Commentary

Administrators Gaming Test- and Observation-Based Teacher Evaluation Methods: To Conform To or Confront the System

Tray J. Geiger, MEd
Doctoral Student
Mary Lou Fulton Teachers College
Arizona State University
Tempe, AZ

Audrey Amrein-Beardsley, PhD
Professor
Mary Lou Fulton Teachers College
Arizona State University
Tempe, AZ

Abstract

In this commentary, we discuss three types of data manipulations that can occur within teacher evaluation methods: artificial inflation, artificial deflation, and artificial conflation. These types of manipulation are more popularly known in the education profession as instances of Campbell's Law (1976), which states that the higher the consequences or stakes surrounding almost any quantifiable event (e.g., one that is based on numerical scores or outcomes), the more likely the scores or outcomes are subject to pressures of corruption and distortion, as directly related to the relative importance or weight of the consequences attached. We examine each type of data manipulation and consider the greater impact of each on practice and policy.

Key Words

teacher evaluation, data manipulation

"Gaming the system," or what can be more popularly identified as instances of Machiavellian, ends-justifying-the-means schema that help advance individuals' careers, frequently occur in all realms of life. Popular press mentions often include accounts of this in the sports world, with international news most recently revealing that 16 professional tennis players, including eight who partook in the 2016 Australian Open—one of the sport's biggest tournaments—were involved in "match fixing," where players purposefully altered the outcomes of matches to make significant sums of money (Blake & Templon, 2016).

Likewise, manipulation of the stock market in terms of insider trading occurs, whereby a person who has insider knowledge makes trades on behalf of him/herself and perhaps others, after which the trades made cause stock prices to artificially inflate or deflate, without any real change to the market-value of the stocks themselves.

Recall in the 1980s when airline executives extended flights' times of arrival to increase the percentages of on-time flight percentages for which airlines were being held accountable. Remember as well when crime rates observed during Richard Nixon's presidency became suspiciously underreported and downgraded to less serious categories to yield the anticipated "less crime" objectives and goals set by the Nixon legislation.

Gaming the System in Education

These occurrences of manipulation are more popularly known in the education profession as instances of Campbell's Law (1976). Campbell's Law states that, in essence, the higher the consequences or stakes surrounding almost any quantifiable event (e.g., one that is based on numerical scores or outcomes), the more likely the scores or outcomes are subject to pressures of corruption and distortion, as directly related to the relative importance or weight of the consequences attached. As one might expect, the effects of Campbell's Law have been prevalent in education for years, predominantly surrounding high-stakes standardized testing and teacher-level accountability policies as based on high-stakes tests.

Instances and reports of teachers helping students with questions on standardized tests, teachers replacing students' incorrect answers with correct answers, teachers excluding or exempting certain low-scoring subgroups from testing, and the like, have been present throughout research and popular press sources (see, for example, Nichols & Berliner, 2007).

These occurrences have been most notable since the passage of former president George W. Bush's No Child Left Behind (NCLB, 2001) Act, but also noted in the research prior (e.g., since the state of Florida first introduced in 1979 what we now know as a high-stakes test). Educators have felt similar pressures to game the system in response to other increased accountability policies and initiatives (e.g., the Race to the Top Act of 2011).

In fact, primarily these two federal education policies (i.e., NCLB and Race to the Top), the latter of which incentivized states with $4.35 billion in federal funds to adopt and implement new and improved teacher evaluation and accountability systems (i.e., as largely reliant upon numerically measuring the extent to which teachers "grow" or "add value" to their students' academic achievement over time using advanced statistical growth or value-added models (VAMs)) mandated and incentivized states, respectively, to theoretically realize educational reform. VAMs

are designed to isolate and measure teachers' alleged contributions to student achievement on large-scale standardized achievement tests as groups of students move from one grade level to the next.

