Accounting for Teachers' Non-Academic Skill Contributions within Teacher Evaluation

Applied Policy Project

Prepared by John Wang for Denver Public Schools

April 2023

DENVER PUBLIC SCHOOLS
Discover a World of Opportunity™

UVA | FRANK BATTEN SCHOOL
of LEADERSHIP *and* PUBLIC POLICY

Acknowledgements

I would like to thank my partners, past and present, from Denver Public Schools – Bridgett Bird, Karen Buelow, Elizabeth Stock, and Leslie Juniel for agreeing to collaborate with me in this endeavor. Furthermore, I would like to thank the Teacher Workforce Collaborative, particularly Allison Atteberry, for facilitating access to data from Denver Public Schools that made these in-depth analyses possible as well as the invaluable feedback provided throughout the process.

I am grateful to Professor Jim Wyckoff and Professor Christopher Ruhm for their helpful feedback and suggestions that undoubtedly strengthened the clarity of the report. I also give my thanks to my APP group members Erica Sachs Langerhans, Arianna Khan, and Johnny Rogers for their support throughout the process along with the other Education Policy PhD students for their feedback on presentations.

Honor Statement

On my honor as a student, I have neither given nor received unauthorized aid on this assignment.

John Wang

Disclaimers

All results and writing here are the product of John Wang. Although the research analyses were prepared using data available from Denver Public Schools through a Memorandum of Understanding between the school district and the Teacher Workforce Collaborative, the content does not reflect the viewpoints of Denver Public Schools or the Teacher Workforce Collaborative. Similarly, the author conducted this study as part of the program of professional education at the Frank Batten School of Leadership and Public Policy, University of Virginia. This paper is submitted in partial fulfillment of the course requirements for the Master of Public Policy degree. The judgments and conclusions are solely those of the author, and are not necessarily endorsed by the Batten School, by the University of Virginia, or by any other agency.

As an "Internal Report" for DPS, John Wang, a member of the CU Core Team, cannot distribute the document in any public context; however, the MOU states that "The CU Core Team may only distribute Internal Reports confidentially to academic peers for informal feedback and graduate students for learning purposes". Furthermore, "Internal Reports will not be posted on any external website or cited by the CU Core Team". The audience is "limited to CU's Academic Peers," so it "may be shared or presented in an academic setting" but the "contents may not be circulated or cited".

Table of Contents

Executive Summary

Research demonstrates that students' academic and non-academic (e.g., *leadership*, *curiosity*, *growth mindset)* skills complementarily shape students' future life trajectories (Lundberg, 2017). Teachers help students improve these skills, but their effectiveness in improving academic and non-academic skills vary greatly (e.g., Hanushek and Rivkin, 2012). Teacher evaluation systems were designed to address the variation in effectiveness by capturing teachers' ability to improve students' academic skills, continue to develop teachers' pedagogical practices, and inform high-stakes personnel decisions like retention or tenure. Unfortunately, preliminary research evidence indicates that teachers' ability to add value to students' non-academic skills are minimally captured with teacher evaluation measures (Kraft, 2019).

Evidence directly from Denver Public Schools (DPS) show that teachers' value-added to the *Behavioral Index*— a latent non-academic skill that is the weighted average of students' Attendance, Suspensions, GPA, and Grade Progression— has no correlations with any of the districts' teacher evaluation measures. The implications result in teachers effective at promoting students' non-academic skills to disproportionately receive lower ratings, be less likely to be promoted, be more likely to leave, and be less likely to be recruited into DPS compared to teachers effective at promoting test scores. Denver Public School's teacher evaluation system, Leading Effective Academic Practice (LEAP), overlooks teachers' contributions to students' non-academic outcomes that are linked to not only students' academic outcomes but also their future career, college, and life outcomes.

Based on a review of scholarly literature and simulations using DPS data, four possible policy pathways were selected to address the teacher evaluation problem that DPS faces:

Alternative 1*: Status Quo*

The *Status Quo* represents Denver Public School's current plan to adjust Student Growth measures: 10% on School Collective, 10% on Student Learning Objectives, and 10% on Individual Statewide Test Growth.

Alternative 2: *Weight Changes*

Weight Changes leverages simulations to identify the Student Growth measure weights that maximize teachers' value-added to both academic and non-academic skills. The optimal weights are 6% on School Collective, 2% on Student Learning Objectives, and 22% on Individual Statewide Test Growth.

Alternative 3: *Replace Measure*

Replace Measure removes the School Collective measure and directly replaces the measure with teachers' Behavioral Index Value-Added (contributions to students' growth on a weighted average of Attendance, Suspensions, and Grade Progression). The Student Growth weights become 10% on Behavioral Index Value-Added, 10% on Student Learning Objectives, and 10% on Individual Statewide Test Growth.

Alternative 4: *Combine Replace Measure and Weight Changes*

The *Combine Replace Measure and Weight Changes* alternative integrates the two prior alternatives by 1) removing the School Collective measure and replacing the measure with teachers' Behavioral Index Value-Added; and 2) employing simulations to identify the Student Growth measure weights that maximize teachers' value-added to both academic and non-academic skills. The Student Growth weights are as follows: 18% on Behavioral Index Value-Added, 1% on Student Learning Objectives, and 11% on Individual Statewide Test Growth.

To evaluate these policy pathways, three criteria were considered: 1) **Effectiveness**, the ability of the policy alternative to capture teachers' academic and non-academic contributions, 2) **Political Feasibility**, the degree to which the alternative may be supported or opposed by key stakeholders, and 3) **Implementation Complexity**, the number of changes that must be made to existing LEAP evaluation processes.

The analyses of these policy pathways conducted later in this report indicate that ***Combine Replace Measure and Weight Changes*** is the most promising. Although the alternative may face some challenges with the support received by key stakeholders and require a number of changes to LEAP, the ability to capture teachers' contributions to students' non-academic skills are *substantially* improved under the alternative.

The implementation of the ***Combine Replace Measure and Weight Changes*** alternative is key for resolving some of the initial concerns with political feasibility. Fortunately, DPS has successfully worked with the Denver Classroom Teachers Association to successfully create the LEAP teacher evaluation system in the past (Jerald, 2013). By adapting key elements of the previous LEAP creation process— a multi-year phase-in process paired with systematic collection of teacher and evaluators' feedback— DPS can move the district forward. Teachers that advance students' non-academic skills will no longer have their contributions overlooked, ultimately improving students' future outcomes.

Introduction to Problem

Motivation

Students' academic skills undoubtedly shape their future life trajectory, a phenomenon extensively documented in the human capital literature (e.g., Murnane, Willett, and Levy, 1995); however, non-academic skills are not only important for students' future long-term outcomes (e.g., Heckman, Stixrud, and Urzua, 2006) but are increasingly rewarded in the U.S. labor market (Deming, 2017). Although students may attain non-academic skills in a variety of settings, a growing body of evidence indicates that teachers add value to a range of students' non-academic skills (e.g., Jennings and DiPrete, 2010)— assignment to teachers that are effective at adding value to these non-academic skills positively predicts future academic outcomes like high school graduation (Jackson, 2018) and college attendance (Backes et al., 2022b) as well as non-academic outcomes like increased earnings (Flèche, 2017). Furthermore, several outcomes like decreased arrests and incarcerations are *exclusively* predicted by teachers' non-academic skill value-added (VA) and not their test score VA (Rose, Schellenberg, and Shem-Tov, 2022). Like teachers' abilities to improve test scores (e.g., Hanushek and Rivkin, 2012), teachers' skill vary greatly on their ability to improve students' non-academic skills (Appendix Figure 1).

Existing teacher evaluation systems were designed with multiple measures to address the variation in teacher effectiveness in improving test scores; however, these evaluation systems *overlook* teachers' contributions to non-academic skills. Rigorous, experimental evidence using data from the Measures of Effective Teaching study find extremely weak correlations (r=0.19) between teachers' evaluation score on a composite measure (35% test score VA, 50% Observations, 5% Student Survey, and 10% Principal Ratings) and teachers' VA on key non-academic skills like growth mindset, grit, and effort (Kraft, 2019). The correlations between teachers' non-academic skill VA and each individual measure fare no better with correlations maxing out at r=0.20 (Kraft, 2019).

Denver Public School's (DPS) teacher evaluation system, Leading Effective Academic Practice (LEAP), *overlooks* teachers' contributions to non-academic skills. The creation of LEAP's Observation, Student Survey, and State Test Score Growth evaluation measures leveraged evidence provided from the Measures of Effective Teaching (DPS, 2022d); as a consequence, these measures similarly overlook teachers' contributions to students' non-academic skills. DPS also has additional measures such as Professionalism, Student Learning Objectives, and School Collective Growth, but there is little reason to believe that teachers' engagement with their colleagues (Professionalism) or students' mastery and growth on the same Colorado Academic Standards (Student Learning Objectives and School Collective Growth) as State Test Score Growth would somehow capture teachers' non-academic skill VA when other measures have failed to do so. Correlations between teachers' non-academic VA and LEAP performance, individual measures, measure indicators using DPS data all essentially have no relationship, corroborating the prior research literature. Overlooking teachers' non-academic contributions result in teachers effective at promoting students' non-academic skills to disproportionately receive lower ratings, be less likely to be selected as mentors (promotions), more likely to leave, and less likely to be recruited into DPS compared to teachers that are effective at promoting test scores. These teachers offer significant gains for students' future career, college, and life outcomes that subsequently do not happen if continually left unaddressed, making the problem even more urgent.

Problem Statement

Denver Public School's (DPS) teacher evaluation system, Leading Effective Academic Practice (LEAP), overlooks teachers' contributions to students' non-academic outcomes that are linked to not only students' academic outcomes but also their future career, college, and life outcomes.

Client Overview

Denver Public Schools Context

Denver Public Schools is a large, urban school district that serves 90,250 students across 207 schools (2022b) with an operating budget of $2,994,719,325 (2022c). In the previous year, DPS had 4,780 classroom teachers with 4,592 of those teachers undergoing teacher evaluation. The core mission of DPS is to "provide all students the opportunity to achieve the knowledge and skills necessary to become contributing citizens in our diverse society" (2022e). The district aspires to do so by cultivating a student experience that provides extended academic opportunities, emphases on the whole child, and the pursuit of passion (DPS, 2022a); DPS has made changes to its curriculum to realize these ambitions (Denver School Board, n.d.). Teachers therefore play a **central role**, so the teacher evaluation system needs to be adapted to match the changing expectations.

Leading Effective Academic Practice (LEAP)

Leading Effective Academic Practice was established in conjunction with Denver Public Schools and the Denver Classroom Teachers Association (DCTA) to 1) set clear expectations to assess teacher performance, 2) ensure excellent teachers in every classroom, and 3) support teachers (DPS, 2022d), and 4) comply with Colorado Law S.B. 10-191 (Colorado General Assembly, 2010). Since LEAP is tied to key personnel decisions such as earning or losing tenure and dismissal (Putnam, Ross, and Walsh, 2018), the LEAP Collaboration Committee, composed of members from both DPS and DCTA, work together to review concerns and make recommendations for future changes to LEAP to ensure transparency and fairness (2021). In particular, the LCC may revise the LEAP Fairness Guide—the explicit guidelines of how teacher evaluation must be implemented in a fair and transparent matter— as well as make *direct* changes to LEAP by consensus with any non-consensus recommendations brought before the Superintendent to decide (DCTA and DPS, 2021).

The creation of LEAP's measures heavily relied on evidence provided from the Measures of Effective Teaching, which was a landmark project that was designed to both identify and measure effective teaching (Gates Foundation, 2015). Through evidence from the Measures of Effective Teaching Project as well as identification of other important aspects of teaching, LEAP culminated in the following measures:

> <u>Observations</u>: Classroom visits by school leaders or peers to collect evidence of the alignment between teachers' classroom learning environment and instructional practices to the Framework for Effective Teaching;
>
> <u>Professionalism</u>: Teachers' actions beyond the classroom by understanding their students culturally and via data, engaging with colleagues and the community, as well as reflecting and developing themselves;
>
> <u>Student Perception Survey</u> (*SPS*): Students' rating of their teachers' ability to facilitate learning, support students, and communicate high expectations;
>
> <u>Student Growth</u>: Teachers' impact on students' academic learning and growth through Individual Statewide Test Growth, Student Learning Objectives, and School Collective Growth.

Changes to LEAP and Problem Interest

LEAP has been continuously evaluated and improved upon since its first implementation as seen in Figure 1. Although the actual weighting and subcomponents sometimes shift on a year-by-year basis due to changes in assessments, introduction as well as removal measures, and Covid-19, LEAP has weighed Observations, Professionalism, and Student Perception Surveys ("Professional Practice") at 50% of the evaluation and Student Growth as the remaining 50% of the evaluation.

The recent passage of S.B. 22-70 reallocates the overall weighting of *Student Growth* to a maximum of 30% and the remaining 70% toward *Professional Practice* measures (Colorado General Assembly, 2022). The law change, although not a major shift from the existing 50/50 split between *Student Growth* and *Professional Practice*, has initiated critical conversations within the LEAP Collaboration Committee and DPS Evaluation team to potentially transform major elements of the existing evaluation system like the removal of discretionary teacher ratings. Similarly, this law change serves as a major opportunity to re-examine LEAP's ability to capture teachers' contributions to students' non-academic skill development.

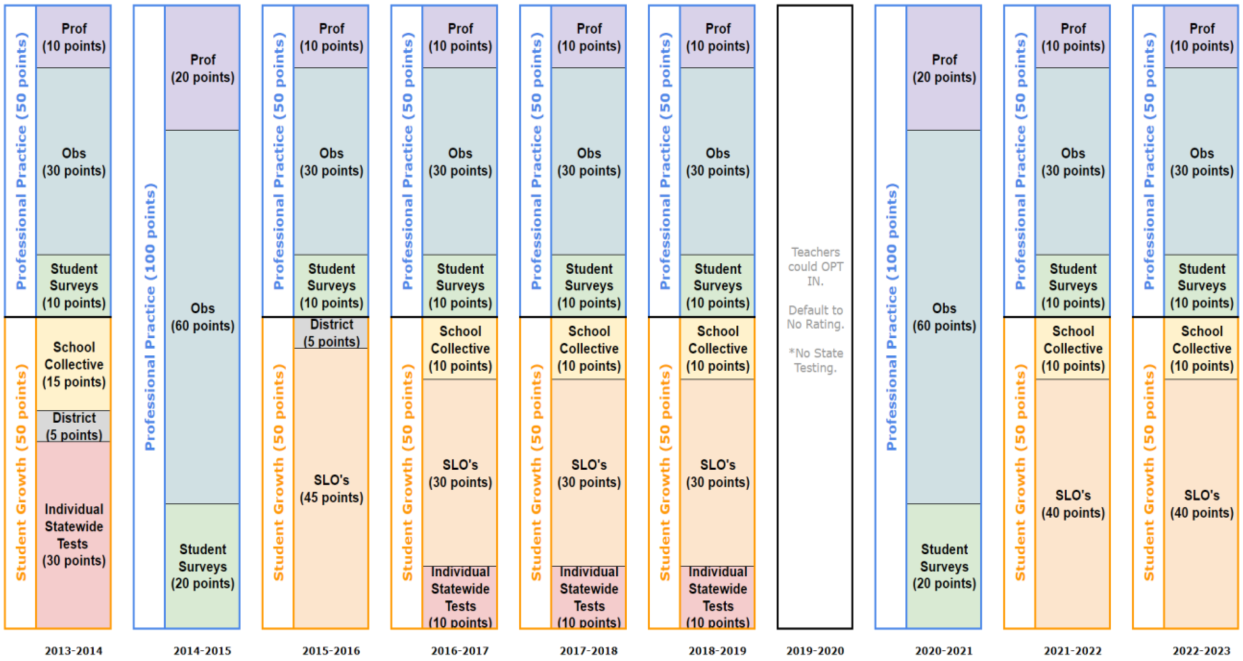


Figure 1: DPS LEAP Year-by-Year Changes

Literature Review

Value-Added

Researchers have been able to *isolate* individual teacher's contributions, or **value-added** (VA), to their students' learning. By controlling for the characteristics of students, classrooms, and schools (Todd and Wolpin, 2003), researchers predict how well a student is expected to perform; the difference between a student's actual performance and the student's predicted performance can be ***causally*** attributed to the teacher (Kane and Staiger, 2008; Kane et al., 2013). While prior research has focused on teachers' value-added to students' academic performance, scholars have begun to leverage information available from student surveys (e.g., Jennings and DiPrete, 2010) or students' observable behaviors like attendance records (e.g., Jackson, 2018) to capture teachers' value-added to students' non-academic skills[1]. The burgeoning research literature has led to additional insights on the importance of teachers.

Although teachers have already been shown to be the most important school-based factor on students' academic achievement (Goldhaber et al., 1999), teachers play a key role in shaping students' long-term outcomes. Students assigned to an effective teacher (84th percentile teacher) on academic value-added significantly improve their SAT Scores (Jackson, 2018; Petek and Pope, 2022; Backes et al., 2022b), AP Courses Taken and Credits Received (Liu and Loeb, 2021; Backes et al., 2022b), Selective College Attendance (Backes et al., 2022b), and Earnings (Chetty et al., 2014b). Similarly, assignment of students to teachers effective at improving the *Behavioral Index*[2]— a latent non-academic skill that is the weighted average of students' Attendance, Suspensions, GPA, and Grade Progression— leads to increased Graduation (Jackson, 2018; Gilraine and Pope, 2021; Rose et al., 2022; Backes et al., 2022b) and Enrollment in a 4-year College (Jackson, 2018; Rose et al., 2022; Backes et al., 2022b) as well as a significant decrease in future arrests and incarceration (Rose et al., 2022). Effective teachers are important, but not all teachers are effective.

Teacher Evaluation

Unfortunately, teachers are far from equally effective at improving students' academic (e.g., Hanushek and Rivkin, 2012) and non-academic skills (Appendix Figure 1), emphasizing the importance of teacher evaluation. While teachers have been evaluated in the past, most teachers received a perfunctory "satisfactory" rating (Weisberg et al., 2009). Contemporary teacher evaluation systems ushered in as part of Race to the Top reforms were intended to better capture differences in teacher effectiveness (McGuinn, 2012). Furthermore, these evaluation systems were theorized to better students' future outcomes by 1) changing the teacher workforce composition (e.g., firing or incentivizing teachers) and 2) improving the current teacher workforce (e.g., professional development). Although quasi-experimental evidence indicates that the average teacher evaluation system failed to meaningfully improve student achievement, graduation, and college enrollment (Bleiberg et al., 2021), the null effects mask success cases. Heterogeneity analysis on exemplary education systems (Dallas, **Denver**, DC, Newark, Tennessee, and New Mexico) conducted by Bleiberg and colleagues show medium-sized positive effects for exemplary teacher evaluation reforms (up to 0.15 SD) on academic achievement. Causal evidence suggests that the mechanisms play a role through either change in composition to the teacher workforce (Kraft et al., 2020; Dee and Wyckoff, 2015; Dee, James, and Wyckoff, 2021) or improvements in retained teachers' performance (Dee and Wyckoff, 2015).

