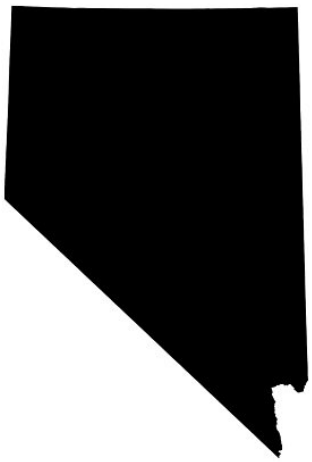
Center for Research, Evaluation, and Assessment

UNLV COLLEGE OF EDUCATION

June 2020

# NEVADA EDUCATOR PEFORMANCE FRAMEWORK: IMPACT AND VALIDITY FINAL REPORT



Presented to the Legislative Committee on Education—June 16, 2020

DISCLAIMER

The Center for Research, Evaluation and Assessment (CREA) at the University of Nevada, Las Vegas is a nonpartisan academic research center dedicated to improving evidence-based decision making in a variety of sectors through the production of high quality design, implementation and interpretation of research-based evaluation and assessment.

The research presented in this report solely represent the analysis and opinions of the authors and are not endorsed by, or reflect the views or positions of the Nevada Department of Education. All remaining errors are our own.

June 2020

# NEVADA EDUCATOR PEFORMANCE FRAMEWORK: IMPACT AND VALIDITY FINAL REPORT


AUTHORS

**Bradley Marianno**, *CREA Director; Assistant Professor of Educational Policy, UNLV*

**Tiberio Garza**, *CREA Associate Director; Assistant Professor of Educational Psychology, UNLV*

**Jonathan Hilpert**, *Director, Office of Learning Analytics, Associate Professor of Learning Analytics, UNLV*

**Adam Kho**, *Assistant Professor of Education, USC*



ACKNOWLEDGMENTS

We gratefully acknowledge the individuals who gave time and effort to support this report. In particular, we are appreciative of the over 6,000 educators who participated in our survey to building administrators and teachers on the Nevada Educator Performance Framework. We also acknowledge the time and effort put in by representatives at the Nevada Department of Education. Specifically, we would like to thank Felicia Gonzales, Jason Dietrich, Kristin Withey, Kathleen Galland-Collins, Alberto Quintero, and KellyLynn Charles for helping coordinate our efforts with those of the Nevada Department of Education. At the University of Nevada, Las Vegas, we thank Monica Johnson, Elizabeth Hofschulte Collins, and Libna Garcia for their assistance with data collection.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# KEY TERMS

**Building Administrator-** An administrator that provides primarily administrative services at the school level.

**Content Validity-** Refers to the match between the items of a measurement tool and the entire domain in purports to measure.

**Construct Validity-** Whether a test actually measures the construct it intends to measure, including the ability to distinguish among types of performance and types of performers.

**Cronbach's Alpha-** A measure of the internal consistency that estimates the average inter-item correlation among items in a single hypothesized measure.

**Dimensionality-** An indication of whether underlying items of a measurement tool capture similar or different concepts.

**Domain-** The primary area of focus for evaluation. The NEPF for teachers defines three domains: Instructional Practice, Professional Responsibilities, and Student Outcomes. The NEPF for administrators defines three domains: Instructional Leadership, Professional Responsibilities, and Student Outcomes. NEPF domains are made up of standards.

**Eigenvalue-** A statistic that captures the amount of total variation explained by a factor in the underlying data.

**Exploratory Factor Analysis-** A statistical procedure used to uncover the underlying relationships between different items of measure. The procedure provides information on the dimensionality of an instrument by locating the smallest number of factors needed to explain correlations among the underlying items.

**Evaluation Cycle-** Consists of the goal-setting and self-assessment processes and a number of supervisory observation cycles with feedback provided to educators during and at the completion of the process.

**Evidence-** Data gathered through the evaluation cycle to support educators' progress on NEPF indicators, standards, and domains. Includes supervisor observation and progress towards meeting the Student Learning Goal.

**Face Validity**- According to those familiar with the measure, measures with high face validity appear to be measuring what they purport to measure.

**Factor Loading**- The relationship between the individual variables and the factor such that scores closer to 1 indicate a stronger relationship with the factor.

**Impact-** An assessment of the contribution of a phenomena to the achieved outcomes.

**Indicator-** The specific activity or process that provides an indication as to an educators' progress on a specific NEPF standards. Indicators are the building block of NEPF standards.

**Internal Consistency-** A measure of whether test items that purport to measure the same thing report similar scores across the same respondent.

**Kurtosis-** A measure of the height and sharpness of the central peak of a distribution relative to a normal distribution where positive values indicate a peak that is higher than typical normal curve and negative values indicate a peak that is lower than a normal curve.

**Nevada Educator Performance Framework-** The statewide system by which teachers' and administrators' performance is measured in Nevada.

**Predictive Validity-** Refers to whether a measurement tool actually predicts scores on another measure that it should theoretically predict.

**Probationary-** A teacher or administrator who is employed on a contract basis for three 1-year periods and has no right to employment after any of the three probationary contract years.

**Post-Probationary-** A teacher or administrator who completed their three-year probationary period, received a designation of "highly effective" or effective" on each of their performance evaluations for 2 consecutive school years, and received a notice of reemployment after the third year of their probationary period.

**Reliability-** A measure of the trustworthiness of a measurement tool—whether the tool consistently yields the same results given similar inputs.

**Skew Statistic-** A measure of the asymmetry of the distribution where negative values over -1 or positive values over 1 typically indicate a strong skew.

**Standard-** The defined statements within NEPF domains that capture what teachers and administrators are expected to know and do. NEPF standards are made up of individual indicators.

**Student Learning Goal-** a pupil-centered goal established in consultation with the supervisor for the duration of the evaluation cycle. Used as the measure of growth in the NEPF Student Outcomes domain.

**Teacher-** a licensed employee the majority of whose working time is devoted to the rendering of direct educational service to pupils of a school district.

**Weight-** The relative importance applied to an NEPF domain in determining an educators' final NEPF rating.

# NEPF TEACHER INSTRUCTIONAL PRACTICE STANDARDS

Standard 1: New Learning is Connected to Prior Learning and Experience

- 1.1. Teacher activates all students' initial understandings of new concepts and skills.
- 1.2. Teacher makes connections explicit between previous learning and new concepts and skills for all students.
- 1.3. Teacher makes clear the purpose and relevance of new learning for all students.
- 1.4. Teacher provides all students opportunities to build on or challenge initial understandings.

Standard 2: Learning Tasks have High Cognitive Demand for Diverse Learners

- 2.1. Tasks purposefully employ all students' cognitive abilities and skills.
- 2.2. Tasks place appropriate demands on each student.
- 2.3. Tasks progressively develop all students' cognitive abilities and skills.
- 2.4. Teacher operates with a deep belief that all children can achieve regardless of race, perceived ability and socio- economic status.

Standard 3: Students Engage in Meaning-Making through Discourse and Other Strategies

- 3.1. Teacher provides opportunities for extended, productive discourse between the teacher and student(s) and among students.
- 3.2. Teacher provides opportunities for all students to create and interpret multiple representations.
- 3.3. Teacher assists all students to use existing knowledge and prior experience to make connections and recognize relationships.
- 3.4. Teacher structures the classroom environment to enable collaboration, participation, and a positive affective experience for all students.

Standard 4: Students Engage in Metacognitive Activity to Increase Understanding of and Responsibility for Their Own Learning

- 4.1. Teacher and all students understand what students are learning, why they are learning it, and how they will know if they have learned it
- 4.2. Teacher structures opportunities for self- monitored learning for all students.
- 4.3. Teacher supports all students to take actions based on the students' own self-monitoring processes.

Standard 5: Assessment is Integrated into Instruction

- 5.1. Teacher plans on-going learning opportunities based on evidence of all students' current learning status.

- 5.2. Teacher aligns assessment opportunities with learning goals and performance criteria.
- 5.3. Teacher structures opportunities to generate evidence of learning during the lesson of all students.
- 5.4. Teacher adapts actions based on evidence generated in the lesson for all students.

# NEPF TEACHER PROFESSIONAL RESPONSIBILITIES STANDARDS

Standard 1: Commitment to the School Community

- 1.1. The teacher takes an active role on the instructional team and collaborates with colleagues to improve instruction for all students.
- 1.2. The teacher takes an active role in building a professional culture that supports school and district initiatives.
- 1.3. The teacher takes an active role in cultivating a safe, learning-centered school culture and community that maintains high expectations for all students.

Standard 2: Reflection on Professional Growth and Practice

- 2.1. The teacher seeks out feedback from instructional leaders and colleagues, and uses a variety of data to self- reflect on his or her practice.
- 2.2. The teacher pursues aligned professional learning opportunities to support improved instructional practice across the school community.
- 2.3. The teacher takes an active role in mentoring colleagues and pursues teacher leadership opportunities.

Standard 3: Professional Obligations

- 3.1. The teacher models and advocates for fair, equitable, and appropriate treatment of all students and families.
- 3.2. The teacher models integrity in all interactions with colleagues, students, families, and the community.
- 3.3. The teacher follows policies, regulations, and procedures specific to role and responsibilities.

Standard 4: Family Engagement

- 4.1. The teacher regularly facilitates two-way communication with parents and guardians, using available tools that are responsive to their language needs, and includes parent/guardian requests and insights about the goals of instruction and student progress.
- 4.2. The teacher values, respects, welcomes, and encourages students and families, of all diverse cultural backgrounds, to become active members of the school and views them as valuable assets to student learning.
- 4.3. The teacher informs and connects families and students to opportunities and services according to student needs.

Standard 5: Student Perception

- 5.1. The students report that the teacher helps them learn.
- 5.2. The students report that the teacher creates a safe and supportive learning environment.

- 5.3. The students report that the teacher cares about them as individuals and their goals or interests.

# NEPF ADMINISTRATOR INSTRUCTIONAL LEADERSHIP STANDARDS

Standard 1: Creating and Sustaining a Focus on Learning

- 1.1. Administrator engages stakeholders in the development of a vision for high student achievement and college and career readiness, continually reviewing and adapting the vision when appropriate.
- 1.2. Administrator holds teachers and students accountable for learning through regular monitoring of a range of performance data.
- 1.3. Administrator structures opportunities to engage teachers in reflecting on their practice and taking improvement actions to benefit student learning and support professional growth.
- 1.4. Administrator systematically supports teachers' short-term and long-term planning for student learning through a variety of means.

Standard 2: Creating and Sustaining a Culture of Continuous Improvement

- 2.1. Administrator sets clear expectations for teacher performance and student performance and creates a system for consistent monitoring and follow-up on growth and development.
- 2.2. Administrator supports teacher development through quality observation, feedback, coaching, and professional learning structures.
- 2.3. Administrator gathers and analyzes multiple sources of data to monitor and evaluate progress of school learning goals to drive continuous improvement.
- 2.4. Administrator operates with a deep belief that all children can achieve regardless of race, perceived ability and socio-economic status.

Standard 3: Creating and Sustaining Productive Relationships

- 3.1. Administrator demonstrates a welcoming, respectful, and caring environment and an interest in adults' and students' well-being to create a positive affective experience for all members of the school community.
- 3.2. Administrator provides opportunities for extended, productive discourse between the administrator and teachers and among teachers to support decision-making processes.
- 3.3. Administrator structures the school environment to enable collaboration between administrators and teachers and among teachers to further school goals.
- 3.4. Administrator has structures and processes in place to communicate and partner with teachers and parents in support of the school's learning goals.

Standard 4: Creating and Sustaining Structures

- 4.1. Administrator implements systems and processes to align curriculum, instruction, and assessment to state standards and college-readiness standards, continually reviewing and adapting when appropriate.

- 4.2. Administrator develops systems and processes to implement a coherent and clearly articulated curriculum across the entire school, continually reviewing and adapting when appropriate.
- 4.3. Administrator allocates resources effectively, including organizing time, to support learning goals.

# NEPF ADMINISTRATOR PROFESSIONAL RESPONSIBILITIES STANDARDS

Standard 1: Manages Human Capital

- 1.1. The administrator collects high quality observation data and evidence of teacher practice in a fair and equitable manner, and utilizes the results of evaluations to provide supports to improve performance.
- 1.2. The administrator uses available data, including teacher effectiveness data, to identify, recognize, support, and retain teachers.
- 1.3. The administrator supports the development of teacher leaders and provides leadership opportunities.
- 1.4. The administrator complies with the requirements and expectations of the Nevada Teacher Evaluation Framework.

Standard 2: Self-Reflection and Professional Growth

- 2.1. The administrator seeks out feedback from colleagues and staff, and uses a variety of data to self-reflect on his or her practice.
- 2.2. The administrator seeks opportunities to increase their professional knowledge in an effort to remain current on educational research and evidence-based practices.
- 2.3. The administrator pursues aligned professional learning opportunities to improve his/her instructional leadership across the school community.

Standard 3: Professional Obligations

- 3.1. The administrator models and advocates for fair, equitable, and appropriate treatment of all personnel, students, and families.
- 3.2. The administrator models integrity in all interactions with colleagues, staff, students, families, and the community.
- 3.3. The administrator respects the rights of others with regard to confidentiality and dignity, and engages in honest interactions.
- 3.4. The administrator follows policies, regulations, and procedures specific to role and responsibilities.

Standard 4: Family Engagement

- 4.1. The administrator involves families and the community in appropriate policy implementation, program planning, and assessment.
- 4.2. The administrator involves families and community members in the realization of vision and in related school improvement efforts.
- 4.3. The administrator connects students and families to community health, human, and social services as appropriate.

# EXECUTIVE SUMMARY

## BACKGROUND

In 2019, the Nevada Legislature enacted SB 475, which required a study assessing the impact and validity of the NEPF. The Center for Research, Evaluation, and Assessment at the University of Nevada, Las Vegas, in coordination with the University of Southern California undertook this study beginning in April 2020. This report represents the culmination of this effort.

Using nine different data sources, this report answers the following research questions:

1. Are the following components of the NEPF appropriate to positively impact teacher and administrator practice and outcomes?:
    a. The content of the NEPF domains.
    b. The internal consistency of the NEPF domains.
    c. The NEPF domain score ranges.
    d. The weighting of each NEPF domain.

2. What is the correlation between NEPF domains for teachers and administrators?

3. What is the year-to-year variation in teacher and administrator NEPF scores?

4. Does school growth in the percentage of teachers/administrators scoring highly on NEPF standards relate to growth in student achievement?

5. Do teachers and administrators believe that growth on the NEPF is related to growth in instructional practices?

## KEY FINDINGS

Our key findings in relation to the validity of the NEPF are summarized in Table 1 and described below. Determining whether a measurement tool is valid is an evaluative judgement based on the weight of evidence in favor of multiple forms of validity as described here (Messick, 1995).

- The NEPF has High Reliability

The NEPF shows strong reliability (i.e. internal consistency) with a Cronbach's alpha of 0.96 for the 10 teacher standards and a Cronbach's alpha of 0.93 for the 8 administrator standards. Cronbach's alpha tells us that the NEPF is yielding similar scores across the same unit—that is if a given school or district is scoring highly on one standard of the teacher or administrator NEPF, they are also scoring highly on another standard of the teacher or administrator NEPF.