VAMs are, accordingly, used to help objectively compute the differences between students' composite test scores from year-to-year, with value-added being calculated as the deviations between predicted and actual growth (including random and systematic error). Differences in growth are to be compared to "similar" coefficients of "similar" teachers in "similar" districts at "similar" times, after which teachers are positioned into their respective and descriptive categories of effectiveness (e.g., highly effective, effective, ineffective, highly ineffective).

Simultaneously, however, Campbell's Law has also since had its way as per the distortion of the very numerical indicators at play, whereby states and many state leaders continue to do whatever it takes to reap or avoid the high-stakes awards and penalties also attached (e.g., significant monetary bonuses paid to superintendents adopting and promoting such policies; see, for example, Amrein-Beardsley, Collins, Holloway-Libell, & Paufler, 2016).

In fact, school administrators in some states and districts have faced incredible pressures to artificially manipulate high-stakes test data for multiple reasons, and they have engaged as a result, all the while evidencing additional instances of Campbell's Law.

Namely, since Race to the Top (2011), and states' and school districts' subsequent foci on teacher level accountability as measured by teachers' levels of growth or value-added, school administrators have taken it upon themselves (or been advised or forcefully persuaded) to:

(1) artificially inflate teachers' observational scores (i.e., rubric-based measures of teachers' in-classroom instructional practice(s)) to protect their teachers against what school administrators often view as the extreme consequences (e.g., teacher termination, the revocation of tenure) attached to what they also often view as unreliable, invalid, or unfair teacher accountability systems;

(2) artificially deflate teachers' observational scores to consciously guard against their own (sub)conscious and "subjective" biases and prejudices, as often charged or accused; and

(3) artificially conflate both teachers' observational scores and growth/VAM scores to guarantee that the two adequately align and correlate as theoretically expected, and also pragmatically required should either or both indicators be used in consequential ways.

Evidence of validity increases as measurement indicators point in the same direction and support the same inferences and conclusions to be drawn (i.e., convergent-related evidence of validity). In this case, if both measures (i.e., the growth or value-added and observational measure) line up, they validate one another, and yield the required evidence needed to support increased confidence in both measures as independent measures of the same construct.

These gaming instances are emphasized herein because school administrators are the educators who are often either encouraging or engaging in these gaming behaviors, again, for a variety of reasons; hence, this is the exact audience that needs to better understand what engaging in primarily these gaming behaviors

means in terms of validity, or the validity of the inferences to be derived via either or both the growth/VAM and observational estimates at play.

Artificial Inflation, Deflation, and Conflation

Artificial inflation

Artificial inflation occurs when school administrators artificially increase, without merit, the ratings of their teachers' in-classroom practices (i.e., via observational rubrics), either covertly or overtly, and most often when administrators want to protect teachers who they deem as "effective" or good-to-great teachers, but whose growth/VAM scores evidence them as significantly less. In these cases, school administrators will often rate these teachers higher than they might rate other teachers of the same quality, simply to offset the typically lower growth/VAM scores.

As Campbell's Law would have it, this is much more likely when there are serious consequences at play, and school administrators aim to protect teachers from what they, again, view as a potential set of inappropriate consequences (e.g., termination after one or two years of poor ratings) to be attached to low composite (i.e., growth/VAM plus observational estimates) scores. Engaging in this practice, while perhaps humanitarian and justified as appropriate or rational, ultimately distorts the validity of the inferences to be drawn by the mere manipulation of one indicator to offset the other.

Artificial deflation

Artificial deflation occurs when school administrators decrease, again without merit, the ratings of their teachers' observational scores. This type of manipulation has been documented much less frequently than artificial inflation; however, it still occurs. Again, in the cases of artificial deflation, school administrators might deliberately rate teachers of equal caliber lower than their comparable peers, to deliberately (and oft-forcedly) guard against their own (sub)conscious and (too) often favorable biases and prejudices when "subjectively" observing and scoring their teachers in practice.