[1] A degree of caution is warranted for interpreting the results of some non-academic skill VA: Backes and Hansen (2018) and Blazar (2018) respectively show that ***some*** individual observable behavior and survey value-added measures may be biased under non-experimental conditions. The focal non-academic skill measure, the *Behavioral Index*, has a robust evidence base that mitigates those concerns.

[2] Unfortunately, teacher VA to most non-academic skills has not yet been tracked into the long-term. Similarly, other non-academic skills may not have fully robust evidence to trust in the results..

Despite the ability for teacher evaluation to be effective in improving students' academic performance, growing evidence suggests that teacher evaluation may overlook teachers' value-added to students' non-academic skills across all measures. The foundational research for many teacher evaluation systems came from the Measures of Effective Teaching Study, which experimentally demonstrated that variations in teacher effectiveness could be captured by multiple measures (Kane et al., 2013). These multiple measures, however, were built and validated around teachers' academic VA, and the same data from the Measures of Effective Teaching study highlights the problem. Kraft leverages the same experimental data to estimate teachers' value-added to *Growth Mindset*, *Grit*, and *Effort*: he finds that these unbiased estimates of teachers' contributions to non-academic skills are very weakly correlated with teachers' evaluation score on a composite measure (35% Test Score VA, 50% Observations, 5% Student Survey, and 10% Principal Ratings) (2019). The correlations (linear relationship of two numerical variables) between teachers' non-academic skill VA and individual evaluation measures (Test Score VA, Observation, Student Survey, and Principal Ratings) fare no better with the highest correlation at r=0.20 (Kraft, 2019). Although another Measures of Effective Teaching study uses the non-experimental estimates of teacher value-added to a composite *Student Survey*, *Happiness in Class*, and *Effort* to show higher correlations with both Observations and a composite Student Survey measure (Mihaly et al., 2013), the discrepancy in results may be explained by evidence that non-experimental estimates of non-academic skill VA on *some* survey measures may be biased (Blazar, 2018). Outside of data from the Measures of Effective Teaching study, a robust evidence base correlates teachers' value-added to test scores, a common measure of contemporary teacher evaluation systems (Ross and Walsh, 2019), and non-academic skills: the consistent results indicate that the two measures are weakly correlated (Jennings and DiPrete, 2010; Gershenson, 2016; Blazar and Kraft, 2017; Jackson, 2018; Backes and Hansen, 2018; Blazar, 2018; Kraft, 2019; Liu and Loeb, 2021; Gilraine and Pope, 2021; Petek and Pope, 2022; Rose et al., 2022; Holt et al., 2022; Blazar and Pollard, 2022; Backes et al., 2022a). Since measures of teacher evaluation systematically miss teachers' non-academic skill contributions, serious issues may arise when considering the mechanisms of teacher evaluation.

The emphasis of teacher evaluation on teachers' contributions to students' academic skills[3] across multiple measures may penalize teachers through the two core mechanisms of teacher evaluation: 1) improving current teachers toward "effective" pedagogical practices, or 2) high-stakes personnel decisions. Under evaluation, teachers are often coached or mentored to improve their instructional practices— experimental and quasi-experimental evidence indicate that "effective" (test score VA) mentors improve their mentees' test score VA (Ronfeldt et al., 2018; Goldhaber et al., 2020), and retained teachers under evaluation exhibit continued gains in their performance (Dee and Wyckoff, 2015). However, the instructional practices deemed "effective" (i.e., promote students' academic skills) may narrow teachers' ability to promote non-academic skills, penalizing teachers that choose to continue promoting students' non-academic skills. In particular, the instructional practices that predict teachers' test score VA often fail to predict teachers' non-academic VA (Jennings and DiPrete, 2010; Flèche, 2017; Blazar and Kraft, 2017); in fact, some instructional practices demonstrate **negative trade-offs**: engaging in one set of instructional practices may improve test scores at the expense of non-academic skills and vice-versa (Bjorklund-Young and Ronda, 2017; Blazar and Pollard, 2022). Since teacher evaluation attempts to improve teachers' skill toward "effective" instructional practices (Donaldson and Firestone, 2021), teachers that want to improve non-academic skills may either not get the opportunity to do so or be penalized in high-stakes personnel decisions.

[3] This may not always be reflected in the actual weighting of VA but is reflected in the construction of priorities and existing choice of measures

The research evidence examining how teacher evaluation systems that overlook teachers' non-academic contributions penalize teachers in high-stakes personnel decisions is sparse, but initial evidence highlights the potential for penalties that impact students' long-term outcomes. Although teacher evaluations may be used for many forms of high-stakes personnel decisions (e.g., recruitment, tenure, pay incentives, promotions), only the firing of teachers has been *simulated* with teachers' non-academic contributions in mind. Several scholars have run simulations on a benchmark policy proposed by Hanushek (2009; 2011) and evaluated previously by Chetty et al. (2014b)— releasing the bottom 5% of the test score value-added distribution[4]. The results of the benchmark simulation using teachers' non-academic skills are illuminating: <u>nearly 85%</u> of teachers in the bottom 5% of test score VA distribution would be <u>retained</u> if the criteria shifted to replacing the bottom 5% of *Behavioral Index*[5] VA distribution (Mulhern and Opper, 2022). A school district with 4,000 teachers—a smaller number of teachers than DPS— may see up to 170 teachers terminated[6] under the benchmark policy who would not have been terminated under a bottom 5% of *Behavioral Index* VA policy. These simulations reveal that teachers are not only potentially penalized, but students outcomes are ultimately affected. Petek and Pope demonstrate that placing equal weight on both test score VA and the *Behavioral Index* VA can significantly reduce the likelihood of dropping out (0.8 P.P.), increase SAT taking (1.2 P.P.) with a negligible decline in students' test scores (~0.002 SD) (2022).The high-stakes personnel decisions, like a policy that replaces the bottom 5% of teachers, that come with teacher evaluation may inadvertently penalize teachers and ultimately impact students' future outcomes.

[4] In practice, 26 states have at most 50% of their teacher evaluation based on student growth (and not necessarily test score VA) and the remainder of their teacher evaluation on other measures

[5] Mulhern's and Opper's Behavioral Index is constructed as a composite of both contemporaneous and future attendance/grades

[6] 0.85*0.05*4000

Overlooking Teachers' Non-Academic Skill Contributions in Denver Public Schools

Limitations of Prior Teacher Evaluation Research

Denver Public Schools not only served as a participant in the Measures of Effective Teaching study but also leveraged the results from the various studies in constructing the LEAP Evaluation system. Many of the previously written insights on teacher evaluation systems overlooking teachers' contributions to students' non-academic skills apply to DPS. Kraft's correlation of teachers' Framework for Effective Teaching performance, the *same* observation framework that DPS uses, and teachers' non-academic VA cap out at a very weak correlation coefficient of 0.13 (2019). Similarly, DPS's data contributes to the weak correlations between teachers' non-academic skill VA and both test score VA and student surveys, which Kraft finds weak correlation coefficients of 0.14 and 0.20 respectively (2019). Although the empirical literature sheds light to the problem of LEAP overlooking teachers' non-academic contributions, there are also clear differences between the research conducted in the Measures of Effective Teaching study and the contemporary teacher evaluation system in Denver.

Teacher evaluation in DPS uses different measures, weights, and high-stakes personnel decisions. The researched measures for evaluating teachers are Test Score VA, Observations, Tripod Student Survey, and Principal Ratings in the Measures of Effective Teaching study; however, DPS has a different student survey and no principal ratings. Furthermore, DPS includes Professionalism, School Collective Growth, and Student Learning Objectives. These measures that are unaccounted for in the Measures of Effective Teaching research form 60% of teachers' evaluation in DPS, so the previous research may not be fully applicable. In addition to different measures and weights, DPS does not have a policy that replaces the bottom 5% of teachers; only teachers that receive a "Not Meeting" (~1% of teachers) are given a Performance Improvement Plan and eventually replaced if conditions are unmet (DPS and DCTA, 2017). Instead, personnel decisions that impact teachers are more likely to be either tenure-related or selection into (senior) team lead roles (Internal Teacher Workforce Collaborative Memo) with some potential for teacher turnover/recruitment impacts. To illustrate the problem in Denver Public Schools and expand beyond the limited research evidence base, data directly from the school district are leveraged to understand how well LEAP accounts for teachers' non-academic contributions and implications on teachers.

Measures and Methods

While DPS is interested in a wide array of non-academic skills, the *Behavioral Index* is prioritized as a key proxy for students' non-academic skills. The *Behavioral Index* essentially consists of a weighted average of students' Attendance, Suspensions, and Grade Progression. The measure (and underlying behaviors of Attendance, Suspensions, and Grade Progression) is theorized to capture latent aspects of students' *agreeableness*, *conscientiousness*, *neuroticism*, *persistence*, *grit*, *self-regulation*, *work ethic*, *passion for learning*, *student engagement*, and *belongingness to a positive and safe classroom* (Monk and Ibrahim, 1984; John et al., 1994; Barbaranelli et al., 2003; Duckworth et al., 2007; Baker et al., 2010; Kelly, 2012; Lounsbury et al., 2014; Ladd and Sorensen, 2017). Despite being a less defined construct, teachers' contributions to the *Behavioral Index* are predictive of desirable long-term outcomes like increased HS Graduation (e.g., Jackson, 2018) and College Enrollment (e.g., Backes et al., 2022b) as well as decreased Criminal Arrests (Rose et al., 2022). Alternative measures are either inaccessible to the research team (e.g., Behavior and Emotional Screening System; Student Perception Survey) or have insufficient coverage rates (e.g., YourVoice) to be used, while the underlying data for the *Behavioral Index* are universally collected and therefore prioritized by the researcher. The construction of the *Behavioral Index* is consistent with best practices from prior research literature (Jackson, 2018; Petek and Pope, 2021; Rose et al., 2022; Backes et al., 2022a;

Backes et al., 2022b) and demonstrates predictive validity on High School Graduation in DPS (Appendix).

Teacher value-added to the *Behavioral Index* was similarly constructed using standard best practices (Chetty et al., 2014a). Teacher value-added provides an unbiased estimate of how much an individual teacher contributes to their students' learning. The method predicts how well a student will perform on the *Behavioral Index* based on their previous Test Scores plus *Behavioral Index* performance along with individual, classroom, and school characteristics: the difference between the students' actual performance on the *Behavioral Index* and the predicted performance is the "value-added" of the teacher. The subsequent analyses align with best practices of using both prior test scores and the *Behavioral Index* for Math and ELA teachers (Jackson, 2018; Petek and Pope, 2021; Rose et al., 2022; Backes et al., 2022a; Backes et al., 2022b)[7]. Years are limited to 2017-2019 to represent *typical* LEAP evaluation (recall Figure 1 shows substantial year-by-year changes). Further technical details are available in the Appendix.

Empirical Evidence

Teachers' Value-Added to the *Behavioral Index* has **no relationship** with teachers' overall LEAP performance. Figure 2 shows scatterplots between teachers' value-added to the *Behavioral Index* (y-axis) and teachers' overall performance on LEAP (x-axis) with ELA teachers on the left and Math teachers on the right. The pattern between the teachers' value-added estimates and LEAP does not appear to trend strongly in any particular direction; the correlation coefficients of 0.02 and 0.04 similarly support the statement. In essence, there is **no relationship** between which teachers receive more or less favorable LEAP evaluations and which teachers have higher or lower Behavioral Index VA, providing supporting evidence that DPS's evaluation system may overlook teachers' non-academic contributions.

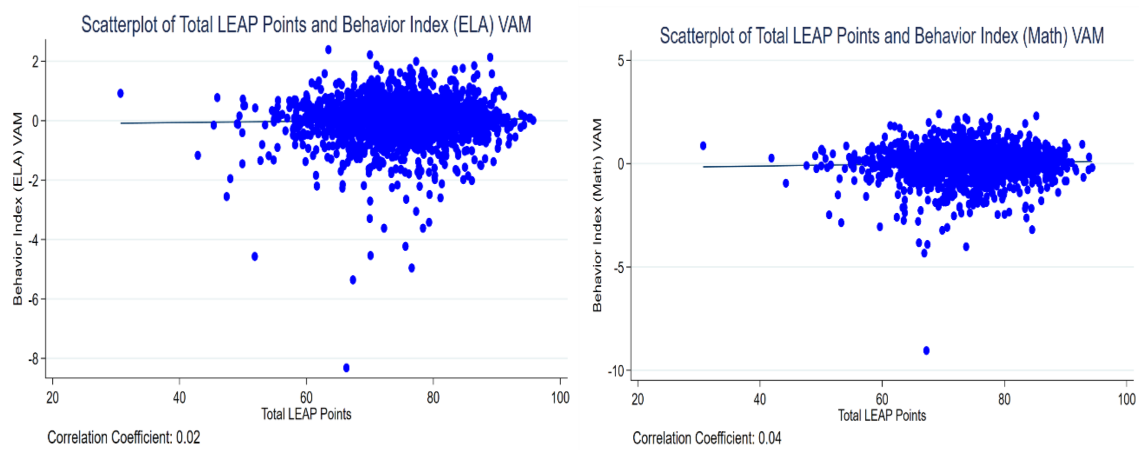


Figure 2: Scatterplots of LEAP Points and *Behavioral Index* VA

Teachers' Value-Added to the *Behavioral Index* has **no relationship** with teachers' performance on individual LEAP measures. Overall LEAP performance may mask how well individual measures capture teachers' non-academic performance due to the differences in weights. Table 1 displays the

[7] Prior test scores have previously been necessary for test score value-added due to concerns of certain students systematically enrolling in certain teacher's classrooms, which would bias estimates of teacher value-added. However, Blazar (2018) provides some suggestive evidence that prior test scores may not be necessary to estimate teachers' value-added to non-academic skills since students may not be fully sorted into classrooms based on their non-academic skills.

correlation coefficients between teachers' value-added (ELA Test, Math Test, *Behavioral Index* – separate for ELA and Math teachers) and their performance on both the overall and individual measures of LEAP[8]. Evidence suggests that it is possible for the Overall LEAP performance to mask how well individual measures capture teachers' value-added—ELA and Math Test scores are weakly correlated with overall LEAP (r=0.30 and r=0.35) but are strongly, positively correlated with Individual Statewide Tests (r=0.68 and r=0.72). However, this is not the case with teachers' *Behavioral Index* value-added: not only is Overall LEAP performance uncorrelated with teachers' non-academic skill value-added but so is every single individual measure. All correlation coefficients are between -0.02 and 0.09, which would be considered **no relationship**. Further robustness checks on individual indicators of Professionalism, Observations, and Student Perception Survey (Appendix Tables 2-4) are consistent with the **no relationship** results.

VA Measure	Overall	Professionalism	Observations	Student Perception Survey	School Collective	Student Learning Objectives	Individual Statewide Tests
ELA Test VA	0.30	0.15	0.18	0.10	0.24	0.08	0.68
Math Test VA	0.35	0.15	0.20	0.14	0.23	0.13	0.72
Beh Indx (ELA) VA	0.03	0.03	0.02	0.02	0.04	-0.02	0.09
Beh Indx (Math) VA	0.04	0.01	0.04	0.00	0.04	0.01	0.04

Table 1: Correlation Coefficients of Value-Added and Individual LEAP Measures

The empirical evidence for Denver Public Schools indicates that LEAP systematically overlooks teachers' contributions to students' non-academic skills. If LEAP somewhat captured these skills, a much stronger positive correlation —or at minimum, a correlation coefficient similar to test score VA —would be expected. Despite the differences in both measures as well as weights, the evaluation system in DPS is consistent with the prior research literature showcasing very weak to no relationship between teachers' non-academic value-added and teachers' performance in evaluation. Teachers' performance on these measures ultimately shape the rating that they receive which subsequently shape how teachers are affected by high-stakes personnel decisions. The next set of figures show the implications of an evaluation system overlooking teachers' contributions in DPS by unveiling what happens to teachers that are neither skilled at promoting academic nor non-academic skills, skilled at improving one or the other, or skilled at improving both. In particular, Math and ELA teachers were separated into quintiles for both academic test score and *Behavioral Index* value-added; teachers were then grouped according to their joint effectiveness—bottom 40% of performance in both test scores and *Behavioral Index*, top 40% in test scores but bottom 40% in

[8] The individual measures are technically "Percent of Possible LEAP Points Received" since some teachers may not have a Student Perception Survey and therefore have a larger weight on Professionalism and Observations.

Behavioral Index, bottom 40% in *Behavioral Index* but top 40% in test scores, and top 40% in both test scores and *Behavioral Index*.

Teachers that are effective in Test Score VA Disproportionately Receive Higher LEAP Ratings (as intended), but teachers that are effective in Behavioral Index VA Disproportionately Receive Lower LEAP Ratings. The immediate impact of an evaluation system that overlooks teachers' non-academic skill contributions is on the ratings that teachers receive in their evaluation. The left column, Rating Distribution, in Figures 3&4 consists of how many teachers (ELA or Math) receive a rating of "Distinguished," "Effective," "Approaching," or "Not Meeting" while each subsequent column shows the rating distribution of teachers in the four groups (Test/BI Bot 40%; Test Top 40% BI Bot 40%; Test Bot 40% BI Top 40%; and Test/BI Top 40%). Teachers that fall into the test top 40% (i.e., Test Top 40% & BI Bottom 40% or Test/BI Top 40%) are overrepresented in the "Distinguished"; those same teachers are underrepresented in the lower ratings of "Approaching" and "Not Meeting". These results indicate that teacher evaluation is working as intended by elevating teachers that promote students' academic skills. However, the teachers that fall in the top 40% of *Behavioral Index* VA receive "Distinguished" ratings at lower rates[9] compared to the baseline rating distribution. Furthermore, these same top 40% teachers in promoting the *Behavioral Index* receive similar levels of unsatisfactory ratings of "Approaching" and "Not Meeting" for Math teachers or substantially more unsatisfactory ratings for ELA teachers[10]. Since these ratings are often used for high-stakes personnel decisions, these rating differences culminate in different opportunities for teachers.