**Table 1. Summary of Validity Findings for the Teacher and Administrator NEPF**

| | Reliability | Validity | | |
|---|---|---|---|---|
| | | Face Validity | Construct Validity | Predictive Validity |
| Teacher NEPF | High | High | Low | Moderate |
| Administrator NEPF | High | High | Low | Moderate |

- The NEPF has High Face Validity

Based on our survey analysis of educator perceptions, most teachers and administrators believe the NEPF is a valid measure of their performance. Face validity captures whether professionals believe a measure is valid. This is consistent with findings on other evaluation systems around the country (Grissom, Blissett, & Mitani, 2018). Observation-based evaluation systems (versus those that rely more heavily on measures of student achievement growth) tend to have higher face validity with educators (Cohen & Goldhaber, 2016).

- The NEPF has Low Construct Validity

Construct validity is achieved when the intended domains (e.g. Instructional Practice and Professional Responsibilities) of a measurement tool are actually distinguishable when they are measured. By design, the NEPF hypothesizes a two factor structure—it groups a series of standards under Instructional Practice and a series of standards under Professional Responsibility. We find evidence the teacher NEPF is best conceived as a single measure of teacher performance versus one that distinguishes between aspects of teacher effectiveness. In other words, evaluators are not distinguishing their ratings between the Instructional Practice domain versus the Professional Responsibilities domain. We find the administrator NEPF distinguishes between two domains, but not along the lines intended by system designers—some Professional Responsibility standards group best with other Instructional Leadership standards and vice versa. This finding is in alignment with a validity study of Washoe County's teacher evaluation system (who use a state-approved alternative to the NEPF) (Lash, Tran, & Huang, 2016) and other evaluation systems around the country (Grissom, Blissett, & Mitani, 2018).

Construct validity is also achieved when a measurement tool is able to distinguish among different types of performers. A valid evaluation tool should be able to tell the difference between high performing and low performing teachers, for example. We find that the teacher and administrator NEPF have low construct validity in that most educators score a final rating of Effective and Highly Effective. During 2018-19, only a tenth of a percent of teachers were classified as Ineffective and only 1.7% as Developing. The vast majority (81.90%) are classified as Effective with another 16% as Highly Effective. During 2018-19, no administrator received a rating of Ineffective in

2018-19 and only 1 percent received a rating of Developing. Most administrators (79%) receive ratings of Effective with another 21% rated as Highly Effective. Evaluation systems around the country, particularly those that rely primarily on observation-based measures of performance, are struggling with a similar result (Grissom & Loeb, 2017; Kraft & Gilmour, 2017; Weisberg et al., 2009).

- The NEPF has Moderate Predictive Validity

Predictive validity captures whether a measurement tool predicts (or is related to) other performance measures that it should be theoretically related to (i.e. student achievement). We find moderate correlations between the teacher NEPF and student achievement. School-level NEPF teacher scores are positively associated with the percentage of students scoring Meets (0.29), Exceeds (0.30), and Proficient (0.33) in math and the percentage of students scoring Meets (0.25), Exceeds (0.23), and Proficient (0.28) in reading. Similar correlations are found between district-level administrator NEPF scores and student proficiency. This is consistent with modest correlations between observation-based teacher evaluation systems and student achievement (Garrett & Steinberg, 2015). However, these results should not be read as evidence that the NEPF is causing higher achievement.

Our key findings in relation to the impact of the NEPF are summarized in Table 2 and described below.

**Table 2. Summary of Impact Findings for the Teacher and Administrator NEPF**

| | Impact | |
|---|---|---|
| | Student Achievement Growth (Statistical) | Teacher Practice (Perceived) |
| Teacher NEPF | Low | High |
| Administrator NEPF | Low | High |

- The NEPF Has No Impact on Achievement Growth

When looking at the impact of NEPF growth on growth in reading and math achievement, we find no significant impact. This is likely due to there being very little detectable growth on the NEPF in the first place. Using the available school- and district-level data, we find that schools and districts grow very little in their NEPF scores over time. The mean year-to-year change in teacher and administrator NEPF scores is essentially zero. The distributions indicate a ceiling effect, where nearly all educators score “Effective” with some scoring “Highly Effective” and very few score “Ineffective” or “Developing.” Consequently, there is very little room on the scoring scale for educators to grow.

- Most Educators Believe the NEPF has an Impact on Practice and Student Outcomes

Based on NDE survey data in 2018-19, 17% of teachers strongly agreed and 50% of teachers agreed that the NEPF is helping them identify areas for growth. A corresponding 24% of teachers disagreed and 9% of teachers strongly disagreed with this statement. In 2018-19, 57% of administrators in Nevada agreed and 11% strongly agreed that the NEPF is benefiting student outcomes whereas 25% disagreed and 6% strongly disagreed.

# RECOMMENDATIONS

- NDE Should Engage in Strategies to Improve Differentiation in Scores Between NEPF Domains

NDE could consider a few strategies to improve evaluator differentiation between the Practice and Responsibilities domains. NDE should look into providing ongoing training in interrater reliability, particularly with regard to the various indicators and sources of evidence used to determine a rating. This could include think-aloud activities where raters are asked to watch a brief observation video and discuss aloud their reasoning for their scoring.

- NDE Should Engage in Strategies to Improve the Distribution in NEPF Final Scores

Here again, additional think-aloud activities could help. NDE could also consider increasing the number of performance levels to create truly inadequate performance levels at the bottom of the scoring range that are rarely used. NDE should also investigate whether school districts, in their implementation of the NEPF, are requiring equal evidence requirements across the rating categories so as to remove the incentive for evaluators (especially those evaluating a high number of educators) to assign ratings of Effective. NDE should investigate the quality of the feedback being provided to educators. In a system with little variation in scoring, the only way to drive student growth is through quality feedback that engages educators in continuous improvement.

- NDE Should Engage in a More Comprehensive and Systematic Data Collection Effort of Individual-Level NEPF Data

The school-level and district-level data used for the bulk of this report likely masks important patterns in the individual-level data. Additionally, aggregated data reduces the power of statistical tests to find meaningful relationships. We recommend that NDE engage in a more comprehensive and systematic effort around the collection of individual-level NEPF data in a way that can track individual educator growth over time while at the same time protect educator privacy.

- NDE Should Improve Its Current NEPF Reporting Process

In order to allow for quick and systematic data reporting and to avoid the prevalence of errors that can occur when using Excel as a data collection tool, we recommend NDE invest in a more comprehensive data management tool—one that can handle individual-level data inputs from school districts, collect data on NEPF indicators, standards, and domains, and allow for streamlined reporting to NDE.

## LIMITATIONS

Our findings are limited by a few factors. We only have individual-level data for a single year (2018-19), requiring us to mainly utilize school- or district-aggregate data over time. These data sources suffer from aggregation bias, or the idea that data that is aggregated to higher-level units can mask important information and patterns in the individual units. Individual data might lead to different conclusions. Second, we only have data on NEPF standard and domain scores. We were unable to assess any patterns in NEPF indicators, and thus are unable to determine which NEPF indicators are performing better or worse. Finally, this analysis was performed on a quick five week timeline to the first presentation, thereby narrowing the scope of the report to focus on the most pressing research questions. A longer time horizon for reporting would allow for an even more detailed exploration of the NEPF and its impact and validity, including detailed interviews with administrators and teachers on NEPF implementation.

# INTRODUCTION

## PURPOSE OF THE REPORT

On January 17th, 2020, the Nevada Department of Education (NDE) in coordination with the State of Nevada Purchasing Division, put forth a request for proposals from vendors to provide an impact and validity study for the Nevada Educator Performance Framework (NEPF) (Solicitation 30DOE-S1026) as required by SB 475 (2019). The Center for Research, Evaluation, and Assessment (CREA) at the University of Nevada, Las Vegas, in coordination with the University of Southern California (USC) undertook this study beginning in April 2020 with a planned completion date of July 2020. This evaluation includes analyses of surveys of Nevada principals and teachers, teacher and administrator NEPF scores, and school achievement data.

The purpose of this report is to provide an overview of the NEPF for teachers and administrators, to provide an assessment on the validity of the NEPF as a mechanism for assessing teacher and administrator performance, and to provide an assessment of the impact of the NEPF on teacher and administrator practice and student performance. As such, this report is divided into two strands—the **validity strand** and the **impact strand**.

The **validity strand** aims to answer the following research questions:

1. Are the following components of the NEPF appropriate to positively impact teacher and administrator practice and outcomes?:
    a. The content of the NEPF domains.
    b. The internal consistency of the NEPF domains.
    c. The NEPF domain score ranges.
    d. The weighting of each NEPF domain.

2. What is the correlation between NEPF domains for teachers and administrators?

3. What is the year-to-year variation in teacher and administrator NEPF scores?

The **impact strand** aims to answer the following research questions:

4. Does school growth in the percentage of teachers/administrators scoring highly on NEPF standards relate to growth in student achievement?

5. Do teachers and administrators believe that growth on the NEPF is related to growth in instructional practices?

# THE NEVADA EDUCATOR PERFORMANCE FRAMEWORK: A DECADE IN REVIEW

In 2009, a highly influential report entitled “The Widget Effect” published by the New Teacher Project criticized existing teacher performance evaluation systems in 12 school districts for failing to distinguish between effective and ineffective teachers in classrooms (Weisberg et al., 2009). The Widget Effect encapsulated a belief that school districts often treat teachers as interchangeable parts, failing to recognize the difference in instructional effectiveness across classrooms. The report coincided with a movement at the federal level to incentivize states to change the way they identify and reward effective teaching. As a part of the American Recovery and Reinvestment Act of 2009, President Barack Obama launched the Race to the Top federal grant competition, providing grant-based support to states willing to institute educational policies that, in part, overhauled performance evaluation systems for teachers and administrators.

States responded with a flurry of legislation aimed at revamping existing evaluation systems for teachers and administrators. During the 2011 state legislative sessions, 19 states enacted comprehensive changes to the way they evaluate teachers and administrators, including Nevada (Marianno, 2015). Two-thousand and eleven commenced Nevada’s 76$^{th}$ Legislative Session (2011), in which the Legislature enacted Assembly Bill 222, creating the 15 member Teachers and Leaders Council (TLC) of Nevada. The bill and the convening of the TLC served as a starting point for the development of the NEPF, the state’s new statewide teacher and administrator performance evaluation system.

The TLC commenced their work on the NEPF with a due date of July 1, 2013 (and implementation during the 2014-15 school year). AB 222 required that this new system recommended by the TLC significantly factor student academic achievement into teachers’ and administrators’ final evaluation rating—50% of the evaluation must be based on student achievement. Additionally, the teacher evaluation system was required to identify whether teachers employ practices that involve and engage parents and families in the classroom. Final evaluation ratings were also required to be on a four point scale of Highly Effective, Effective, Minimally Effective, and Ineffective. The TLC received $32,000 in appropriations for this effort.

**Figure 1. NEPF Goals and Purposes**

| Goals | Purposes |
|---|---|
| 1. Foster student learning and growth.<br>2. Improve educators' instructional practices.<br>3. Inform human capital decisions based on a professional growth system.<br>4. Engage stakeholders in the continuous improvement and monitoring of a professional growth system. | To identify effective instruction and leadership and to establish criteria to determine whether educators:<br><br>1. Are helping students meet achievement targets and performance expectations.<br>2. Are effectively engaging families.<br>3. Are collaborating effectively.<br>4. Are growing through targeted professional development and support.<br>5. Have information on which to base human capital decisions including rewards and consequences.<br>6. Are using data to inform decision-making. |

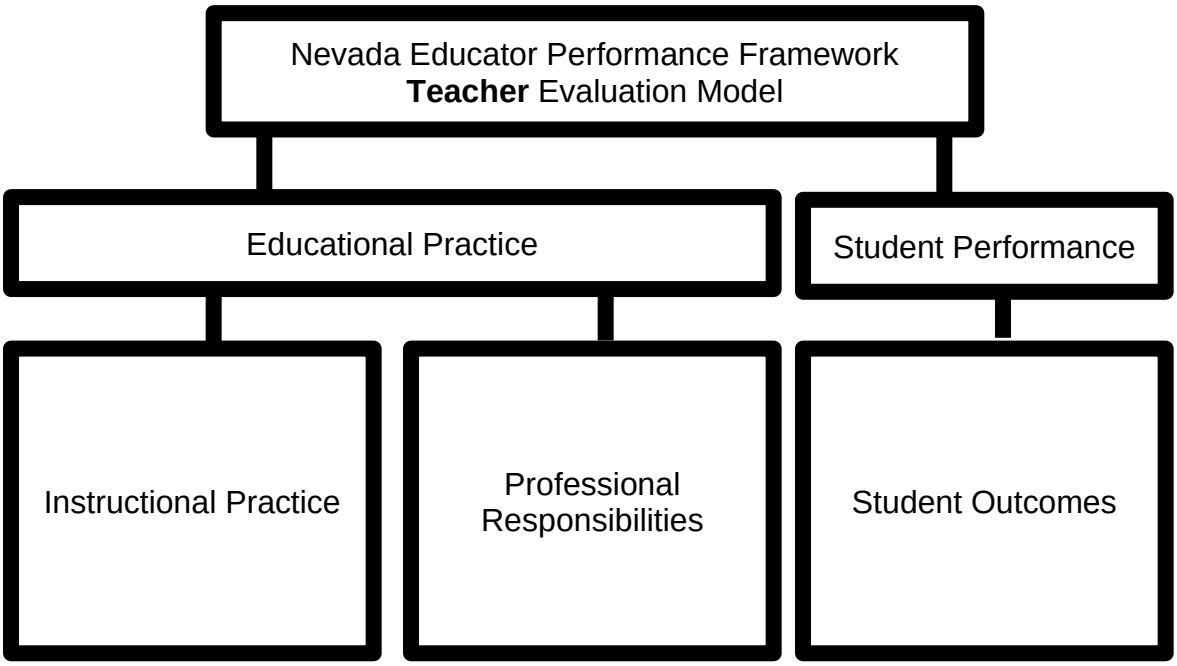
## 2013: THE NEPF IS PRESENTED AND IMPLEMENTATION IS DELAYED

The TLC issued recommendations in 2012, effectively establishing the goals and purposes for the new NEPF performance evaluation system (see Figure 1) (Fitzpatrick & Salazar, 2012; Nevada Teachers and Leaders Council, 2013). Next, the TLC set forth the components of the NEPF teachers' and administrators' evaluation systems shown in Figures 2 and 3. Importantly, they recommended that the new performance evaluation systems be made up of three domains that fall into two overarching categories: Educational Practice and Student Performance.

For teachers, the Educational Practice category is made up of two of the three domains- Instructional Practice and Professional Responsibilities. The Student Performance category is made up of the Student Outcomes domain. For administrators, the Educational Practice category is made up two of the three domains- Instructional Leadership and Professional Responsibilities. Like the teacher system, the Student Performance category is made up of the Student Outcomes domain.

The TLC did not recommend the weighting for the Educational Practice domains. The Student Outcomes domain would account for 50% of both teacher and administrator final evaluation scores, per the guidelines set forth in Assembly Bill 222. The TLC further recommended that implementation of the NEPF be delayed until a validation study could be completed. Additionally, the council recommended that the student achievement portion of the NEPF focus on student academic growth and be appropriately adjusted for teachers in non-tested grades.

**Figure 2. NEPF Teacher Evaluation Model**

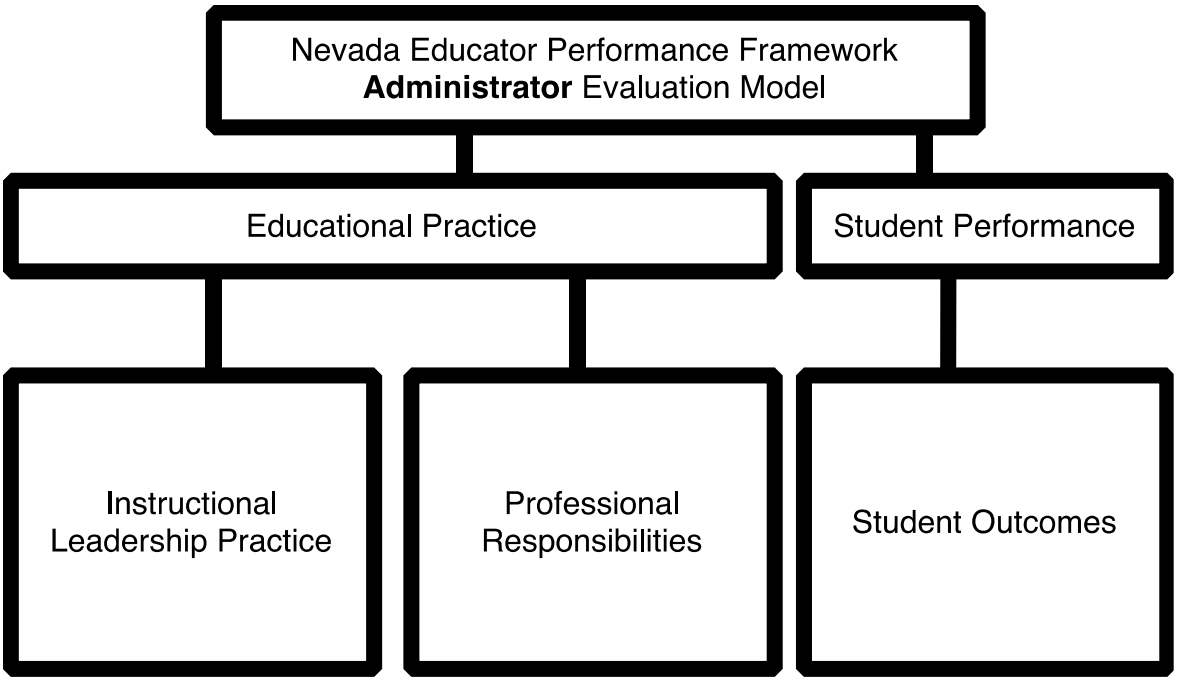



During Nevada's 77th Legislative Session (2013), the Legislature enacted Senate Bill 407, which delayed the implementation of the NEPF performance evaluation system from the 2014-15 school year to the 2015-16 school year. Instead, SB 407 required that school districts participate in a validation study during the 2014-15 school year in preparation for full implementation the following year.

**2014-15: THE NEPF IS STUDIED**

During the 2014-15 school year, the Regional Professional Development Programs provided training to school districts on how to implement the NEPF evaluation systems for teachers and administrators. One of the key tasks of the researchers completing the validation study was to determine the appropriate weighting for the educational practice and professional responsibilities domains (since the Student Outcomes domain was already set at 50%). During the 2014-15 school year in which the validation study occurred, the NEPF teacher and administrator evaluation systems relied on the preliminary weighting for teacher and administrators shown in Table 3.

**Figure 3. NEPF Administrator Evaluation Model**




West Ed completed a two-year validation study in June of 2015 (WestEd, 2015). They found fairly low interrater reliability (the percent of agreement between expert raters and school administrator raters) in the teacher observation ratings, particularly in regards to Instructional Practice Standards 4 and 5 (Students engage in metacognitive activity to increase their own learning and assessment is used to integrate instruction). Focus groups with superintendents yielded positive opinions of the NEPF training but concern that educators may not be adequately prepared to roll out the NEPF during the 2015-16 school year. This concern was mirrored in the survey to teachers and site administrators and phone call interviews with site administrators. Importantly, the study provided no guidance on the appropriate weighting for the NEPF on the Educational Practice and Professional Responsibilities domains nor did it provide any assessment of the validity of the NEPF beyond educators' own judgements (which were high—educators' believe the NEPF identifies practices and responsibilities associated with quality instruction and professionalism).

**Table 3. Initial Domain Weights During NEPF Study Year (2014-15)**

| Teachers | | | Administrators | | |
|---|---|---|---|---|---|
| Domain | Weight | Evidence | Domain | Weight | Evidence |
| Instructional Practice | 35% | Observation Cycles | Instructional Leadership | 35% | Observation Cycles |
| Professional Responsibilities | 15% | Observation Cycles | Professional Responsibilities | 15% | Observation Cycles |
| Student Outcomes | 50% | School growth (35%)<br>School proficiency (5%)<br>Achievement gap reduction (10%) | Student Outcomes | 50% | School growth (35%)<br>School proficiency (5%)<br>Achievement gap reduction (10%) |

## 2015-2017: THE NEPF IS ROLLED OUT

Following the completion and presentation of the NEPF validation study, the Legislature again revisited the NEPF during Nevada's 78th Legislative Session (2015). In particular, the Legislature enacted Assembly Bill 447, which made further changes to the NEPF. The bill specifically prohibited the use of the student outcome data in the evaluation of teachers and administrators during the 2015-16 school year due to testing irregularities. Instead, for 2015-16, the Practice (80%) and Professional Responsibilities (20%) domains were given primary weight, as shown in Table 4.

During the 2016-17 school year, the student outcomes domain made a return to the NEPF, as shown in Table 5. For the first time, NDE introduced the Student Learning Goal (SLG), which reduced the Student Outcome domain's reliance on student testing data for both teachers and administrators (Nevada Department of Education, 2016). In addition, school-aggregate student proficiency accounted for 10% of a teachers' or administrators' final rating. Progress towards an individually determined SLG accounted for the other 10%. The weighting of the Instructional Practice (for teachers) and Instructional Leadership (for administrators) were reduced from 80% to 60% to provide room for the reemergence of the Student Outcomes Domain.

In terms of the SLG, using a goal-setting and planning tool, teachers and administrators work with their supervisors to review objectives, standards, and student performance to determine the most important area for student learning and set an attainable goal for student progress in that area. Further, they determine the assessments used to measure progress on the goal, collect baseline data, and monitor progress during the school year. Finally, the educator and supervisor review the results and the supervisor

assigns a one through four rating based on the SLG rubric and the amount of achievement growth experienced by students in the area articulated by the SLG.

**Table 4. Domain Weights During NEPF Rollout Years (2015-16)**

| Teachers | | | Administrators | | |
|---|---|---|---|---|---|
| Domain | Weight | Evidence | Domain | Weight | Evidence |
| Instructional Practice | 80% | Observation Cycles | Instructional Leadership | 80% | Observation Cycles |
| Professional Responsibilities | 20% | Observation Cycles | Professional Responsibilities | 20% | Observation Cycles |
| Student Outcomes | 0% | | Student Outcomes | 0% | |

**Table 5. Domain Weights During NEPF Rollout Years (2016-17)**

| Teachers | | | Administrators | | |
|---|---|---|---|---|---|
| Domain | Weight | Evidence | Domain | Weight | Evidence |
| Instructional Practice | 60% | Observation Cycles | Instructional Leadership | 60% | Observation Cycles |
| Professional Responsibilities | 20% | Observation Cycles | Professional Responsibilities | 20% | Observation Cycles |
| Student Outcomes | 20% | School proficiency (10%) Student Learning Goal (SLG) (10%) | Student Outcomes | 20% | School proficiency (10%) Student Learning Goal (SLG) (10%) |

## 2017-2019: THE NEPF IS ADJUSTED AGAIN

During the 2017 legislative session (Nevada's 79th Legislative Session) lawmakers officially codified the SLG as the primary mechanism to evaluate educators' impact on the Student Outcomes domain. Assembly Bill 320 required that any educator who provides direct instructional services to students develop learning goals. Further, the law requires that the SLG account for 20% of an educators' final evaluation. The practice and professional responsibility weights remained unchanged from 2016-17 (see Table 6).

The 2018-19 school year brought further changes to the weighting of the practice, responsibilities, and outcomes domains. The Student Outcomes domain was scaled up

to represent 40% of an educators' final evaluation rating (per Assembly Bill 320 (2017)), comprised completely of an educators' progress towards their SLG. Additionally, the Instructional Practice/Leadership domain was scaled down to represent 45% of the final rating and the Professional Responsibilities domain changed to comprise 15% of the final rating (Nevada Department of Education, 2019). These weights are shown in Table 7.

When lawmakers met again during Nevada's 80th Legislative Session they once again revisited the NEPF (having done so in every legislative session since the TLC was first convened in 2011). Lawmakers enacted Senate Bill 475 which reduced the weight on the Student Outcomes domain from 40% to 15% during the 2019-20 school year (Nevada Department of Education, 2020). The Practice and Responsibilities domains subsequently received new weighting—65% and 20%, as shown in Table 8. Further, the bill required NDE to enter into a contract to study the impact and validity of the NEPF, thereby forming the impetus for this study.

**Table 6. Domain Weights During NEPF Adjustment Years (2017-18)**

| Teachers | | | Administrators | | |
|---|---|---|---|---|---|
| Domain | Weight | Evidence | Domain | Weight | Evidence |
| Instructional Practice | 60% | Observation Cycles | Instructional Leadership | 60% | Observation Cycles |
| Professional Responsibilities | 20% | Observation Cycles | Professional Responsibilities | 20% | Observation Cycles |
| Student Outcomes | 20% | Student Learning Goal (SLG) | Student Outcomes | 20% | Student Learning Goal (SLG) |

**Table 7. Domain Weights During NEPF Adjustment Years (2018-19)**

| Teachers | | | Administrators | | |
|---|---|---|---|---|---|
| Domain | Weight | Evidence | Domain | Weight | Evidence |
| Instructional Practice | 45% | Observation Cycles | Instructional Leadership | 45% | Observation Cycles |
| Professional Responsibilities | 15% | Observation Cycles | Professional Responsibilities | 15% | Observation Cycles |
| Student Outcomes | 40% | Student Learning Goal (SLG) | Student Outcomes | 40% | Student Learning Goal (SLG) |

**Table 8. Domain Weights During NEPF Adjustment Years (2019-20)**

| Teachers | | | Administrators | | |
|---|---|---|---|---|---|
| Domain | Weight | Evidence | Domain | Weight | Evidence |
| Instructional Practice | 65% | Observation Cycles | Instructional Leadership | 65% | Observation Cycles |
| Professional Responsibilities | 20% | Observation Cycles | Professional Responsibilities | 20% | Observation Cycles |
| Student Outcomes | 15% | Student Learning Goal (SLG) | Student Outcomes | 15% | Student Learning Goal (SLG) |

## FINAL SCORE RANGES FOR TEACHERS AND ADMINISTRATORS

Notwithstanding changes to the weighting of the NEPF domains over the history of the NEPF, some elements have remained constant over time. In particular, the cutoff scores for determining teachers' or administrators' final rating as adopted by the State Board of Education have remained the same since 2015.

NRS 391.460 requires that the State Board of Education adopt regulations that require an employee's overall performance to be determined as 1) Highly Effective; 2) Effective; 3) Developing; or 4) Ineffective. Additionally, the State Board of Education has to include criteria for making the final rating designation.

Once the proper weighting is applied, teachers' and administrators' scores on each domain are combined to generate a final evaluation score. Teachers and administrators are then assigned a final rating category based on their final evaluation score. Beginning in the 2015-16 school year, the State Board of Education set the following cutoff scores for a teachers' and administrators' final rating (see Table 9):

**Table 9. Cutoff Scores for Educators' Final Rating (2015-16 to 2019-20)**

| Teachers | | Administrators | |
|---|---|---|---|
| Highly Effective | 3.6-4.0 | Highly Effective | 3.6-4.0 |
| Effective | 2.8-3.59 | Effective | 2.8-3.59 |
| Developing | 1.91-2.79 | Developing | 1.91-2.79 |
| Ineffective | 1.0-1.9 | Ineffective | 1.0-1.9 |

For teachers or administrators to receive a rating of Effective, they must have scored a 2, 3, or 4 on the Student Outcomes domain. For them to receive a rating of Highly Effective, they must have received a 3 or 4 on the Student Outcomes domain.

## SUPPORT AND DISMISSAL BASED ON FINAL RATINGS

Nevada law does not prescribe in detail the types of supports that must be provided to teachers and administrators who do not reach a final rating of Highly Effective, Effective, or Developing (i.e. are rated as Ineffective). Probationary teachers and administrators who receive a rating of Ineffective must be notified that their contract may be nonrenewed for the next school year and that they may have, at their request, a different evaluator for the next school year (NRS 391.725). Additionally, upon request of the teacher or administrator, an evaluator must make a "reasonable effort" to assist the educator in improving their performance (NRS 391.725) using the Educator Assistance Plan Tool.

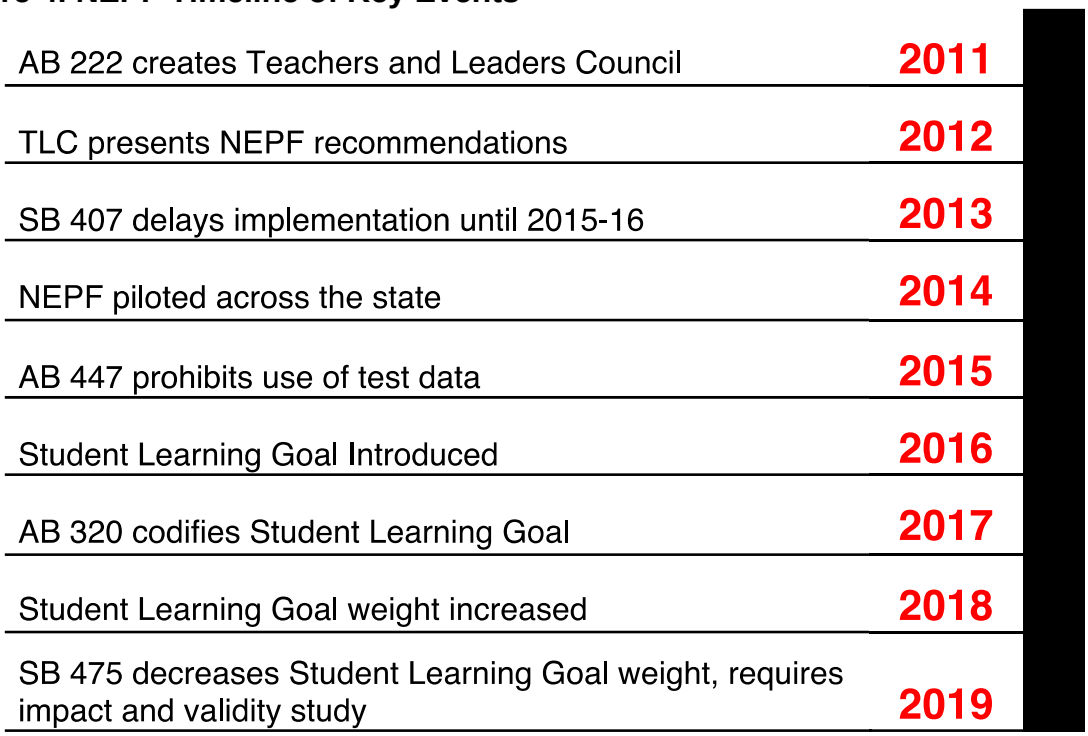
Teachers and administrators who successfully complete the three year probationary period and receive a rating of Highly Effective or Effective for two consecutive school years can move to post probationary status (NRS 391.820). A post probationary teacher or administrator who subsequently receives a rating of Ineffective or receives a rating of Developing and Ineffective during a two year consecutive period is required to serve an additional three year probationary period and may subsequently be dismissed during this probationary phase (NRS 391.730).