For example, in the now famous *Widget Report* (Weisberg, Sexton, Mulhern, & Keeling, 2009), researchers reported that only 1% of teachers were rated as "unsatisfactory," which they deemed as nonsensical given the US as a whole is still merely performing around average as compared to other comparable industrialized nations. Indeed, "subjective" school administrators were to blame; hence, this became one of the key policy reports that convinced federal policymakers to move forward with the Race to the Top (2011) competition to entirely reform states' "subjective" teacher evaluation systems.

Accordingly, school administrators have since been asked or forced to artificially deflate teachers' observational scores to essentially ensure that all scores, when taken together, fit a normal bell curve, which will illustrate to others that there is indeed a normal distribution of teachers as per their effectiveness, versus a skewed distribution demonstrating "too many" effective teachers.

While an entirely arbitrary venture, this is being perpetually encouraged to counter critics' aforementioned claims. For example, the National Institute for Excellence in Teaching (NIET) that sponsors and promotes state and district use of their TAP System for Teacher and Student Advancement, evidently encourages TAP evaluators to generously distribute average scores (i.e., 3 = at expectations) and to use high scores

"sparingly" (i.e., 5 = significantly above expectations). Teachers are to start at "a rock solid 3—4s are to be given out sparingly and [teachers] are not to or rarely receive a 5." Likewise, teachers being evaluated "should strive to score a 3 on the TAP rubric as scoring a 5 should be nearly impossible" (anonymous teachers, personal communications, 2016). This is to help guarantee that those who adopt (and pay for) the TAP system realize what TAP markets: that TAP users' observational scores will improve states' prior *Widget Effect* results and reduce state's prior *Widget Effect* tendencies (see Figure 1).

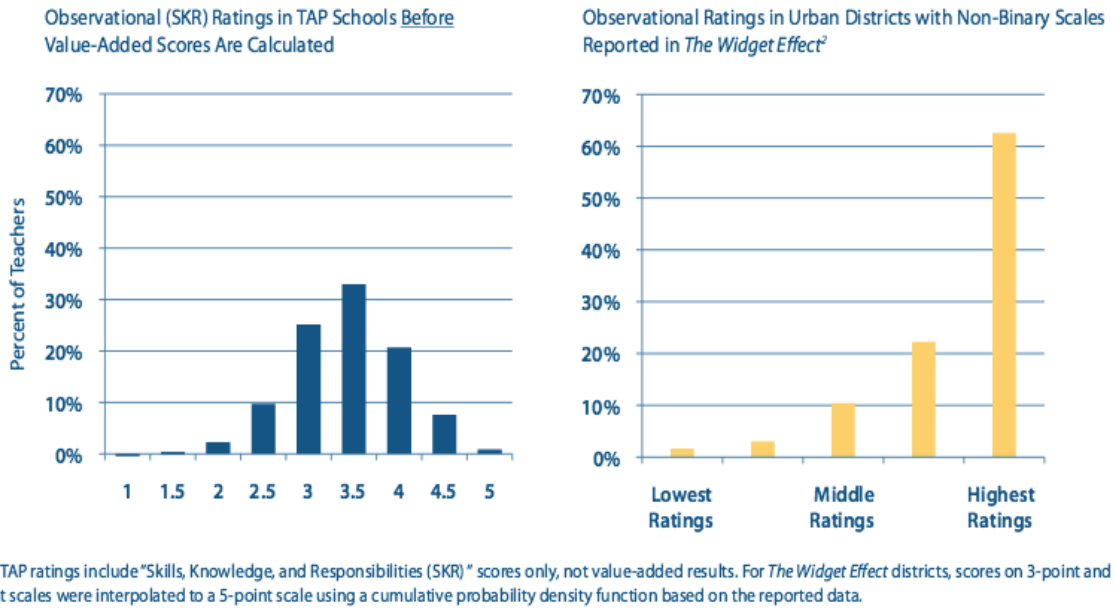

Figure 1. Observational ratings in TAP schools versus urban districts with traditional evaluation systems [as titled in the original, Jerald & Van Hook, 2011, p. 1].

