APA Division 15
Policy Brief Series
Vol. 1, No. 2
Jan. 2020

DIVISION 15 OF THE AMERICAN PSYCHOLOGICAL ASSOCIATION PRESENTS

ADDRESSING TEACHER EVALUATION APPROPRIATELY

A RESEARCH BRIEF FOR POLICYMAKERS

BY DR. ALYSON L. LAVIGNE & DR. THOMAS L. GOOD

Teachers have a clear and important impact on student learning.[1] We have long known that teachers impact student achievement.[2] Thus, considerable policy attention, time, and effort have focused on teacher evaluation. Our recommendations for teacher evaluation include placing more emphasis upon **improving teaching** rather than stratifying teachers in terms of their effectiveness.

Note: This is an official statement of Division 15 (Educational Psychology) of the American Psychological Association, and does not represent the position of American Psychological Association or any of its other Divisions or subunits.

CURRENT LANDSCAPE

We all believe that our students should be taught by effective teachers, and a focus on evaluation is one way to achieve this end. But unfortunately, most believe that teacher evaluation is ineffective because all teachers are rated the same (as good) and more and less effective teachers are not identified.[3] Recent efforts to improve teacher evaluation have been **costly** (for example, Race to the Top (RTTT) alone cost 4 billion dollars) and remain **ineffective** as massive reforms intended to improve practice were unsuccessful.[4] Further, research by educational psychologists, and other social scientists on these new reforms has identified many unintended and harmful effects upon teachers and the teaching profession.[5,6]

In today's schools, two practices dominate evaluation models as prescribed by local and national policies:

Statistical Approaches:

Using standardized test scores in sophisticated statistical models such as value-added modeling (VAM) to identify good teachers (i.e., those who obtain more student achievement than other teachers teaching similar students under similar circumstances).

&

Observation of Teachers:

Using supervisors' (usually principals') ratings of teachers on rubrics to identify good teachers (i.e., what good teachers do in classrooms).

Both are problematic and have not improved student achievement [7], particularly when used in high-stakes teacher evaluations, in part because they fail to recognize the complexity of teaching or how to measure it.

WHAT RESEARCH INDICATES

Sufficient evidence of the limitations of VAM has led some professional associations to raise caution about its use. Further, VAM scores have been challenged in court.[8] In short, research indicates that:

1. **Statistical approaches are flawed.** Notable deficiencies include:

   a. **VAM scores do not adequately compare teachers.** VAM scores are unable to account for teachers' disparate contexts and all of the factors that explain student achievement outcomes.[9]
   b. **A teacher's *effectiveness* often varies.[10] Thus, it is difficult to achieve appropriate reliability to justify its use in high-stakes teacher evaluation.** Even if abundant data were available (e.g., 10 years), decisions based upon student achievement would be wrong 12% of the time.[11]
   c. **Many teachers do not have VAM scores.** Typically, nearly a third of teachers do not have VAM scores. Thus, assigning a VAM score to these teachers based upon the average score of other teachers (a school-level value-added score) is unfair.[12]
   d. **VAM scores do not help teachers improve their instruction.[13]** This is, in part, because teachers' VAM scores are weakly—and sometimes not at all—related to what teachers actually *do* in classrooms (e.g., observational data). Furthermore, VAM scores do not identify for teachers areas in need of improvement (e.g., concepts that were particularly difficult, student misconceptions).
   e. **The use of individual value-added scores discourages collegial exchange** and the sharing of ideas and resources across the school.[14]

2. **Current observation systems are problematic for high-stakes evaluation.** Research indicates that:

   a. **Classroom teaching is complex, dynamic, and contextual.[15]** No single observation system measures all aspects of good teaching. One ramification of the selectivity of observation systems is that they measure only how the teacher interacts with the class as a whole and ignore how teachers interact with individual students. This can be a serious omission as research shows that in some classrooms the frequency and quality that some students share with their teacher is much more favorable than the pattern other students enjoy. [16, 17]
   b. **Observing and providing teachers with feedback receives too little time.[18]** It is often secondary to principals' more immediate, urgent, and unpredictable tasks.
   c. **Principals have not been prepared well to observe or provide useful feedback to teachers.** This may be in part because this aspect of the job has traditionally accounted for a small proportion of how principals spend their time in schools.
   d. **Teacher *practice* varies, and often should.[19]** Good teaching does not mean that teachers should necessarily teach similarly from lesson to lesson and day to day. On any given day, a teacher may need to adjust the lesson (e.g., student absences, lawnmower roars outside the window). The common practice of three formal observations/year/teacher does not adequately account for these fluctuations and, therefore, appropriate reliability cannot be achieved to justify the use of observational measures for high-stakes teacher evaluation.

IMPLICATIONS FOR POLICY

Instead, policy should reflect an approach to teacher evaluation that prioritizes teachers--their expertise and their growth and development, such as:

- **Eliminate high-stakes teacher evaluations based only on student achievement data and limited observation.** Teaching is complex and difficult. Accordingly, the evaluation of teaching requires many observations and multiple data sources. Observation data should primarily be used in combination with other sources of data to inform reflective conversations among teachers or among teachers and supervisors.
- **Provide opportunities for teachers to be heard** (e.g., provide input on what is measured and how). Key stakeholders should be engaged in initial conversations regarding any new teacher evaluation policies and models and from development through implementation, as well as post-implementation evaluations. Teacher evaluation models that reflect collective values will likely have greater success.
- **Acknowledge and act upon the potential for improving the system in fundamental ways** using technology, collaboration, and other innovations to transform extant practice.

IMPLICATIONS FOR PRACTICE[20]

Likewise, practice should reflect an approach to teacher evaluation that prioritizes teachers' professional growth and development, such as:

- **Emphasize formative feedback.** Teachers benefit most from and have more opportunities to act upon information that is provided at the beginning and middle of the year than feedback that is provided at summative end-of-year evaluations.
- **Consider if and how additional observers—beyond the principal—can be leveraged.** Most principals do not have sufficient time to collect enough evidence to have meaningful discussions about practice and its improvement.[21] However, subject- or grade-specific peers and coaches (who are well grounded in findings from research on teaching) may be equally if not more helpful in providing feedback that improves teaching.
- **Encourage principals, knowledgeable peers, and coaches to provide feedback to improve teaching and learning**—feedback that: emphasizes practices that are consistently associated with student achievement, promotes teacher self-reflection and action, and aligns with professional development. Focus on the day-to-day and weekly gains that teachers achieve not just end-of-year summative scores.
- **Promote observation measures that reflect the context and complexity of teaching as well as teacher needs**—measures may be sensitive to the student population or content area or are flexible enough to attend to, address, measure, and support feedback on specific teacher concerns (e.g., depth of knowledge questioning, quality of student-teacher talk).
- **Encourage more valid observations of teachers.**[22] Knowing that teaching practices vary, evaluation procedures might include observing an entire unit or range of a teacher's classes (e.g., remedial, advanced), lessons (e.g., student-directed, problem-based, especially difficult concepts, newly developed lesson), and content areas.
- **Encourage teacher engagement in opportunities that have the potential to improve their efficacy and effectiveness** (e.g., lesson study, peer observation and coaching, Professional Learning Communities, high-quality professional development).
- **Approach accountability as an opportunity for growth.** Districts can use the increased emphasis on holding teachers accountable as required by policymakers as an opportunity to revise their teacher evaluation models to respond to the challenges summarized here and to enact the belief that **improving instruction should be at least as important as evaluating instruction**.

This brief is based primarily on a recent publication by Alyson L. Lavigne and Thomas L. Good (2019): Enhancing Teacher Education, Development, and Evaluation: Lessons Learned from Educational Reform. New York, NY: Routledge.

Please direct questions and media inquiries to:

Alyson L. Lavigne
Utah State University
alyson.lavigne@usu.edu

or

Thomas L. Good[23]
University of Arizona
goodt@email.arizona.edu

NOTES

1. Aaronson, Barrow, & Sanders, 2007; Goldhaber & Brewer, 1999; Konstantopoulus, 2014; Rubie-Davies, 2014. Estimates of teacher effects on student achievement vary with a robust estimate being 1-21% (Konstantopoulus, 2014). Although these effects are important, it must be understood that teacher effects are small relative to other factors that impact student achievement (e.g., poverty).
2. Brophy & Good, 1986; Good, Biddle, & Brophy, 1975.
3. Weisberg, Sexton, Mulhern, & Keeling, 2009.
4. Garet et al., 2017; Good & Lavigne, 2018; Stecher et al., 2018.
5. Collins, 2014; Ford, Van Sickle, Clark, Fazio-Brunson, & Schween, 2015; Lavigne, 2014.
6. Media associated with teacher evaluation and the need to remove ineffective teachers has inadvertently created the impression that many of our teachers are incompetent. Berliner (2018) has explored the extent to which bad teachers exist. We, like him, believe that popular opinions about the prevalence of bad teachers has been woefully inflated. And, in part, too much attention of current evaluation practices has arguably been placed on identifying and removing poor teachers and too little attention placed on improving normative practice.
7. See Lavigne and Good (2019) and Stecher et al. (2018) in the case of RTTT, Jensen et al. (2019) in the case of the Measures of Effective Teaching project, and in the case of standardized achievement measures to measure effective teaching, see Lavigne (2014), Lavigne and Good (2015), and Nichols and Berliner (2007).
8. The American Educational Research Association (2015) and the American Statistical Association (2014) has recommended against the use of VAM scores in teacher evaluation. Further, the legal implications of high-stakes teacher evaluation is significant and have been explored (see Hazi, 2017, and Page, Amrein-Beardsley, & Close, 2019, for reviews, and for a specific example, see the federal lawsuit of seven teachers and the Houston Federation of Teachers against the Houston Unified School District in which a settlement was reached that required the District to pay Texas AFT $237,000 for attorney fees and expenses related to the lawsuit and to not use value-added scores to terminate a teacher as long as the teacher is unable to independently test or challenge the score).
9. According to the American Statistical Association (2014), VAM scores are calculated at the level of the classroom. The validity of VAM scores as measures of teacher effectiveness depends on how well the particular model adopted adjusts for other factors that might systematically affect, or bias, a teachers' VAM score. These include classroom-level differences (contextual factors) that may be due in part to other factors that are not included in the model (e.g., class size, teaching "high need" students, or having students who receive extracurricular tutoring)(p. 4). Given that teachers explain a small amount of the variance in student achievement outcomes, it is nearly impossible for VAM scores to account for all of the other factors that explain the majority of variance in student achievement outcomes.
10. See Good and Lavigne (2015) for a review of research on the instability of teacher effectiveness.
11. Schochet and Chiang (2010).
12. Those teachers who are not teaching in tested subject areas (e.g., music, or social studies) do not have standardized achievement tests that are necessary for computing an individual VAM score.
13. For more discussion of the use of value-added assessments in K-12 education, see Amrein-Beardsley and Collins (2012), and for more information about value-added assessment in teacher education programs see Worrell et al. (2014).
14. When teachers' performance is relative—determined, in part, by the performance of their peers—as it is when value-added scores are used, it is less likely that teachers would readily share ideas and resources as they would in schools that encourage collegial relations. For more information about the importance of allowing for teacher cooperation, see Johnson (2019) and Rosenholtz (1989).
15. The complexity of classroom teaching and learning has long been established (Dunkin & Biddle, 1974; Jackson, 1968; Shulman, 1987). More recently, researchers are becoming more sensitive to linguistic and cultural differences that need to be reflected in observational systems (see Jensen, Grajeda, & Haertel, 2018; Lavigne & Oberg De La Garza, 2015).
16. Brophy & Good (1970); Good & Lavigne (2018); McCaslin & Good (1996); Rubie-Davies (2014); Weinstein (2009); Snell & Lefstein (2018); Timmermans, Rubie-Davies, & Rjosk (2018).
17. We noted that many extant observation systems may be insensitive to cultural differences in classrooms. In addition to this issue, there is also the problem that observers (independent of the observation instrument) may hold biases that influence scoring decisions. Campbell and Ronfeldt (2018) have noted that teachers teaching classes with more males, students of color, and low-performing students sometimes receive lower ratings than warranted. Such ratings may result in the unwarranted conclusion that their students are less capable. Research has illustrated that there is strong evidence that notable learning gains are found for students in some schools that primarily enroll students from low income homes especially when emphasis is placed upon relative rather than absolute student achievement (Owens, Reardon, & Jencks, 2016). For more recent evidence, see: https://edopportunity.org/discoveries/affluent-schools-are-not-always-best/. Elsewhere, McCaslin and Good (2008) and their colleagues presented several articles summarizing their studies of school reform in Arizona. Their analysis, based on classroom observations and student surveys, illustrated that students were engaged in academic tasks, and that teachers had created warm and productive learning environments. However, teachers alone cannot compensate for the effects of student mobility and poverty density (McCaslin & Good 2008).
18. Lavigne & Good (2019).
19. Reliably assigning teachers to a category (e.g., needs improvement, exemplary; correct 90% of the time) would require 10 observations/per teacher/per year—which is not feasible given the other demands placed on principals.
20. Research has shown some fundamental problems with the way teacher evaluation is commonly conducted. The research is less clear on how these difficulties can be addressed. This section illustrates promising possibilities that may help to make teacher evaluation and development activities more transparent and more actively involve teachers in the design and improvement of instruction.
21. Given the range of teaching context it is difficult to offer specific recommendations, however, if we were to understand teaching it makes sense to examine a sequence of instruction to see how teachers plan, implement, and assess the effectiveness of a unit. As one moves more from evaluating to developing talent, more ambitious observational strategies need to be used. Similarly, for teachers that are making frequent use of small group instruction or project-based learning, it seems imperative that evaluators adjust their protocols to match the form of instruction that is being presented.
22. Lavigne & Good (2019).
23. Good is a Professor Emeritus of the Department of Educational Psychology and a Co-Director of the Center for Research on Classrooms at The University of Arizona.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of labor Economics, 25*(1), 95–135.

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

American Statistical Association. (2014, April 8). *ASA statement on value-added models for educational assessment*. Retrieved from https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS EVASS) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12).

Berliner, D. C. (2018). Between Scylla and Charybdis: Reflections on and problems associated with the evaluation of teachers in an era of metrification. *Education Policy Analysis Archives, 26*(54). http://dx.doi.org/10.14507/epaa.26.3820

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillian.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than what we bargained for? *American Educational Research Journal, 55*(6), 1233–1267.

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). http://dx.doi.org/10.14507/epaa.v22.1594

Dunkin, M. J., & Biddle, B. J. (1974). *The study of teaching*. New York: Holt, Rinehart & Winston.

Ford, T. G., Van Sickle, M. E., Clark, L. V., Fazio-Brunson, M., & Schween, D. C. (2015). Teacher self-efficacy, professional commitment, and high-stakes teacher evaluation policy in Louisiana. *Educational Policy, 31*(2), 202–248.

Garet, M.S., Wayne, A.J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). *The impact of providing performance feedback to teachers and principals, executive summary* (NCEE 2018-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Goldhaber, D., & Brewer, D. (1999). Teacher licensing and student achievement. In C. Finn & M. Kanstoroom (Eds.), *Better teachers, better schools* (pp. 83–102). Washington, D.C.: Thomas B. Fordham Institute.

Good, T., Biddle, B., & Brophy, J. (1975). *Teachers make a difference*. New York: Holt, Rinehart, and Winston.

Good, T. L., & Lavigne, A. L. (2015). Issues of teacher performance stability are not new: Limitations and possibilities. *Education Policy Analysis Archives, 23*(2). http://dx.doi.org/10.14507/epaa.v22n95.2014

Good, T. L., & Lavigne, A. L. (2018). *Looking in classrooms* (11th ed.). New York: Routledge.

Hazi, H. M. (2017). VAM under scrutiny: Teacher evaluation litigation in the states. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 90* (5-6), 184–190. doi:10.1080/00098655.2017.1366803

Jackson, P. W. (1968). *Life in classrooms*. New York: Holt, Rinehart and Winston.

Jensen, B., Grajeda, S., & Haertel, E. (2018). Measuring cultural dimensions of classroom interactions. *Educational Assessment, 23*(4), 250–276. doi: 10.1080/10627197.2018.1515010

Jensen, B., Wallace, T. L., Steinberg, M. P., Gabriel, R. E., Dietiker, L., Davis, D. S., …Rui, N. (2019). Complexity and scale in teaching effectiveness research: Reflections from the MET study. *Education Policy Analysis Archives, 27*(7). http://dx.doi.org/10.14507/epaa.27.3923

Johnson, S. M. (2019). *Where teachers thrive: Organizing schools for success*. Cambridge, MA: Harvard Education Press.

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record, 116*(1), 1–21.

Lavigne, A. L. (2014). Exploring the implications of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*(1).

Lavigne, A. L., & Oberg De La Garza, T. (2015). The practice and evaluation of culturally responsive literacy for English Language Learners in the 21st century. In R. Allington and R. Gabriel (Eds.), *Evaluating literacy instruction: Principles and promising practices* (pp. 58–78). New York: Routledge.

Lavigne, A. L., & Good, T. L. (2015). *Improving teaching through observation and feedback: Going beyond state and federal mandates*. New York: Routledge.

Lavigne, A. L., & Good, T. L. (2019). *Enhancing teacher education, development, and evaluation: Lessons learned from educational reform*. New York: Routledge.

McCaslin, M., & Good, T. (1996). The informal curriculum. In D. Berliner and R. Calfee (Eds.), *The handbook of educational psychology, First Ed.,* pp. 622–670. New York: American Psychological Association/Macmillan.

McCaslin, M. & Good, T. (2008). Special issue: School reform matters. *Teachers College Record, 110*(11),2317-2496

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

Owens, A., Reardon, S., & Jencks, C. (2016) Income segregation between schools and districts. *American Educational Research Journal, 53*(4), 1159-1197.

Rosenholtz, S. (1989). *Teachers' workplace: the social organization of schools*. New York, NY: Longman.

Rubie-Davies, C. (2014). *Becoming a high expectation teacher: Raising the bar*. London: Routledge.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and student performance based on student test score gains*. Washington, DC: IES National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Shulman, L. S. (1987). Knowledge and teaching: Foundations of a new reform. *Harvard Educational Review, 57*(1), 1–23.

Snell, J., & Lefstein, A. (2018). "Low ability," participation, and identify in dialogic pedagogy. *American Educational Research Journal, 55*(1), 40–78.

Stecher B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J…Chambers, J. (2018). *Improving teaching effectiveness. Final report. The intensive partnerships for effective teaching through 2015–2016*. Retrieved from https://www.rand.org/content/dam/rand/pubs/research_reports/RR2200/RR2242/RAND_RR2242.pdf

Timmermans, A. C., Rubie-Davies, C. M., Rjosk, C. (2018). Pygmalion's 50th anniversary: the state of the art in teacher expectation research [Special issue]. *Educational Research and Evaluation, 24*(3-5).

Weinstein, R. S. (2009). *Reaching higher*. Cambridge, MA: Harvard University Press.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on difference in teacher effectiveness*. Brooklyn, NY: The New Teachers Project. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Worrell, F. Brabeck, M., Dwyer, C., Geisinger, K., Marx, R., Noell, G., and Pianta R. (2014). *Assessing and evaluating teacher preparation programs*. Washington, DC: American Psychological Association.