COMMENTARY

Christopher H. Tienken, Editor
AASA Journal of Scholarship and Practice

Pay for Performance Based on Standardized Test Scores: Twenty Questions

Proposals for administrator and teacher evaluation schemes are not in short supply. Pay for performance systems based on students' results from state mandated standardized tests is a policy idea gaining traction in the halls of the United States (U.S.) Congress and state legislatures.

The Elementary and Secondary Education Act (ESEA) renewal proposal includes rewarding teachers and administrators for increasing student standardized test scores.

The Race to the Top (RTTT) federal grant program requires states to link the evaluations and pay of school administrators and teachers to student performance. States such as Colorado, Texas, New Jersey, Missouri, Florida, Tennessee, Nevada, Idaho, Illinois, Indiana, and others have passed legislation or have bills under consideration to link administrator and/or teacher pay to student performance as measured in part or totally by results on summative state-mandated standardized tests of academic skills and knowledge.

Will this be another policy example of data-less decision making?

Above the Law

Professor and economist Charles Goodhart (1975) is credited with demonstrating that "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes" (p.116). This principle is known as Goodhart's Law.

Some questions and concerns arise when one applies that law to performance pay for teachers and school administrators based on student results on statewide tests. In the case of performance pay based on a student test score, it is the test score that becomes the "observed statistical regularity." Since the inception of the No Child Left Behind Act ([NCLB] 2002), Goodhart's Law has clearly been observed.

The validity of state test results became unstable as a result of the high-stakes regulatory consequences attached to them. For example, the size of many states' gains on the National Assessment of Educational Progress

(NAEP) did not keep pace with gains on their mandated state tests of skills and knowledge. States like Texas reported large gains in the percentage of students who scored proficient on the Texas Assessment of Knowledge and Skills, but demonstrated a smaller percentage gain of students rated proficient on the NAEP.

In response to the myopic focus placed on test scores by state education officials, school district leaders, in some cases, resorted to test-gaming practices. Among these are holding back low-achieving students instead of promoting them into a grade level with an important mandated test (Remember the Texas Miracle?), and counseling large numbers of students to drop out and pursue a GED prior to a high school exit exam.

A growing practice includes targeting school resources to those students close to passing the state test, known as "bubble kids," at the expense of other students. Students deemed as "almost or just-proficient" receive additional instruction while those more needy or gifted do not.

These well-documented practices illustrate how some districts are trying to raise scores, but ultimately are decreasing the validity of the results and impoverishing the educational experience for all students (Booher-Jennings 2005).

In some school districts, the results from state standardized tests provide little real information about student learning. The results are skewed because they are produced through intensive test preparation, lax truancy enforcement during testing cycles, yearly changes to state proficiency cut points, increased dropout rates in urban areas, moving and shifting of students among schools so their scores do not count, enrolling students to "home school" status at test time, and other practices that have little to do with quality education practices, but have been known to raise aggregate test scores.

In essence, any links to student learning and the quality of teachers and school administrators are tenuous because the scores are produced, in part, by manipulation of the system and a focus on test scores causes a manipulation of the system (McNeil et al. 2008; Stroup 2009; Ravitch 2010). Issues of moral, professional, and ethical pollution will increase after more teachers and administrators are subjected to these types of performance schemes. The test scores and processes that produce them will continue to become corrupted (Nichols and Berliner 2005).

What we are now seeing in greater frequency is the confluence of Goodhart's Law and Campbell's Law. Campbell (1976) stated, "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p.46).

Consider how body counts, school quality rankings in the newspaper, George H. W. Bush's war on drugs, and end of the month police ticket quotas (Rothstein, Jacobson, and Wilder, 2008) are easily corrupted. The body count analogy made by Rothstein et al. (2008) is useful. It is well known that U.S. commanders and civilian policy makers in the Department of Defense used quantitative data to make battlefield decisions during the Vietnam War (McNamara and VandeMark, 1996).

Quantitatively speaking, the U.S. won the Vietnam War by a landslide with less than 60,000 casualties compared to an estimated 1.17 million North Vietnamese (Soames, 2005). But as we now know the quantitative measures used by policy makers during the

Vietnam era to monitor success turned out to be imperfect, incorrect, and corrupted indicators.

A similar thing takes place when people make important decisions about students, teachers, administrators and schools based on student results from one statewide test. Koretz (2008, 236) called this phenomenon "corruption of measures" in educational testing policy.

Educators and policy makers who support pay for performance need to step back, slow down, ask more questions, and not accept the superficial answers coming from governors, state legislators, and others who neither understand the statistical intricacies nor in some cases care to learn.

Recent Research

School administrators need to move beyond the noise and corporate marketing of pay-for-performance schemes based on student test results and educate themselves on recent empirical evidence on the subject. Information gleaned from studies and reports provide some clarity on the issue.

First, very few white-collar private sector professionals receive performance pay based on a single or very narrow set of indicators. In fact, only six percent of private-sector employees received direct, output-based cash payments according to the 2005 National Compensation Survey (Adams et al. 2009; Springer et al. 2010). Most of those workers were in commission-based fields like used-car salesmen, penny-stock brokers, and real estate agents; hardly comparable professions to that of raising children to be productive, ethical, and moral citizens.

Results from the longitudinal Project on Incentives in Teaching (POINT) conducted by researchers at Vanderbilt University's Peabody School of Education suggested that performance pay did not have a significant impact on student achievement in mathematics in Grades 5-8 (Springer et al. 2010) for students of teachers eligible for bonuses from between $5,000 to $15,000 compared to teachers not eligible.

The researchers stated, "… there were no significant differences for students in Grades 6-8 when separate effects were estimated for each grade level" (p. 43). A positive effect was found only in Grade 5 and it did not persist in Grade 6 or other grade levels. The researchers stated, "To conclude, there is little evidence that POINT incentives induced teachers to make substantial changes to their instructional practices or their level of effort …" (p.45).

Similar results were found from another experimental study conducted in New York City (Fryer, 2011). "Surprisingly, all estimates of the effect of teacher incentives on student achievement are negative in both elementary and middle school …" (p. 18).

The impact of performance pay on student achievement in elementary school and middle school in the area of language arts and mathematics, as measured by state standardized tests in NYC, was negative with effect sizes ranging from -0.02 to -0.05. Furthermore, the pay system in the NYC experimental study did not improve student attendance, grade point average, or achievement on alternative measures of achievement such as other standardized tests taken by students.

Results were similar for high school students. "Similar to the analysis of elementary and middle schools, there is no evidence that teacher incentives had a positive effect on achievement. Estimates of the effect of teacher incentives on high school achievement are all small and statistically non-significant" (p. 18).

Why?

So why would we, as a country, want to pursue another policy that has not been fully vetted, tested, or modeled to identify and address all the possible negative unintended consequences to children and education professionals? Evidence suggests that pay for performance based solely, or to a large degree, on standardized test scores is not universally effective and could be detrimental to achievement (Adams, Heywood, and Rothstein 2010; Buzik & Laitusis, 2010; Springer et al. 2010).

Twenty Questions

Before we launch ourselves off of yet another reform precipice without a parachute for children, those who are proposing the policy should at least have evidence-based answers for the following questions:

1. Why expose children and education professionals to yet another unproven intervention? (Think high school exit exams, Reading First, charter schools, vouchers, high stakes standardized testing in Grades 3-8, etc.)

2. Why, if only approximately six percent of professionals in the private sector have their pay tied directly to quantitative indicators, are we so quick to implement such plans in schools without further study or attention to the unintended consequences raised in recent studies on the topic (Adams et al. 2009; Springer et al. 2010)?

3. How do proponents of pay for performance based on student test results reconcile the scheme with theories such as Hertzberg's (1968) Two-Factor Theory of Motivation, Maslow's Hierarchy of Needs (1954), Reactance Theory, and the work of Pfeffer and Sutton (2006), among others, which suggest that long-term effects will be detrimental to the system and not result in improved student learning?

4. What protections will be put in place in the pay for performance schemes to protect against the narrowing of the curriculum that occurs when test results become the ultimate outcome variable to determine the quality of the education processes (see Au 2007)?

5. According to UNICEF (2005), the United States is second only behind Mexico in the percentage of children living in poverty in the industrialized world. How will pay for performance programs account for the debilitating effects poverty has on achievement (Coleman et al. 1966; Hart and Risley 1995; Sirin 2005; Emerson 2009)?

6. Student prior achievement has an effect size of 0.67 on later achievement. That is the difference between scoring at the 50th percentile compared to scoring at the 73rd percentile on a nationally norm-referenced test (Feinstein 2003; Duncan et al. 2007). How will pay schemes based on test results account for prior achievement?

7. Without mandated random assignment of students to classes how will policymakers ensure that classes are balanced in terms of student prior achievement, disabilities, and other demographic characteristics that effect student achievement on statewide standardized tests?

8. The effect size difference in achievement for students who attend a high-quality preschool program compared to those who do not is about 0.44, or equal to the difference between scoring at the 63rd percentile versus the 50th percentile. How will performance pay systems account for the influence of children having attended a high-quality, low-quality, or no preschool program at all on student achievement (Jones 2002; Loeb et al. 2004)?

9. How will pay schemes account for the effects of low birth weight on academic achievement? Low birth weight—more prevalent for African American babies and babies born into poverty—has a direct effect on IQ if medical and educational interventions are not in place during the early years of a child's life (Bhutta et al. 2002). The effect size difference between low birth weight babies who did not receive appropriate interventions during the early years and babies born within normal weight ranges is about 0.54, or the difference between scoring at the 50th percentile and the 65th percentile.

10. How will pay schemes account for changes in achievement caused by students going through divorce or a death of a parent? Although small, the achievement differences averaged 0.17 or about six percentile points on norm-referenced tests (Kunz 1995; Jeynes 2006).

11. How will the schemes separate the influence on student achievement that the Grade 8 language arts teacher has on Grade 8 math performance? For example, a review of the nation's high school and Grade 8 tests reveals that there is about a 0.50 to 0.75 correlation between language arts and math scores on state tests (Tienken 2008). How do the current policy proposals disentangle the interrelatedness of the education process that takes place in schools and outside of the school walls? Subject area learning does not occur in a vacuum.

12. How will pay systems that are linked to student standardized test scores account for the standard error of measurement (SEM)? SEM is similar to the margin of error in a political poll and it is inherent in all standardized test results. The reported score is not the student's true score (Tienken 2008). The amount of error on the Grade 8 state tests ranges from 3 scale-score points to 85 scale-score points nationally. In New Jersey, there are about 10 scale-score points of error in student test scores. If a student receives a 200 scale score, the true score can be anywhere from a 190 to a 210. That range could mean the difference between receiving a raise or not. No state education agency mediates SEM at the student level (Tienken 2011).

13. How will pay for performance schemes account for differences in access to resources within and among classes within schools in the same district?

14. How will pay schemes account for having to work for a school or district administrator or school board that does not understand the research regarding evidence-based practice and mandates negative or educationally bankrupt practices?

15. Are pay for performance policy initiatives just Trojan horses for union busting and under-paying teachers and administrators?

16. Why are some school administrators and their organizations actively supporting pay for performance schemes when they lack answers to the above questions?

17. Should school administrators who willingly implement pay for performance schemes linked to student results on standardized tests without strong empirical evidence lose their licenses due to educational malpractice?

18. Does implementing an untested intervention on children who are compelled to participate violate any of the *Interstate School Leaders Licensure Consortium* (ISLLC) standards? If not, why?

19. Would a child be compelled to be part of a medical experiment in which the prior results were negative and/or unknown? If not, then why are some school leaders allowing students in their schools to be subjected to this unknown system?

20. If the private sector cannot get pay for performance schemes correct and most private sector managers do not think they are a good idea (Pfeffer & Sutton, 2006), why is the education field willing to support these ideas?

School leaders—and, more importantly, teachers—have very little control over the answers to these questions. Schooling does not dictate the processes or environments that cause poverty, divorce, low birth weight, or academic experiences prior to entering school. Nor can it mediate fully their effects using resources currently available. Therein resides the problem: The proposed policies on pay for performance do not account for or mediate the main factors that affect performance on state standardized tests.