In this case, it is not that the actual qualities of teachers have changed, likely whatsoever; rather, what has changed is the scale and the scale scores that are emphasized, which means nothing more than a scale-and-switch scheme of sorts, as a method of artificial deflation.

Related, Charlotte Danielson—architect of the Danielson's Framework for Teaching—was recently quoted as saying that "teachers should live in the 'effective' and only [occasionally] visit [the] 'highly effective" zones within her teacher observational system (Ramaswamy, 2014).

See also the recent claim made by the president of the National Council on Teacher Quality (NCTQ), that "If I were a superintendent and I didn't see a fairly good

distribution curve within my district [as per teachers' effectiveness ratings], I'd be suspicious about what was going on" (Amar, 2016).

Accordingly, it is oftentimes superintendents who on their own accord or are often (ill)advised by naive edu-philanthropists like this, who are forcing their school administrators to artificially suppress their observational ratings of teachers, again, to force such socially-Darwinian illusions of normality, via more symmetrically distributed teacher effectiveness curves.

This too, of course, has serious implications for the validity of the inferences to be drawn, upon which high-stakes decisions are to be made, in that what is "true" is being forcibly distorted by socially constructed definitions of what "truth" is supposed to look like.

Artificial conflation

Lastly, artificial conflation occurs when school administrators guarantee or ensure that teachers' growth/VAM estimates are adequately aligned or correlated with their observational scores and ratings. Reports of artificial conflation have been reported, more specifically, in Alabama, Georgia, and Tennessee, and most recently, Texas. In Tennessee, the state's Board of Education (2012) actually made it state policy that teachers' observational scores be forcibly aligned with their growth/VAM scores, regardless of what it took to reach the increased levels of alignment externally mandated and desired.

State leaders even provided guidelines to help school administrators check their own levels of "subjectivity," and consequently artificially manipulate teachers' observational scores (typically downwards) when the alignment between teachers' observational and growth/VAM scores fell outside of an (arbitrarily defined) "acceptable" range. Similarly, state level policies in both Alabama and Georgia assert that the multiple measures (i.e., observational and growth/VAM scores) used to evaluate teachers should also be positively correlated, with similar emphases on charging those with the authority to manipulate teachers' observational scores (i.e., school administrators) to match teachers' more "objective" growth/VAM counterparts.

In the Houston Independent School District (HISD), one of the nation's largest urban public school districts in the nation, many school principals have also reported that they were under significant pressure from district administrators to ensure that their teachers' observational and growth/VAM scores were also satisfactorily "aligned."

Further, these school principals reported actually manipulating teachers' observational scores to match their growth/VAM scores, to not be officially identified as "at risk for misalignment" and in need of intervention and improvement themselves as school administrators with supervisory/observational roles. Teachers also reported being aware that their school principals were doing this, noting also that they knew their principals were being forced to do so by, in this case, the district's superintendent (Collins, 2014; Paufler, under review; see also Amrein-Beardsley et al., 2016).

Apparently, it is around these more "objective" indicators that all other more "subjective" indicators are to revolve, although current research suggests that neither or these two indicators should be so privileged, or trusted (see, for example, American Statistical Association, 2014; American Educational Research Association, 2015). This also has

serious implications for the validity of the inferences to be drawn and used for decision-making purposes. But perhaps more importantly, doing this or engaging in and encouraging such behaviors negates the entire enterprise, as well as the entire purpose for doing and financing all of this in the first place.

Conclusions

What school administrators need to know is that they are unequivocally remiss if they believe artificially manipulating teachers' observational scores is a beneficial or warranted practice.

Worse would be if school administrators continue to engage in such practices, without fighting back (and often upwards in education's oft-hierarchial systems) in that this is, simply put, very bad educational measurement and professional practice.

