores. But these generally positive findings are too small, too confounded, come in too short a time span and leave too many unexplored questions. Despite this wealth of detail, the reader is not much wiser regarding the issues and impact of performance pay by the end.

Based on their responsible caveats and generally modest conclusions, the authors acknowledge the study's limitations. They sampled widely but not deeply, conducted thorough analyses of the

data they collected, and raised cautions as to specific conclusions. If their study is used to summarize a massive state experiment as a prelude to a more in-depth study and analysis, it adds to the conversation. But if it is touted as proof that performance pay can work, it is being misused. It is far too early in the D.A.T.E. program, and the gaps in data are far too great, to rely on this report to support any conclusion that performance pay "works." Unlike the reports referenced above from Nashville, Chicago and Denver, this study does not delve deeply enough to address this causal question. It describes the details of D.A.T.E. but cannot yet demonstrate that the concept of incentivizing teachers has positive or lasting results. Its findings are interesting, but not sufficient for guiding policy.

One final point may be the most significant: the study does not question the underlying premises of performance awards for teachers: Is TAKS a legitimate measure of student achievement? Can test scores alone define good teaching, or might D.A.T.E. be encouraging unintended consequences and undesirable ends such as narrowing the curriculum, emphasizing test-taking over thinking, encouraging student passivity, or ignoring students' social and emotional growth? Since D.A.T.E. districts serve primarily low-income, urban students, such a result might increase the knowledge gap even as the test-based "achievement" gap decreases. An evaluation that considered these basic questions of program impact would provide important guidance for policymakers.

Notes and References

1 Springer, M.G., Lewis, J.L., Ehlert, M.W., Podgursky, M.J., Crader, G.D., Taylor, L.I., Gronberg, T.J., Jansen, D.W., Lopez, O.S., Stuit, D.J. (December 2010). *District Awards for Teacher Excellence Program: Final Report*. National Center for Performance Incentives. Retrieved December 29, 2010, from http://www.performanceincentives.org/data/files/news/BooksNews/FINAL_DATE_REPORT_FOR_NCPI_SITE.pdf.

2 Stutz, T. (2010, December 6). Merit pay helps keep staff, raises scores, Texas study finds. *Education Week*, Teacher. Retrieved December 29, 2010, from http://edweek.org/tm/articles/2010/12/06/mct_txmeritpay.html?tkn=TUTDADd5R.

3 Springer, M. G., Ballou, D. H., Laura, Le, V.-N., Lockwood, J., McCaffrey, D., Pepper, M. F. et al. (2010, September 21). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Retrieved September 22, 2010, from http://www.performanceincentives.org.

4 Glazerman, S., & Seifullah, A. (2010, May 17). *An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report*. Retrieved August 18, 2010, from http://mathematica-mpr.com/education/.

5 Slotnik, W. J., Smith, M., Helms, B. J., & Glass, R. (2004, January). *Catalyst for Change: Pay for Performance in Denver, Final Report*.Boston: Community Training & Assistance Center.

6 Gratz, D. B. (2010, May). Looming Questions in Performance Pay. *Phi Delta Kappan, 91*(8), 16-21.

7 See, for example:

Berliner, D. C. & Nichols, S. L. (2007, 12 March). High Stakes Testing is Putting the Nation At Risk. *Education Week, 26*(27), 36, 48.

Heubert, J. P., & Hauser, R. M., editors. (1999). *High Stakes: Testing for Tracking, Promotion, and Grading*. Washington, DC: National Academy Press.

8 McNeil, L. M. (2000). *Contradictions of School Reform: Educational Costs of Standardized Testing*. New York: Routledge.

9 McNeil, L. M. (2000). *Contradictions of School Reform: Educational Costs of Standardized Testing*. New York: Routledge.

10 See, for example:

Johnson, S. M., Donaldson, M. L., Munger, M. S., Papay, J. P., & Qazilbash, E. K. (2007, June). *Leading the Local: Teachers Union Presidents Speak on Change, Challenges*. Washington, DC: Education Sector Reports.

Dolton, P., McIntosh, S., & Chevalier, A. (2003). *Teacher Pay and Performance*. London: Institute of Education, University of London.

11 Gratz, D. B., Slotnik, W. J., & Helms, B. J. (2001). *Pathway to Results: Pay for Performance in Denver*. Boston: Community Training & Assistance Center.

DOCUMENT REVIEWED:	**District Awards for Teacher Excellence Program: Final Report**
AUTHORS:	M.G. Springer *et al*
PUBLISHER/THINK TANK:	National Center for Performance Incentives
DOCUMENT RELEASE DATE:	December 2010
REVIEW DATE:	March 3, 2011
REVIEWER:	Donald B. Gratz, Curry College
E-MAIL ADDRESS:	dgratz@curry.edu
PHONE NUMBER:	(617) 333-2130

SUGGESTED CITATION:

Gratz, D.B. (2011). *Review of "District Awards for Teacher Excellence Program: Final Report."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-district-awards.