**Unpacking IMPACT’s Impact, or the Lack Thereof:**

**Fatal Flaws and Limitations with the National Bureau of Economic Research (NBER)’s IMPACT Study**

When Michelle Rhee sat as the DC Public Schools Chancellor, she implemented a controversial teacher-evaluation system called IMPACT to help guarantee that every child in DC would have access to a “highly effective” teacher. Accordingly, since 2009-2010 IMPACT has been used across DC’s public schools, more recently with federal Race to the Top (2011) funds in support, to help 6,500 DC teachers become more effective or be gone. IMPACT has thus far led to teacher bonuses (professedly up to $25,000 per year) as well as terminations of more than 500 teachers (≈ 8%) throughout DC’s Public Schools.

Financial award and termination decisions have been determined by rating DC teachers as per their “effectiveness ratings” as based on their “teacher performance.” Teacher performance ratings are calculated using, most importantly in terms of weight, indicators of student learning and achievement as measured by teachers’ individual value-added scores on DC’s large-scale standardized tests, *or* “on other assessments if [teachers do not] teach a DC [tested] grade or subject” (DC Public Schools, n.d.). The performance effectiveness ratings also include other “teacher performance” indicators (based on classroom observations,[[1]](#footnote-1) teachers’ support for school initiatives,[[2]](#footnote-2) *other* teacher-assessed student achievement data,[[3]](#footnote-3) additional indicators of teacher professionalism,[[4]](#footnote-4) and school-level value-added [the least valued system indicator weighted among the indicators at 5%]), although the priority area is academic achievement. As per the IMPACT system “a teacher’s most important responsibility is to ensure that her or his students learn and grow” (DC Public Schools, n.d.).

Whether this sweeping reform has had the impact perspicuous in its name, however, has been the source of contention and dispute, until now. A recent study led by researchers at the University of Virginia and Stanford University (Dee & Wyckoff, 2013) put forth “evidence” that IMPACT has indeed had the impact desired: that the incentives (i.e., financial bonuses and base salary increases) and discincentives (i.e., actual or threats of termination) tied to teachers’ effectiveness ratings impacted both teacher retention and “teacher performance.” However, there are fatal flaws in this study that *had this study undergone peer review before it was released to the public and before it was hailed in the media*, would not have created the media hype that ensued.

Shame (again) on the National Bureau of Economic Research (NBER) for releasing to the media yet another “working paper” that was not peer-reviewed. Nor was it “subject to the review by the NBER Board of Directors that accompanies official NBER publications” (Dee & Wyckoff, 2013). This happened in January 2012 as well, when NBER featured another “working paper,” this time conducted by three Ivy League economists claiming that fourth graders taught by high value-added teachers would go on to lead uber-healthy and wealthy lives years later (Chetty, Friedman, & Rockoff, 2012a). Afterwards, researchers who represented the types of peer reviewers this paper demanded prior to its release (see, for example, Adler, 2013; Baker, 2012b; Ravitch, 2012; see also Winerip, 2012) found in it fatal flaws (likely) preventing the paper, to date, from being published in any reputable, scholarly, educational or economics-based outlet. Besides a condensed version of the paper published in the conservative, non-peer reviewed *Education Next* (Chetty, Friedman, & Rockoff, 2012b)this paper has yet to be judiciously endorsed by the authors’ academic peers.

Something that is becoming all too familiar has occurred again here. Had NBER held off on releasing this IMPACT “working paper” until the study’s economists submitted their working report to a scientific journal for peer review, the public would have been spared yet another misleading pitch for staying the course with value-added and similar Race to the Top-type initiatives. NBER and the journalists covering NBER’s “working papers” (see, for example, Brown, 2013; News Desk, 2013; Toppo, 2013b) must learn to hold their applause until after the scientific community does its job.

Here are some of the fatal flaws that putting this “working paper” through the peer review process would have certainly uncovered. This would have also likely prevented both the thrills and objections that followed this study’s “official” release.

**Fatal Flaws**

**Bias**

Both of the study’s authors have advanced degrees in economics, and they conduct educational reform studies focusing on the uses of statistical and econometric methods to inform policies targeted at competition and their related (dis)incentives. Had this study undergone blind peer review, whereby the authors’ identities would not have been disclosed, their research backgrounds would not have been as evident. However, their inclined tendencies made transparent throughout this piece would have otherwise caused pause. In the first paragraph alone, the authors set forth a series of assumptions that would have permitted readers to correctly predict the study’s findings without reading any further, much less through to the end of this manuscript. This would not have been permitted had the scientific community been involved.