## SUMMARY

Since its conceptualization by the TLC in 2012, the singular constant with the NEPF has been change. In early phases of the NEPF (2013-2015), or the "presentation" and "study" years, the Student Outcomes domain and its reliance on state assessments was met with great skepticism by educators and their employee associations/unions. As designed, the NEPF required that 50% of an educators' final evaluation be based on student performance on standardize exams (growth, proficiency, and achievement gap reduction).

After the completion of validation study by WestEd (2014-15), the Student Outcomes Domain was not factored into educators' evaluations (during the 2015-16 school year). Instead, lawmakers planned to slowly phase in the domain over time, eventually reaching the full weight called for in statute. However, during the "rollout" and "adjustment" years of the NEPF, the Student Outcomes Domain morphed into what it is today. During the 2016-17 school year, administrative regulations introduced the SLG, which over time replaced any reliance on state assessments in teacher and administrator evaluations in Nevada. Accounting for 40% of educators' final rating in 2018-19 and 15% of educators' final rating in 2019-20, the SLG is based on local approved assessments and is now the primary way by which educators' contributions to student learning are assessed by the NEPF.

**Figure 4. NEPF Timeline of Key Events**



# DATA AND METHODS

## DATA SOURCES

To evaluate the validity and impact of the NEPF, we employed multiple methods of data collection, with each data collection method tailored to the research question being answered. Table 10 lists the nine different data sources used in this report.

**Table 10. Data Sources**

| Data | Variables of Interest | Source | Year(s) | Sample Size (n) |
|---|---|---|---|---|
| Teacher NEPF Scores | Teacher scores on NEPF domains and standards | Nevada Department of Education | 2018-2019 | n=20,813 |
| Administrator NEPF Scores | Administrator scores on NEPF domains and standards | Nevada Department of Education | 2018-2019 | n=1,229 |
| School-Aggregate Teacher NEPF Scores | School-average scores on NEPF domains and standards | Nevada Department of Education | 2015-2016-2018-2019 | 2015-2016: n= 540<br>2016-2017: n= 565<br>2017-2018: n= 567<br>2018-2019: n= 568 |
| District-Aggregate Administrator NEPF Scores | District-Average Scores on NEPF Domains and Standards | Nevada Department of Education | 2015-2016-2018-2019 | 2015-2016: n= 12<br>2016-2017: n= 11<br>2017-2018: n= 11<br>2018-2019: n= 11 |
| Nevada Report Card Data | School and district characteristics | Nevada Department of Education | 2015-2016–2018-2019 | Schools:<br>2015-2016: n= 664<br>2016-2017: n= 662<br>2017-2018: n= 684<br>2018-2019: n= 691<br>Districts:<br>2015-2016: n= 17<br>2016-2017: n= 17<br>2017-2018: n= 17<br>2018-2019: n= 17 |

**Table 8. Data Sources (Continued)**

| Data | Outcomes of Interest | Source | Year(s) | Sample Size (N) |
|---|---|---|---|---|
| NCES Common Core of Data | School and district characteristics | National Commission on Education Statistics | 2015-2016 to 2018-2019 | Schools:<br>2015-2016: N= 664<br>2016-2017: N= 662<br>2017-2018: N= 684<br>2018-2019: N= 691<br>Districts:<br>2015-2016: N= 17<br>2016-2017: N= 17<br>2017-2018: N= 17<br>2018-2019: N= 17 |
| Teacher Surveys | Perceptions about impact and validity of the NEPF | Nevada Department of Education, CREA developed survey | 2017-2018-2018-2019, 2019-2020 | NDE Survey<br>2017-2018: N=4,523<br>2018-2019: N=6,358<br>CREA Survey<br>2019-2020: N=5,840 |
| Administrator Surveys | Perceptions about impact and validity of the NEPF | Nevada Department of Education, CREA developed survey | 2017-2018-2018-2019, 2019-2020 | NDE Survey<br>2017-2018: N=448<br>2018-2019: N=665<br>CREA Survey<br>2019-2020: N=484 |
| Review of State Evaluation Systems | Characteristics of other state evaluation systems | CREA developed rubric | 2019-2020 | N=50 |

## TEACHER AND ADMINISTRATOR NEPF SCORES

NDE provided individual teacher and administrator NEPF scores for the 2018-19 school year. These individual scores were reported by each school district in the state. To protect the identity of teachers, any grade-level, school, or subject identifiers were removed from the data before being provided to NDE.

The teacher and administrator NEPF Score file contained the name of a teacher/administrators' district, their raw Practice score, their raw Responsibilities score, their raw SLG score, and their raw summative score. The file does not contain teacher and administrator scores on individual standards or on the indicators within standards.

The Teacher NEPF Score file contained information from 20,813 teachers across 16 Nevada school districts (a rural district was missing) for the 2018-19 school year. The Administrator NEPF Score file contained information from 1,229 administrators across 15 Nevada school districts (two rural districts were missing).

## SCHOOL-AGGREGATE TEACHER AND DISTRICT-AGGREGATE ADMINISTRATOR NEPF SCORES

Before 2018-19, NDE collected NEPF scores in the aggregate (at the school-level for teachers and at the district-level for administrators) and did so for the 2015-16 through the 2018-19 school year.

The School-Aggregate Teacher NEPF Scores file contains the name of the district, the name of the school, the total number of teachers at the school, the total number of teachers exempt from an evaluation that school year, the number of teachers given a final rating of Ineffective, Developing, Effective, and Highly Effective, the percentage of teachers given a final rating of Ineffective, Developing, Effective, and Highly Effective, the school average score on Instructional Practice standards 1, 2, 3, 4, and 5, the school average score on Professional Responsibilities standards 1, 2, 3, 4, and 5, the school average SLG score, and the school average final score.

The District-Aggregate Administrator NEPF Scores file contains the name of the district, the total number of administrators in the district, the total number of administrators exempt from an evaluation that school year, the number of administrators given a final rating of Ineffective, Developing, Effective, and Highly Effective, the percentage of administrators given a final rating of Ineffective, Developing, Effective, and Highly Effective, the district average score on Instructional Leadership standards 1, 2, 3, and 4, the district average score on Professional Responsibilities standards 1, 2, 3, and 4, the district average SLG score, and the district average final score.

In total, the School-Aggregate Teacher NEPF Scores file contains information for teachers at 540 schools (81%) in 2015-16, 565 schools (85%) in 2016-17, 567 schools in 2017-18 (83%), and 568 schools (82%) in 2018-19. The District-Aggregate Administrator NEPF Scores file contains information for 12 school districts (71%) in 2015-16 and 11 school districts (65%) in 2016-17, 2017-18, and 2018-19. The dataset does not contain information from the now defunct Achievement School District or the State Public Charter School Authority (who is not required to use the NEPF).

## NEVADA REPORT CARD AND NCES COMMON CORE OF DATA

We obtained information on demographic and other characteristics of Nevada schools and districts from the Nevada Report Card data and from the National Center for Education Statistics (NCES) Common Core of Data.

For a given school or district, from the Nevada Report Card Data file, we obtained information on the number of students enrolled at a school or district, the number tested in mathematics, the percent proficient in mathematics and reading, the percent developing in mathematics and reading, the percent approaching standard in mathematics and reading, the percent meeting standards in mathematics and reading, and the percent exceeding standards in mathematics and reading. We also obtained, for a given school or district, the percentage of American Indian/Alaskan Native

students, the percentage of Asian students, the percentage of Black students, the percentage of Hispanic students, the percentage of White students, the Percentage of Pacific Islander students, the percentage of students of two or more races, the percentage of male and female students, the percentage of English Language Learner (ELL) students, the percentage of Free and Reduced Price Lunch (FRL) students, and the percentage of migrant students.

From the NCES Common Core of Data file, we obtained information on the grade-level of schools, schools' Title I status, and the number of teachers within the school or district.

In total, these data contain information on 664 schools and 17 school districts for the 2015-16 school year, 662 schools and 17 school districts from the 2016-17 school year, 684 schools and 17 school districts from the 2017-18 school year, and 691 schools and 17 schools districts from the 2018-19 school year. We did not maintain information in the dataset from the State Public Charter School Authority.

## TEACHER AND ADMINISTRATOR SURVEYS

Since 2015-16, NDE has sent out an annual survey on the NEPF to building administrators and classroom teachers. The questions on this survey have changed considerably over time, with only four items remaining relatively consistent across all of the survey administrations. In 2018-19, the Department adjusted the survey and intends to maintain a relatively similar battery of questions in subsequent survey administrations which will help in judging trends in educators' opinions on the NEPF over time. For the purposes of our analyses, we leveraged data from the 2017-18 and 2018-19 survey administrations as these data tend to be the most complete. Questions on these surveys captured educators' perceptions on the impact and implementation of the NEPF.

The 2017-18 annual Department survey contained responses from 4,523 (42%) teachers and 448 administrators (40.2%). The 2018-19 annual Department survey contained responses from 6,358 teachers (58.4%) and 665 administrators (59.6%).

We supplemented the annual Department survey with a survey of our own, administered to teachers and building administrators during May 2020. The survey focused primarily on capturing educators' perceptions on the quality and amount of feedback they receive from their supervisor during their NEPF evaluation cycle on each of the Practice and Responsibility standards. We also asked a series of questions about the SLG process. Due to the ongoing disruptions of COVID-19, we asked that educators respond with a typical evaluation cycle in mind. A full list of the survey questions is provided in Appendix A and Appendix B.

In total, we received 5,840 (21%) responses from teachers and 484 responses from administrators (30%).

### REVIEW OF STATE EVALUATION SYSTEMS

We also performed a review of teacher and administrator evaluation systems in all 50 states, assessing other systems for their similarity with the NEPF. We generated a similarity index by scoring other states' educator evaluation systems based on the following 8 categories (scored for the teacher and administrator systems separately).

1. Does the state have an evaluation system that is locally-developed, state-developed, or locally-developed and state-aligned?
2. Does the system measure student growth by progress on a SLG, on summative state standardized tests only, or on local standardized assessments?
3. Does the evaluation system require probationary educators to be evaluated annually?
4. Does the evaluation system require at least 3 observations of probationary educators?
5. Does the evaluation system have a Practice domain?
6. Does the evaluation system have a Responsibilities domain?
7. Is feedback required to be provided?
8. Does the evaluation system have at least four final rating categories?

We developed the following 8 categories iteratively based on available information in the National Council on Teacher Quality (NCTQ) State Teacher Policy Database and our own review of evaluation systems. The State Teacher Policy Database tracks how states' measure student growth and professional practice. Additionally, the database has information on the frequency of evaluation and observations. This information is collected through a regular survey to state education administrators. We supplemented this information with our own internet search for states' evaluation system domains. In total, 32 states clearly provide evaluation system domains and standards for teachers on their agency websites. In total, 36 states clearly provide evaluation system domains and standards for administrators on their agency websites.

## METHODS

Table 11 lists the research questions with the corresponding data sources used to address each question. In what follows, we provide more detail on the methods by which we leverage each of the data sources to answer these research questions.

## ASSESSING THE VALIDITY OF THE NEPF

The **validity strand** aims to answer the following research questions:

1. Are the following components of the NEPF appropriate to positively impact teacher and administrator practice and outcomes?:

    a. The content of the NEPF domains

To understand whether the content of the NEPF domains were appropriate to positively impact teacher and administrator practice and outcomes we relied on our systematic review of state teacher evaluation systems around the country. We first explored how similar or different other evaluation systems are around the country. As mentioned above, we first generated a similarity index by scoring other states' educator evaluation systems based on 8 criteria (scoring states' teacher and administrator systems separately).

We also relied on educator perceptions of the fairness and validity of the NEPF utilizing NDE surveys from the 2017-18 and 2018-19 school year and our own CREA survey administered during the 2019-20 school year. The perceptions of educators allow for an assessment of the face validity of the NEPF, or the degree to which the NEPF actually measures what it purports to measure according to those most familiar with it.

    b. The internal consistency of the NEPF domains

To assess whether the internal consistency of the NEPF domains were appropriate to positively impact teacher and administrator practice and outcomes, we conducted two separate analyses using the School-Aggregate Teacher NEPF Scores file and the District-Aggregate Administrator NEPF scores file. We included data from 2016-17 to 2018-19, and we excluded 2015-16 due to the prevalence of missing data. We treated each school-year and district-year observation as independent. We also ran the analysis on each year separately and found very similar results so we report all years together.

First, we calculated Cronbach's alphas for the Instructional Practice and Professional Responsibility domains for teachers and the Instructional Leadership and Professional Responsibility domains for administrators. Cronbach's alpha provides a measure of the internal consistency of the NEPF domains by estimating the average inter-item correlation among the domain standards. An internally consistent test is one in which test items that purport to measure the same thing report similar scores across the same respondent. Thinking of the NEPF domains and standards like items on a test, Cronbach's alpha tells us whether a given educator is scoring similarly on the different NEPF standards within a domain. If the standards within a given NEPF domain (say Instructional Practice) are highly correlated with one another (as they should be, if they are truly capturing information on a given teacher's Instructional Practice), then we

would expect a high Cronbach's alpha score (above 0.70 on a scale between 0 and 1), and we could conclude that the Instructional Practice domain of the NEPF is internally consistent and reliable.