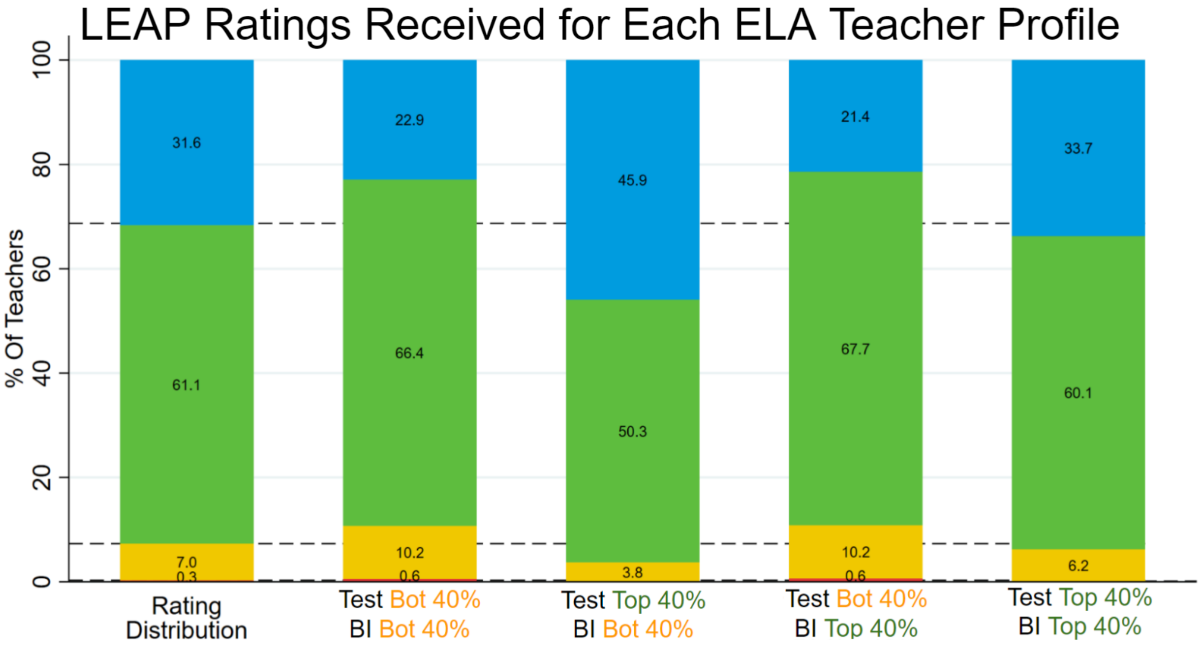


Figure 3: LEAP Ratings Received for ELA Teacher Profiles

[9] ELA: (21.4%+33.7%)/2 = 27.55% [Lower than 31.6% baseline of the ELA teacher rating distribution]; Math: (16.3%+31.8%)/2 = 24.05% [Lower than 25% baseline of Math teacher rating distribution]

[10] ELA: (10.2%+0.6%+6.2%+0.0%)/2 =8.5% [Greater than 7.3% baseline of Approaching/Not Meeting ELA teacher rating distribution]; Math: (13.3%+1.1%+3.6%+0.0%) = 9% [Similar to 9.2% baseline of Approaching/Not Meeting Math teacher rating distribution]

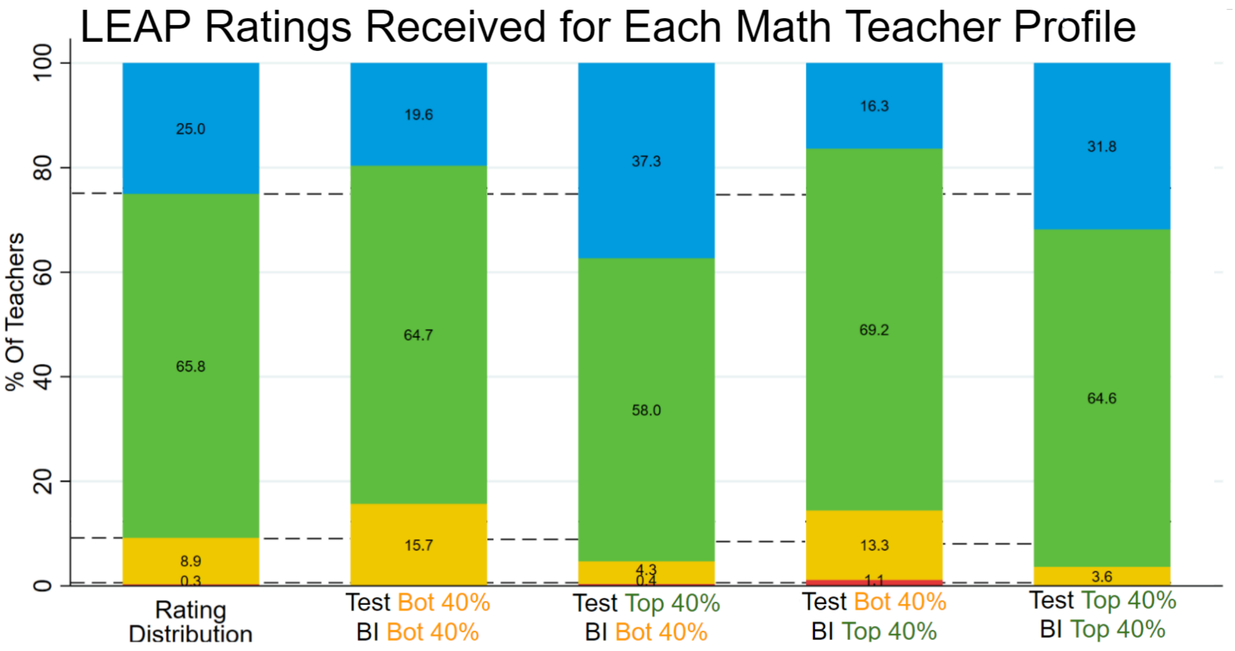


Figure 4: LEAP Ratings Received for Math Teacher Profiles

Teachers selected as Mentors are more likely to be in the top 40% of test scores than *Behavioral Index*, raising concerns of the instructional practices prioritized and teachers that receive those opportunities. Mentors, or (Senior) Team Leads, play an instrumental role in teachers' evaluation especially with respect to improving teachers' instructional practices (LEAP HANDBOOK). Since mentors are two to three times more likely to be "Distinguished" compared to mentees (Internal Teacher Workforce Collaborative Memo) and previous evidence indicates that teachers in the top 40% of test score VA are more likely to be "Distinguished," there is a concern that teachers effective at promoting students' non-academic skills receive fewer promotion opportunities. Furthermore, these mentors skilled at promoting students' academic skills may prioritize instructional practices that improve academic skills at the expense of students' non-academic skills. Although mentees appear to be closer balanced in their value-added to test scores and the *Behavioral Index* performance[11], the evidence in Figure 5&6 suggests that mentor teachers are more likely to be in the top 40% of test score VA and less likely to come from the top 40% of *Behavioral Index* VA[12]. It appears that promotions are disproportionately given to teachers effective at promoting test scores to the detriment of teachers effective at improving students' non-academic skills. The larger concentration of mentors that are in the top 40% of test score VA over *Behavioral Index* VA may subsequently lead to mentees becoming more effective at instructional practices that add value to students' academic skills, which may culminate in negative effects on students' non-academic skill development.

[11] Top 40% test [29.2% ELA; 30.9% Math] and top 40% BI [29.5% ELA; 29.1% Math]

[12] Top 40% of tests [37.4% ELA; 44.2% Math] than Top 40% of BI [33.7% ELA; 40.1% Math]

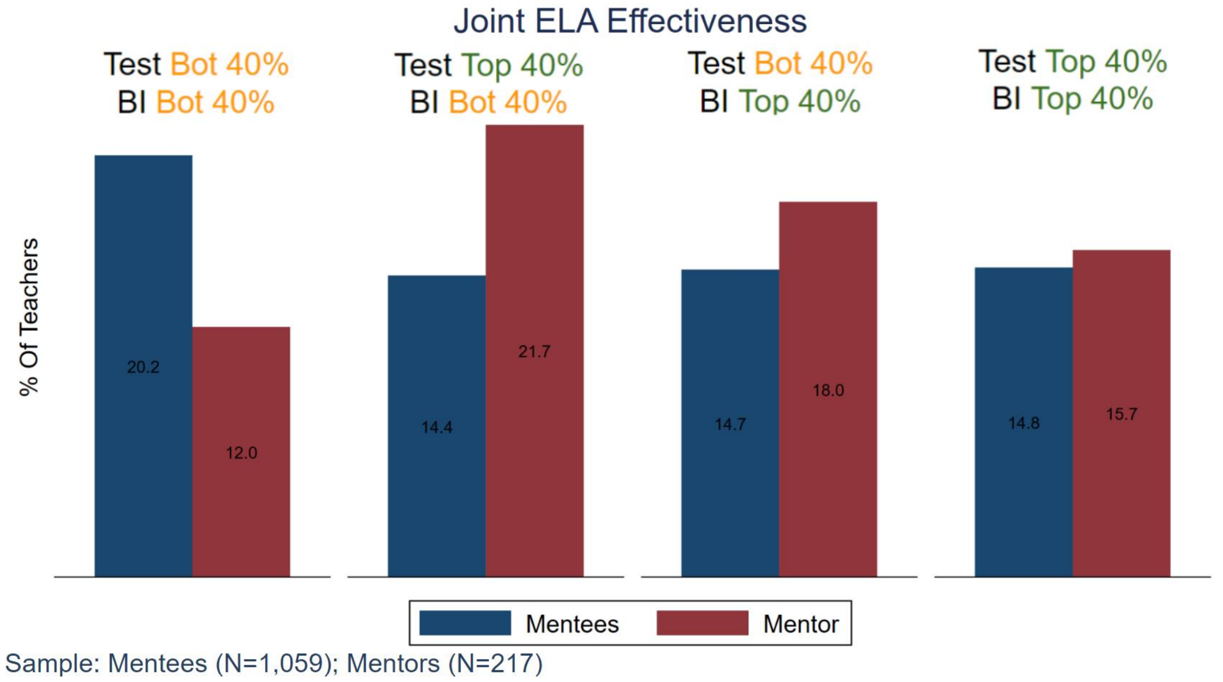


Figure 5: Teacher Profile of ELA Mentor and Mentee Teachers

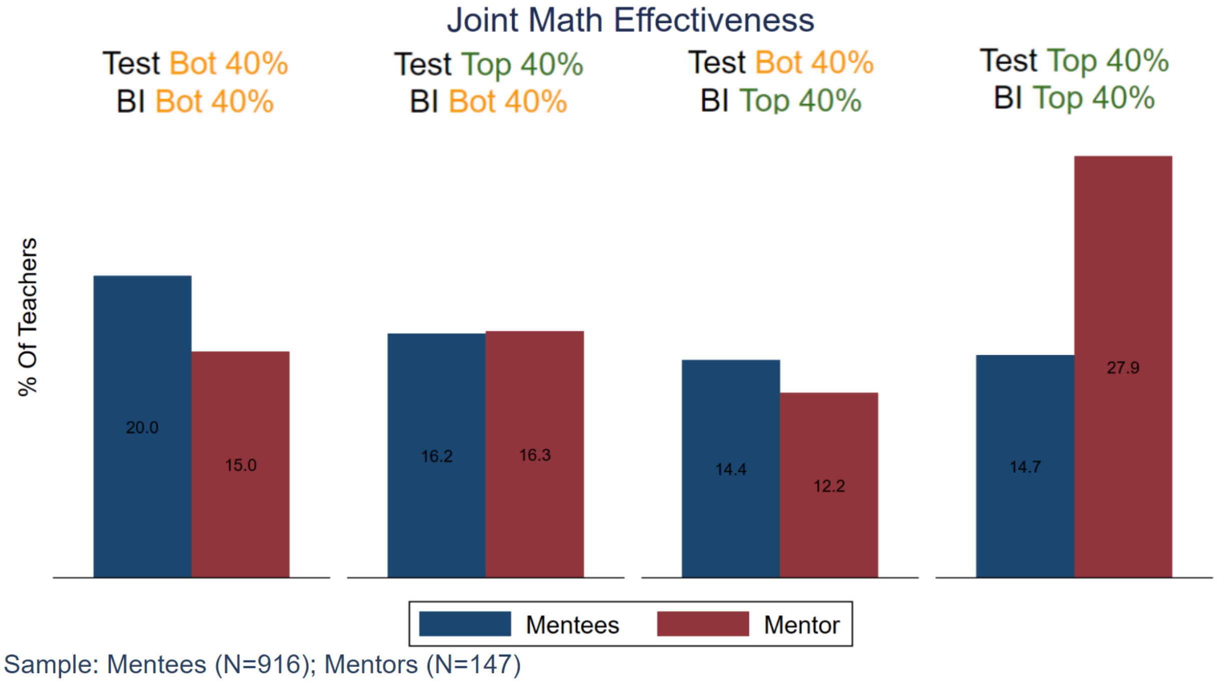
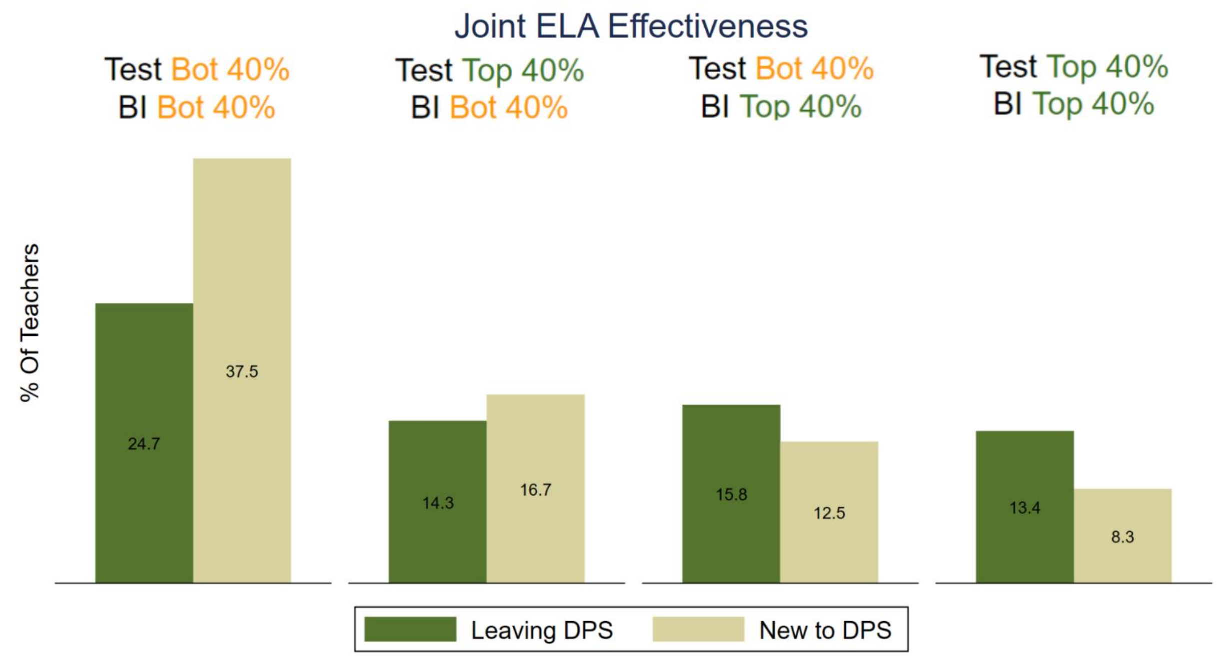


Figure 6: Teacher Profile of Math Mentor and Mentee Teachers

Teachers that leave are more likely to be in the top 40% of *Behavioral Index* VA and the teachers that are new to DPS are more likely to be in the top 40% of test score VA. Although DPS policies rarely dismiss teachers (i.e., Performance Improvement Plan for the ~1% "Not Meeting" before dismissal), teachers may still leave if they are dissatisfied with accountability (Sutcher et al., 2016), and teachers recruited in their place may reflect who is perceived to be an "effective" teacher by school leaders (Jacob and Lefgren, 2008). It is important to note that the evidence is ultimately descriptive since teachers can leave for a multitude of reasons completely unrelated to evaluation (e.g., retirement); similarly, the teachers recruited do not reveal underlying shifts of teachers within the district (hence the much smaller sample size). The preliminary evidence in Figure 7&8 suggests that that teachers that leave DPS are more likely to be in the top 40% of the *Behavioral Index* VA compared to the top 40% of the test score VA[13]. Some of the teachers that replaced the leaving teachers are much more likely to be in the top 40% of test score VA compared to the top 40% of *Behavioral Index* VA[14]. The disproportionality of teachers leaving and being recruited into the district may reflect the prioritization of DPS's evaluation and a small degree of penalties imposed for teachers that are effective at promoting students' non-academic skills.

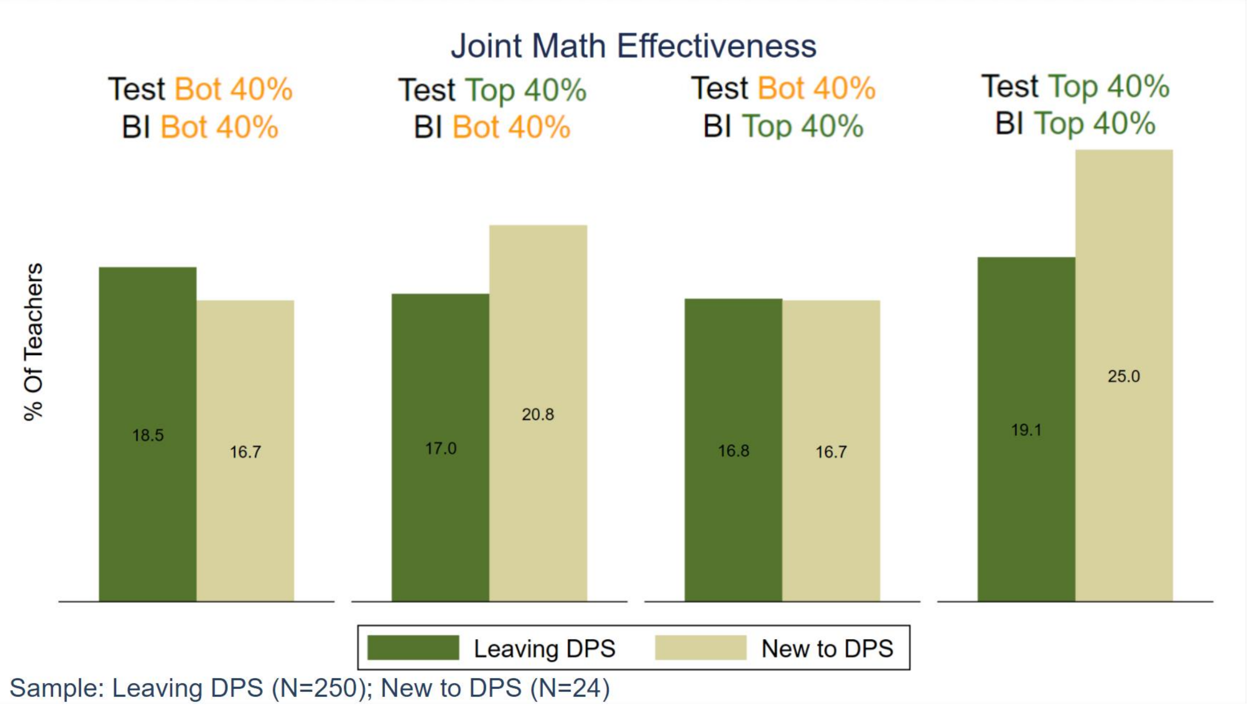


Sample: Leaving DPS (N=312); New to DPS (N=24)

Figure 7: Teacher Profile of ELA Leavers and New to DPS Teachers

[13] Teachers that leave – top 40% of tests [27.7% ELA; 36.1% Math] versus top 40% of BI [29.2% ELA; 35.9% Math] – BI more likely to leave

[14] New Teachers – top 40% of tests [25% ELA; 45.8% Math] versus top 40% of BI [20.8% ELA; 41.7% Math]



Sample: Leaving DPS (N=250); New to DPS (N=24)

Figure 8: Teacher Profile of Math Leavers and New to DPS Teachers

Summary of Problem in DPS

Denver Public Schools' teacher evaluation system overlooks teachers' contributions to students' non-academic skills. Evidence suggests that not only does teachers' value-added to the *Behavioral Index* have no relationship with teachers' overall LEAP performance, but there is also no relationship with individual LEAP measures or underlying indicators of individual measures. The implications result in teachers effective at promoting students' non-academic skills to disproportionately receive lower ratings, be less likely to be selected as mentors, more likely to leave, and less likely to be recruited into DPS compared to teachers that are effective at promoting test scores. Given the DPS's desire to emphasize students' non-academic skills, the teacher evaluation system needs to be improved to account for teachers' non-academic contributions.

Policy Pathways and Evaluation Criteria

Possible Policy Pathway Evidence

DPS has the following broad policy pathways that the district can pursue instead of the *Status Quo*: 1) *Change Weights of Existing Teacher Evaluation Measures*; 2) *Replace Measure with New Measure*; or 3) *Combine Replace Measure and Weight Changes*. Since LEAP has undergone many changes in weighting of teacher measures in the past decade (Figure 1), a natural policy option would be to simply change the existing weights of LEAP to better capture teachers' non-academic skills. Unfortunately, the research evidence (e.g., Kraft, 2019) and data from DPS (Tables 1/ Figures 2-8) suggests that current teacher evaluation measures overlook teachers' contributions to students' non-academic skill development, placing limitations on the *Change Weights* option. The limitations of *Changing Weights* may be resolved through the introduction of a new measure that directly replaces an existing measure to better capture teachers' non-academic skills. In particular, the *School Collective Growth*—a measure that provides all teachers at the same school with the same score and minimally distinguishes between teachers— could be replaced with the *Behavioral Index* VA, which has extensive research supporting its predictive validity for desirable long-term outcomes. The final option could consider the combination of the two by directly replacing *School Collective Growth* with the *Behavioral Index* VA and changing the weights of the remaining measures. Consistent research demonstrates that larger weights on measures directly assessing teachers' VA will best capture teachers' effectiveness on that measure (Mihaly et al., 2013; Kane and Staiger, 2012), so the inclusion of a measure directly assessing teachers' non-academic skill contributions as well as subsequently changing the weights will likely improve LEAP's ability to capture teachers' effectiveness in promoting these non-academic skills. However, DPS is not interested in only promoting students' non-academic skills at the expense of students' academic skills, so a policy pathway must account for both.

Simulations were conducted to identify the combination of *Student Growth* measure weight changes that optimized teachers' effectiveness at promoting academic and non-academic skills. The LEAP evaluation system consists of two broad categories of measures: Professional Practice (70% weight) and Student Growth (30% weight). The first simulation iterated through every possible combination of weight changes to the Student Growth measure[15]. Each iteration predicted the test score and *Behavioral Index* value-added of a teacher that performed perfectly on LEAP (earns all 100 points); the weights that maximized both test score and *Behavioral Index* VA were stored (Appendix). The best combinations informed the creation of the *Weight Changes* policy alternative. Another simulation directly replaced the School Collective Growth measure with teachers' points earned from *Behavioral Index* VA (point conversion detailed in Appendix but parallels Individual Statewide Test Growth point conversion) to similarly iterate through every possible combination of Student Growth weight change and capture the combination that maximized both test score and *Behavioral Index* VA predictions for teachers that earned all 100 points. The simulation results also inform the evaluations of alternatives for the *Status Quo* and *Replace Measure* solutions.

[15] Another version iterated through both Professional Practice and Student Growth measure changes – total of 1,155,525 iterations– and attained nearly identical results.

Policy Alternatives

Alternative 1*: Status Quo*

The *Status Quo* represents Denver Public School's current plan to adjust Student Growth measures: 10% on School Collective, 10% on Student Learning Objectives, and 10% on Individual Statewide Test Growth.

Alternative 2: *Weight Changes*

Weight Changes leverages simulations to identify the Student Growth measure weights that maximize teachers' value-added to both academic and non-academic skills. The optimal weights are 6% on School Collective, 2% on Student Learning Objectives, and 22% on Individual Statewide Test Growth.

Alternative 3: *Replace Measure*

Replace Measure removes the School Collective measure and directly replaces the measure with teachers' Behavioral Index Value-Added (contributions to students' growth on a weighted average of Attendance, Suspensions, and Grade Progression). The Student Growth weights become 10% on Behavioral Index Value-Added, 10% on Student Learning Objectives, and 10% on Individual Statewide Test Growth.

Alternative 4: *Combine Replace Measure and Weight Changes*

The *Combine Replace Measure and Weight Changes* alternative integrates the two prior alternatives by 1) removing the School Collective measure and replacing the measure with teachers' Behavioral Index Value-Added; and 2) employing simulations to identify the Student Growth measure weights that maximize teachers' value-added to both academic and non-academic skills. The Student Growth weights are as follows: 18% on Behavioral Index Value-Added, 1% on Student Learning Objectives, and 11% on Individual Statewide Test Growth.

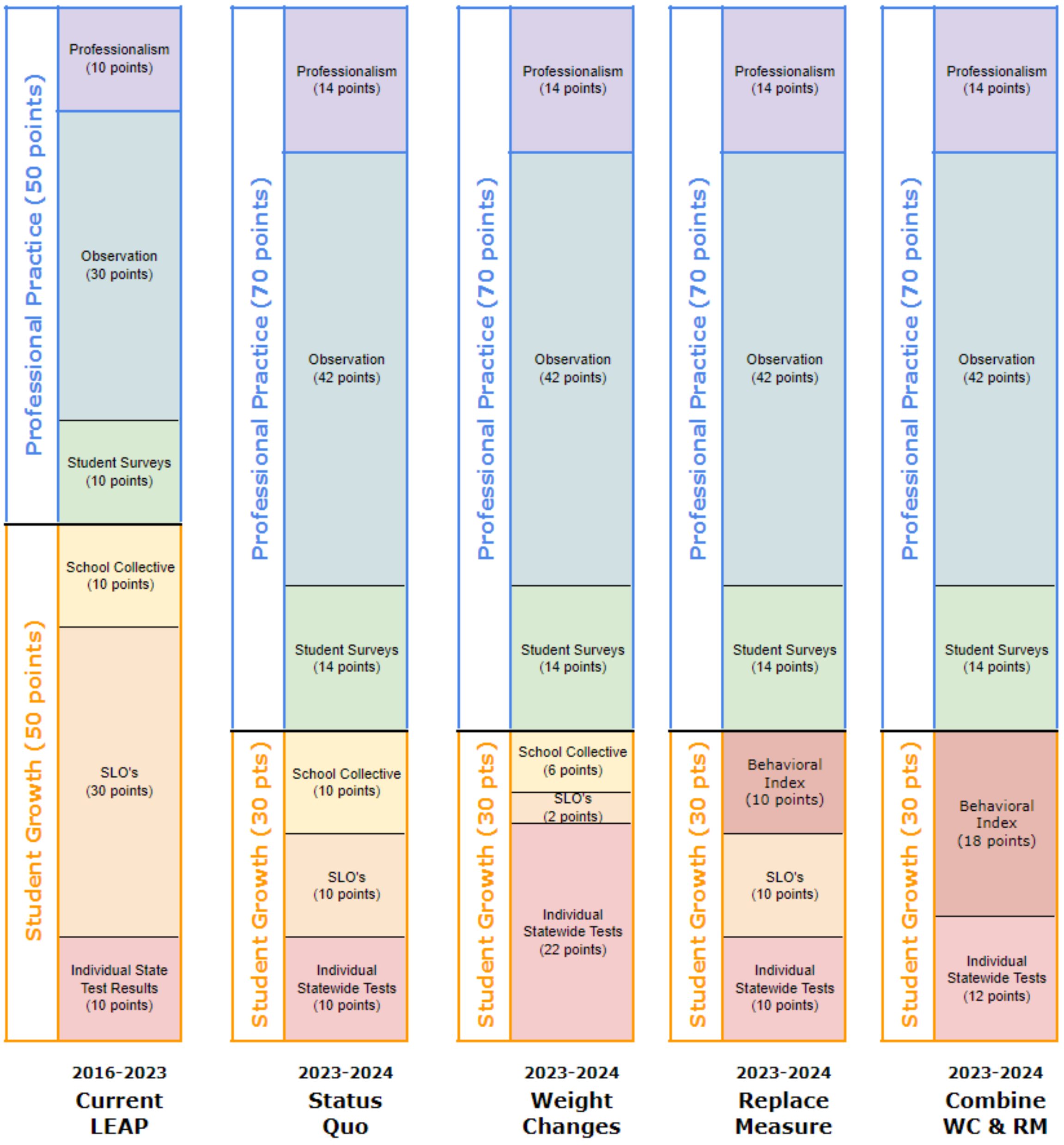


Figure 9: Current LEAP and Potential Policy Pathways

Evaluation Criteria

Effectiveness (50/100 Points)

The Effectiveness criterion evaluates teachers' predicted value-add to students' academic and non-academic skills for each alternative of a teacher earns all 100 points derived from the simulations described above. The predicted academic and non-academic value-added are converted to percentiles, averaged, and subtracted by 50[16] to compute the number of points earned. Further details and calculations are available in the Appendix.

Political Feasibility (25/100 Points)

The Political Feasibility criterion examines how key stakeholders (Teachers Union, School Board, Superintendent, DPS Evaluation Team, and the Public) are projected to respond to each alternative based on historical evidence, internal documents, conversations, surveys, and contemporary goals. Stakeholders projected responses will be categorized as "Very Opposed (1 point)," "Slightly Opposed (2 points)," "Neither Opposed Nor Favored" (3 points)," "Slightly Favored (4 points)," or "Very Favored (5 points)" for each alternative and summed to evaluate the alternative's Political Feasibility. Details are available in the Appendix.

Implementation Complexity (25/100 Points)

The Implementation Complexity criterion counts the number of changes DPS must make to existing LEAP Evaluation processes (e.g., LEAP Handbook rubrics), Accuracy/Fairness of Evaluations (e.g., data collection processes or conditions for valid evaluations), and Training (e.g., professional development of teaches and evaluators). Details of each can be found in the Appendix. Points are awarded according to the following system: 0-5 changes receive 25 points; 6-10 changes receive 20 points; 11-15 changes receive 15 points; 16-20 changes receive 10 points; and 21-25 changes receive 5 points. The full list of possible changes can be found in the Appendix.

[16] A teacher who scores perfectly on LEAP may be predicted to have a *Behavioral Index* value-add of 1 SD above the average teacher's *Behavioral Index* value-add. Converted to a percentile, that teacher who scores perfectly on LEAP would be in the 84th percentile compared to the average teacher, who is by definition in the 50th percentile. The subtraction of 50 from the percentile supports the ability to predict how a teacher performs relative to an average teacher.

Evaluating Alternatives

Alternative 1: *Status Quo*

Effectiveness: 26 Points

The *Status Quo* alternative is effective at predicting teachers' academic value-added but does not differentiate teachers' non-academic value-added. A teacher who receives a perfect score (100 points) on the LEAP evaluation using the Status Quo weights and measures is predicted to be in the 97.55th percentile of ELA/Math value-added, but only in the 54.77th percentile of Behavioral Index value-added. Therefore, the *Status Quo* captures and differentiates teachers' academic effectiveness but not their non-academic effectiveness. After taking the average of the predicted percentiles (~76th) for perfect-performing teachers on academic and behavioral index value-added, and subtracting it from 50, the *Status Quo* alternative receives 26 points for the Effectiveness criterion.

Political Feasibility: 20 Points

Most stakeholders, with the exception of the DPS Evaluation Team, either favor the *Status Quo* or do not oppose it. The Public will be "Neither Opposed Nor Favored" because the alternative, like previous LEAP changes, is unlikely to generate much public discussion or controversy, leading to no reaction. The Teachers Union, School Board, and Superintendent are all expected to be "Very Favored" towards the *Status Quo*. The Teachers Union has historically opposed value-added (McGuinn, 2012) and still does not want them in evaluation (NEA, 2020), and the School Board and Superintendent are backed by the Teachers Union and are likely to follow its lead or risk removal (see Appendix for "Denver Public Schools Context" and "Teacher Union's Influence on School Board and Superintendent Political Feasibility"). The DPS Evaluation Team is the only stakeholder that is "Slightly Opposed" to the *Status Quo*, as they are interested in better capturing teachers' contributions to students' non-academic skills. However, they still believe that LEAP works well overall. Based on these factors, the *Status Quo* earns 20 points for the Political Feasibility criterion.

Implementation Complexity: 10 Points

The *Status Quo* alternative is essentially a replication of DPS's current teacher evaluation system, so DPS would only need to make a few adjustments to LEAP Evaluation Processes, Accuracy/Fairness of Evaluations, and Training. Specifically, DPS would need to make seven changes to LEAP Evaluation Processes (e.g., LEAP Collaboration Committee Conversations or LEAP Ratings Cut-Offs) to adhere to agreements with the Teacher Union (DPS and Denver Classroom Teachers Association, 2017; DPS and Denver Classroom Teachers Association, 2021), integrate two changes to LEAP Fairness Guides to ensure the accuracy/fairness of evaluations (e.g., Refine Minimum Conditions Before Evaluation and Adjust "Redress" Processes), and incorporate seven changes to existing training (e.g., Identify New Training for Teachers on Instructional Practices) due to the lowered emphasis on Student Learning Objectives compared to the current state of LEAP. Overall, the total number of changes required for the Status Quo is 16, earning 10 points for the Implementation Complexity criterion.

Alternative 2: *Weight Changes*

Effectiveness: 27 Points

Weight Changes demonstrate negligible improvements for predicting teachers' academic and Behavioral Index value-added compared to the *Status Quo*. The predicted percentiles for teachers that earn 100 points on each academic (98.36th) and behavioral index (55.17th) value-added, maintaining the inability to differentiate teachers' non-academic effectiveness. After taking the average of the predicted percentiles (~77th) for perfect-performing teachers on academic and behavioral index value-added, and subtracting it from 50, the *Weight Changes* alternative receives 27 points for the Effectiveness criterion.

Political Feasibility: 8 Points

Due to the emphasis of *Weight Changes* on Individual Statewide Test Growth (22% from 10%), most stakeholders are projected to oppose *Weight Changes*. The Public will still be "Neither Opposed Nor Favored" since the alternative, like prior LEAP changes, will not appear in public discourse (see "Public and Political Feasibility" in Appendix), culminating in no reaction. However, the Teachers Union is projected to be "Very Opposed" as suggested by prior reluctance to have weights above 10% on Individual Statewide Test Growth (see "Teacher Union's Political Feasibility" in Appendix). The union-backed School Board and Superintendent are projected to match the Teachers Union in being "Very Opposed" as well (see Appendix for "Denver Public Schools Context" and "Teacher Union's Influence on School Board and Superintendent Political Feasibility"). The DPS Evaluation Team is still "Slightly Opposed" due to the group's desire to better capture teachers' contributions to students' non-academic skills in its evaluation system. As a result, *Weight Changes* receives 8 points toward Political Feasibility

Implementation Complexity: 10 Points

Weight Changes is similar to the *Status Quo* with the need for many of the same modifications to the LEAP Evaluation Processes (7) and Training (7). However, the added emphasis on Individual Statewide Test Growth (22% vs. 10%) for the Weight Changes alternative heightens the need for robustness checks on growth measures to ensure that these measures can be validly used for teachers' evaluation when the stakes have increased. As a result, the number of changes to Accuracy/Fairness of Evaluation increases to 3, bringing the total number of changes to 17. The Implementation Complexity score for the Weight Changes alternative is therefore 10 points.

Alternative 3: *Replace Measure*

Effectiveness: 38 Points

The *Replace Measure* alternative captures teachers' academic contributions while enhancing the ability to capture teachers' non-academic contributions. Replacing the School Collective Growth measure with teachers' *Behavioral Index* Value-Added preserves teachers' academic effectiveness. The predicted percentiles for teachers that earn 100 points on each academic (97.35th) and behavioral index (79.10th) value-added; after taking the average of the predicted percentiles (~77th) for perfect-performing teachers on academic and behavioral index value-added, and subtracting it from 50, *Replace Measure* receives 38 points for the Effectiveness.

Political Feasibility: 12 Points

Although stakeholders care about students' non-academic skill development, the incorporation of a value-added measure will still be met with opposition for all stakeholders expect the Public and DPS Evaluation Team. The Teachers Union has demonstrated resistance to value-added measures (McGuinn, 2012; NEA, 2020) and will continue to be "Very Opposed". Similarly, the union-backed School Board and Superintendent, despite the explicit goals to improve students' non-academic skills (DPS School Board, n.d.; Asmar, 2022), is still projected to follow the response of the Teachers Union or risk replacement. However, the unprecedented incorporation of non-academic value-added would likely generate public discourse, as the introduction of test score value-added had done previously (McGuinn, 2012), prompting a reaction by the Public. The Public is projected to be "Slightly Favored" due to the interest in supporting students' social/emotional needs (Harstad, 2021). The DPS Evaluation Team also becomes "Very Favored" in response to the evaluation shift since the alternative still captures the elements of LEAP that works well while capturing teachers' non-academic contributions. The number of points earned by the *Replace Measure* alternative is 12.

Implementation Complexity: 5 Points

Replace Measure overhauls teacher evaluation, making changes to the LEAP Evaluation Process, the Accuracy/Fairness of Evaluations, and Training. The implementation for LEAP Evaluation Processes and Training—similar to other alternatives— results in 7 changes each. Unlike other alternatives, significant modifications occur for Accuracy/Fairness of Evaluations: DPS would need to enhance Data Collection Processes (e.g., obtain granular data for students at classroom level and add data audits) as well as make clarifications to the LEAP Fairness Guide (e.g., how is Discipline defined and attributed to teachers?) that culminates in 11 changes. Given the 25 changes required to implement the *Replace Measure* alternative, the Implementation Complexity score is 5 points.

Alternative 4: *Combine Replace Measure and Weight Changes*

Effectiveness: 43 Points

Combine Replace Measure and Weight Changes strengthens LEAP's ability to predict teachers' non-academic contributions while upholding the ability to capture teachers' academic contributions. A teacher that earns a perfect LEAP performance is predicted, on average, to be in the ~89th percentile of Behavioral Index value-added and ~97th percentile of ELA/Math value-added. The *Combine Replace Measure and Weight Changes* receives 43 points after taking the average of the predicted percentiles (~93rd) for perfect-performing teachers on academic and behavioral index value-added and subtracting it from 50. The enhancement of LEAP's ability to capture teachers' non-academic contributions without substantial trade-offs on capturing teachers' academic contributions culminate in *Combine Replace Measure and Weight Changes* as the best alternative on the Effectiveness criterion.

Political Feasibility: 11 Points

The *Combine Replace Measures and Weight Changes* is still likely to be met with opposition with mostly similar support from the Public and DPS Evaluation Team. The resistance to value-added by the Teachers Union and subsequent support by both the union-backed School Board and Superintendent culminate in "Very Opposed" ratings. However, the incorporation of non-academic value-added would likely generate public discourse, and the Public is projected to be "Slightly Favored" because of the interest in supporting students' social/emotional needs (Harstad, 2021). The DPS Evaluation Team becomes "Slightly Favored" in response to the evaluation shift since the alternative does not capture all the previous elements of LEAP that "worked well" to the same extent (i.e., Student Learning Objectives have 1% weight), but still supports the alternative because it captures teachers' non-academic contributions. Therefore, the number of points earned by the *Combine Replace Measure and Weight Changes* is 11.

Implementation Complexity: 5 Points

Combine Replace Measure and Weight Changes overhauls teacher evaluation similarly to the *Replace Measure* alternative, making changes to the LEAP Evaluation Process, the Accuracy/Fairness of Evaluations, and Training, so the alternative receives 0 points. While some of the nuances of implementation are unique to the *Combine Replace Measure and Weight Changes* alternative, the Implementation Complexity mirrors the *Replace Measure* alternative. As a result of the 25 changes made, *Combine Replace Measure and Weight Changes* receives 5 points for its implementation complexity.

Recommendation and Implementation

Total Points Earned for Each Alternative

Alternative	*Status Quo*	*Weight Changes*	*Replace Measure*	*Combine Replace Measure & Weight Changes*
Effectiveness	26	27	38	43
Political Feasibility	20	8	12	11
Implementation Complexity	10	10	5	5
Total Points	56	45	55	59

Table 2: Policy Pathway Total Points Summary

Recommendation and Trade-Offs

The optimal recommendation according to the Effectiveness, Political Feasibility, and Implementation Complexity criteria is *Combine Replace Measure and Weight Changes*. The alternative sufficiently captures teachers' non-academic contributions, predicting that teachers who earn perfect LEAP scores are in the ~89th percentile in Behavioral Index value-added, while still accounting for teachers' academic contributions (~97th percentile). Although the alternative is relatively similar to both *Weight Changes* and *Replace Measure* in terms of Political Feasibility and *Replace Measure* in terms of Implementation Complexity, *Combine Replace Measure and Weight Changes* faces significant trade-offs compared to the *Status Quo*. In particular, the *Status Quo* has the highest Political Feasibility: it is well aligned to the desires of the Teachers Union (and hence the School Board and Superintendent). Similarly, the *Status Quo* also minimizes the number of changes to the existing teacher evaluation system, so there is also the lowest amount of Implementation Complexity. In contrast, *Combine Replace Measure and Weight Changes* faces significant opposition from the Teachers Union, School Board, and Superintendent while also featuring many changes. The superior ability to capture teachers' non-academic contributions without neglecting teachers' academic contributions exceeds the Political Feasibility and Implementation Complexity challenges, culminating in *Combine Replace Measure and Weight Changes* as the best alternative.

It is important to acknowledge that the recommendation would shift to the *Status Quo* if Political Feasibility were weighed more. To make *any changes* to LEAP requires either the entire LEAP Collaboration Committee (composed of both DPS and Teachers Union members) to unilaterally agree or, in the absence of consensus, be decided by the Superintendent (DPS and DCTA, 2017). Given the opposition to value-added and influence of the Teachers Union on the Superintendent, the *Combine Replace Measure and Weight Changes* alternative might never happen. The implementation therefore plays a key role to build buy-in for teachers in LEAP changes with the option to simply **revert** LEAP back to the *Status Quo*.

Implementation Plan

Combine Replace Measure and Weight Changes, substantially improves LEAP's ability to capture teachers' non-academic contributions, but resistance from the Teachers Union, School Board, and Superintendent may thwart the actualization of the alternative by the DPS Evaluation Team. However, successful implementation is possible: replication of the collaborative process between DPS and the Teachers Union that led to the creation LEAP and its sustained impact is key for the incorporation of *Behavioral Index* Value-Added. In particular, implementation should occur over multiple years: Year 1) Data-gathering and Small Pilot Phase; Year 2) 30-school Pilot and Refinement Phase; and Year 3) District-wide Implementation Phase. Systematic feedback collection is embedded in each of the three phases paired with the ability to **revert** LEAP back to the *Status Quo*. The commitment to seriously engage stakeholders' viewpoints in the design, development, and implementation of *Combine Replace Measures and Weight Changes* maximizes buy-in and minimizes Political Feasibility threats.

Groundwork

The DPS Evaluation Team should propose a multi-year implementation plan (detailed below) that starts small and gradually increases the number of schools affected; it should also include a clause that allows LEAP to **revert** back to the *Status Quo* after *any phase*. Furthermore, the DPS Evaluation Team should emphasize the importance of systematic feedback from teachers and their evaluators (teacher leaders and principals) to ensure that teacher evaluation works for those directly impacted by changes. The systematic feedback will be collected on the LEAP Website, a dedicated DCTA LEAP Liaison, Surveys, and Focus Groups. Since the plan mostly parallels the *successful* creation of LEAP in DPS (Jerald, 2013), start-up resistance to *Combine Replace Measures and Weight Changes* by the Teachers Union will be minimized with the potential for enhanced buy-in over the course of implementation.

Assuming initial agreement to implementation, the LEAP Collaboration Committee (LCC) would need to make *initial* decisions on the Accuracy/Fairness of Evaluations as well as LEAP Evaluation Processes and Training for the Behavioral Index Value-Added. The guiding principle for the Accuracy/Fairness of Evaluation: teachers' evaluations should solely reflect the *behaviors their students display during their classroom period* that can be *controlled by the teacher*— in essence, an excused absence nor spontaneous fight outside of the classroom that culminates in a suspension should penalize teachers' evaluation. The Data Collection Processes (e.g., collection of granular data or data audits) and LEAP Fairness Guide should be adjusted to match up with this principle. The other changes to LEAP Evaluation Processes and Training should match how Individual Statewide Test Growth is done, from the translation of measure into points to the use of high-quality instructional material for teachers' development. All initial decisions are subject to change after each implementation phase.

Year 1: Data-gathering and Small Pilot Phase

The Small Pilot Phase will feature, at minimum, 5 schools that will inform the development of LEAP Evaluation Processes, Accuracy/Fairness of Evaluation, and Training. To ensure engagement, school faculty will vote to volunteer to implement the changes to LEAP evaluation. To ensure that their critical feedback is heard, the DPS Evaluation Team will conduct Focus Groups with teachers and evaluators in addition to soliciting feedback via the LEAP Website, DCTA LEAP Liaisons, and Surveys. The cumulative information collected will be used to advise the LEAP Collaboration Committee on the necessary changes for LEAP Evaluation

Processes and the Accuracy/Fairness of Evaluation. Simultaneously, the DPS Evaluation Team will identify and closely observe teachers that have previously been "Effective" (84th percentile and above) in *Behavioral Index* Value-Added for insights on their classroom instructional practices. These insights will form the basis for the development of Teacher training material and best practices for teacher leaders/principals to coach teachers on in the next phase.

Year 2: 30-School Pilot and Refinement Phase

The Small Pilot Phase schools will continue with the LEAP Evaluation Changes, but an additional 30-Schools, per the Memorandum of Understanding (2021), will opt into the Year 2 Phase. Following the identification of 30 schools, the DPS Evaluation Team will collaborate with Principals at each school to detail the *Combine Replace Measures and Weight Changes* alternative and provide updates on changes from the Small Pilot Phase. The DPS Evaluation Team will then leverage training material created from the Small Pilot Phase to better prepare teachers and evaluators for the changes to LEAP.

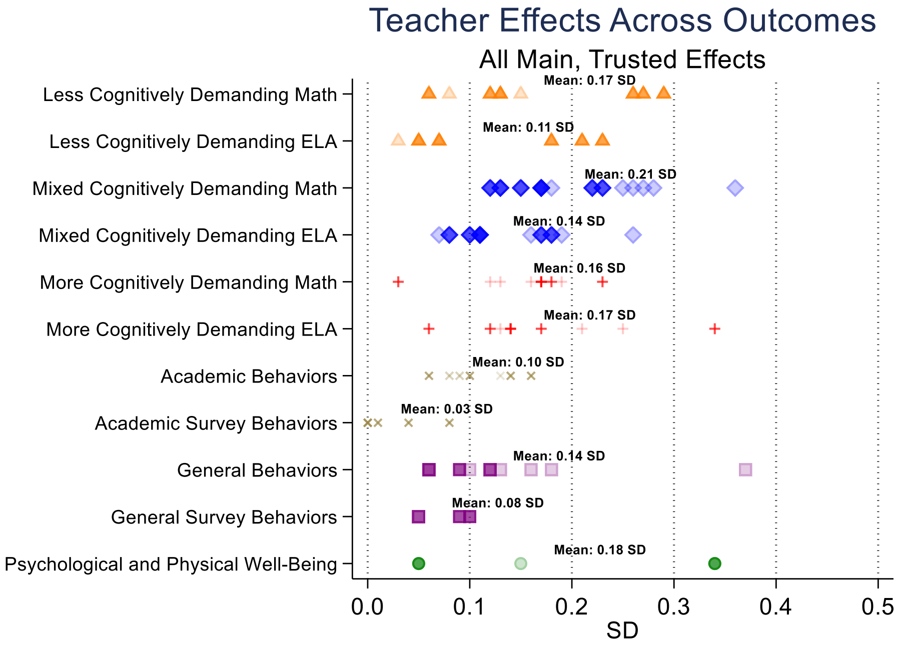
Systematic collection of feedback of teachers from the LEAP Website, DCTA LEAP Liaisons, and Surveys will be used to further refine the evaluation processes. The LEAP Collaboration Committee will again take the information and make any changes for LEAP Evaluation Processes and the Accuracy/Fairness of Evaluation. In addition, feedback on the formal Training experience using material created from the previous year will be solicited to inform future delivery of Training. Teachers from the Small Pilot Phase that have demonstrated success across multiple years will be asked (and compensated) to create additional training material. Successful Evaluators (i.e., improved teachers' Behavioral Index Value-Added from Year 1 to Year 2) from the Small Pilot Phase will be asked to train future evaluators on how to better coach teachers.

Year 3: District-wide Implementation Phase

The final District-Wide Implementation Phase takes the cumulative changes and adjustments to Training from Year 1 and Year 2 and places all teachers in DPS under the new evaluation system. Feedback will still be collected and presented to the LEAP Collaboration Committee to ensure continuous improvement, but the changes should be relatively minimal after two years of pilots. Training and development of high-quality instructional material will continually be improved to reflect the best pedagogical practices.

Appendix

Variation of Teachers' Ability to Improve Students' Skills



Appendix Figure 1: Teacher Effects across Academic and Non-Academic Skills

Behavioral Index Creation

To create the *Behavioral Index*, I specifically took three variables (% of Max Enrollment Time Attended, Suspensions, and Grade Progression) for all students in DPS and did a factor analysis with the Bartlett method to identify the 'behavioral' factor that underlies all three of these behaviors (we use the first factor in line with previous literature like Jackson, 2018).

The first factor can then be constructed as the weights of each variable:
Behavioral Index = -0.11146(Total Suspensions) + 1.08726(Max Enrollment Attendance Percentage) + 1.05128(Grade Progression)

The *Behavioral Index* is subsequently standardized to have a mean of 0 and a SD of 1.

Teacher Value-Added Estimation

To isolate teachers' contributions to the *Behavioral Index* (and test scores), the data are restricted to the years of 2017-2019 for Math and ELA teachers – the years of typical evaluation in DPS. The estimation followed best practices for calculating teachers' value-added.

$$Y_{icjt} = X_{icjt}\beta + \varphi + \varepsilon_{icjt}$$, where $$\varepsilon_{icjt} = \theta_{ij} + \epsilon_{cjt} + \epsilon_{icjt}$$

A regression model of the outcomes (Y) using data from individual students within a classroom for a particular year are run on observable characteristics (X) and grade-year fixed effects (Phi) for each

ELA and Math teachers. In particular the observable characteristics consist of a third-order polynomial using lagged achievement (Math and ELA) and lagged *Behavioral Index* along with individual student characteristics (e.g., gender and race) as well as classroom level characteristics (English Language Learner %, Special Education %, Gifted and Talented %, and Free-Reduced Price Lunch %). The grade-year fixed effects estimate effects for students that have the same grade-year, which helps resolve concerns of transitory shocks to students' *Behavioral Index*. The residuals can be decomposed into teachers' effects (Theta) as well as transitory classroom shocks and noise.

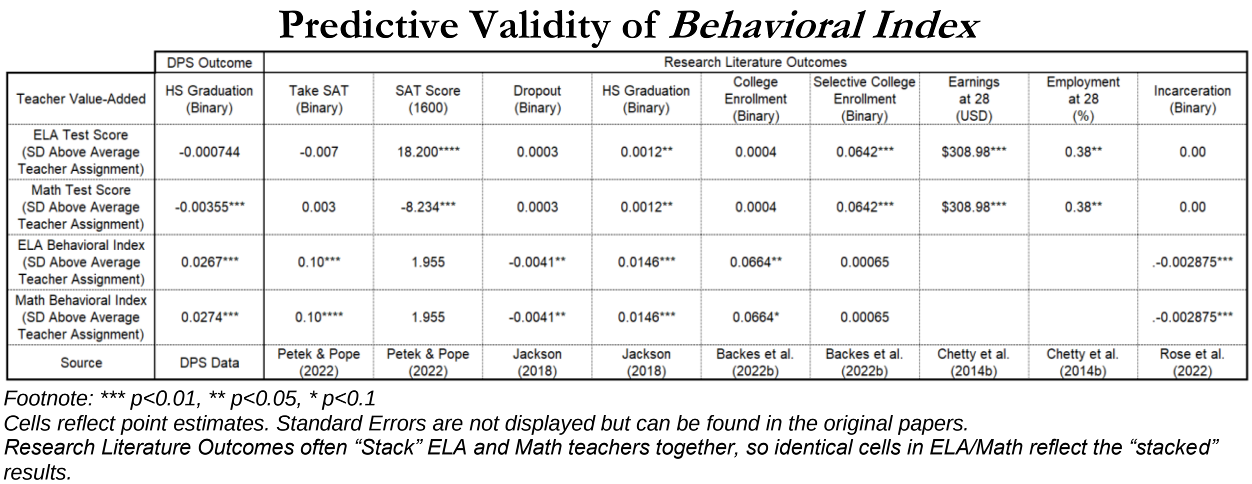
$$\varepsilon_{icjt} = Y_{icjt} - X_{icjt}\beta - \varphi$$

The remainder left unexplained after accounting for observable characteristics, assuming the identification assumptions hold (i.e., students are essentially randomly assigned to teachers after conditioning on observables characteristics), is attributable to teachers. To find that amount attributable to teachers, the portion left unexplained, or the student-level residuals, for each teacher is computed.

$$\widehat{\theta_{\imath\jmath,-t}} = \bar{\varepsilon}_{ij,-t}$$

However, because there is a need to avoid 'mechanical endogeneity,' or the problem of various common shocks like sampling variation or classroom shocks, to be part of the teacher's effectiveness measure, Chetty, Friedman, and Rockoff (2014a) jackknife approach is employed. By predicting teachers' effectiveness using different years than the one taught, the teachers' value-added without mechanical endogeneity can be estimated. The value-added estimate used is Theta Hat.

Note: While it is very common for scholars to employ shrinkage techniques to reduce the noise of estimated teachers' value-added, it is less useful for this analysis purpose. In particular, the analyses will rarely feature the teachers' actual value-added estimate and instead focus on the percentiles that teachers occupy. As Guarino et al. (2015) demonstrate, the use of shrinkage techniques does not meaningfully change the rank-order of teachers, so for the purposes of assessing teacher evaluation, it is largely unnecessary.



Predictive Validity of *Behavioral Index*

	DPS Outcome	Research Literature Outcomes								
Teacher Value-Added	HS Graduation (Binary)	Take SAT (Binary)	SAT Score (1600)	Dropout (Binary)	HS Graduation (Binary)	College Enrollment (Binary)	Selective College Enrollment (Binary)	Earnings at 28 (USD)	Employment at 28 (%)	Incarceration (Binary)
ELA Test Score (SD Above Average Teacher Assignment)	-0.000744	-0.007	18.200****	0.0003	0.0012**	0.0004	0.0642***	$308.98***	0.38**	0.00
Math Test Score (SD Above Average Teacher Assignment)	-0.00355***	0.003	-8.234***	0.0003	0.0012**	0.0004	0.0642***	$308.98***	0.38**	0.00
ELA Behavioral Index (SD Above Average Teacher Assignment)	0.0267***	0.10***	1.955	-0.0041**	0.0146***	0.0664**	0.00065			.-0.002875***
Math Behavioral Index (SD Above Average Teacher Assignment)	0.0274***	0.10****	1.955	-0.0041**	0.0146***	0.0664*	0.00065			.-0.002875***
Source	DPS Data	Petek & Pope (2022)	Petek & Pope (2022)	Jackson (2018)	Jackson (2018)	Backes et al. (2022b)	Backes et al. (2022b)	Chetty et al. (2014b)	Chetty et al. (2014b)	Rose et al. (2022)

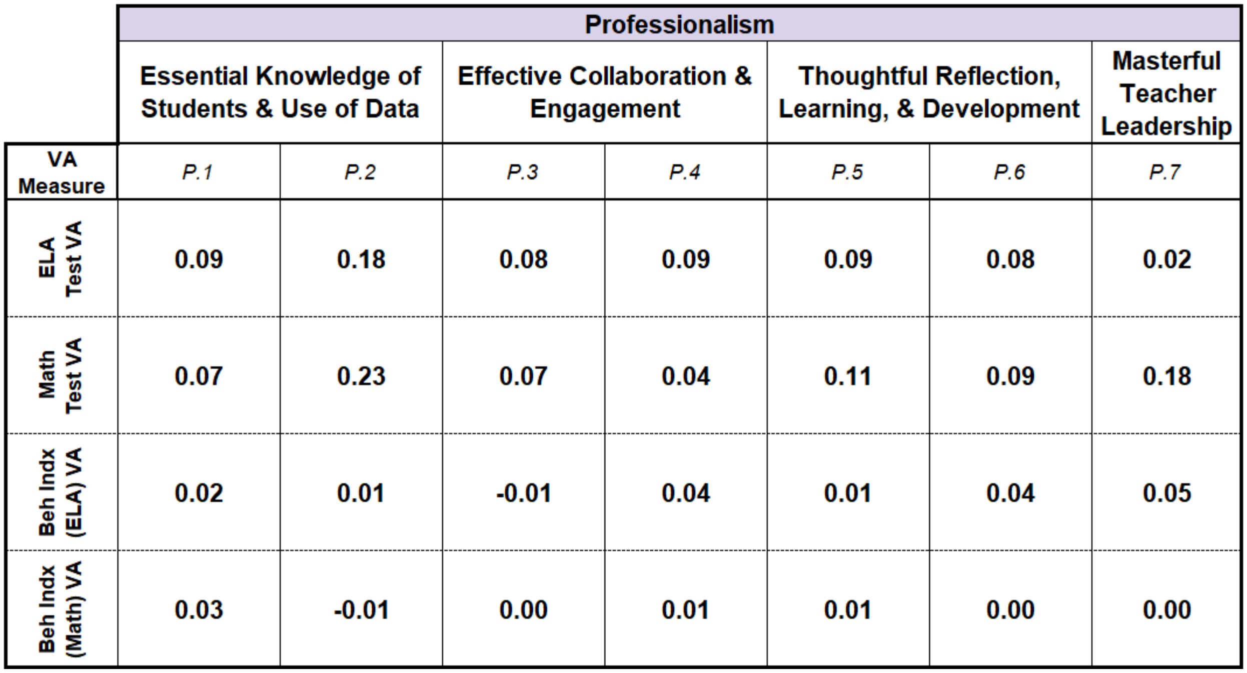
*Footnote: *** p<0.01, ** p<0.05, * p<0.1*
Cells reflect point estimates. Standard Errors are not displayed but can be found in the original papers.
Research Literature Outcomes often "Stack" ELA and Math teachers together, so identical cells in ELA/Math reflect the "stacked" results.

Appendix Table 1: Predictive Validity of *Behavioral Index* in DPS and Research Literature

This table documents how well the assignment of a student to a teacher 1 SD above average predicts future outcomes. Some outcomes can be improved by both *Behavioral Indices* and test scores, while others are exclusive to one or the other. In short, these skills are complementary. The behavioral index not only positively predicts students' future academic outcomes (e.g., HS Graduation and College Attendance) but also their future life outcomes like decreased incarceration rates (Rose et al., 2022).

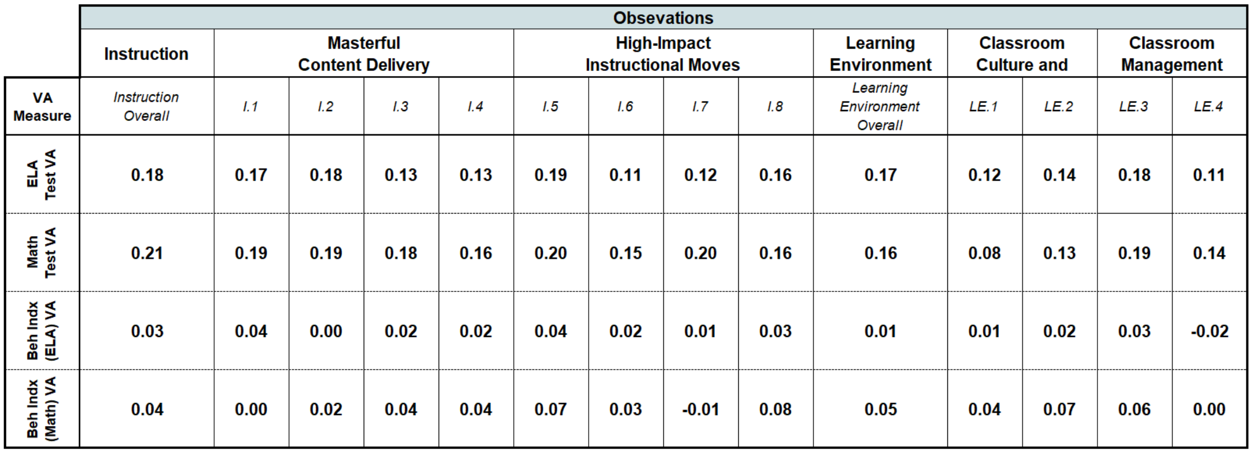
The results for DPS will inevitably be different than other research literature because this analysis attempts to estimate teachers' contributions for as many teachers as possible. Most scholars restrict their data to a particular range of grade levels (e.g., 3rd-5th or 9th). While there should definitely be caution in the overinterpretation of this particular version of calculating teachers' value-added, the one outcome that can be currently computed (HS Graduation) seems to align with prior research literature. Teachers' contributions to test scores have slightly close point estimates with HS Graduation though they are not statistically significant at the 0.05 level, but Jackson uses ONLY 9th graders, while our estimates are done using all DPS teachers. The *Behavioral Index* point estimates are both statistically significant but slightly lower than Jackson's estimate. This provides reassurance that the value-added computation is not completely off base and may be potentially used to investigate how well LEAP currently evaluates teachers' non-academic contributions.

Correlations of LEAP and Individual Indicators



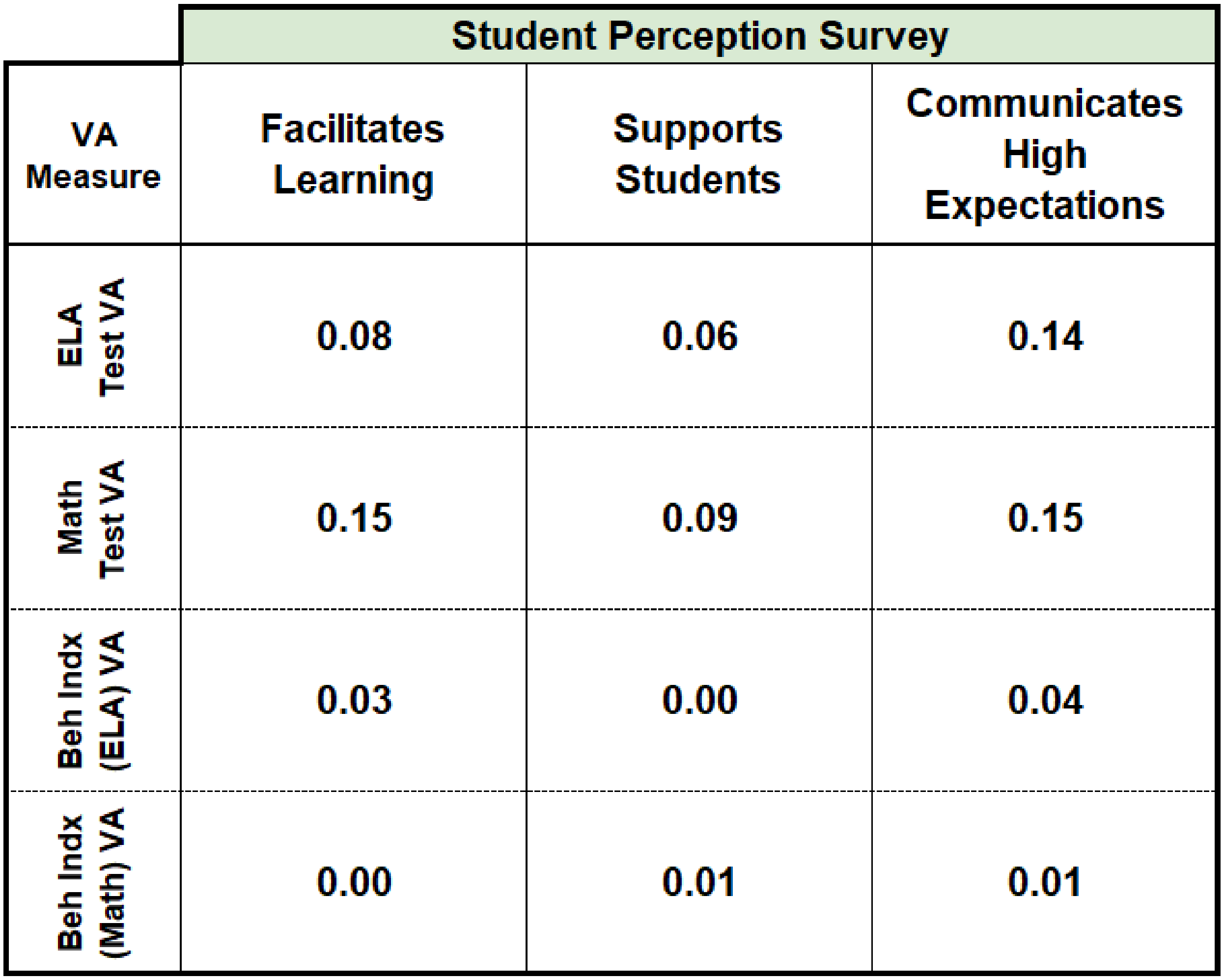
	Professionalism						
	Essential Knowledge of Students & Use of Data		Effective Collaboration & Engagement		Thoughtful Reflection, Learning, & Development		Masterful Teacher Leadership
VA Measure	*P.1*	*P.2*	*P.3*	*P.4*	*P.5*	*P.6*	*P.7*
ELA Test VA	**0.09**	**0.18**	**0.08**	**0.09**	**0.09**	**0.08**	**0.02**
Math Test VA	**0.07**	**0.23**	**0.07**	**0.04**	**0.11**	**0.09**	**0.18**
Beh Indx (ELA) VA	**0.02**	**0.01**	**-0.01**	**0.04**	**0.01**	**0.04**	**0.05**
Beh Indx (Math) VA	**0.03**	**-0.01**	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**

Appendix Table 2: Correlations of Teacher Value-Added and Professionalism Indicators



	Obsevations													
	Instruction	Masterful Content Delivery				High-Impact Instructional Moves				Learning Environment	Classroom Culture and		Classroom Management	
VA Measure	*Instruction Overall*	*I.1*	*I.2*	*I.3*	*I.4*	*I.5*	*I.6*	*I.7*	*I.8*	*Learning Environment Overall*	*LE.1*	*LE.2*	*LE.3*	*LE.4*
ELA Test VA	0.18	0.17	0.18	0.13	0.13	0.19	0.11	0.12	0.16	0.17	0.12	0.14	0.18	0.11
Math Test VA	0.21	0.19	0.19	0.18	0.16	0.20	0.15	0.20	0.16	0.16	0.08	0.13	0.19	0.14
Beh Indx (ELA) VA	0.03	0.04	0.00	0.02	0.02	0.04	0.02	0.01	0.03	0.01	0.01	0.02	0.03	-0.02
Beh Indx (Math) VA	0.04	0.00	0.02	0.04	0.04	0.07	0.03	-0.01	0.08	0.05	0.04	0.07	0.06	0.00

Appendix Table 3: Correlations of Teacher Value-Added and Observation Indicators

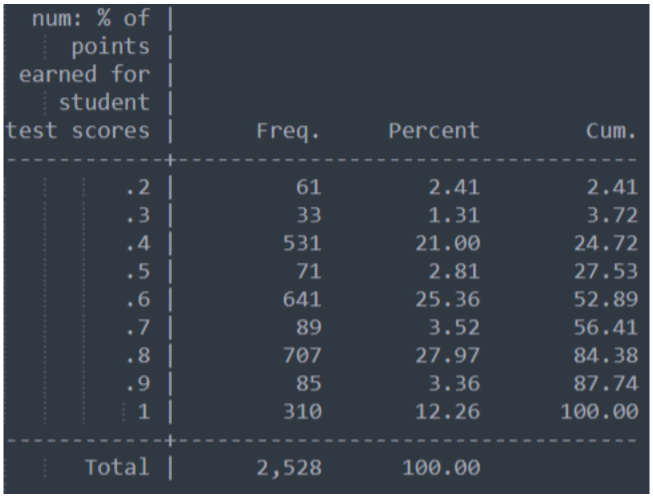


	Student Perception Survey		
VA Measure	Facilitates Learning	Supports Students	Communicates High Expectations
ELA Test VA	0.08	0.06	0.14
Math Test VA	0.15	0.09	0.15
Beh Indx (ELA) VA	0.03	0.00	0.04
Beh Indx (Math) VA	0.00	0.01	0.01

Appendix Table 4: Correlations of Teacher Value-Added and Student Perception Survey Indicators

Point Conversion from Behavioral Index Value-Added

To weigh the behavioral index values in teacher evaluation measures, there is a need to develop some sort of conversion between a teacher's behavioral index VA and points. While the researcher does not have access to the exact formula for how test scores convert to points, the distribution of points may be approximated. In particular, this is the distribution of points for individual statewide test growth:



```
  num: % of |
     points |
 earned for |
    student |
test scores |      Freq.     Percent        Cum.
------------+-----------------------------------
         .2 |         61        2.41        2.41
         .3 |         33        1.31        3.72
         .4 |        531       21.00       24.72
         .5 |         71        2.81       27.53
         .6 |        641       25.36       52.89
         .7 |         89        3.52       56.41
         .8 |        707       27.97       84.38
         .9 |         85        3.36       87.74
          1 |        310       12.26      100.00
------------+-----------------------------------
      Total |      2,528      100.00
```

Appendix Table 5: Individual Statewide Test Growth Distribution

A similar distribution can be generated by taking teachers' Behavioral Index VA and converting the score into percentiles. Teachers whose percentile performance is 2.41 and below would receive 0.2 or 20% of the points earned for the Behavioral Index. The teachers that fall between the 2.41 and 3.72 (2.41 + 1.31) percentile would earn 0.3 points and so forth. Since some teachers may have multiple estimates of Behavioral Index VA (recall that the Behavioral Index VA was separately estimated for Math and ELA teachers), the points earned are averaged.

Simulation Details

Goal: a teacher that earns all 100 points on their LEAP evaluation should be, on average, among the most effective teachers (i.e., helping maximize their students' growth). With that being said, there were three maximization criteria identified:

1) Maximize "Overall" Effectiveness → The projected effectiveness of a teacher that earns all 100 LEAP points in their test and behavioral index value-added is maximized → Sum up their on average Math effectiveness SD units + their behavioral index effectiveness SD units.

2) Maximize "Test Score" Effectiveness → Projected test-based value-added of a teacher that earns all 100 LEAP points is maximized

3) Maximize "Behavioral Index" Effectiveness → Projected behavioral index value-added of a teacher that earns all 100 LEAP points is maximized.

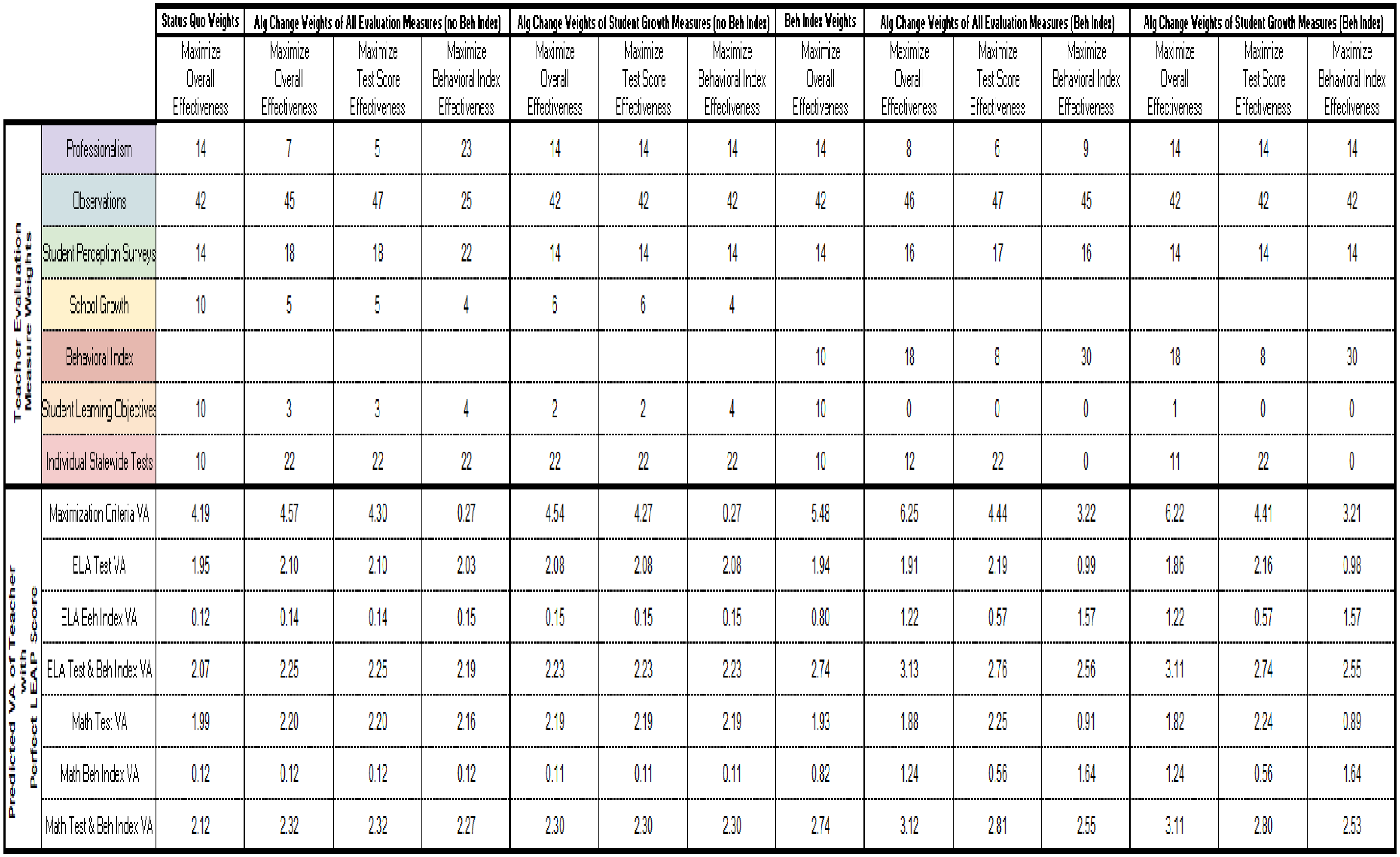
In accordance with Colorado S.B. 22-070, the point allocations are constrained to 70 possible points for Professional Practice and 30 possible points for Student Growth. Using 6 measures (3 in PP and 3 in SG), there are 1,155,525 possible combinations of weights. All weight combinations (e.g., 64 pts Obs, 3 pts Prof, 1 pt SPS; 23 pts SG, 4 pts SLO, 1 pt IST) were iterated through. In particular, the simulation starts by calculating each teacher's "Points Possible" on a measure— under the existing teacher evaluation system, a teacher may earn 28 / 35 Observation points. The "Points Possible" (28/35) is multiplied by the weight (64 Obs pts) for each measure. The products are then summed together to form teachers' overall LEAP score for that weight combination.

In the simulation process, we identify the weights that show the largest differentiation of teacher effectiveness among teachers that earn all 100 LEAP points (on that weight combination). A Multivariate regression of Test Score + Behavioral Index VA on LEAP Evaluation points (done separately for Math and ELA) was estimated. The LEAP Evaluation points coefficient is then used to predict the effectiveness of a teacher that earns all 100 LEAP Points. For example, one weight combination might show that a teacher who earns 100 points has a z-score of ~0.025 (51st percentile) in value-added, while another weight combination might show that teacher who earns 100 points has a z-score of ~1.645 (95th percentile) in value-added.

For each Maximization criteria (Test+BI, Test, or BI), identify the weight combination that would maximize the effectiveness of teachers. The largest total of the predicted VA for Math and ELA according to the maximization criteria is calculated.

> Test+BI: Predicts VA for a teacher that earns 100 LEAP points using a particular weight combination (e.g., 64 pts Obs, 3 pts Prof, 1 pt SPS; 23 pts SG, 4 pts SLO, 1 pt IST) and sums the Predicted ELA Test VA, Predicted ELA Behavioral Index VA, Predicted Math Test VA, Predicted Math Behavioral Index VA.
>
> For Test: Predicts VA for a teacher that earns 100 LEAP points using a particular weight combination (e.g., 64 pts Obs, 3 pts Prof, 1 pt SPS; 23 pts SG, 4 pts SLO, 1 pt IST) and sums the Predicted ELA Test VA and Predicted Math Test VA
>
> For BI: Predicts VA for a teacher that earns 100 LEAP points using a particular weight combination (e.g., 64 pts Obs, 3 pts Prof, 1 pt SPS; 23 pts SG, 4 pts SLO, 1 pt IST) and sums the Predicted ELA Behavioral Index VA and Predicted Math Behavioral Index VA.

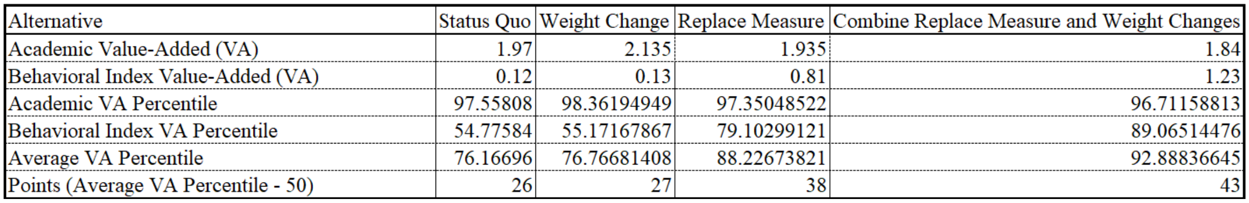
NOTE: The final recommendations take a modified version of the simulation that strictly iterates through Student Growth measures (School Collective / Behavioral Index were substituted depending on the alternative). The optimization including Professional Practice is qualitatively similar (see Figure below). In addition, the simulation results were adjusted to keep one point for SLO's in lieu of completely eliminating the measure for the ***Combine Replace Measure and Weight Changes*** alternative. The results are essentially *identical*.



		Status Quo Weights	Alg Change Weights of All Evaluation Measures (no Beh Index)			Alg Change Weights of Student Growth Measures (no Beh Index)			Beh Index Weights	Alg Change Weights of All Evaluation Measures (Beh Index)			Alg Change Weights of Student Growth Measures (Beh Index)		
		Maximize Overall Effectiveness	Maximize Overall Effectiveness	Maximize Test Score Effectiveness	Maximize Behavioral Index Effectiveness	Maximize Overall Effectiveness	Maximize Test Score Effectiveness	Maximize Behavioral Index Effectiveness	Maximize Overall Effectiveness	Maximize Overall Effectiveness	Maximize Test Score Effectiveness	Maximize Behavioral Index Effectiveness	Maximize Overall Effectiveness	Maximize Test Score Effectiveness	Maximize Behavioral Index Effectiveness
Teacher Evaluation Measure Weights	Professionalism	14	7	5	23	14	14	14	14	8	6	9	14	14	14
	Observations	42	45	47	25	42	42	42	42	46	47	45	42	42	42
	Student Perception Surveys	14	18	18	22	14	14	14	14	16	17	16	14	14	14
	School Growth	10	5	5	4	6	6	4							
	Behavioral Index								10	18	8	30	18	8	30
	Student Learning Objectives	10	3	3	4	2	2	4	10	0	0	0	1	0	0
	Individual Statewide Tests	10	22	22	22	22	22	22	10	12	22	0	11	22	0
Predicted VA of Teacher with Perfect LEAP Score	Maximization Criteria VA	4.19	4.57	4.30	0.27	4.54	4.27	0.27	5.48	6.25	4.44	3.22	6.22	4.41	3.21
	ELA Test VA	1.95	2.10	2.10	2.03	2.08	2.08	2.08	1.94	1.91	2.19	0.99	1.86	2.16	0.98
	ELA Beh Index VA	0.12	0.14	0.14	0.15	0.15	0.15	0.15	0.80	1.22	0.57	1.57	1.22	0.57	1.57
	ELA Test & Beh Index VA	2.07	2.25	2.25	2.19	2.23	2.23	2.23	2.74	3.13	2.76	2.56	3.11	2.74	2.55
	Math Test VA	1.99	2.20	2.20	2.16	2.19	2.19	2.19	1.93	1.88	2.25	0.91	1.82	2.24	0.89
	Math Beh Index VA	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.82	1.24	0.56	1.64	1.24	0.56	1.64
	Math Test & Beh Index VA	2.12	2.32	2.32	2.27	2.30	2.30	2.30	2.74	3.12	2.81	2.55	3.11	2.80	2.53

Appendix Table 6: Simulation Results

Effectiveness Criterion Calculations



Alternative	Status Quo	Weight Change	Replace Measure	Combine Replace Measure and Weight Changes
Academic Value-Added (VA)	1.97	2.135	1.935	1.84
Behavioral Index Value-Added (VA)	0.12	0.13	0.81	1.23
Academic VA Percentile	97.55808	98.36194949	97.35048522	96.71158813
Behavioral Index VA Percentile	54.77584	55.17167867	79.10299121	89.06514476
Average VA Percentile	76.16696	76.76681408	88.22673821	92.88836645
Points (Average VA Percentile - 50)	26	27	38	43

Appendix Table 7: Effectiveness Points Calculation

Out of the simulation results above, the analyses of Effectiveness focus specifically on the "Maximize Overall Effectiveness" columns for Status Quo, Alg Change Weights of Student Growth Measures (no Beh Index), Beh Index Weights, and Alg Change Weights on Student Growth Measures (Beh Index). The predicted VA for ELA Test VA and Math Test VA are averaged (e.g., 1.95 z-score and 1.99 z-score respectively for *Status Quo*) to estimate the "Academic Value-Added (VA)" cell of the table for a particular alternative (1.97 z-score for *Status Quo*). The predicted VA for ELA Beh Index VA and Math Beh Index VA (e.g., 0.12 z-score and 0.12 z-score respectively for *Status Quo*) are also averaged to estimate the "Behavioral Index Value-Added (VA)" cell of the table for a particular alternative (0.12 z-score for *Status Quo*). The averaged predicted VA z-scores for each Academic and Behavioral Index are subsequently converted into percentiles using excel's NORM.S.DIST function ("Cumulative" option specified as "TRUE"). For example, Status Quo has a 97.56 percentile for Academic VA and a 54.78 percentile for Behavioral Index VA. The percentiles are averaged together and subtracted by 50 to calculate the number of points earned for each alternative. To minimize perceptions of false precision, the number of points earned is rounded to the nearest whole number.

Political Feasibility Criterion Justification

Denver Public Schools Context

Denver Public Schools had been at the forefront of school reform since 2005 while the teacher union, the Denver Classroom Teachers Association (DCTA), has continually been in opposition. Superintendents Michael Bennet and Tom Boasberg, paired with a pro-reform school board, embraced a range of reforms such as school choice and accountability for student performance; the DCTA initiated lawsuits against these reforms that were ultimately unsuccessful (Baxter and Gottlieb, 2022). Baxter and Gottlieb then detail the gradual shift in political favor: frustrations with standardized testing along with the association of charter schools and the Trump administration culminated in organized opposition to school reform by the Denver Classroom Teacher Association (2022). Eventually, the Denver Public School's Board flipped in 2019 with five out of seven board members backed by the union (Asmar, 2019), setting the stage for current political feasibility of changes to teacher accountability.

The Superintendent at the time of the DPS School Board flip, Susana Cordova, had been instrumental in continuing the legacy of school reform and had previously helped create the reforms in her prior roles such as the teacher evaluation system, LEAP (Asmar, 2015). The majority union-backed school board immediately initiated efforts to roll back major reforms (Baxter and Gottlieb, 2021) with reported efforts to undermine Cordova (Asmar, 2020b) that is rumored to have led to her mid-school year departure (Asmar, 2020a). The Superintendent that replaced Cordova, Dr. Alex Marrero, was supported by the union-backed school board (Brundin, 2021) and explicitly endorsed by DCTA (Choi and Gould, n.d.).

The subsequent DPS School Board November 2021 election further entrenched DCTA's position in the school board: all members (7/7) were union-backed with one candidate later resigning prior to his term for another job opportunity (Meltzer, 2022). One School Board member was criticized to "be driven by ideology, blind fealty to Denver Classroom Teacher Association, or some unhealthy mixture of both" (Gottlieb, 2022) and another made explicit commitments to "partner with the union to help fix the broken system (LEAP) we currently grade our educators by" (Quattlebaum, 2021). It appears that these union-backed candidates implicitly must adhere to DCTA's desired policies or otherwise risk their election: one school board member stated in response to one DCTA desired policy: "I would rather lose my reelection and lose support from DCTA than violate my moral compass and vote for something I do not believe we have had the proper time to engage with the community on" (Mullen, 2022).

Teacher Union's Influence on School Board and Superintendent Political Feasibility

As the previous section on Denver Public School's Context demonstrates, the teacher union maintains a powerful role in shaping the political backdrop of the school district. School Board members that are non-union backed have been systematically organized out of their elected office and replaced by a union-backed School Board member. These union-backed School Board members have then undermined even the Superintendent of the district until the previous Superintendent left. The current School Board and Superintendent therefore occupy a precarious situation: failing to adhere to the desires of the teacher union may pose a risk to their elected office or job. Although both the Superintendent and School Board care deeply about students' non-academic skill development as outlined by the End Policies (n.d.), their goals are not *contingent* on teacher evaluation changes. Therefore, these two stakeholders are projected to closely mirror the perspectives of the teacher union for assessing alternatives.

Teacher Union's Political Feasibility (Implicitly School Board and Superintendent)

The teacher union has historically opposed the use of student achievement data for teacher evaluation (McGuinn, 2012) and maintains that opposition even today. The Denver Classroom Teachers Association is under the National Education Association, which "opposes using student test scores to assess teacher effectiveness [VAM] … the Association believes such tests may be used only to provide non-evaluative formative feedback rather than to support high-stakes decision, such as termination" (2020). In addition, the state-level equivalent of the National Education Association that DCTA is under, the Colorado Education Association, continues to advocate for the reduced use of student growth measures in teacher evaluation: the recent reduction of student growth measures from 50% to 30% of a teacher's evaluation from S.B. 22-070 "does not go as far as CEA [Colorado Education Association] and many members would like" (2022). Furthermore, at the local level, the Teacher Union was very involved in constructing LEAP (Jerald, 2013) and may have provided the feedback [no explicit evidence] that led to the decrease of Individual Statewide Test from a 30% weight in 2013-14 to a 10% weight in all future years. The continued opposition to the use of student achievement data, particularly value-added, shapes DCTA's response to alternatives.

In an internal document between the DPS Evaluation Team and DCTA representatives on the LEAP Collaboration Committee that ultimately led to the establishment of the *Status Quo* alternative, DCTA rejected all proposed changes that increased the weight on teachers' individual measures (Student Learning Objectives and Individual Statewide Test Growth). Instead, DCTA pushed for the *Status Quo* alternative weights keeping Individual Statewide Test Growth— the closest to a "value-added" measure— the same 10% weight. Incorporating any increase in weighting for value-added (Individual Statewide Test Growth) has been met with resistance and opposition. Similarly, incorporating *another* value-added measure is likely going to face strong opposition by DCTA since it is making an even larger portion of teacher evaluation on value-added.

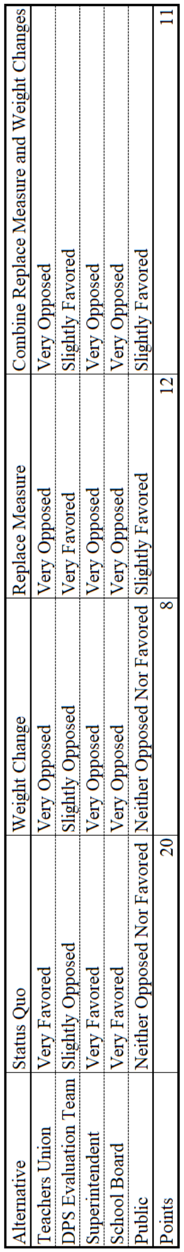
DPS Evaluation Team and Political Feasibility

The "formal evidence" for the DPS Evaluation Team is non-existent in publicly available formats. Instead, I rely on direct conversations with members of the DPS Evaluation Team and internally shared documents to provide their viewpoints on various alternatives. After presenting initial results to the DPS Evaluation team (e.g., Behavioral Index VA correlations with LEAP measure performance), I found that the evaluation team believes that many aspects of LEAP evaluation are successful, but the inability of LEAP to capture teachers' non-academic contributions were deemed particularly concerning. In our conversation, the DPS evaluation team recognizes the growing emphasis on students' socio-emotional learning, agency, and a range of other non-academic skills being emphasized in the district (DPS School Board, n.d.) and subsequently sees the need to eventually incorporate teachers' non-academic contributions into teacher evaluation. The political feasibility may slightly shift depending on the degree of weight changes: in one internal document that helped establish the Status Quo alternative, the evaluation team appears to be open to *small weight changes* (few %'s) but not overly weighting any particular measure. However, the DPS Evaluation was open to the removal of the School Collective measure since the new law, S.B. 22-070, no longer requires a School Collective measure (previously, the School Collective measure was *required* by S.B. 10-191) and the School Collective measure simply does not distinguish between teachers (every teacher in the same school gets the same score with minimal variation of overall school performance).

Public and Political Feasibility

While the Public undeniably has a stake in what happens in public schools, the Public must know about the changes to even care about them. In particular, the Status Quo and Weight Changes are unlikely to ever fall under the radar of the Public: LEAP has changed many times (including Weight Changes) since its inception in 2012 (LEAP Handbooks, 2012-2022) with documentation on the district's website, but these small changes have rarely been reported in the news— a google search of "Leading Effective Academic Practice" reveals 2 news articles with one in 2014 from the Denver Post and another in 2022 in D.C. Policy Center. As a consequence of minimal reporting on the topic, the Public is unlikely to care about the Status Quo and Weight Changes alternatives.

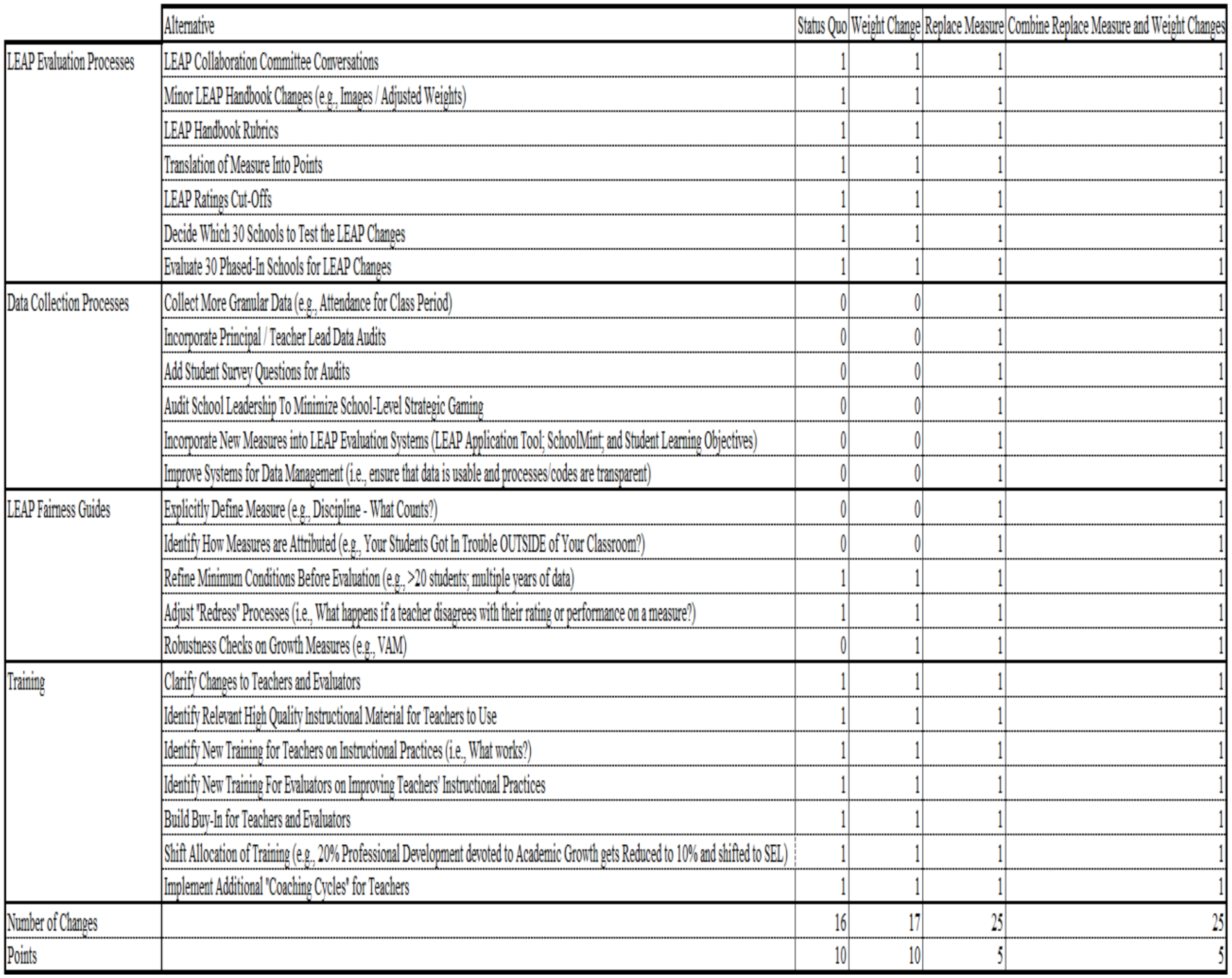
However, the Public has responded to more controversial educational topics like standardized testing (Erdahl, 2014) and, of course, voted in School Board meetings to express their preferences on Denver Public Schools education. If the Denver Classroom Teacher Association were to leverage its advocacy ability against the Replace Measure (or Combine Replace Measure and Weight Changes) alternative, the Public may take notice and respond. But, the outcome may not solely be aligned with the teacher union: the Public does seem to think that there is too much emphasis on standardized tests (PDK, 2015) but still believe standardized tests are important as long as students' social/emotional needs are supported (Harstad, 2021).



Alternative	Status Quo	Weight Change	Replace Measure	Combine Replace Measure and Weight Changes
Teachers Union	Very Favored	Very Opposed	Very Opposed	Very Opposed
DPS Evaluation Team	Slightly Opposed	Slightly Opposed	Very Favored	Slightly Favored
Superintendent	Very Favored	Very Opposed	Very Opposed	Very Opposed
School Board	Very Favored	Very Opposed	Very Opposed	Very Opposed
Public	Neither Opposed Nor Favored	Neither Opposed Nor Favored	Slightly Favored	Slightly Favored
Points	20	8	12	11

Appendix Table 7: Political Feasibility Point Summary

Implementation Complexity Criterion Calculations



	Alternative	Status Quo	Weight Change	Replace Measure	Combine Replace Measure and Weight Changes
LEAP Evaluation Processes	LEAP Collaboration Committee Conversations	1	1	1	1
	Minor LEAP Handbook Changes (e.g., Images / Adjusted Weights)	1	1	1	1
	LEAP Handbook Rubrics	1	1	1	1
	Translation of Measure Into Points	1	1	1	1
	LEAP Ratings Cut-Offs	1	1	1	1
	Decide Which 30 Schools to Test the LEAP Changes	1	1	1	1
	Evaluate 30 Phased-In Schools for LEAP Changes	1	1	1	1
Data Collection Processes	Collect More Granular Data (e.g., Attendance for Class Period)	0	0	1	1
	Incorporate Principal / Teacher Lead Data Audits	0	0	1	1
	Add Student Survey Questions for Audits	0	0	1	1
	Audit School Leadership To Minimize School-Level Strategic Gaming	0	0	1	1
	Incorporate New Measures into LEAP Evaluation Systems (LEAP Application Tool; SchoolMint; and Student Learning Objectives)	0	0	1	1
	Improve Systems for Data Management (i.e., ensure that data is usable and processes/codes are transparent)	0	0	1	1
LEAP Fairness Guides	Explicitly Define Measure (e.g., Discipline - What Counts?)	0	0	1	1
	Identify How Measures are Attributed (e.g., Your Students Got In Trouble OUTSIDE of Your Classroom?)	0	0	1	1
	Refine Minimum Conditions Before Evaluation (e.g., >20 students; multiple years of data)	1	1	1	1
	Adjust "Redress" Processes (i.e., What happens if a teacher disagrees with their rating or performance on a measure?)	1	1	1	1
	Robustness Checks on Growth Measures (e.g., VAM)	0	1	1	1
Training	Clarify Changes to Teachers and Evaluators	1	1	1	1
	Identify Relevant High Quality Instructional Material for Teachers to Use	1	1	1	1
	Identify New Training for Teachers on Instructional Practices (i.e., What works?)	1	1	1	1
	Identify New Training For Evaluators on Improving Teachers' Instructional Practices	1	1	1	1
	Build Buy-In for Teachers and Evaluators	1	1	1	1
	Shift Allocation of Training (e.g., 20% Professional Development devoted to Academic Growth gets Reduced to 10% and shifted to SEL)	1	1	1	1
	Implement Additional "Coaching Cycles" for Teachers	1	1	1	1
Number of Changes		16	17	25	25
Points		10	10	5	5

Appendix Table 8: Implementation Complexity Criterion Calculations

References

Asmar, M. (2015, December 1). *Susana Cordova named acting superintendent of Denver Public Schools*. Chalkbeat Colorado. https://co.chalkbeat.org/2015/12/1/21103242/susana-cordova-named-acting-superintendent-of-denver-public-schools

Asmar, M. (2019, November 7). *Why the Denver school board "flipped" and what might happen next*. Chalkbeat Colorado. https://co.chalkbeat.org/2019/11/7/21109184/why-the-denver-school-board-flipped-and-what-might-happen-next

Asmar, M. (2020a, November 13). *Superintendent Susana Cordova is leaving Denver Public Schools*. Chalkbeat Colorado. https://co.chalkbeat.org/2020/11/13/21564534/superintendent-susana-cordova-leaving-denver-public-schools

Asmar, M. (2020b, December 3). *Changing priorities shaped Susana Cordova's relationship with the Denver school board*. Chalkbeat Colorado. https://co.chalkbeat.org/2020/12/3/22151381/denver-school-board-susana-cordova-relationship

Asmar, M. (2022, May 23). *Denver superintendent's goals include dismantling "oppressive systems."* Chalkbeat. https://co.chalkbeat.org/2022/5/23/23138733/denver-alex-marrero-superintendent-goals-school-board

Backes, B., Cowan, J., Goldhaber, D., & Theobald, R. (2022a). Teachers and School Climate: Effects on Student Outcomes and Academic Disparities. *Caldercenter.org*. https://caldercenter.org/publications/teachers-and-school-climate-effects-student-outcomes-and-academic-disparities

Backes, B., Cowan, J., Goldhaber, D., & Theobald, R. (2022b). Teachers and Students' Postsecondary Outcomes: Testing the Predictive Power of Test and Nontest Teacher Quality Measures. *Caldercenter.org*. https://caldercenter.org/sites/default/files/CALDER%20Working%20Paper%20270-1022.pdf

Backes, B., & Hansen, M. (2018). The impact of Teach for America on non-test academic outcomes. *Education Finance and Policy*, *13*(2), 168–193.

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diance Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper# 278. *Economic Policy*. https://eric.ed.gov/?id=ED516803

Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the Big Five in late childhood. *Personality and Individual Differences*, *34*(4), 645–664.

Baxter, P., & Gottlieb, A. (2022). Dismantling Denver: The city was a national model for education reform. Then union-backed candidates took over the school board. *Education Next*, *22*, 26+.

Bjorklund-Young, A. V., & Ronda, V. (2017). *The multidimensionality of teacher quality: Teaching skills and students' noncognitive skills †*. paa.confex.com.

https://paa.confex.com/paa/2018/mediafile/ExtendedAbstract/Paper22091/Bjorklund_Ronda_Noncogs.pdf

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, *13*(3), 281–309.

Blazar, D., & Kraft, M. A. (2017). Teacher and Teaching Effects on Students' Attitudes and Behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146–170.

Brundin, J. (2021, June 3). *Denver Public Schools Board Votes 6-1 To Appoint Alex Marrero As New Superintendent*. Colorado Public Radio. https://www.cpr.org/2021/06/03/denver-public-schools-alex-marrero-new-superintendent/

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *The American Economic Review*, *104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Choi, T., & Gould, R. (n.d.). *DPS Selects Dr. Alex Marrero As the New Superintendent*. Denver Classroom Teachers Association. Retrieved March 2, 2023, from https://denverteachers.org/dps-selects-dr-alex-marrero-as-the-new-superintendent/

Colorado Education Association. (2022, May 13). *The Highs and Lows of the 2022 Legislative Session*. Colorado Education Association. https://coloradoea.org/news-updates/the-highs-and-lows-of-the-2022-legislative-session/

Colorado General Assembly. (2010). *Senate Bill 10-191*. Colorado Department of Education. https://www.cde.state.co.us/educatoreffectiveness/overviewofsb191

Colorado General Assembly. (2022). *SB22-070*. Colorado General Assembly. https://leg.colorado.gov/bills/sb22-070

Dee, T. S., James, J., & Wyckoff, J. (2021). Is effective teacher evaluation sustainable? Evidence from District of Columbia Public Schools. *Education Finance and Policy*, *16*(2), 313–346.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management: [the Journal of the Association for Public Policy Analysis and Management]*. https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.21818

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*. https://academic.oup.com/qje/article-abstract/132/4/1593/3861633

Denver Public Schools and Denver Classroom Teacher Association. (2017). *DCTA Agreement 2017*. Denver Teachers. https://denverteachers.org/wp-content/uploads/DCTA-Agreement-

2017-2022-with-Financial-Agreement.pdf

Denver Public Schools and Denver Classroom Teachers Association. (2021, March 26). *SY21-22 LEAP Pilot MOU*. https://hr.dpsk12.org/wp-content/uploads/sites/37/SY21-22-LEAP-PILOT-MOU.pdf

Denver Public Schools School Board. (n.d.). *Denver Public Schools End Policies*. Denver Public Schools Board of Education. Retrieved March 2, 2023, from https://board.dpsk12.org/policy/

Donaldson, M. L., & Firestone, W. (2021). Rethinking teacher evaluation using human, social, and material capital. *Journal of Educational Change*, *22*(4), 501–534.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101.

Erdahl, K. (2014, March 29). *Parents protest against standardized testing by opting kids out*. Fox 31 News. https://kdvr.com/news/parents-protest-against-standardized-testing-by-opting-kids-out/

Fleche, S. (2017). Teacher Quality, Test Scores and Non-Cognitive Skills: Evidence from Primary School Teachers in the UK. CEP Discussion Paper No. 1472. *Centre for Economic Performance*. https://eric.ed.gov/?id=ED583855

Gates Foundation. (2015, October 6). *Measures of Effective Teaching (MET) Project*. Gates Foundation. https://usprogram.gatesfoundation.org/news-and-insights/articles/measures-of-effective-teaching-project

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125–149.

Gilraine, M., & Pope, N. G. (2021). *Making Teaching Last: Long-Run Value-Added* (No. 29555). National Bureau of Economic Research. https://doi.org/10.3386/w29555

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A Three-way Error Components Analysis of Educational Productivity. *Education Economics*, *7*(3), 199–208.

Goldhaber, D., Krieg, J., & Theobald, R. (2020). Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics*, *63*, 101792.

Gottlieb, A. (2022, June 15). *Dysfunction paralyzes Denver school board*. The Gazette. https://gazette.com/dysfunction-paralyzes-denver-school-board/article_bdb72af4-eb55-11ec-81b7-9bda61161784.html

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *40*(2), 190–222.

Hanushek, E. A. (2009). Teacher deselection. *Creating a New Teaching Profession*, *168*, 172–173.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466–479.

Hanushek, & Rivkin. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*. https://hanushek.stanford.edu/sites/default/files/publications/Hanushek+Rivkin%202012%20AnnRevEcon%204.pdf

Harstad, P. (2021, February 1). *Colorado Survey on Education*. Colorado Education Association. https://web.archive.org/web/20210210160011/https:/coloradoea.org/wp-content/uploads/2021-Colorado-Survey-on-Education-Standardized-Tests.pdf

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, *24*(3), 411–482.

Holt, S., Vinopal, K., Choi, H., & Sorensen, L. (2022). Strictly speaking: Examining teacher use of punishment and student outcomes. In *Annenberg Institute at Brown University* (EdWorkingPaper: 22-563). https://doi.org/10.26300/meqn-w550

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *The Journal of Political Economy*, *126*(5), 2072–2107.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, *26*(1), 101–136.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher Effects on Social and Behavioral Skills in Early Elementary School. *Sociology of Education*, *83*(2), 135–159.

Jerald, C. D. (2013). *Beyond Buy-In: Partnering with practitioners to build a professional growth and accountability system for Denver's educators*. https://policycommons.net/artifacts/1302034/beyond-buy-in/1905325/

John, O. P., Caspi, A., Robins, R. W., Moffitt, T. E., & Stouthamer-Loeber, M. (1994). The "little five": exploring the nomological network of the five-factor model of personality in adolescent boys. *Child Development*, *65*(1), 160–178.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Research Paper. MET*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.638.2716

Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation* (No. 14607). National Bureau of Economic Research. https://doi.org/10.3386/w14607

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project.

Bill & Melinda Gates Foundation. https://eric.ed.gov/?id=ED540960

Kelly. (2012). Understanding teacher effects: Market versus process models of educational improvement. *Teacher Quality: Understanding Teacher …*.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *The Journal of Human Resources*. http://jhr.uwpress.org/content/54/1/1.short

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics*, *188*, 104212.

Ladd, H. F., & Sorensen, L. C. (2017). Returns to teacher experience: Student achievement and motivation in middle school. *Education Finance and Policy*, *12*(2), 241–279.

Liu, J., & Loeb, S. (2021). Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School. *The Journal of Human Resources*, *56*(2), 343–379.

Lounsbury, J. W., Steel, R. P., & Loveland, J. M. (2004). An investigation of personality traits in relation to adolescent school absenteeism. *Journal of Youth and Adolescence*. https://doi.org/10.1023/B:JOYO.0000037637.20329.97

Lundberg, S. (2017). Non-cognitive skills as human capital. *Education, Skills, and Technical Change: Implications for Future US GDP Growth*, 219–243.

McGuinn, P. (2012). Stimulating Reform: Race to the Top, Competitive Grants and the Obama Education Agenda. *Educational Policy* , *26*(1), 136–159.

Meltzer, E. (2022, March 13). *Denver school board member Brad Laurvick resigning*. Chalkbeat Colorado. https://co.chalkbeat.org/2022/3/13/22976175/denver-school-board-member-brad-laurvick-resigning

Mihaly, K., Mccaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of Effective Teaching*. sites.dartmouth.edu. https://sites.dartmouth.edu/dstaiger/files/2019/06/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Monk, D. H., & Ibrahim, M. A. (1984). Patterns of Absence and Pupil Achievement. *American Educational Research Journal*, *21*(2), 295–310.

Mulhern, C., & Opper, I. (2022). *Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness* (EdWorkingPaper: 21-451). Annenberg Institute at Brown University. https://doi.org/10.2139/ssrn.3912373

Mullen, D. (2022, March 24). *Denver school board votes in favor of executive limitation*. Denver Gazette. https://denvergazette.com/news/education/denver-school-board-votes-in-favor-of-executive-limitation/article_89c1ebb0-abc4-11ec-8520-87bf0a4acfe9.html

Murnane, R. J., Willett, J. B., & Levy, F. (1995). *The Growing Importance of Cognitive Skills in Wage Determination* (No. 5076). National Bureau of Economic Research. https://doi.org/10.3386/w5076

National Education Association. (2020, July). *NEA Teacher Evaluation and Accountability Toolkit*. National Education Association. https://www.nea.org/resource-library/nea-teacher-evaluation-and-accountability-toolkit

Petek, N., & Pope, N. G. (2022). *The multidimensional impact of teachers on students*. nathanpetek.com. http://www.nathanpetek.com/uploads/1/2/0/1/120192201/multidimensionalteachers.pdf

Phi Delta Kappa. (2015). *Testing doesn't measure up for Americans*. Phi Delta Kappa. https://web.archive.org/web/20160115095242/http:/pdkpoll2015.pdkintl.org/highlights

Putnam, H., Ross, E., & Walsh, K. (2018). Making a difference: Six places where teacher evaluation systems are getting results. *National Council on Teacher Quality*. http://files.eric.ed.gov/fulltext/ED590763.pdf

Quattlebaum, M. (2021, August 14). *Michelle Quattlebaum DCTA Endorsement*. Twitter. https://twitter.com/MichelleForDPS/status/1426584176753479680

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher* , *47*(7), 405–418.

Rose, E., Schellenberg, J., & Shem-Tov, Y. (2022). *The Effects of Teacher Quality on Adult Criminal Justice Contact*. https://www.lowe-institute.org/wp-content/uploads/2022/04/RoseSchellenbergShemtov2022.pdf

Ross, E., & Walsh, K. (2019). State of the states 2019: Teacher and principal evaluation policy. *National Council on Teacher Quality*.

Schools, D. P. (2022a). *DPS Thrives Strategic Roadmap*. https://www.dpsk12.org/about/dps-Thrives/. https://issuu.com/dpscommunications/docs/dps_thrives_strategic_roadmap

Schools, D. P. (2022b). *Facts and Figures*. Denver Public Schools. https://www.dpsk12.org/about/facts-figures/#students

Schools, D. P. (2022c). *Financial Transparency*. Denver Public Schools. https://financialservices.dpsk12.org/financialtransparency/

Schools, D. P. (2022d). *LEAP Handbook 2022-2023*. http://thecommons.dpsk12.org/. http://thecommons.dpsk12.org/cms/lib/CO01900837/Centricity/Domain/103/LEAPHandbook_2022-23_Aug%202022.pdf

Schools, D. P. (2022e). *Mission and Vision*. Denver Public Schools. https://www.dpsk12.org/about/mission-and-vision/

Sutcher, L., Darling-Hammond, L., & Carver-Thomas, D. (2016). A coming crisis in teaching? Teacher supply, demand, and shortages in the U.s. *Learning Policy Institute*. http://files.eric.ed.gov/fulltext/ED606666.pdf

Todd, P. E., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, *113*(485), F3–F33.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*. https://eric.ed.gov/?id=ED515656