For example, in the first sentence of the “working paper," the authors note that, “research consensus has coalesced around the notion that teacher quality is a critically important determinant of student development and achievement” (Dee & Wyckoff, 2013, p. 1). They do state this with research citations in support, but they do not disentangle the research-based truth that, rather, the teacher effect (i.e., that accounts for 10-20% of the variance in test scores) is a strong school-level factor. But the teacher effect is not a strong enough factor to supersede all else that occurs outside of schools that more substantially impacts student development and achievement (i.e., that accounts for 80-90% of the variance in test scores; Berliner, 2012; Good, 2014; Mathis, 2012; Schochet & Chiang, 2010, 2013).

The authors note that “decades of empirical research have provided relatively little evidence on observed teacher traits that can consistently predict teacher quality,” when just last month Mathematica Policy Research released yet another non-peer-reviewed study that after critically reviewed yielded evidence that, despite their best value-added intentions, National Board Certified Teachers (NBCTs) and teachers with about four more years of teaching experience (i.e., two “observable teacher traits”) had greater effects on student achievement in high-needs schools than the comparison teachers to whom they were compared (Glazerman, Protik, Teh, Bruch, & Max, 2013; see also Amrein-Beardsley, 2013).

The authors also note that professional development has no evidence that warrants supporting expenditures in this area mainly because “professional development is [not] driven by rigorous assessment” (Dee & Wyckoff, 2013, p. 1).[[5]](#footnote-5) Yet they make similar flippant statements when pointing fingers at “rigid single-salary schedules” (p. 5) and teacher unions (p. 6) at the alleged roots of many educational problems, but they make such claims without any “rigorous” research support themselves.

Throughout, and all the while as well, they uphold DC’s IMPACT system, both implicitly and explicitly as the most “seminal” (i.e., original, groundbreaking, influential, inspirational) system in the country (Dee as cited in Toppo, 2013b). Such instances of bias would certainly have prevented publication of this piece as peer reviewers would have unquestionably taken issue with their underlying enthusiasm for IMPACT, and ultimately questioned to what extent research bias slanted the findings they uncritically advanced.

**Real-World Incentives**

Related, it is implicitly celebrated throughout the text of this study, that exceptionally large financial rewards were made available and this also caused the claimed effects “at-scale” (Dee & Wyckoff, 2013, p. 3 & 8) using “real-world” monies (p. 3 & 47) that can and are still being sustained over time. The authors add additional conjectures (without empirical support) that this might be why the DC IMPACT program is allegedly working. In reality, while “these one-time bonuses *could* [emphasis added] amount to as much as $25,000” (p. 10), bonuses of this size were made available only to teachers who taught in a poor school (i.e., where the percentage of free and reduced price lunch eligible students was at least 60%), to teachers who had individual value-added scores (i.e., 17% of the teachers in DC – see more about this forthcoming), and to teachers who taught a high-need subject area (i.e., special education, bilingual education, English as a Second Language [ESL], or secondary mathematics and science) which rarely (if ever) included teachers for whom individual value-added scores were available.

In other words, while the possibility to yield the prized $25,000 bonuses may have existed (as revered by the authors and the media; see, for example, Brown, 2013; News Desk, 2013, Toppo, 2013b), it existed only in the oddest of circumstances. The $25,000 bonuses were likely rare if ever present, as the qualifying conditions for such awards were almost entirely mutually exclusive (e.g., only high school teachers of mathematics would have had individual value-added scores). It seems the authors advanced erroneous findings that large bonuses increased output, when in fact a very small fraction of DC teachers were eligible for the bonuses revered. In addition, no actual numbers of awardees or other related descriptives (e.g., how many teachers received different awards) were provided.[[6]](#footnote-6)

**Fairness**

Related (again, and as mentioned), what was not advertised or stated as a major limitation was that only 17% of the teachers included in this study (i.e., teachers of reading and mathematics in grades 4 through 8) were actually evaluated under the IMPACT system for that which they contributed to the system’s most valued indicator: student achievement. Rather, 83% of the teachers *did* *not have* student test scores available to determine if they were indeed effective using individual value-added scores.[[7]](#footnote-7) This figure is significantly larger than current figures estimating that 60-70% of all public school teachers cannot be held directly accountable using value-added scores for the lack of having them (Gabriel & Lester, 2013; Harris, 2011; see also Thomas, 2012).

Instead, these teachers were evaluated on their “performance” using almost exclusively (except for the 5% school-level value-added indicator) the same subjective measures integral to many traditional evaluation systems as well as student achievement/growth on *teacher-developed* and *administrator-approved* classroom-based tests.[[8]](#footnote-8) Again, this is particularly interesting as a teacher’s “value-added” was to be the key indicator of “teacher performance” as part of the IMPACT system (see also DC Public Schools, n.d.). Yet researchers in this study evaluated the impact of a teacher evaluation system that for 83% of DC’s teachers was entirely subjective and, for the lack of a better term, “old school.” For 83% of the DC IMPACT teachers, their overall weighted scores captured nothing different from that which schools have been doing for decades past in terms of their teacher evaluation systems – the traditional systems being accosted and used as rationales for why new and improved systems are needed (Weisberg, Sexton, Mulhern, & Keeling, 2009; see also Brill, 2009; Gates, 2012). Yet this is the system that, as asserted by the authors of this study, caused astounding effects in DC.