**Table 11. Research Questions with Corresponding Data Sources**

| Research Question | Source |
|---|---|
| Validity Strand | |
| 1. Are the following components of the NEPF appropriate to positively impact teacher and administrator practice and outcomes?:<br>a. The content of the NEPF standards.<br>b. The internal consistency of the indicators of the NEPF domains.<br>c. The NEPF standard score ranges.<br>d. The weighting of each NEPF domain. | • Teacher NEPF Scores<br>• Administrator NEPF Scores<br>• School-Aggregate Teacher NEPF Scores<br>• District-Aggregate Administrator NEPF Scores<br>• Review of State Evaluation Systems<br>• Teacher Surveys<br>• Administrator Surveys |
| 2. What is the correlation between NEPF standards for teachers and administrators? | • Teacher NEPF Scores<br>• Administrator NEPF Scores<br>• School-Aggregate Teacher NEPF Scores<br>• District-Aggregate Administrator NEPF Scores |
| 3. What is the year-to-year variation in teacher and administrator NEPF scores? | • School-Aggregate Teacher NEPF Scores<br>• District-Aggregate Administrator NEPF Scores |
| Impact Strand | |
| 4. Does school growth in the percentage of teachers/administrators scoring highly on NEPF standards relate to growth in school student achievement? | • School-Aggregate Teacher NEPF Scores<br>• District-Aggregate Administrator NEPF Scores<br>• Nevada Report Card Data<br>• NCES Common Core of Data |
| 5. Do teachers/administrators believe that growth on the NEPF is related to growth in instructional practices? | • Teacher Surveys<br>• Administrator Surveys |

While Cronbach's alpha can give us some indication of the internal consistency of the NEPF domains, the test does not speak to the dimensionality of the NEPF, or the extent to which the Practice and Responsibility domains, with their respective standards and indicators are measuring different elements of teacher and administrator effectiveness (or whether they are so highly correlated that they are essentially measuring the same effectiveness qualities). To measure the dimensionality of the NEPF we employed Exploratory Factor Analysis (EFA). The objective of EFA is to uncover the underlying relationships between different indicators of a variable. In our case, we treat the NEPF Practice and Responsibility domains as variables measured by teacher and administrator scores on the underlying standards. Mechanically, EFA partitions the interrelationship (or variance) of the NEPF standards into two components—that which is *shared* among a set of NEPF standards and that which is *unique* among a set of NEPF standards. For the NEPF to have strong dimensionality, we should uncover that the Practice standards have a large common variance and the Responsibilities standards share a large common variance. To take into account that scores on the individual standards are generated from the same evaluator (and other aspects of the evaluation process that might lead to a correlation between standards within domains), we use Promax rotation of the factor solution derived from the EFA.

c. The NEPF domain score ranges

To determine whether the NEPF domain score ranges are appropriate to positively impact practice, we utilize the School-Aggregate Teacher NEPF Scores file and the District-Aggregate Administrator NEPF scores file to explore the distribution of educator performance on each NEPF domain and standard across all years. If the NEPF is doing a good job at distinguishing educator performance, then each NEPF domain and its respective standards would show substantial variation and scoring that follows an approximate normal distribution as shown in Figure 5, much like we expect from our summative tests of student performance. In other words, we should find that on any given standard and domain, very few educators are performing at the top and bottom of the distribution, with a mean and median score right in the middle of the distribution.

Alternatively, if the NEPF is not appropriately calibrated to current educator performance, and is instead too challenging, then we might expect to see a distribution that is skewed right (the tail of the distribution is to the right), in which case we observe most of our educators performing poorly, with very few reaching the upper echelons of the performance distribution (as shown in Figure 6).

Finally, we could imagine a scenario where the NEPF is not appropriately calibrated to current educator performance, but in the alternative direction, where scoring highly is too easy. In this case, we might expect to see a distribution that is skewed left (the tail of the distribution is to the left), in which case we observe most of our educators performing highly, with very few performing inadequately. This is commonly termed a "ceiling effect" where most of the participants on a test score the maximum score, thereby making it difficult to distinguish between performers at the upper end of the performance distribution (as shown in Figure 7).

**Figure 5. Graphical Depiction of a Normal Distribution**

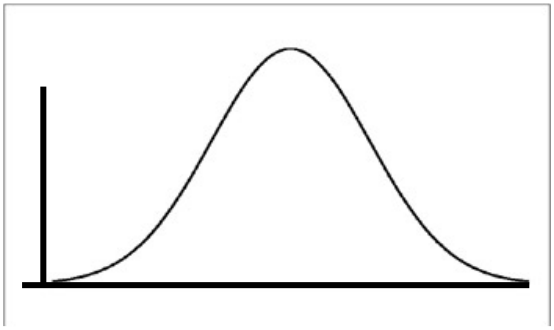


**Figure 6. Graphical Depiction of Skewed-Right Distribution**

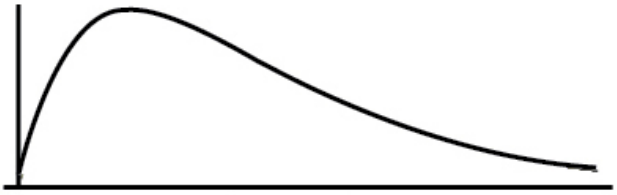
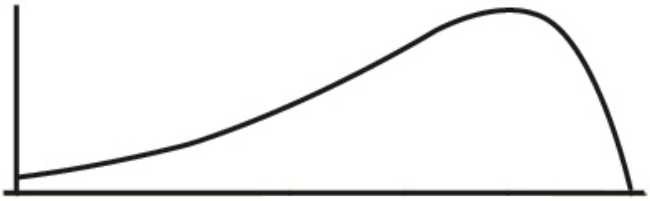


**Figure 7. Graphical Depiction of a Skewed-Left Distribution**



In addition to showing the distributions, we present the minimum and maximum scores, standard deviations, skew statistics (a measure of the asymmetry of the distribution where negative values over -1 or positive values over 1 typically indicate a strong skew), and kurtosis statistics (a measure of the height and sharpness of the central peak of a distribution relative to a normal distribution where positive values indicate a peak that is higher than a typical normal curve and negative values indicate a peak that is lower than a normal curve).

d. The weighting of each NEPF domain

To judge whether the weight of each NEPF domain is appropriate to positively impact practice we ran through a few simulations of weighting and then assessed the bivariate correlations between the final NEPF score under the different weighting simulations with student reading and math proficiency. If we observed a strong correlation between the final NEPF score and student proficiency, then we are more confident in the predictive validity of the NEPF. In other words, that the weighting of NEPF is set in a manner that produces final scores that relate to student outcomes. In addition, we explored whether the distribution of educator effectiveness produced under the different weight scenarios follows a normal distribution in which educator effectiveness can be appropriately distinguished using the current cutoff scores for determining the final effectiveness ratings.

Because of the availability of multiple years of data, we first performed analysis on the School-Aggregate Teacher NEPF Scores file and the District-Aggregate Administrator NEPF scores files. We explored the impact of the weighting of the NEPF domains by recalculating the final school-average and district-average NEPF scores in each year using the 2019-20, 2018-19, and 2017-18 weights for Practice, Responsibility, and Student Outcomes. That is, in our first run through the data, we gave an educator's score on the Practice domain a weight of 65%, an educator's score on the Responsibilities domain a weight of 20%, and an educator's score on the Student Outcomes domain a weight of 15% in each year, consistent with the 2019-20 weights. We subsequently performed two other analyses of the data, one where we reweighted based on the 2018-19 weights (45% Practice, 15% Responsibilities, and 40% Student Outcomes) and again on the 2016-17 weights (60% Practice, 20% Responsibilities, and 20% Student Outcomes). This exercise allowed us to examine the effects of the various weighting systems used over the years on educators' final ratings and on student proficiency

We run similar analyses on the Teacher NEPF Scores File and the Administrator NEPF Scores file, which contain individual level scores on each domain and overall but just for the 2018-19 school year. We compare the results from the teacher-level and administrator-level analysis to the results we obtain at the school- and district-levels.

2. What is the correlation between NEPF domains for teachers and administrators?

To assess the correlation between NEPF domains for teachers and administrators, we leverage the School-Aggregate Teacher NEPF Scores file and the District-Aggregate Administrator NEPF scores files. To make the teacher and administrator scores comparable, we have to aggregate the scores to the same level, in this case the district level. Consequently, we assess the correlation between district-average NEPF domain scores for teachers and administrators using a bivariate correlation matrix.

3. What is the year-to-year variation in teacher and administrator NEPF scores?

To assess the year-to-year variation, we take the school-average score for teachers in the present year and subtract it from the school-average score for teachers in the year

prior (e.g. the school average in 2017-18 is subtracted from the school average in 2016-17). Doing so creates a year-to-year change measure that captures the relative difference in this year's score from last year's score. A positive result on this measure means that a school experienced growth in their average NEPF score. A negative result on this measure means that a school regressed in their average NEPF score between years. We then plot the distribution of change and run summary statistics on the year-to-year change measure which provides detail on the average school-level change in NEPF scores. We do the same for administrators at the district level. This exercise helped us determine how much growth educators experience on the NEPF over time. Because we do not have individual-level NEPF scores over time, we have to rely on school-average growth and district-average growth from the School-Aggregate Teacher NEPF Scores file and the District-Aggregate Administrator NEPF scores files.

## ASSESSING THE IMPACT OF THE NEPF

The **impact strand** aims to answer the following research questions:

4. Does school growth in the percentage of teachers/administrators scoring highly on NEPF standards relate to growth in school student achievement?

A valid impact analysis must be able to disentangle the impact of the NEPF on student growth from the variety of other concurrent programs, policies, out-of-school factors, and inside-of-school factors that also impact the same outcome. In other words, a valid approach must account for school, child, teacher, and district characteristics/demographics that are also correlated with student growth. The easiest way to do this is to randomly assign some schools to use the NEPF for evaluation and other schools to not use the NEPF. Random assignment on a large enough sample ensures that both observed and unobserved characteristics of students, teachers, and schools are balanced between the treatment and control groups such that any difference in the outcomes between the two groups can only be attributed to the provision of the NEPF as the evaluation framework (i.e., the treatment).

Given that the NEPF was not randomly assigned to some schools and not others (as is often the case in other policy evaluations), we must employ alternative methodologies to assess the impact of the NEPF. The methodology described here leverages the longitudinal nature of the NEPF data (i.e. we observe school-level and district-level NEPF scores over multiple years) in a series of fixed effect models in an effort to control for unobserved and observed differences in Nevada schools and districts. This methodology meets ESSA evidence Tier 2- Quasi-Experimental.

A fixed effect approach allows each school or district to serve as its own control group. In the absence of an easily defined control group that did not receive "treatment" from the NEPF, perhaps the most valid comparison to be made is within district or within school. A within school comparison involves comparing a given school's student performance today (relative to prior years) in relation to its NEPF score today (relative to prior years). Similarly, a within district comparison involves comparing a given district's

student performance today (relative to prior years) in relation to its NEPF score today (relative to prior years). The fixed effect approach easily accounts for fixed (i.e., time invariant) differences between schools and identifies changes over time. The model is formally estimated as follows:

$$y_{st} = \beta_0 + \beta_1 NEPF_{st} + X_{st}\beta_2 + \tau_t + S_s + e_{st} \qquad (1)$$

$y_{st}$ is a measure of student achievement for school *s* in year *t*, as measured on the annual Smarter Balanced Assessment (SBAC). In particular, we utilize a commonly used uncoarsening procedure to translate frequency counts of students scoring in each performance category on the SBAC (Emerging, Approaching, Meets, Exceeds) into standardized scores (Reardon, Kalogrides, & Ho, 2017; Reardon, Shear, Castellano, & Ho, 2016; Shear & Reardon, 2019). $NEPF_{st}$ represents the percentage of teachers scoring Effective or Highly Effective in the respective school. $\beta_1$ is the parameter of interest and represents the marginal effect of a percentage point increase in the average school NEPF performance on school achievement growth. In alternate models, we also use school average NEPF scores, which are continuous measures from 1 to 4, rather than levels (Effective and Highly effective versus Developing and Ineffective) to assess how student achievement growth is related to NEPF performance along the entire spectrum of NEPF scores rather than only at the threshold of Developing and Effective.

We control for various time-varying school characteristics using $X_{st}$, a vector that includes the percentage of male students, students of color, students eligible for free or reduced-price meals (a proxy for students' socioeconomic status), English language learner students, and students with an individualized education plan (IEP). $\tau_t$ represents a year fixed effect to account for changes in school growth that are common to all schools in Nevada. $S_s$ represents a school fixed effect and accounts for variation in school achievement that is constant over time. To account for multiple observations per school (from different school-by-years), we cluster our standard errors at the school level.

Given the smaller number of administrators, to understand how administrators' growth on NEPF standards relate to growth in student achievement, we conduct an analysis similar to equation 1 but at the district level:

$$y_{dt} = \beta_0 + \beta_1 NEPF_{dt} + X_{dt}\beta_2 + \tau_t + S_d + e_{dt} \qquad (2)$$

The student achievement growth measure ($y_{dt}$) is aggregated to the district level. $NEPF_{dt}$ represents the percentage of administrators scoring Effective or Highly Effective in district *d* in year *t*. The control variables are also aggregated to the district level. We still include year fixed effects and use district fixed effects to account for variation in district achievement growth that is constant over time. We conduct a similar alternate model as with the school-level analysis for teachers using continuous NEPF scores. To account for multiple observations per district (from different district-by-years), we cluster our standard errors at the district level.

5. Do teachers/administrators believe that growth on the NEPF is related to growth in instructional practices?

We first leveraged existing perception data from the annual NDE surveys on the NEPF. We explored the survey for overlapping items between the 2017-18 and 2018-19 administrations (as the items have changed over time). We identified two teacher indicators that capture perceptions on the impact and validity of the NEPF. The items were worded slightly differently between the 2017-18 and 2018-19 surveys. The items are as follows: *My evaluation was fair* and *My evaluation helped me identify my areas of growth as an educator*. In 2017-2018, these items were worded as *The NEPF scores you received are fair* and *Using NEPF Standards and Protocols have helped you to identify your areas of growth as an educator*.

We also identified two overlapping items related to impact and validity from the administrator surveys. The administrator items in 2018-19 were: *The implementation of NEPF is positively impacting student learning at my school(s)* and *My evaluation was fair*. These same items in 2017-18 were worded as *At your school, implementation of NEPF Standards and Protocols is positively impacting student learning* and *The NEPF scores you received are fair*.

We supplemented this survey data with our own survey. As mentioned above, this survey focused primarily on capturing educators' perceptions on the quality and amount of feedback they receive from their supervisor during their NEPF evaluation cycle on each of the Practice and Responsibility standards. We also asked a series of questions about the SLG process. We provide summary statistics for the items from the NDE survey and our own survey in the results section.

**Limitations**

As mentioned above, our findings are limited by a few factors. We only have individual-level data for a single year (2018-19), requiring us to mainly utilize school- or district-aggregate data over time. These data sources suffer from aggregation bias, or the idea that data that is aggregated to higher-level units can mask important information and patterns in the individual units. Individual data might lead to different conclusions. Second, we only have data on NEPF standard and domain scores. We were unable to assess any patterns in NEPF indicators, and thus are unable to determine which NEPF indicators are performing better or worse. Finally, this analysis was performed on a quick five week timeline to the first presentation, thereby narrowing the scope of the report to focus on the most pressing research questions. A longer time horizon for reporting would allow for an even more detailed exploration of the NEPF and its impact and validity, including detailed interviews with administrators and teachers on NEPF implementation.

# VALIDITY OF THE NEPF

## RESEARCH QUESTIONS 1-3

### THE CONTENT OF THE NEPF DOMAINS HAVE HIGH FACE VALIDITY

We explored face validity on two fronts: First, we sought to understand the extent to which the NEPF, with its domains and standards, look similar or different from other evaluation systems used around the country. We created a similarity index based on 8 criteria (mentioned in the Data section of this report). These criteria are shown in the first row of Table 12. The *Authority* column represents whether the state's teacher evaluation system is locally-developed but state aligned (L/S), is state developed (S), or is locally-developed (L). The *Student Learning Goal* column documents whether student growth is measured by progress towards an SLG. The *Annual Prob. Eval* column captures whether the state's system requires that probationary teachers be evaluated at least annually. The *> 1 Prob. Observ.* column measures whether the evaluation requires greater than one observation cycle for probationary teachers. We also tracked whether the state's evaluation system has a Practice and Responsibilities domain. The *Feedback* column demarcates whether the state's evaluation system required feedback to be provided to the educator. The next column, *> 3 Rating Catg.*, tracks whether the state's evaluation requires greater than three final rating categories. A perfect similarity score of 8 indicates that the state's system closely resembles the NEPF on each of the 8 categories. Each state's score is provided in the Score column. We collected the same information for administrators as shown in Table 13.

In terms of teacher evaluation systems (Table 12), we find very close similarity between the NEPF and Minnesota's, West Virginia's, and Rhode Island's teacher evaluation systems. Eight states require student growth to be measured utilizing an SLG and most systems require an annual evaluation of probationary teachers (45) with more than one observation cycle (34). We also find that the majority of states have similar domains as the NEPF—37 states have a Practice domain and 26 states have a Responsibilities domain. We expected more state evaluation systems to, like the NEPF, require feedback to be provided to educators during or after the evaluation cycle. This is the case in only 20 states. Thirty-eight states, like the NEPF, require greater than three final rating categories. Altogether, this evidence suggests that the teacher NEPF is not an outlier on our established criteria when compared to evaluation systems from other states and actually closely resembles evaluation systems in a handful of states.

## Table 12. Comparison of State Teacher Evaluation Systems to the NEPF

| State | Authority | Student Learning Goal | Annual Prob. Eval | > 1 Prob. Observ. | Practice Domain | Resp. Domain | Feedback | > 3 Rating Catg. | Score |
|---|---|---|---|---|---|---|---|---|---|
| **NV** | **S** | X | X | X | X | X | X | X | 8 |
| **MN** | **L/S** | X | X | X | X | X | X | X | 7 |
| **WV** | **S** | | X | X | X | X | X | X | 7 |
| **RI** | **L/S** | **X** | X | X | X | X | X | X | 7 |
| **IN** | **L/S** | | X | X | X | X | X | X | 6 |
| **DE** | **S** | | X | X | X | X | | X | 6 |
| **PA** | **S** | | X | X | X | X | | X | 6 |
| **WA** | **S** | | X | X | X | | X | X | 6 |
| **KS** | **L/S** | | X | X | X | X | X | X | 6 |
| **NE** | **L/S** | | X | X | X | X | X | X | 6 |
| **NM** | **L/S** | | X | X | X | X | X | X | 6 |
| **TX** | **L/S** | X | X | X | X | X | | X | 6 |
| **NC** | **S** | | X | X | X | | X | X | 6 |
| **WI** | **S** | X | X | | X | X | | X | 6 |
| **GA** | **S** | X | X | | X | X | | X | 6 |
| **SC** | **S** | X | | X | X | | X | X | 6 |
| **MD** | **L/S** | | X | X | X | X | X | | 5 |
| **LA** | **S** | | X | X | X | | | X | 5 |
| **CT** | **L/S** | | X | X | X | X | | X | 5 |
| **KY** | **L/S** | | X | X | X | X | | X | 5 |
| **ME** | **L/S** | | X | X | X | X | | X | 5 |
| **MA** | **L/S** | | X | | X | X | X | X | 5 |
| **NH** | **L** | | X | X | X | X | | X | 5 |
| **AL** | **S** | | X | X | X | | | X | 5 |
| **HI** | **S** | | X | X | | | X | X | 5 |
| **MS** | **S** | | X | | X | X | | X | 5 |
| **OK** | **S** | | X | | X | | X | X | 5 |
| **ID** | **L/S** | | X | X | | | X | X | 4 |
| **MT** | **L/S** | | X | | X | X | | X | 4 |
| **OH** | **L/S** | | X | | X | X | | X | 4 |
| **TN** | **L/S** | | X | X | X | | X | | 4 |
| **UT** | **L/S** | | X | X | X | | | X | 4 |
| **IA** | **L** | | X | X | X | X | | | 4 |
| **OR** | **L** | X | X | X | | | | X | 4 |
| **AZ** | **L/S** | | X | X | | | | X | 3 |
| **MI** | **L/S** | | X | | | | X | X | 3 |
| **MO** | **L/S** | | | | X | X | | X | 3 |
| **NJ** | **L/S** | | X | X | | | X | | 3 |
| **NY** | **L/S** | | X | | X | X | | | 3 |
| **ND** | **L/S** | | X | X | | | X | | 3 |
| **SD** | **L/S** | | | | X | X | | X | 3 |
| **CO** | **L** | | X | X | | | | X | 3 |
| **VA** | **L/S** | | X | | X | | | X | 3 |
| **WY** | **L/S** | | X | | X | | | X | 3 |
| **IL** | **L/S** | | X | X | | | | | 2 |
| **AK** | **L/S** | | X | X | | | | | 2 |
| **FL** | **L/S** | | X | X | | | | | 2 |
| **VT** | **L** | | | | X | | | X | 2 |
| **CA** | **L/S** | | X | | | | | | 1 |
| **AR** | **L/S** | | | | | | | | 0 |

In terms of administrator evaluation systems (Table 13), we find relatively close similarity to the evaluation systems in Kansas, Nebraska, Georgia, Hawaii, Louisiana, South Carolina, and Wisconsin. One of the major differences in the NEPF for administrators relative to systems in other states is that the NEPF requires the SLG as a measure of student growth whereas only two other states require SLGs for administrators. Most systems require an annual evaluation of probationary administrators (38) though only 16 require more than one observation cycle of these educators. We also find that the majority of states have similar domains as the NEPF for administrators—38 states have a Practice domain and 21 states have a Responsibilities domain. Twenty-one states require feedback to be provided to educators during or after the evaluation cycle and 42 have greater than three final rating categories. Again, taken together, this evidence suggests that the administrator NEPF is not an outlier in its design when compared to evaluation systems from other states, though its provision of the SLG and the requirement of more than one observation cycle is somewhat unique.

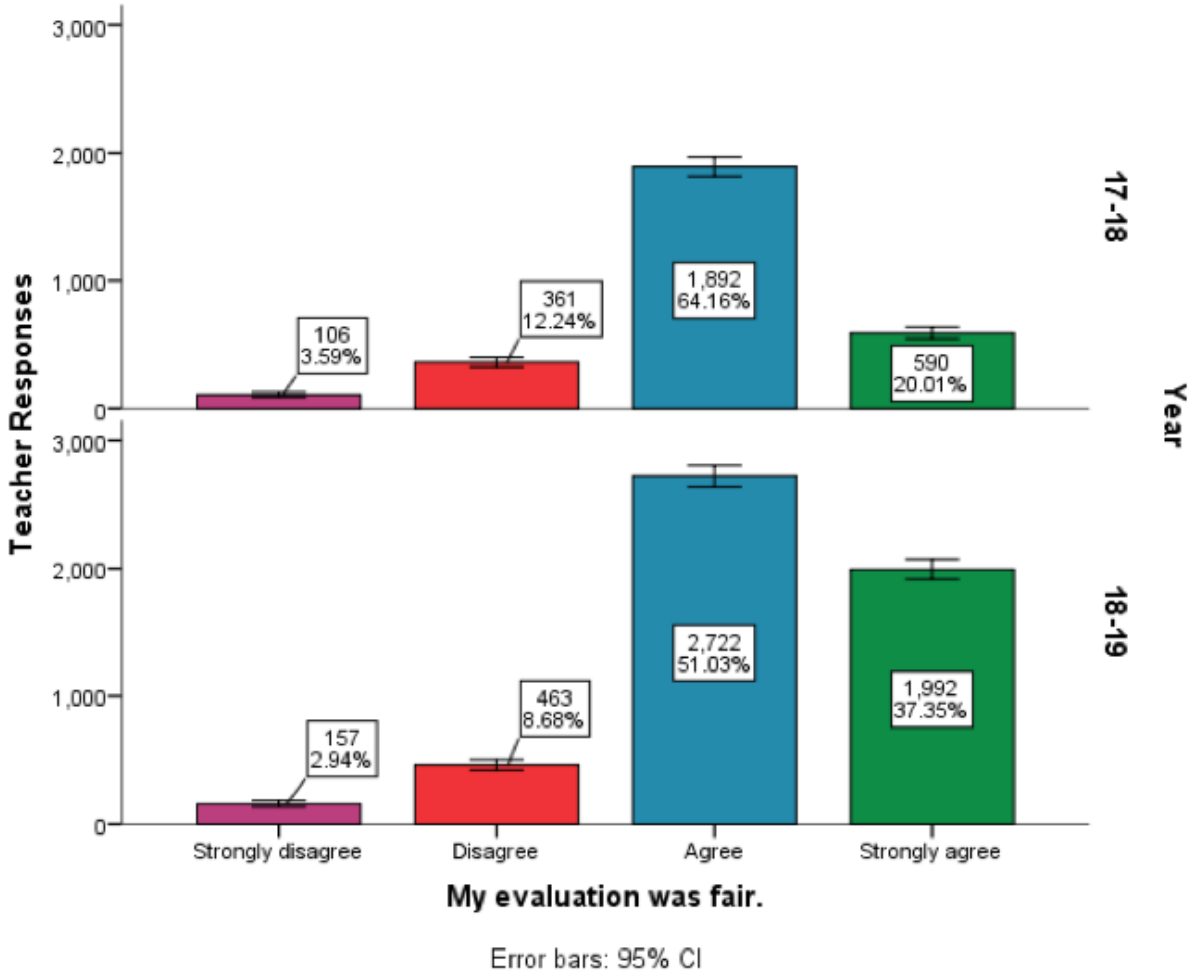
We also made an effort to gauge whether educators believed the NEPF was a valid measure of their performance (i.e. face validity). We first assessed whether educators believed their evaluation was fair using data from the NDE 2017-18 and 2018-19 annual surveys (Figures 8 and 9). Among the 2,949 teacher responses on the question in 2017-18, 84% of teachers agreed or strongly agreed that their evaluation was fair and only 16% disagreed or strongly disagreed. In 2018-19, among the 5,264 teacher responses on the question, 88% of teachers agreed or strongly agreed that their evaluation was fair and 12% disagreed or strongly disagreed.

Administrators also consider their NEPF evaluation fair. For example, among the 326 administrators that responded on the question in 2017-18, 88% strongly agreed and agreed their NEPF evaluation was fair. Only 12% disagreed or strongly disagreed that their NEPF evaluation was fair. The 2018-19 survey yielded even higher percentages of administrators who believe their evaluation was fair (93.3%).

## Table 13. Comparison of State Administrator Evaluation Systems to the NEPF

| State | Authority | Student Learning Goal | Annual Prob. Eval | > 1 Prob. Observ. | Practice Domain | Resp. Domain | Feedback | > 3 Rating Catg. | Score |
|---|---|---|---|---|---|---|---|---|---|
| NV | S | X | X | X | X | X | X | X | 8 |
| KS | L/S | | X | X | X | X | X | X | 6 |
| NE | L/S | | X | X | X | X | X | X | 6 |
| GA | S | | X | X | X | X | | X | 6 |
| HI | S | | X | | X | X | X | X | 6 |
| LA | S | | X | X | X | | X | X | 6 |
| SC | S | | X | | X | X | X | X | 6 |
| WI | S | | X | | X | X | X | X | 6 |
| MA | L/S | | X | | X | X | X | X | 5 |
| MN | L/S | | X | | X | X | X | X | 5 |
| NJ | L/S | | X | X | X | | X | X | 5 |
| OR | L | X | X | | X | X | | X | 5 |
| AL | S | | X | X | X | | | X | 5 |
| MS | S | | X | | X | X | | X | 5 |
| NC | S | | X | | X | X | | X | 5 |
| WV | S | X | X | | X | | | X | 5 |
| UT | L/S | | X | X | X | | | X | 4 |
| CT | L/S | | X | X | X | | | X | 4 |
| IN | L/S | | X | X | | | X | X | 4 |
| MD | L/S | | X | | X | X | | X | 4 |
| MO | L/S | | | | X | X | X | X | 4 |
| MT | L/S | | X | | X | X | | X | 4 |
| NM | L/S | | X | | X | | X | X | 4 |
| TX | L/S | | X | | X | | X | X | 4 |
| RI | L/S | | X | X | X | | | X | 4 |
| IA | L | | X | | X | X | X | | 4 |
| DE | S | | X | | | | X | X | 4 |
| PA | S | | X | | | | X | X | 4 |
| WA | S | | X | | X | | | X | 4 |
| TN | L/S | | X | X | X | | | | 3 |
| AZ | L/S | | X | | X | | | X | 3 |
| AR | L/S | | | | X | X | | X | 3 |
| ID | L/S | | X | | X | | | X | 3 |
| IL | L/S | | X | X | | | X | | 3 |
| ME | L/S | | | | X | X | | X | 3 |
| MI | L/S | | X | | | | X | X | 3 |
| NY | L/S | | X | X | | | X | | 3 |
| SD | L/S | | | | X | X | | X | 3 |
| CO | L | | X | | X | | | X | 3 |
| NH | L | | | | X | X | | X | 3 |
| OK | S | | X | | | | | X | 3 |
| CA | L/S | | | | X | | | X | 2 |
| FL | L/S | | X | | | | X | | 2 |
| KY | L/S | | | | X | X | | | 2 |
| OH | L/S | | | X | | | | X | 2 |
| VA | L/S | | | | X | | | X | 2 |
| WY | L/S | | | | X | | | X | 2 |
| AK | L/S | | X | | | | | | 1 |
| ND | L/S | | | X | | | | | 1 |
| VT | L | | | | | | | X | 1 |

**Figure 8. Teacher Responses to Survey Item: My Evaluation was Fair**




We further asked on our CREA survey whether teachers and administrators believed the final score they receive on the NEPF is a valid measure of their performance. The results are shown in Table 14. We find that similar percentages of teachers (60%) and administrators (58%) agree and strongly agree that the NEPF is a valid measure of their performance. In terms of disagreement, 21% of teachers and 17% of administrators disagree or strongly disagree that the NEPF is a valid measure of their performance.

**Figure 9. Administrator Response to Survey Item: My Evaluation was Fair**

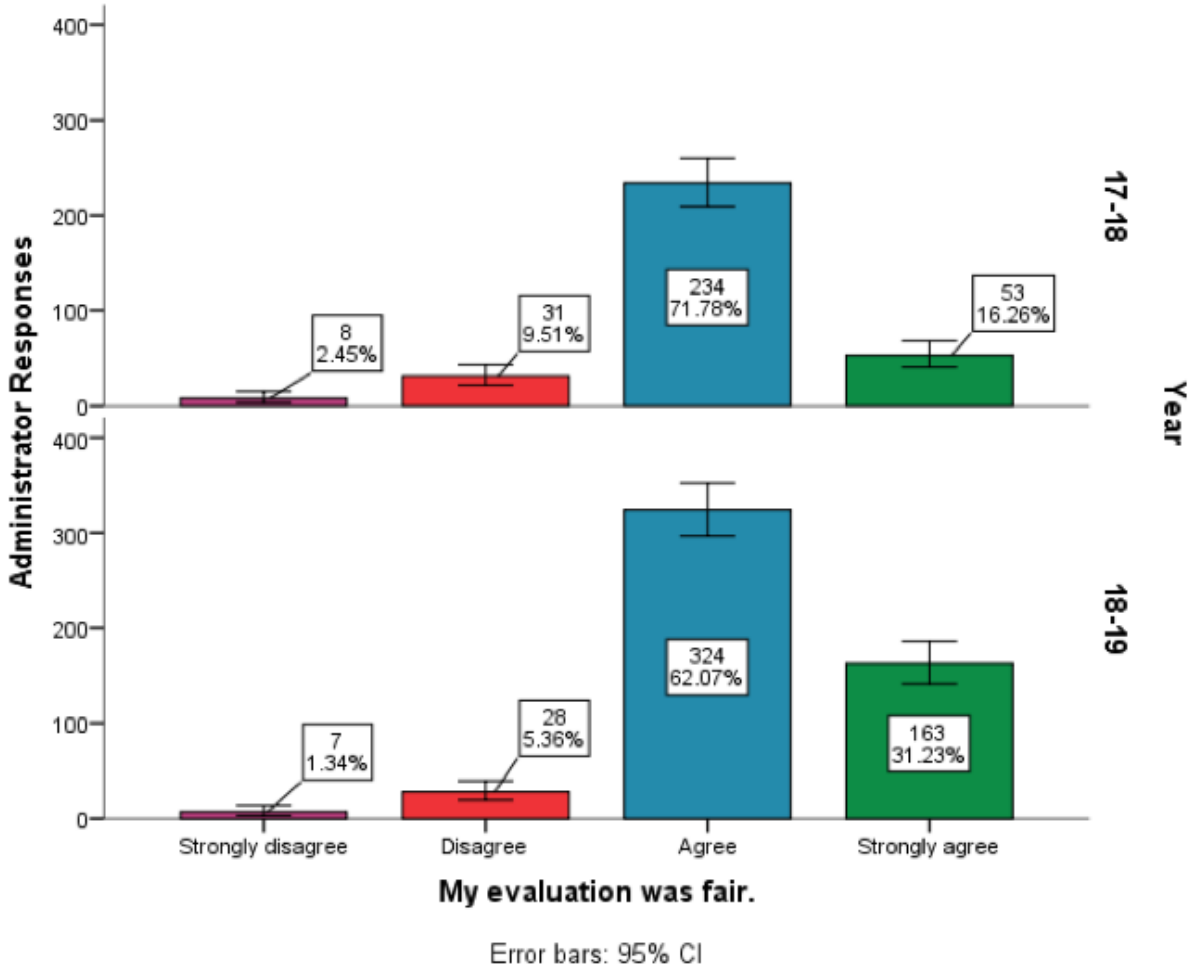



**Table 14. Educator Response to Survey Item: The Final Score From My Evaluation is a Valid Measure of My Performance**

| | Mean | SD | Strongly Disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| **Teachers** | 3.51 | 1.18 | 8.3% | 12.2% | 19.1% | 40.8% | 19.6% |
| **Administrators** | 3.49 | 1.09 | 7.2% | 9.4% | 25.8% | 42.1% | 15.4% |

## THE NEPF DOMAINS ARE INTERNALLY CONSISTENT (HIGH RELIABILITY)

As mentioned above, we calculate Cronbach's alpha as a measure of the internal consistency (reliability) of the full NEPF teacher framework (i.e. ratings on all standards regardless of domain) as well as the internal consistency of the NEPF Instructional Practice and the NEPF Professional Responsibilities domains for teachers. For this analysis we used the School-Aggregate Teacher NEPF Scores file and the District-Aggregate Administrator NEPF scores file. The results are shown below in Table 15.

**Table 15. Cronbach's Alpha for NEPF Standards**

| Teachers | | Administrators | |
|---|---|---|---|
| Domain | Cronbach's alpha | Domain | Cronbach's alpha |
| Overall (All 10 Standards) | 0.96 | Overall (All 8 Standards) | 0.93 |
| Instructional Practice | 0.95 | Instructional Leadership | 0.90 |
| Professional Responsibilities | 0.92 | Professional Responsibilities | 0.79 |

As a reminder, a Cronbach's alpha above 0.90 is considered excellent, a Cronbach's alpha above 0.80 is considered good, and a Cronbach's Alpha above 0.70 is considered acceptable. Practically, a large Cronbach's Alpha means that an educator scoring highly on one NEPF standard within the domain is also scoring highly on other NEPF standards within the same domain.

The Cronbach's alpha for the overall NEPF teacher framework utilizing all 10 standards was excellent (α = .96). Inter-item correlations ranged from 0.60 to 0.90 and all standard ratings contributed positively to the internal consistency of the framework. Looking specifically at the Instructional Practice domain, the alpha coefficient for the five Instructional Practice standards was also excellent (α = 0.95). Inter-item correlations ranged from 0.65 to 0.80 and again all standards contributed to th e internal consistency of the domain. In terms of the Professional Responsibilities domain, the alpha coefficient for the five Professional Responsibility subset of standards was also excellent (α = 0.90). Inter-item correlations ranged from 0.62 to 0.83 and all standards contributed to internal consistency.

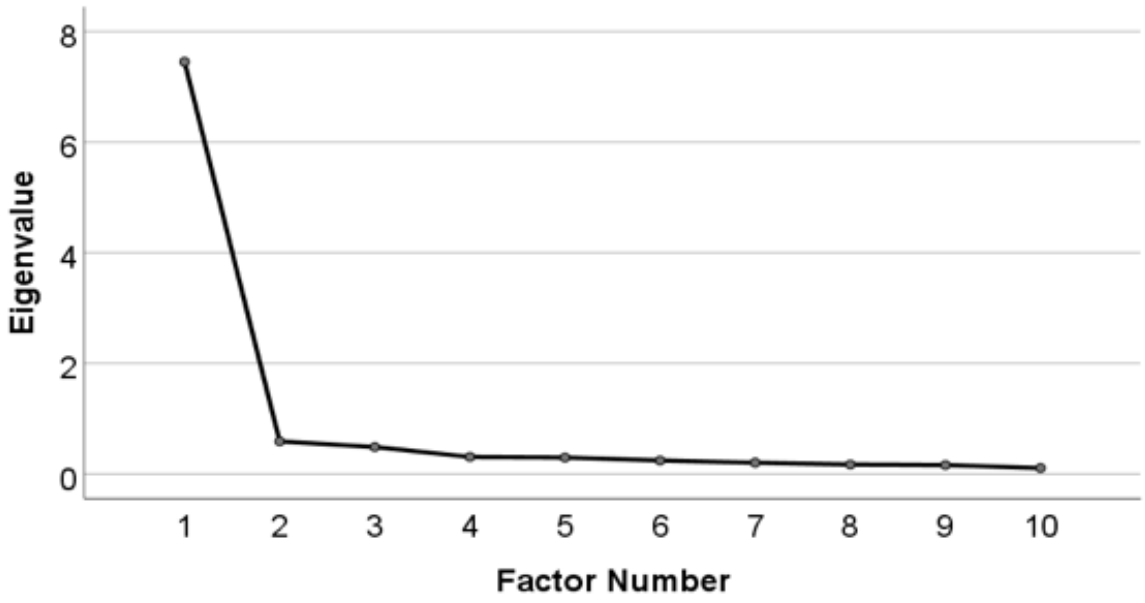
The Cronbach's alpha for the overall NEPF administrator framework utilizing all 8 standards was excellent (α = 0.93). Inter-item correlations ranged from 0.34 to 0.86 and all standard ratings, except Standard 2 of the Professional Responsibility Domain—Self Reflection and Professional Growth, contributed to the internal consistency of the framework. The alpha coefficient for the Instructional Leadership domain was also

excellent (α = .90). Inter-item correlations ranged from 0.52 to 0.86 and all standards contributed to internal consistency. The alpha coefficient for the Professional Responsibility domain was acceptable (α = 0.79). Inter-item correlations ranged from 0.37 to 0.65 and all standards contributed to internal consistency.

## THE NEPF HAS LOW DIMENSIONALITY AND LITTLE VARIATION IN SCORES (LOW CONSTRUCT VALIDITY)

As an indicator of construct validity, it is also useful to explore the dimensionality of the NEPF. As a reminder, dimensionality has to do with whether the NEPF domains and standards are measuring similar or different things regarding educator performance. By design, the NEPF hypothesizes a two factor structure—it groups a series of standards under Instructional Practice and a series of standards under Professional Responsibility. In short, these are considered two important dimensions of the NEPF, each measured by progress on the respective standards and indicators as determined by supervisor observation.

**Figure 10. Scree Plot of EFA Results on the 10 NEPF Teacher Standards**



Exploratory factory analysis was used to examine whether the hypothesized two factor structure consisting of the two NEPF teacher domains of Instructional Practice and Professional Responsibilities best fit the data. We performed the EFA on the School-Aggregate Teacher NEPF Scores file and the District-Aggregate Administrator NEPF scores file. Based on the school-aggregate teacher scores on the 10 NEPF teacher standards, the scree plot indicated that the two factor hypothesis was incorrect. A scree plot is a plot of the eigenvalues generated from EFA, ordered from largest to smallest. Eigenvalues capture the amount of total variation explained by the factor in the underlying data. High Eigenvalues (usually Eigenvalues over 1) entail that the factor does a good job explaining the data (or that the factor is capturing a unique dimension of the data). The scree plot (Figure 10) indicates that a single factor solution was the best fit to the data, producing a single factor with an eigenvalue greater than one accounting for approximately 75% of the variation in the data.

Table 16 shows the factor loadings (the relationship between the individual NEPF standard and the factor such that scores near 1 indicate a stronger relationship with the factor). The standards are ordered based on the magnitude of their relationship with the factor. Note because of the Promax rotation of the factors, the factor loadings are regression coefficients and not standardized correlations (and thus, can be over 1).

Table 16 makes clear that all of the NEPF teacher standards from both the Practice and Responsibilities domain load on to factor 1 with a correlation of at least 0.76. We found the highest loading for Instructional Practice Standard 3- Students Engage in Meaning-Making through Discourse and Other Strategies at 0.91. We found the lowest loading for Professional Responsibilities Standard 5: Student Perception at 0.76. The Instructional Practice standards comprised three of the top five factor loadings while the Professional Responsibilities standards comprised three of the five bottom factor loadings. However, these three Professional Responsibilities factor loadings were still correlated with factor 1 at between 0.76 and 0.84. Ultimately, these results lend support to the idea that the NEPF teacher performance framework is best conceived of as a unidimensional measure of teacher effectiveness—schools that score highly on the Practice domain also score highly on the Responsibilities domain.

Turning our attention to the administrator NEPF standards, we similarly ran EFA using the district-aggregate administrator scores on the 8 NEPF administrator standards. The scree plot in Figure 11 yields some support for the two factor hypothesis for administrators in that we could identify two factors with Eigenvalues over 1. Again Eigenvalues over 1 entail that the factor is capturing a unique element of the data. However, we also note the large size of the factor 1 Eigenvalue, which accounts for 68% of the variation in the data.

**Table 16. Factor Loadings on Factor 1 for the 10 NEPF Teacher Standards**

| | Factor 1 |
|---|---|
| **Instructional Practice Standard 3:** Students Engage in Meaning-Making through Discourse and Other Strategies | 0.91 |
| **Instructional Practice Standard 2:** Learning Tasks have High Cognitive Demand for Diverse Learners | 0.90 |
| **Professional Responsibilities Standard 3:** Professional Obligations | 0.87 |
| **Instructional Practice Standard 5:** Assessment is Integrated into Instruction | 0.85 |
| **Professional Responsibilities Standard 2:** Reflection on Professional Growth and Practice | 0.85 |
| **Instructional Practice Standard 1:** New Learning is Connected to Prior Learning and Experience | 0.85 |
| **Professional Responsibilities Standard 4:** Family Engagement | 0.84 |
| **Instructional Practice Standard 4:** Students Engage in Metacognitive Activity to Increase Understanding of and Responsibility for Their Own Learning | 0.82 |
| **Professional Responsibilities Standard 1:** Commitment to the School Community | 0.81 |
| **Professional Responsibilities Standard 5:** Student Perception | 0.76 |
| Extraction Method: Principal Axis Factoring.<br>Rotation Method: Promax with Kaiser Normalization. | |

When we explore the factor loadings (shown in Table 17), we find that the 8 NEPF standards do not load on factors 1 and 2 cleanly. In other words, we find that factor 1 is actually best made up of two Instructional Leadership standards (standards 1 and 2) and two Professional Responsibilities standards (standards 1 and 2). Similarly, factor 2 is also made up of two Instructional Leadership standards (standards 3 and 4) and two Professional Responsibilities standards (standards 3 and 4). Given the cross-loadings across the NEPF administrator domains (and the large factor 1 eigenvalue) it's difficult to justify a two factor solution. Instead, it appears that the data best support the idea that the NEPF administrator framework is unidimensional measure of building administrator effectiveness—districts that score highly on the Instructional Leadership domain also score highly on the Professional Responsibilities domain.