Editor's note: Portions adapted from Tienken, C.H. (2011). Pay for performance: Whose performance? *Kappa Delta Pi Record, 47*(4), 152-154.

References

Adams, S. J., J. S. Heywood, and R. Rothstein. (2009). Teachers, performance pay, and accountability: What education should learn from other sectors. *Washington, DC: Economic Policy Institute.*

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher 36*(5), 258–67.

Bhutta, A. T., Cleves, M. A., Casey, P. H., Cradock, M. M. , & Anand, K. J. S. (2002). Cognitive and behavioral outcomes of school-aged children who were born preterm: A meta-analysis. *Journal of the American Medical Association 288*(6), 728–37.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal 42*(2), 231–68.

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 9*(7), 537-544.

Campbell, D. (1976). *Assessing the impact of planned social change.* Occasional paper series, paper #8. Kalamazoo, MI: Evaluation Center, Western Michigan University.

Coleman, J. S., et al. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Dept. of Health, Education, and Welfare, Office of Education.

Emerson, E. (2009). Relative child poverty, income inequality, wealth, and health. *Journal of the American Medical Association 301*(4), 425–26.

Fryer, R. G. (2011). *Teacher Incentives and Student Achievement: Evidence from New York City Public Schools*. NBER Working Paper Series. Retrieved from: http://ssrn.com/abstract=1776785 national bureau of economics research

Goodhart, C. A. E. (1975). *Monetary relationships: A view from Threadneedle Street.* Papers in Monetary Economics. Sydney, New South Wales, Australia: Reserve Bank of Australia.

Hart, B., and T. R. Risley. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore, MD: Paul .H. Brookes Pub. Co.

Herzberg, F. (1968). One more time: How do you motivate employees? *Harvard Business Review, 46*(1), 53–62.

Jeynes, W. H. (2006). The impact of parental remarriage on children: A meta-analysis. *Marriage and Family Review 40*(4), 75–102.

Jones, S. S. (2002). The effect of all-day kindergarten on student cognitive growth: A meta-analysis. Unpublished Ed.D. diss. University of Kansas, Lawrence, KS.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Kunz, J. (1995). The impact of divorce on children's intellectual functioning: A meta-analysis. *Family Perspective, 29*(1), 75–101.

Loeb, S., B. Fuller, Kagan, S, & Carrol, B. (2004). Child care in poor communities: Early learning effects of type, quality, and stability. *Child Development 75*(1), 47–65.

McNamara, R. S. & VanDeMark, B. (1996). I*n retrospect: The tragedy and lessons of Vietnam*. New York: Vintage.

McNeil, L. M., Coppola, E., Radigan, J., & Heilig, J. V. (2008). Avoidable losses: High-stakes accountability and the dropout crisis. *Education Policy Analysis Archives, 16*(3), 1–48.

Nichols, S. L. & Berliner, D. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. Tempe, AZ: Education Policy Studies Laboratory, Arizona State University.

No Child Left Behind Act. (2002). Public Law 107–110. Washington, DC: U.S. Congress. Available at: www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

Pfeffer, J. & Sutton, R. L. (2006). *Hard facts: Dangerous half-truths and total nonsense*. Boston, MA: Harvard Business School Press.

Ravitch, D. (2010). *The death and life of the great American school system*. New York: Basic Books.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3): 417–53.

Soames, J. (2005). *A history of the world*. New York: Routledge.

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., & Stecher, B. (2010). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Stroup, W. M. (2009, March 18). What Bernie Madoff can teach us about accountability in education. *Education Week,*

Tienken, C. H. 2008. The characteristics of state assessment results. *Academic Exchange Quarterly, 12*(3): 34–39.

Tienken, C. H. (2011). Structured inequity: The intersection of socioeconomic status and the standard error of measurement of state-mandated high school test results. *NCPEA Yearbook*. Ypsilanti, MI: NCPEA Publications.

UNICEF. (2005). Child poverty in rich countries, 2005. Innocenti Report Card No. 6. Florence, Italy: UNICEF Innocenti Research Centre. Available at: www.unicef.org/irc.