It is important to note here as well that throughout this study no mention is made about the 17% of teachers who actually had “hard data,” that is, in terms of the growth their students demonstrated over time on standardized test scores.[[9]](#footnote-9) This is also a fatal flaw, as this group should have been analyzed as the unique and distinctly separate group, as erroneously assumed to be (by those in the media and elsewhere) the group of teachers who yielded many of the effects claimed.

**Classification Errors**

Also troubling, particularly for the 17% of teachers with individual value-added data, is that awards were made based on two consecutive years of “growth” while most if not all value-added researchers agree that at least three-years of data should be used to measure growth, given these measures are *highly* unstable and *do not currently* warrant enough good decisions, even given three years of data (Glazerman & Potamites, 2011; Goldschmidt, Choi, & Beaudoin, 2012; Harris, 2011; Ishii & Rivkin, 2009; Sanders as cited in Gabriel & Lester, 2013). The most current estimates indicate that even with three years of data, 25% to 50% of the teachers with individual value-added data are still at risk of being misclassified (Au, 2011; Schochet & Chiang, 2010, 2013; Shaw & Bovaird, 2011).

Of concern here, mainly, is the prevalence of false positive or false discovery errors (i.e., Type I errors), whereas an ineffective teacher is falsely identified as effective, or (i.e., Type II errors), whereas an ineffective teacher might go unnoticed instead. While the district’s decision to use fewer than the recommended three-years of data was probably out of the control or purview of this study’s authors, that Type I and II errors certainly distorted study findings, yet they were not discussed either.[[10]](#footnote-10) They should have been discussed, however, given such errors likely impacted why teachers “rated under the IMPACT system remained employed by DCPS through the next academic year or left for whatever reason” (Dee & Wyckoff, 2013, p. 12). “Whatever reason” likely included teachers who should not have stayed but did due to Type I errors and teachers who should not have left but did due to Type II errors, especially around the thresholds researchers used (see more about this forthcoming), but this was not accounted for or discussed as a consideration, much less a limitation,[[11]](#footnote-11) although this certainly impacted the authors’ findings about teacher retention and “teacher performance” conditional on retention.

**Multiple Measures**

One thing the authors did correctly was applaud DC’s efforts given their system’s use of a more holistic approach to evaluating teacher effectiveness using multiple measures (and multiple observers for the observational component of the IMPACT system). However, the authors falsely (and solely) credited the final Bill & Melinda Gates Foundation (2013) study for the idea and for empirically advancing such an approach. Doing this, instead, is in line with the current *Standards for Educational and Psychological Testing* that (while currently being revised) were collaboratively written more than ten years ago by the finest academics in educational measurement: members of the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (2000; see also Baker et al., 2010; Capitol Hill Briefing, 2011).

Likewise, contrary to what the authors advance, there is not “a growing consensus [that] underscores the importance of [such] a balanced approach” (Dee & Wyckoff, 2013, p. 5), mainly because these “emerging best-practice design principles” (p. 5) are, in reality, greatly restrained *if* systems based on multiple measures *include within them value-added estimates*. While, again, this system was almost entirely subjective for 83% of the teachers being evaluated, for whom teacher-level value-added estimates were not available,other such systems continue to present major issues with reliability and validity.

Validity is an essential of any measurement, and reliability (e.g., the classification errors mentioned) is a necessary or qualifying condition for validity (Brennan, 2006, 2013; Kane, 2006, 2013; Messick, 1975, 1980, 1989, 1995). Put differently, without consistency one cannot achieve validity; hence, such systems are limited in their validity by the weakest of the measure’s reliability, or lack thereof (Baker, 2012b; Baker, Oluwole, & Green 2013; Papay, 2010; Rothstein & Mathis, 2013). Again, this problem pertains here to 17% of the teachers teaching in DC’s public schools, but reliability and validity also impact all systems based on multiple measures, regardless of the measures used. As such, these statistics should have also been disclosed and discussed, and this should have been done well before authors advanced additional unsubstantiated statements noting, for example, that multiple measures “have been shown to be more reliable than single measures” with one reference (i.e., Bill & Melinda Gates Foundation, 2013) in support. Peer reviewers would not have permitted authors to leave such complex issues like this just at that.

**Cheating**

The authors also dismiss the 2011 cheating scandal in DC (Gillim & Bello, 2011) and the cheating allegations that still surround DC’s public schools (Strauss, 2013; Toppo, 2013a), noting with confidence (without hard data) that whatever data distortions that came about due to “cheating” did not significantly impact their analyses. First, they casually remind readers that “these test-based measures of teacher performance were only relevant for [DC] teachers [with individual value-added scores] under IMPACT and these teachers constitute[d] less than 20 percent [i.e., 17%] of the analytical samples” (Dee & Wyckoff, 2013, p. 14). While their forthrightness is appreciated, this reminds readers about what this study actually did (and did not do) given the limited sample for which “teacher performance” could be objectively measured.

For the other 17%, they claimed the cheating instances were immaterial (see, for example, p. 14). Authors claimed (without any data) that cheating instances declined once the district hired external contractors to investigate cheating instances (e.g., Caveon Test Security). While it is important to note that the many ways by which cheating occurs cannot be identified, for example, via analyses of erasures on Scantron sheets – the types of cheating investigations such companies typically conduct (U.S. Department of Education, 2013) – the authors further argued that the cheating scandal “pre-date[d] the introduction of IMPACT,” and hence predated their analyses. Again, however, it seems instances of cheating are still alive and well in DC’s public schools (Strauss, 2013; Toppo, 2013a).

**Manipulation and Artificial Inflation**

As mentioned, a major study limitation was that the indicators used to define and observe changes in “teacher performance” (for the 83% of DC teachers) were based almost entirely on highly subjective, highly manipulable, and highly volatile indicators of “teacher performance” – a term used throughout the study that implies a (false) sense of certainty and objectivity as “teacher performance” is presumably based on teachers’ contributions to student achievement. Here, however, given the socially constructed definitions and numerical indicators capturing “teacher performance,” score bias by manipulation and artificial inflation were also present if teachers were able to influence their ratings across the evaluative indicators used (Haladyna, Nolen, & Haas, 1991; Nichols & Berliner, 2007; Rothstein, 2011). Although the threat of this was also banally dismissed by the authors (Dee & Wyckoff, 2013, p. 16), and also dismissed as “theoretically [not really] reasonable” (p. 24), evidence authors presented elsewhere indicates that the authors should not have been so indifferent.

In 2009-2010 “some teachers (i.e., 85 out of the 436) [19.5%] were able to appeal successfully their initial IMPACT rating as an ME [Minimally Effective] teacher in IMPACT’s first year…[hence]…the resulting causal estimands may only be defined for teachers who “complied” with their initial assignment[s or classifications]” (p. 18). While this number decreased in 2010-2011, the authors claim that this purported decline occurred because, again without evidence or support, the district was less “flexible” the year following (p. 19). No other potential causes (e.g., score inflation to *realize the positive* or *avoid the negative* consequences attached to teachers’ IMPACT scores, to avoid further appeals, or to sidestep other unnecessary compliance work) were given or explicitly considered as reasons for such positive changes in “teacher performance” scores over time.

Telling as well was that “no teacher in the HE [Highly Effective] analytical sample changed the IMPACT rating implied by an initial score” (p. 19). Teachers who were satisfied with their scores (or perhaps considered themselves safe) had less of an incentive to overtly protest or covertly skew particular indicators within or their overall scores. Likewise, “retention during this period was noticeably higher among the higher-performing teachers” (p. 20). No surprise here either. This provides additional evidence of score inflation due to manipulation, as score deflation would not have taken “teacher performance” scores in the desired directions (i.e. up). While the authors claimed that, “the opportunity to earn performance-based financial rewards increased retention (p. 20)” as well, another plausible reason for this over time was that these teachers stayed because they were not under direct and imminent threats (i.e., deflation was less likely than its inverse).

When using tests, and especially subjective indicators to measure “teacher performance,” one must exercise caution to ensure that those being measured do not engage in manipulation and inflation techniques known to effectively increase the scores derived and valued, particularly within such high-stakes accountability systems (Haladyna et al., 1991; Nichols & Berliner, 2007; Rothstein, 2011). Again, because for 83% of the teachers their performance indicators were manipulable, evidence of artificial inflation was also evident here, albeit dismissed, again.

**Jumping Performance Thresholds**

Related, the authors claim that in IMPACT’s first year, “the dismissal threat implied by an ME [Minimally Effective] rating did not induce detectable changes in teacher performance” (Dee & Wyckoff, 2013, p. 20). One year later, however, the data suggested “a sizable jump” in “teacher performance” (i.e., in excess of 10 points) amid teachers whose initial scores “*placed them* [emphasis added to underscore the implied objectivity]” (p. 20) as Minimally Effective (ME). While the authors assert that this evidences that previously low-performing teachers undertook steps to “*meaningfully improve* [emphasis added] their performance” (p. 20), no mention was made here, again, about the real possibility that scores were artificially inflated at the same time.

Only is it in the final section of the study that the authors wrote the following:

“principals [were] likely to have been aware when one of their teachers faced a dismissal threat due to a prior ME [Minimally Effective] rating or the possibility of a base-salary gains [sic] due to a prior HE [Highly Effective] rating. This awareness may have positively influenced [i.e., artificially inflated] how teachers were rated.” (p. 25)

This “positive influence” or inflation is likely, again, because for 83% of the teachers these subjective indicators were the only indicators (besides school-level value-added at 5%) on which they were measured and held accountable. As Campbell’s Law (1976) demonstrates, instruments over which people have complete or partial control can be easily distorted, and the probability of distortion increases the higher the stakes attached to output (see also Haladyna et al., 1991; Nichols & Berliner, 2007; Rothstein, 2011). In this study a certain yet unknown number of indicators or overall scores were certainly, albeit to varying degrees, inflated. This likely occurred in the positive direction to avoid the negative or reap the positive consequences attached to teachers’ IMPACT scores, and this was likely done by the teachers themselves (having so much control over their own data) or the administrators in charge (having so much control over teachers’ outcome data).

Interestingly, “teacher performance” also increased for highly effective (HE) teachers as well. While the authors claim this occurred because “teachers [had] a powerful financial incentive to continue to perform well” (Dee & Wyckoff, 2013, p. 21), this also supports the general positive trend indicating that, because the scores were based on subjective indicators subject to manipulation, they had nowhere to go but up.[[12]](#footnote-12) This also likely explains some of the notable (albeit notably artificial) effect sizes[[13]](#footnote-13) the authors also reported (e.g., an effect size of 0.27 [p. 23] which is significantly larger than the average effect size [0.16] found in educational research; (Borman, Hewes, Overman, & Brown, 2003). Such effect sizes are much easier to obtain if scores are susceptible.

Finally, a final fatal flaw that likely would have also prevented this piece from being published was that the authors examined (keeping in mind the aforementioned issues) only one-year effects. Claimed effects, although not always statistically significant (see, for example, Dee & Wyckoff, 2013, p. 22), were observed mainly in the second year, “when the policy was *more clearly credible* [emphasis added]” (p. 22). Whether the claimed effects disappeared one year later is unknown, as is whether any of the claimed (although likely erroneous) effects were sustained over a period of time longer than one year.

**Conclusions**

Of utmost concern with this piece was that the overarching conclusions advanced by the authors and perpetuated by the media (using direct quotes from the authors and others [e.g., Michelle Rhee[[14]](#footnote-14)]) were faulty, unfounded, and for the most part untrue. The improvements in “teacher performance” observed did not indicate that IMPACT was/is having its intended and, as inter-textually positioned, amazing effects. The positive “high-powered incentives” (Dee & Wyckoff, 2013, p. 27) and negative disincentives did not “substantially improve the *measured performance* [emphasis added to underscore the implied objectivity] of the teaching workforce” (p. 28), or more specifically improve the *performance* of high-performing teachers and previously low-performing teachers who remained teaching in DC’s public schools.

Did teachers "either [leave] or [get] better," as per the direct quote of one of the study’s authors – Stanford’s Thomas Dee (Toppo, 2013b)? Was it, and if so, why was it that low-performing teachers left the system? And did all of this improve the skills of those who stuck around? The answers to all of these questions, that were advanced as findings, are emphatically no. Only Brown (2013), however, was onto this, writing from a more acute, critical stance that, “the study [was] silent about whether the incentives [had] translated into improved student achievement.” This is exactly right, and this is exactly why this study, very likely, would *not* have been published in a scientific, academic, peer-reviewed journal. This is likely why we may not ever see it in a scholarly outlet as well.

The larger lesson here, perhaps, is about the ways in which these faulty findings were used to further promote DC’s IMPACT system by flippantly blasting and then broadcasting them without adherence to scientific norms and professional standards. Unfortunately, in this case and one other similar case mentioned prior (Chetty et al., 2012a), doing this seems to have had quite an impact regardless. Thanks go out again to NBER for releasing yet another “working paper” for the media to uncritically consume and transmit to an even less informed and more highly assuming public. It would have been more scientifically fair and accurate had the headlines read something about researchers finding weak, preliminary, error-prone, and subjective evidence that in one year post-IMPACT, some things might have happened while many things did not. But nobody really has a clue whether findings are real and whether they might sustain themselves over time. Further, this study has really nothing to do with “teacher performance” as presumed and inter-textually positioned as objectively captured throughout the study. Rather, student achievement and teachers’ purported impacts on growth in achievement over time was nearly (for 83% of the population) impossible to capture.

**References**

Adler, M. (2013). Findings vs. interpretation in “The Long-Term Impacts of Teachers” by Chetty et al. *Education Policy Analysis Archives, 21*(10).Retrieved from http://epaa.asu.edu/ojs/article/view/1264.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2000). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Amrein-Beardsley, A. (2013, November 7). Mathematica study: Value-added and teacher experience. *VAMboozled! A blog by Audrey Amrein-Beardsley.* Retrieved from http://vamboozled.com/?p=351

Au, W. (2011, Winter). Neither fair nor accurate: Research-based reasons why high-stakes tests should not be used to evaluate teachers. *Rethinking Schools*. Retrieved from http://www.rethinkingschools.org/archive/25\_02/25\_02\_au.shtml

Baker, B. D. (2012a, January 17). Fire first, ask questions later? Comments on recent teacher effectiveness studies. *School Finance 101.* Retrieved from http://schoolfinance101.wordpress.com/2012/01/07/fire-first-ask-questions-later-comments-on-recent-teacher-effectiveness-studies/

Baker, B. D. (2012b, March 31). Firing teachers based on bad (VAM) versus wrong (SGP) measures of effectiveness: Legal note. *School Finance 101.* Retrieved from http://schoolfinance101.wordpress.com/2012/03/31/firing-teachers-based-on-bad-vam-versus-wrong-sgp-measures-of-effectiveness-legal-note

Baker, B. D., Oluwole, J. O., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Education Policy Analysis Archives, 21*(5), 1-71*.* Retrieved from http://epaa.asu.edu/ojs/article/view/1298

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, D.C.: Economic Policy Institute. Retrieved from http://www.epi.org/publications/entry/bp278

Berliner, D. C. (2012, October 16). Effects of inequality and poverty vs. teachers and schooling on America’s youth. *Teachers College Record, 116*(1). Retrieved from http://www.tcrecord.org/Content.asp?ContentID=16889

Bill & Melinda Gates Foundation. (2013, January 8). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project’s three-year study.* Seattle, WA. Retrieved from http://www.gatesfoundation.org/press-releases/Pages/MET-Announcment.aspx

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research,* 73(2), 125-230. doi:10.3102/00346543073002125

Brennan, R. L. (2006) *Perspectives on the evolution and future of educational measurement*. In R. L. Brennan (Ed.) 2006. Educational Measurement (4th ed.), pp. 1-16. Westport, CT: American Council on Education/Praeger.

Brennan, R. L. (2013).  Commentary on "Validating interpretations and uses of test scores." *Journal of Educational Measurement, 50*(1), 74-83. doi: 10.1111/jedm.12001

Brill, S. (2009, August 31). The rubber room. *The New Yorker.* Retrieved from http://www.newyorker.com/reporting/2009/08/31/090831fa\_fact\_brill

Brown, E. (2013, October 16). Study: D.C.’s teacher evaluation system affects workforce. *Washington Post.* Retrieved from http://www.washingtonpost.com/local/education/study-dcs-teacher-evaluation-system-affects-workforce/2013/10/16/c850b9ce-3684-11e3-80c6-7e6dd8d22d8f\_story.html

Campbell, D. T. (1976). *Assessing the impact of planned social change.* Hanover, NH: The Public Affairs Center, Dartmouth College. Retrieved from https://www.globalhivmeinfo.org/CapacityBuilding/Occasional%20Papers/08%20Assessing%20the%20Impact%20of%20Planned%20Social%20Change.pdf

Capitol Hill Briefing. (2011, September 14). *Getting teacher evaluation right: A challenge for policy makers.* A briefing by E. Haertel, J. Rothstein, A. Amrein-Beardsley, and L. Darling-Hammond. Washington, D.C.: Dirksen Senate Office Building (research in brief). Retrieved from http://www.aera.net/Default.aspx?id=12856

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2012a, January 26). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood.* National Bureau of Economic Research (NBER). Working Paper 17699. Retrieved from http://www.nber.org/papers/w17699

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2012b, Summer). Great teaching. *Education Next, 12*(3). Retrieved from http://educationnext.org/great-teaching/#

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Dee, T., & Wyckoff, J. (2013, October). *Incentives, selection, and teacher performance: Evidence from IMPACT.* National Bureau of Economic Research (NBER). Working Paper 19529. Retrieved from http://www.nber.org/papers/w19529

District of Columbia (DC) Public Schools. (n.d.). *An overview of IMPACT.* Retrieved from http://www.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+%28Performance+Assessment%29/An+Overview+of+IMPACT

Gabriel, R. & Lester, J. N. (2013). Sentinels guarding the grail: Value-added measurement and the quest for education reform. *Education Policy Analysis Archives, 21*(9), 1-30. Retrieved from http://epaa.asu.edu/ojs/article/view/1165

Gates, B. (2012, February 22). Shame is not the solution. *New York Times – Opinion Pages*. Retrieved from http://www.nytimes.com/2012/02/23/opinion/for-teachers-shame-is-no-solution.html?\_r=3&smid=FB-nytimes&WT.mc\_id=OP-E-FB-SM-LIN-SIN-022312-NYT-NA&WT.mc\_ev=click

Gillim, J., & Bello, M. (2011, March 30). When standardized test scores soared in D.C., were the gains real? *USA Today.* Retrieved from http://usatoday30.usatoday.com/news/education/2011-03-28-1Aschooltesting28\_CV\_N.htm

Glazerman, S. M., & Potamites, L. (2011, December). *False performance gains: A critique of successive cohort indicators.* Mathematica Policy Research. Retrieved from www.mathematica-mpr.com/publications/pdfs/.../False\_Perf.pdf

Glazerman, S., Protik, A., Teh, B., Bruch, J., & Max, J. (2013, November). *Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment.* Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/publications/pdfs/education/tti\_high\_perform\_teachers.pdf

Goldschmidt, P., Choi, K., & Beaudoin, J. B. (2012, February). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Technical Issues in Large-Scale Assessment State Collaborative on Assessment and Student Standards. Council of Chief State School Officers.Retrieved from http://www.ccsso.org/Resources/Publications/Growth\_Model\_Comparison\_Study\_Practical\_Implications\_of\_Alternative\_Models\_for\_Evaluating\_School\_Performance\_.html

Good, T. L. (2014). What do we know about how teachers influence student performance on standardized tests: And why do we know so little about other student outcomes? *Teachers College Record,* 116(1). Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17289

Haladyna, T. M., Nolen, N. S., & Haas, S. B. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher*, *20*(5), 2-7. doi:10.2307/1176395

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value added. *Education Finance and Policy, 4,* 520-536. doi:10.1162/edfp.2009.4.4.520

Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17-64). Washington, D.C.: The National Council on Measurement in Education & the American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:10.1111/jedm.12000

Mathis, W. (2012, September). *Research-based options for education policy making: Teacher evaluation*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/options-teacher-evaluations

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-66.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), Educational Measurement, 3rd ed. (pp. 13-103.) New York: American Council on Education and Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

News Desk. (2013, October 17). Study concludes Rhee's 'impact' improved D.C. teachers. PBS Newshour. Retrieved from http://www.pbs.org/newshour/rundown/2013/10/study-concludes-rhees-impact-improved-dc-teachers.html

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

Papay, J. P. (2010). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal.* doi: 10.3102/0002831210362589

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

Ravitch, D. (2012, January 17). What value did the Chetty study add? *Education Week*. Retrieved from http://blogs.edweek.org/edweek/Bridging-Differences/2012/01/dear\_deborah\_just\_days\_ago.html

Ravitch, D. (2013, November 14). Fred Smith: The errors of state testing in New York. *Diane Ravitch’s Blog.* Retrieved from http://dianeravitch.net/2013/11/14/fred-smith-the-errors-of-state-testing-in-new-york/

Rothstein, J. (2011, January). *Review of learning about teaching*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/thinktank/review-learning-about-teaching

Rothstein, J., & Mathis, W. J. (2013, January). *Review of two culminating reports from the MET Project*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

Schochet, P. Z. & Chiang, H. S. (2010, July). *Error rates in measuring teacher and school performance based on student test score gains*. U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20104004/

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*, 142-171. doi:10.3102/1076998611432174

Shaw, L. H. & Bovaird, J. A. (2011, April). *The impact of latent variable outcomes on value-added models of intervention efficacy*. Paper presented at the Annual Conference of the American Educational Research Association (AERA), New Orleans, LA.

Strauss, V. (2013, April 12). Why not subpoena everyone in D.C. cheating scandal — Rhee included? *The Washington Post*. Retrieved from http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/04/12/why-not-subpoena-everyone-in-d-c-cheating-scandal-rhee-included/

Thomas, J. R. (2012, January 25). Coming soon: Teacher report cards based on student performance. *The CT Mirror*. Retrieved from http://ctmirror.org/story/15172/coming-soon-teacher-report-cards-based-student-performance

Toppo, G. (2013a, April 11). Memo warns of rampant cheating in D.C. public schools. *USA Today.* Retrieved from http://www.usatoday.com/story/news/nation/2013/04/11/memo-washington-dc-schools-cheating/2074473/

Toppo, G. (2013b, October 17). Program in D.C. removed bad teachers: Researchers found that the program also improved the skills of the teachers who stayed. *USA Today.* Retrieved from http://www.usatoday.com/story/news/nation/2013/10/17/impact-washington-dc-shools/2993233/

U.S. Department of Education. (2013). Testing integrity symposium: Issues and recommendations for best practice. Washington DC: Institute of Education Sciences and the National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2013/2013454.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). “The Widget Effect.” *Education Digest, 75*(2), 31–35.

Winerip, M. (2012, January 15). Study on teacher value uses data from before teach-to-test era. *The New York Times.* Retrieved from http://www.nytimes.com/2012/01/16/education/study-on-teacher-value-uses-data-from-before-teach-to-test-era.html?\_r=0