While it might seem like an easy ace or safe play in the game to avoid being deemed as "too subjective," to dodge any sort of "misalignment" issues, or rather engage in a perceptibly necessary act to protect one's teachers, engaging in any of the three behaviors detailed prior can be incredibly dangerous as any of these practices ultimately distort the validity of both measures of teacher effectiveness, as well as the validity of the inferences to be drawn as based on both measures combined, in all cases and regardless of the degree, to levels that results and outcomes can no longer be trusted, used, or supported with evidence.

This, accordingly, must be stopped.

Author Biographies

Tray Geiger is a doctoral student in educational policy and evaluation in the Mary Lou Fulton Teachers College at Arizona State University. His research interests include teacher evaluation systems, education accountability policy, and Critical Race Theory. E-mail: tjgeiger@asu.edu

Audrey Amrein-Beardsley is currently a professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research interests include educational policy, educational measurement, research methods, and more specifically, high-stakes tests and value-added methodologies and systems. E-mail: audrey.beardsley@asu.edu

References

Amar, M. (2016, July 15). Denver Public Schools set to strip nearly 50 teachers of tenure protections after poor evaluations. *Chalkbeat Colorado*. Retrieved from http://www.chalkbeat.org/posts/co/2016/07/14/denver-public-schools-set-to-strip-nearly-50-teachers-of-tenure-protections-after-poor-evaluations/#.V5Yryq47Tof

American Educational Research Association (AERA). (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher, X(Y), 1-5. doi:10.3102/0013189X15618385 Retrieved from http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html

American Statistical Association (ASA). (2014). ASA statement on using value-added models for educational assessment. Alexandria, VA. Retrieved from http://VASboozled.com/wp-content/uploads/2014/03/ASA_VAS_Statement.pdf

Amrein-Beardsley, A., Collins, C., Holloway-Libell, J., & Paufler, N. A. (2016). Everything is bigger (and badder) in Texas: Houston's teacher value-added system. [Commentary]. *Teachers College Record.* Retrieved from http://www.tcrecord.org/Content.asp?ContentId=18983

Blake, H., & Templon, J. (2016, January 17). The tennis racket. *BuzzFeed News/BBC*. Retrieved from https://www.buzzfeed.com/heidiblake/the-tennis-racket?utm_term=.krym92EzO#.hmznB7x0z

Campbell, D. T. (1976). *Assessing the impact of planned social change*. Hanover, NH: The Public Affairs Center, Dartmouth College. Retrieved from http://portals.wi.wur.nl/files/docs/ppme/Assessing_impact_of_planned_social_change.pdf

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS Education Value-Added Assessment System (EVAAS®). *Education Policy Analysis Archives, 22*(98), 1-42. doi:http://dx.doi.org/10.14507/epaa.v22.1594 Retrieved from http://epaa.asu.edu/ojs/article/view/1594

Jerald, C. D., & Van Hook, K. (2011). More than measurement: The TAP system's lessons learned for designing better teacher evaluation systems. Santa Monica, CA: National Institute for Excellence in Teaching (NIET). Retrieved from http://files.eric.ed.gov/fulltext/ED533382.pdf

Nichols, S. L. & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge, MA: Harvard Education Press.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425. (2002). Retrieved from http://www.ed.gov/legislation/ESEA02/

Paufler, N. A. (under review). Principal reflections on the implementation of a teacher evaluation system. *Teachers College Record.*

Race to the Top (RttT) Act of 2011, S. 844--112th Congress. (2011). Retrieved from http://www.govtrack.us/congress/bills/112/s844

Ramaswamy, S. V. (2014). Teacher evaluations: Subjective data skew state results. *The Journal News*. Retrieved from http://www.lohud.com/story/news/education/2014/09/12/state-teacher-evals-skewed/15527297/

Tennessee State Board of Education (TSBE). (2012). Teacher and principal evaluation policy. Nashville, TN. Retrieved from https://www.tn.gov/assets/entities/sbe/attachments/7-27-12-II_C_Teacher_and_Principal_Evaluation_Revised.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The Widget Effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York, NY: The New Teacher Project (TNTP). Retrieved from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf