Fair Is Fair: Outcome Assessment, Constitutive Luck, and Teacher Evaluation

Matthew Hayden
Teachers College, Columbia University

The influence of the so-called "accountability movement" in education in the United States has elevated the perceived importance of student assessments to predominance over most other educational issues. The emphasis has been so great that the two most recent national-level educational policy initiatives, No Child Left Behind and Race to the Top, placed student assessment outcomes at the center of their policies. As a result, schooling reform movements have increasingly focused on the use of student assessments to evaluate teachers. In this essay, I will attempt to show not only that such practices are flawed, but that they also undermine the stated goals of their advocates.[1]

The Problem

In his 2009 contribution to a book that focused on the work of Steven Cahn, Randall Curren examines the use of assessments of student work, particularly assessments he calls *outcome assessments* (OA), in order to reveal issues of justice and fairness concerning students.[2] As an example, imagine a class in which an exam is given and in general, the students did very poorly. Further examination shows that all of the students answered a certain subset of questions incorrectly. It is then deduced that the teacher did not teach that material well. Since the students' poor grades appear to be the result of the teacher's failure, Curren asks whether it would be fair to let the scores stand. He answers that it would not and that among various other steps that could be taken, the teacher could simply "throw out" the offending questions so the students were not held accountable for the teacher's failure.[3] Curren produces the bad scores due to teacher failure as one of many effects that could come to bear, through no fault of the students, upon their performances on an assessment, and he calls such effects *constitutive luck* (CL). Other similar factors might be poverty, parental obstruction, adult-like responsibilities at home, abusive or unhealthy home life, or even temporary influences such as illness, or family or community trauma. Curren's argument is that students should not be held responsible for results that are caused by that which lies beyond their control, including contexts and situations that influence their test performances and academic achievements.

At the conclusion of his chapter, Curren offers some suggestions for assessment reform: (1) "Grades should only be used in schools to promote learning and make decisions conducive to equal educational attainment, and not released for use as credentials." (2) "Assessments of talent for use in labor markets and admissions to higher education should not occur until the completion of high school, when the effort to produce equal educational attainments as a basis for fair equality of opportunity will have run its course." (3) OA might be useful in "encouraging productive and motivating supervision of students' intellectual development, and by

PHILOSOPHY OF EDUCATION *2012* | Claudia W. Ruitenberg, editor

providing exit exams or other forms of evaluation that could serve as credentials." (4) OA could be used to help teaching align with intended curricular outcomes.[4]

In the context of Curren's argument — fairness to students — these conclusions are appropriate and well reasoned. There are interesting issues to explore in light of Curren's suggestions specifically related to OA and student achievement, but I will instead focus on the application of his argument about fairness in student assessment to teacher evaluation. While fairness to students is important, they are not alone in deserving justice. I would like to apply Curren's thinking about the ethical importance of mitigating the undeserved negative effects of influences beyond the students' control — Curren's CL — to the same conditions for teachers. If we can use OA for the evaluation of a teacher's alignment with intended curricular outcomes, then what factors might contribute to a negative evaluation of such alignment for which the teacher is not responsible? To what degree might CL obligate us to reconsider our use of OA in evaluating teachers particularly when teacher jobs are at stake?

Constitutive Luck and Teacher Evaluation

In the current climate of "accountability" teachers are under pressure to "prove" their effectiveness through the results of OA, and primarily through student results on tests. While OA can be represented by assessments other than tests, all OA essentially rely on student outcomes or products. I am not opposed to the use of periodic testing or other assessments as formative tools to guide improvement in teacher practices, but we need to think more carefully about their use as summative tools for teacher effectiveness in the same way that Curren suggests caution regarding student achievement and OA.[5] Regardless of the kind of assessment used, two basic assumptions are in play, both of which require examination: (a) that OA measure the effects of teacher inputs, and (b) that the inputs of teaching can actually be measured.

Research has shown that the efficacy of OA in measuring teacher effectiveness is muddled. The problems of standardized OA are clear, however, in regard to CL; socioeconomic status and other forms of CL are more highly correlated to achievement than any other factors. Value-added modeling (VAM) has been formulated to mitigate these influences and it has identified variations in student achievement correlated to certain teachers, but it also reveals that so-called "effective" teachers have less influence on increasing student achievement than "ineffective" teachers do in limiting it.[6] Thus, "effective" teachers face more difficulty demonstrating their effectiveness through increased achievement, particularly with high-achieving students where measurable effects will be almost invisible. Additionally, such models have "proven to be unstable across statistical models, years, and classes that teachers teach."[7] More troubling is that even though teacher inputs are the most influential of school effects on student achievement, they still only account for 7.5 to 8.5 percent of the variations in student achievement and, of those teacher effects, only 3 percent are readily observable. Thus 97 percent of the teacher effects on student achievement elude measurement through student OA.[8]

10.47925/2012.331

Researchers will quibble about the specific numbers, but the general trend is clear: non-school factors, which constitute the bulk of CL, have substantially larger effects on student performance and achievement in school than teacher effectiveness.[9] Since CL is not something students can get rid of at will, and teachers cannot erase it for them, it will be a factor in any assessment of students. Thus, the use of OA for teacher evaluation means that teachers are held accountable for the CL of the students. Fairness, then, compels us to either accurately account for CL in student OA or not use OA at all for these purposes.

What Do Students Owe?

The focus on what the teacher does or does not do and its impact on the student's academic and professional future is important and necessary. It is reasonable to explore how a teacher's mistake or incompetence might unfairly and adversely affect a student's academic and professional future. However, one may wonder about a contrary, though just as likely, scenario in which the teacher has done all that he should or could be expected to do, but the class scores are low because the students are wholly or mostly to blame. Perhaps they were lazy, refused to do their homework and readings, or failed to pay attention in class, essentially "earning" the low grades they received. Whose responsibility is this? Should the teacher bear the brunt of accountability for the students' own choices about their educational behavior? If OA are used for teacher evaluation in cases like this, the teacher will be "punished" for the students' errors, or if one prefers, choices and behaviors. In today's accountability climate teachers are under threat of sanctions ranging from substantial (public reporting of class scores, for example) to severe (reassignment) to potentially and personally catastrophic (being terminated) if their students do not perform well on certain tests or collections of OA. Curren's analysis recognizes that students, even older ones, should not be held accountable for their academic performances and behaviors while they are still developing their rational, intellectual, and judgment capacities, at least until they have reached an "age of reason."[10] If we accept this, then we must also accept that such mitigating factors should be integrated into the evaluation of teachers via OA.

Is it fair for teachers to be sanctioned because some students choose not to do their utmost in learning the material assessed? This exposes a problem that makes teachers vulnerable even under Curren's excellent recommendations. Students would know that the assessments in eighth grade are merely formative and the results will not irreparably harm their academic and future professional lives. This could easily compel a student to study less and take less care in his education even though such actions will make the teacher's evaluation suffer. What other profession grounds the employment stability of an adult on the judgments, behaviors, or whims of fourteen-year-olds? Curren's recommendations laudably appear to err on the side of protecting the students from the long-term negative effects of CL, but do so at the risk of enabling potentially negative and habit-forming behaviors.[11] This is not a criticism — I have no remedy — but rather another observation to include in consideration of the problem of using student outcomes to evaluate teachers.

10.47925/2012.331

Constitutive luck is an influence on OA that is as substantial for evaluating teacher competency as it is for student achievement and if there is any genuine interest in getting assessment and evaluation right, for students and for teachers, we must account for CL in all its inconvenient complexity. If not, it is time to admit that we undertake evaluation practices to satisfy a bureaucratic need to demonstrate measurement activity more than to actually measure. I do not suggest that teachers have no effect on student OA performance or the behavior that leads to academic choices, only that such teacher effects are limited and overwhelmed by the same CL that afflicts students. The next section highlights an important connection in this regard, one that is difficult to measure and made worse in the attempt to do so.

Shooting Our Own Feet

Curren writes that "[w]e can't, in fairness, simply give children the choice to learn or not — to develop intellectually or not — then expect them to suffer the consequences if they weren't adequately motivated."[12] Positive motivation for learning does not simply happen. However, if teachers and school administrators are held accountable for student academic achievement and OA results to the extent that their jobs depend upon it, one can be certain that both will adopt methods that give them more control over the assessment outcomes. In a study of the effects of "controlling" measures in assessment and teacher evaluations on teaching practices, Luc Pelletier and Elizabeth Sharp discovered that the use of such measures administratively encourages the use of controlling measures pedagogically.[13] Schools under pressure to raise student performance on tests apply pressure to their teachers to get the results desired.

Edward Deci et al. conducted a study in which two groups of teachers were asked to help students improve their problem solving abilities.[14] One of the groups was also asked to make sure their students met high standards. The researchers found that the teachers in this group exhibited more controlling teaching behavior than the other teachers, such as more criticism of students, the use of more hints, and language that is more directing. The students of the "controlling" teachers also performed worse on an assessment than the other students. Teachers admitted that pressure exerted from "above" caused them to be more controlling in their teaching due the compelling notion that doing so would allow them to exert more influence over the results of the assessments.[15] Unfortunately, this assumption *does* produce more influence over the results, but in the opposite way intended, and the effect increases with the stakes.

Richard Ryan and Netta Weinstein examined the effects of testing on the motivation of students and found that the administration of high stakes and standardized tests has negative effects on all areas connected to student motivation and results in decreased performance on tests and student achievement.[16] In particular, assessments that are "controlling" and put pressure on students to achieve specific and predetermined outcomes using a reward and punishment system lead to decreased performance and to the exertion of the least amount of effort to obtain a given result. Such "controlling" assessments are often the norm for standardized assessments, and particularly those that are summative, high stakes, and increasingly

10.47925/2012.331

used for teacher evaluation.[17] However, even if OA are not high stakes for students, if they are used for evaluating teachers they are likely to elicit similar controlling efforts since they are de facto high stakes measures.

Why adopt these controlling measures? Teachers may recognize, either tacitly or implicitly, that they possess limited control over student outcomes and are paradoxically held accountable for them. They know all too well how CL intervenes to disrupt even the most skilled and deliberate of teaching efforts. Therefore, teachers under pressure to guarantee outcomes feel compelled to wield more influence on the results. This compulsion is the result of a "trickle down" effect of controlling behaviors. In response to public pressure or mandates school administrators demand improved scores or a specific outcome from teachers, and may delineate various procedures and steps, or at the very least, schedule assessments to determine if their demands have been met. These activities leave little room for a teacher's curricular and subject-area autonomy, and may be interpreted as a reflection on competency. Adopting controlling pedagogies may seem logical to a teacher given externally imposed constraints and narrowed objectives. The controlling pedagogy will then begin to erode his motivation by distancing him from his students, whom he will begin to see as a means to an end (higher scores, not necessarily more learning), and thus ignore the connectedness required for both motivation and good teaching. Of course, the students will concurrently experience the process undergone by the teacher: loss of autonomy, slighted competence, and loss of connectedness to both teacher and subject matter (the latter seen as a means to an end).

Recommendations

What can we use to assess teachers? Instead of focusing on the results, we should focus on the practices or process that produce them. Using OA to evaluate teachers is like assessing student achievement based on family income. Certainly, students can wield some influence on how they or their parents earn money and manage expenses, but they are mostly subject to the income-earning and financial-management capacities of others. Teachers may have limited influence on student behaviors in their classrooms and may positively influence student motivation for learning and studying, but the combination of CL and the compulsion to controlling behaviors contribute more significantly and negatively to those motivations and behaviors and go further in determining the results of OA than the teacher's positive inputs.

To use OA effectively in teacher evaluation would require an examination of not only the results of OA, but also the specific practices used in preparation for OA. Analyzing OA without knowing what was done can neither inform teaching practice nor offer an accurate evaluation of the teacher. What has worked before cannot necessarily be relied upon to work again.[18] The conditions of teaching and learning are different each year as new combinations of CL are introduced. We cannot make the mistake of assuming that there is a straight, linear, one-to-one correlation between OA results and teaching quality — nor from year to year, as VAM would have it. Knowing *that* something has gone wrong in a teacher's practice is entirely

10.47925/2012.331

different from, and woefully less adequate than, knowing *what* specifically has gone wrong and why. We need to know what the intermediate actions of the teacher and the conditions of learning were that influenced those results before we can know if the teacher has been effective or not. Further, if we recognize that non-school effects and CL make the dominant and significant contribution to student OA results, we cannot ignore this fact when assessing teachers using those results. At minimum we should significantly reduce the influence of OA in our evaluation of teachers and limit the scope of our punitive responses.

To that end, and with a reluctant acceptance of the reality of the continued practice of using OA in teacher evaluations, I offer the following suggestions for teacher assessment using a revised version of Curren's suggestions regarding students. Curren's student-specific words and phrases have been replaced in order to apply the substance of his suggestions to teachers; these changes are italicized below.

(1) *Scores* should only be used in schools to promote learning and make decisions conducive to equal *teaching* attainment, and not released for use as credentials. OA should only be used to inform teaching practices and improve specific teaching problems as indicated by specific items or subjects assessed. There would be no summative record of these scores to construct a permanent or cumulative record. In thinking about the results of OA, Curren suggests that a teacher should ask the question, "How well did the students learn what they *could have been reasonably expected to learn,* given the limitations of the instruction they received?" I believe it is only fair to ask the similar question of teacher evaluators. How well did the teacher teach what he could have been reasonably expected to teach given the limitations present? "Limitations" would refer to the CL that the students bring with them as well as other non-teacher influences in play.

(2) Assessments of talent for use in labor markets and *use in personnel decisions* should not occur until the completion of *a specified period of time*, when the effort to produce equal *teaching* attainments as a basis for fair equality of opportunity will have run its course. Most educational professionals agree that the first years of teaching contain a challenging learning curve. Some studies have found new teacher attrition rates to be as high as fifty percent over the first five years.[19] Others have shown a positive correlation of student test scores to years of teacher experience.[20] Thus, some categorization or norm referencing would be required in teacher comparisons. Comparing a third-year teacher with a twelfth-year teacher would be as unfair as comparing a ninth-grade student with a twelfth-grade student. We need to keep this in mind in evaluating teachers, especially if the results of such evaluations are used for credentialing purposes (such as license renewal) or employment decisions. In such cases, it might be necessary to create norm-referenced groups based on years of teaching experience, separating early career teachers from experienced teachers or creating three- to five-year cohort spreads. Given the high attrition rates for new teachers that extend two years beyond most tenure probationary periods of three years, it might be reasonable to extend the probationary period to five years to cover this crucial period of what amounts to "on

10.47925/2012.331

the job training." The loss of tenure security for teachers in years four and five might be mitigated by a concurrent increased investment in professional development for early career teachers and thus be an incentive for schools to offer continuing contracts to those they have worked hard to train.

(3) OA might be useful in encouraging productive and motivating supervision of *teacher professional development*, and by providing *progress* exams *with* other forms of evaluation that could *be used for re-licensure*. Current pre-service testing for licensure is handled at the state level in the United States, and often requires passing scores on exams such as PRAXIS I and/or II or on some state-level proprietary test such as the Minnesota Teacher Licensure Examination (MLTE) for basic skills, content knowledge, and pedagogy. Current re-licensure requirements typically require additional education coursework, from master's degree programs to online study programs that earn continuing education credits. It might be possible to use OA to create targeted professional development interventions that could be integrated into re-licensure approval processes, thus attempting to satisfy the demands of increased teaching proficiency by multiple stakeholders, but without the career-damaging effects of summative use.

(4) OA could be used to help teaching align with *professional standards*. Rather than using OA only to align teaching with curricular outcomes for students, why not use them to align professional development with professional standards. For instance, it has been found that National Board Certified Teachers (NBCT) are more positively correlated to increased student achievement than non-NBCT (those who have applied for but been denied certification).[21] The National Board's assessment criteria for certification include video recordings of teaching, student work samples, demonstration of content knowledge, and proof of effective interaction with the community outside the classroom. Clearly OA would speak to only a small portion of such criteria, and this is precisely the point. OA can offer us only a part of the teacher evaluation solution.

The ultimate priority is student learning. If bad scores are the result of poor teaching and if content matters, then we need to direct our resources toward re-training and re-teaching. To deny this opportunity implies a weaker commitment to teaching and learning than is otherwise expressed by schools, teachers, policymakers, and the public. What matters most is determining not where the students and teacher are, but how they got there. If the primary aim of OA in evaluating teachers is to improve student learning, then we cannot ignore the fundamental need of time and resources to do that, nor can we ignore the influence of CL. If the primary aim is to identify the "bad" teachers, then OA are a flawed and incomplete instrument for doing so and may cause more harm in learning than they prevent. If fairness to students matters, then we need to be able to teach them when we know we have not. Is it fair to this year's students to only use OA to improve the learning for next year's students? Of course it is not, but the predominant interest in OA as a tool for summative purposes only obstructs the use of OA for more efficacious processes and for the primary tasks of teaching and learning.

10.47925/2012.331

Conclusion

If students should not be held accountable for what lies beyond their control, neither should teachers. Unfortunately, students are held accountable for influences beyond their control and so are teachers, resulting in diminished returns for both. Some steps teachers might take to mitigate the effects of CL are laudable, but ineffective in maintaining fairness to all students, standards of proficiency, and effectively evaluating teachers. OA may be effective for demonstrating assessment practices, but it is not always clear what exactly the assessment tells us about the role of the teacher in the results. Bowing to pressure to produce narrowly defined academic achievement will actually decrease teacher effectiveness and student achievement, thus proving counterproductive to the aims the pressure is designed to attain. The sad fact is, however, that both students and teachers are routinely punished (and rewarded) for the results of processes influenced substantially by that which lies beyond their control, amounting to a singularly unjust system of educational deserts. I endorse the reforms that protect students, as suggested by Curren, and ask only that they be applied to teachers, too.

---

1. I note that there is a difference between "assessment" and "evaluation." Assessment is reflective, diagnostic, flexible, and cooperative. Evaluation is prescriptive, judgmental, fixed, and competitive. It is very likely that there are intrinsic and conceptually based problems in using an assessment as an evaluation, but that is an investigation for a future essay.

2. Randall Curren, "Academic Standards and Constitutive Luck," in *A Teacher's Life: Essays for Steven M. Cahn*, eds. Robert Talisse and Maureen Eckert (Lanham, MD: Lexington Books, 2009). Note that an OA could be any tool used by a teacher to assess student learning, from mere observation to oral questioning to a multiple choice test.

3. This example is not meant to be definitive. It is used due to its ease in illustrating the effects of changing or altering assessments after the fact.

4. Curren, "Academic Standards and Constitutive Luck," 29.

5. Admittedly, using an exam for external credentialing purposes, as described in suggestion (b), is a summative assessment.

6. William Sanders and Sandra Horn, "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research," *Journal of Personnel Evaluation in Education* 12, no. 3 (1998): 247–256. This is the main drawback of focusing on gains each year as a student moves through the grade levels.

7. Eva Baker, Paul Barton, Linda Darling-Hammond, Edward Haertel, Helen Ladd, Robert Linn, Diane Ravitch, Richard Rothstein, Richard Shavelson, and Lorrie Shepard, *Problems with the Use of Student Test Scores to Evaluate Teachers (*Washington, DC: Economic Policy Institute, 2010).

8. Dan Goldhaber, "The Mystery of Good Teaching," *Education Next* 2, no. 1 (2002): 50-55. Also see Steven Rivkin, Eric Hanushek, and John Kain, "Teachers, Schools, and Academic Achievement," *Econometria* 73, no. 2 (2005).

9. Goldhaber, "Mystery of Good Teaching." Goldhaber found that only 21 percent of the variation in student test scores could be attributed to school effects (including teachers), but 60 percent of the variation was attributable to non-school effects such as individual and family characteristics.

10. See Curren, "Academic Standards and Constitutive Luck," 18–20. Curren cites an argument by John Rawls in regard to fair equality of opportunity.

11. It is also the case that only an individual's performance in schooling is measured. Examples abound of people for whom success in schooling was elusive, but who, upon their graduation (or expulsion), were able to succeed quite well in business or life projects.

10.47925/2012.331

12. Curren, "Academic Standards and Constitutive Luck," 29.

13. See Luc Pelletier and Elizabeth Sharp. "Administrative Pressures and Teachers' Interpersonal Behavior in the Classroom," *Theory and Research in Education* 7, no. 2 (2009): 174–183.

14. Edward Deci, Nancy Spiegel, Richard Ryan, Richard Koestner, and Manette Kauffman, "Effects of Performance Standards on Teaching Styles: Behavior of Controlling Teachers," *Journal of Educational Psychology* 74, no. 6 (1982): 852–859.

15. Ian Taylor and Nikos Ntoumanis, "Teacher Motivational Strategies and Student Self-Determination in Physical Education," *Journal of Educational Psychology* 99, no. 4 (2007): 747–760.

16. Richard Ryan and Netta Weinstein, "Undermining Quality Teaching and Learning: A Self-Determination Theory Perspective on High-Stakes Testing," *Theory and Research in Education* 7, no. 2 (2009): 224–233.

17. This includes the increasing use of "value-added" models. Recent studies have shown that despite early optimism, such models are as yet ineffective in identifying effective or ineffective teachers. See Baker, *Problems with the Use of Student Test Scores,* 2.

18. Gert Biesta, "Why 'What Works' Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research," *Educational Theory* 57, no. 1 (2007): 1–22.

19. Thomas Smith and Richard Ingersoll, "What Are the Effects of Induction and Mentoring on Beginning Teacher Turnover?," *American Educational Research Journal* 41 (2004): 682. See also Karen DeAngelis and Jennifer Presley, "Toward a More Nuanced Understanding of New Teacher Attrition," *Education and Urban Society* 43, no. 5 (2011): 599.

20. Though this correlation is found mostly in the early years of teaching. See Rivkin, Hanushek, and Kain, "Teachers, Schools, and Academic Achievement," 449, .

21. Dan Goldhaber and Emily Anthony, *Can Teacher Quality Be Effectively Assessed* (Washington, DC: Urban Institute, 2004).

10.47925/2012.331