1. The central component of the IMPACT systems for most teachers (i.e., 83%) is based on “rigorously scored classroom observations tied to the district’s Teaching and Learning Framework (TLF). The TLF specifies the criteria by which DCPS defines effective instruction and structures [itself around] a scoring rubric. The TLF includes multiple domains such as leading well-organized, objective-driven lessons, checking for student understanding, explaining content clearly, and maximizing instructional time” (Dee & Wyckoff, 2013, p. 8-9). [↑](#footnote-ref-1)
2. Via the Commitment to the School Community (CSC) measure, teachers are “assessed by their administrators on a rubric that measures their support of school initiatives, efforts to promote high expectations, and partnerships with students’ families and school colleagues” (Dee & Wyckoff, 2013, p. 9). [↑](#footnote-ref-2)
3. For teachers who do not have an individual value-added score, they “instead receive a Teacher-Assessed Student-Achievement (TAS) score. At the beginning of each academic year, teachers choose (and administrators approve) learning goals based on non-CAS assessments. At the end of the year, administrators rate the teacher’s success in meeting these goals using a rubric that emphasizes student learning or content mastery” (Dee & Wyckoff, 2013, p. 9). [↑](#footnote-ref-3)
4. The Core Professionalism (CP) indicator is used to rate teachers on the basis of their attendance, punctuality, adherence to policies and procedures, and their levels of respect. “Teachers are assumed to be professionals, and, therefore, CP scores can only reduce a teacher’s overall IMPACT score” (Dee & Wyckoff, 2013, p. 9-10). [↑](#footnote-ref-4)
5. They note elsewhere, however, that “teacher professional development can be effective, though there are far too few to discern patterns in the characteristics of successful programs (Yoon et al. 2007)” (Dee & Wyckoff, 2013, p. 28). [↑](#footnote-ref-5)
6. Similar claims were made about the salary base pay increases for teachers rated as “highly effective” two years in a row who could have realized monies “worth up to $27,000 *per year* [italics in the original]*”* (p. 11). [↑](#footnote-ref-6)
7. All teachers regardless of whether they had individual value-added scores were held accountable for school-level value-added that counted for 5% of their overall effectiveness ratings. [↑](#footnote-ref-7)
8. Teacher effectiveness ratings were based on, in order of importance by the proportion of weight assigned to each indicator: (1) the notoriously subjective scores derived via the district-created and purportedly “rigorous” (Dee & Wyckoff, 2013, p. 5) yet invalid (i.e., not having been validated) observational instrument with which teachers are observed five times per year by different folks, but about which no psychometric data were made available (e.g., Kappa statistics to test for inter-rater consistencies among scores); (2) teacher-created tests that were administrator-approved and for which teachers’ self-reported outcomes were evaluated by administrators given whether teachers reported meeting their predetermined and pre-approved goals and objectives; (3) what teachers had done to contribute to their school communities; and (4) what teachers purportedly contributed to their school’s value-added (the least weighted indicator at 5%). [↑](#footnote-ref-8)
9. At only one point is the group of teachers with individual value-added scores mentioned as distinctly different (i.e., in Table 5 referenced in Dee & Wyckoff, 2013, p. 25), yet no numbers of teachers included in this data display are included, which is yet another major flaw. [↑](#footnote-ref-9)
10. In the study’s limitation section at the end of the paper, the authors write that, “[a]ny teacher-evaluation system will make some number of objectionable errors in how teachers are rated and in the corresponding consequences they face” (Dee & Wyckoff, 2013, p. 29), but that is all that they write in terms of the errors prevalent and relevant in the research literature. [↑](#footnote-ref-10)
11. Readers should attempt to read through the sections in which the authors discuss who they included and excluded to examine DC IMPACT’s effects (Dee & Wyckoff, 2013, p. 11-14). Readers might come to their own conclusions, then, regarding whether this study warranted publication given all of the idiosyncratic decisions that were made to include/exclude different types of teachers in these analyses accordingly. This further limited the practicality, as well as the comprehensibility of this study, especially considering that in the end they found “that the mean teacher retention rate [was] somewhat lower in the ‘minimally effective’ [sample] (i.e., 84 percent) than in the ‘highly effective’ [sample] (i.e., 88 percent)” (Dee & Wyckoff, 2013, p. 13). This hardly represents a miracle in teacher retention. [↑](#footnote-ref-11)
12. Read about a recent case of this in which an experienced testing expert testified before a committee of the New York State Senate about the state’s test scores that “went up, up, up until 2010, when the state admitted that the previous dramatic gains were illusory, a consequence of artful adjustments of the “cut score” (passing rate). Then the scores began to rise again, until this past year’s Common Core tests, when the state scores fell deep into the basement, and three quarters of the state’s children were marked as failures” (Ravitch, 2013). [↑](#footnote-ref-12)
13. An effect size is a measure of the strength or magnitude of, in this case, a statistical change over time, from a pre-observation to a post-observation. It is used to statistically describe, literally, the “size” of the “effect” of whatever treatment might have occurred, again in this case over time (e.g., Cohen’s d; Cohen, 1988). [↑](#footnote-ref-13)
14. Michelle Rhee lauded the findings from this study on Twitter, writing "This study affirms: It is possible to identify, support & improve teacher effectiveness” (see: https://twitter.com/MichelleRhee/status/390918631246155776). [↑](#footnote-ref-14)