**Figure 11. Scree Plot of EFA Results on the 8 NEPF Administrator Standards**

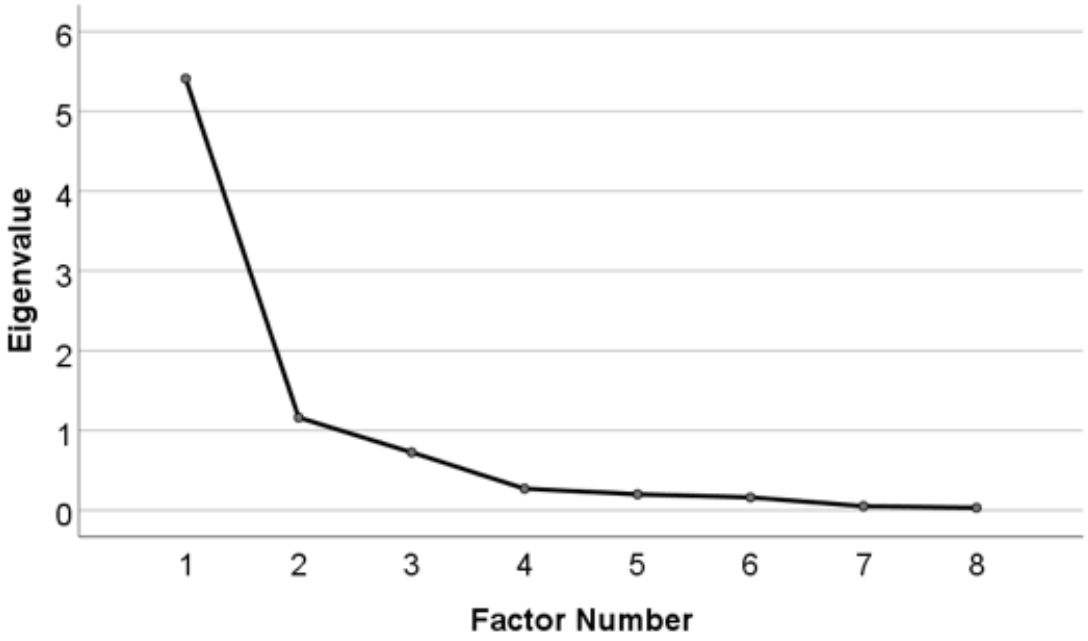



**Table 17. Factor Loadings on Factor 1 and 2 for the 8 NEPF Administrator Standards**

| | Factor 1 | Factor 2 |
|---|---|---|
| **Professional Responsibilities Standard 2:** Self-Reflection and Professional Growth | 0.91 | |
| **Instructional Leadership Standard 1:** Creating and Sustaining a Focus on Learning | 0.90 | |
| **Professional Responsibilities Standard 1:** Manages Human Capital | 0.83 | |
| **Instructional Leadership Standard 2:** Creating and Sustaining a Culture of Continuous Improvement | 0.66 | |
| **Instructional Leadership Standard 3:** Creating and Sustaining Productive Relationships | | 1.06 |
| **Instructional Leadership Standard 4:** Creating and Sustaining Structures | | 0.90 |
| **Professional Responsibilities Standard 4:** Professional Obligations | | 0.78 |
| **Professional Responsibilities Standard 3:** Family Engagement | | 0.66 |
| Extraction Method: Principal Axis Factoring.<br>Rotation Method: Promax with Kaiser Normalization. | | |

Another indication of construct validity is whether the NEPF, as a measure of teacher performance, can distinguish between high and low performers. One way to explore this is to look at the amount of variation in the scores. We first explored the amount of variation within each NEPF domain (i.e. Practice, Responsibility, and Student Outcomes) for teachers (aggregated at the school-level) and administrators (aggregated at the district-level). Below we show the distributions (Figures 12 and 13, all years together) on each NEPF standard and domain, first for teachers and then for administrators. Again, we are only able to do this at the school-level (for teachers) and at the district-level (for administrators) due to the availability of data. In addition to showing the distributions, we also show (Table 18) the minimum and maximum scores, standard deviations, skew statistics (again, a measure of the asymmetry of the distribution where negative values over -1 or positive values over 1 typically indicate a strong skew), and kurtosis statistics (again, a measure of the height and sharpness of the central peak of a distribution relative to a normal distribution where positive values indicate a peak that is higher than typical normal curve and negative values indicate a peak that is lower than a normal curve).

**Table 18. Summary Statistics for School-Level NEPF Teacher Domains and Standards (All Years)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **Instructional Practice Average** | 3.26 | 0.20 | 2.60 | 4.00 | 0.65 | 0.46 |
| **Instructional Practice Standard 1** | 3.26 | 0.24 | 1.00 | 4.00 | 0.82 | 14.03 |
| **Instructional Practice Standard 2** | 3.33 | 0.23 | 2.58 | 4.00 | 0.40 | 0.02 |
| **Instructional Practice Standard 3** | 3.31 | 0.22 | 2.68 | 4.00 | 0.64 | 0.71 |
| **Instructional Practice Standard 4** | 3.17 | 0.21 | 2.33 | 4.00 | 0.84 | 2.06 |
| **Instructional Practice Standard 5** | 3.23 | 0.22 | 2.29 | 4.00 | 0.88 | 1.31 |
| **Professional Responsibilities Average** | 3.29 | 0.22 | 2.63 | 4.00 | 0.77 | 0.20 |
| **Professional Responsibilities Standard 1** | 3.35 | 0.22 | 2.33 | 4.00 | 0.54 | 0.62 |
| **Professional Responsibilities Standard 2** | 3.27 | 0.22 | 2.64 | 4.00 | 0.87 | 0.95 |
| **Professional Responsibilities Standard 3** | 3.27 | 0.26 | 2.65 | 4.00 | 0.86 | -0.04 |
| **Professional Responsibilities Standard 4** | 3.25 | 0.23 | 2.00 | 4.00 | 0.99 | 1.59 |
| **Professional Responsibilities Standard 5** | 3.34 | 0.31 | 1.87 | 4.00 | 0.35 | -0.50 |
| **Student Outcomes (SLG score)** | 3.30 | 0.37 | 1.00 | 4.00 | 0.45 | 0.21 |

Figure 12 shows the distribution of school-level NEPF teacher Instructional Practice scores for all years. The figure makes clear that the school average NEPF Instructional Practice scores follow a roughly normal distribution that is skewed right, with the individual standard distributions looking very similar to the overall Instructional Practice domain distribution. The positive skew for the Instructional Practice domains and standards derives from the fact that more teachers score on the top end of the distribution (between 3.80 and 4) than at the bottom end of the distribution (below 2.60).

**Figure 12. Distribution of School-Level NEPF Teacher Instructional Practice Scores (All Years)**

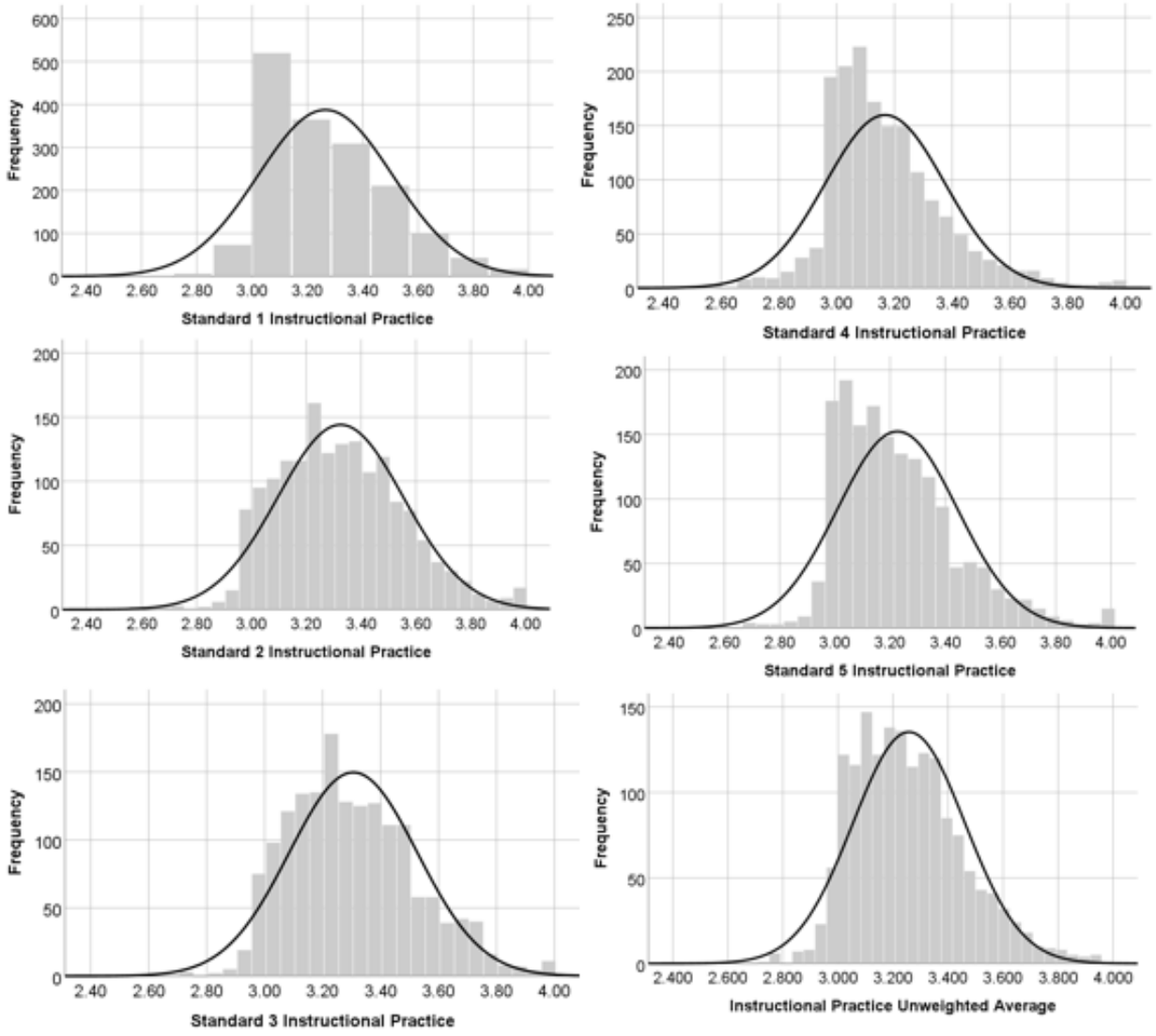



Table 18 reveals the mean, standard deviation, min, max, skewness, and kurtosis for the Instructional Practice distributions shown in Figure 12. The mean school average Instructional Practice score is 3.26, with a standard deviation of 0.20. Again, the skewness statistic (0.65) denotes a positive right skew (due to the presence of schools scoring highly, on average, but not poorly) and the kurtosis statistic (0.46) indicates a fairly normal looking distribution with somewhat short tails, where most of the observations cluster in the middle around the average of 3.26. The summary statistics for the individual Instructional Practice standards look very similar to those of the overall domain score.

**Figure 13. Distribution of School-Level NEPF Teacher Professional Responsibilities Scores (All Years)**

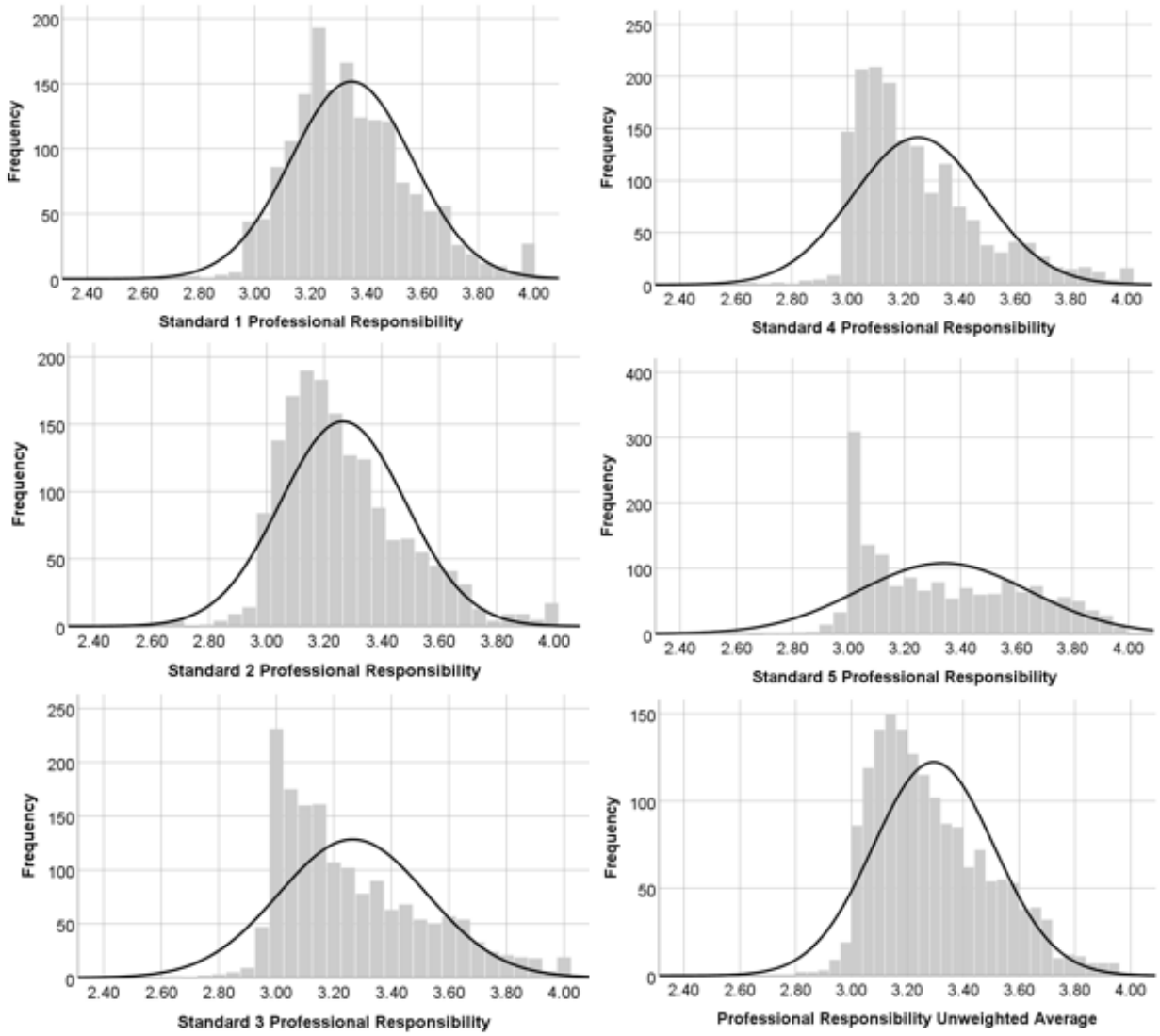



Figure 13 shows the distribution of school-level NEPF teacher Professional Responsibility scores for all years. Again, these distributions have a slightly positive skew (due to most school averages not falling on the bottom tail of the distribution). The Professional Responsibility standards distributions look very similar to the overall domain score distribution, with the exception of Professional Responsibilities Standard 5 (Student Perception), which has more variation around the mean than any of the other standards (with a standard deviation of 0.31. Note, that this standard also had the lowest factor loading in our assessment of internal consistency above. Table 16 confirms findings from the visual distributions. It shows that the school-average Professional Responsibilities score is 3.29 with a standard deviation of 0.22, a skew statistic of 0.77 and a kurtosis of 0.20. With the exception of Professional Standard 5, most of the standards are similar in their summary statistics.

Figures 14 and 15 and Table 19 provide this same information for the administrator Instructional Leadership and Professional Responsibilities domains, aggregated at the district level. Like the teacher NEPF results, we find that the district-level scores on the domains and standards follow a roughly normal distribution, with a mean just over 3, and positive skew (because very few districts perform at the bottom of the distribution) and kurtosis. There are two exceptions—both Instructional Leadership Standard 1 (Creating and Sustaining a Focus on Learning) and Instructional Leadership Standard 2 (Creating and Sustaining a Culture of Continuous Improvement) have small, negative skews (as indicated by the negative skew statistics in Table 19) due to the presence of some district averages at the bottom end of the distribution on these standards. Note from Table 19 that both of these standards have the small minimum scores (with the exception of Student Outcomes).

**Table 19. Summary Statistics for District-Level Administrator Domains and Standards (All Years)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **Instructional Leadership Average** | 3.26 | 0.21 | 2.86 | 3.78 | 0.50 | 0.54 |
| **Instructional Leadership Standard 1** | 3.35 | 0.23 | 2.57 | 3.78 | -0.80 | 2.22 |
| **Instructional Leadership Standard 2** | 3.30 | 0.29 | 2.66 | 4.00 | -0.46 | 1.10 |
| **Instructional Leadership Standard 3** | 3.20 | 0.30 | 2.62 | 4.00 | 0.57 | 0.80 |
| **Instructional Leadership Standard 4** | 3.28 | 0.23 | 2.79 | 3.78 | 0.12 | 0.02 |
| **Professional Responsibilities Average** | 3.24 | 0.22 | 2.91 | 3.83 | 0.75 | 0.50 |
| **Professional Responsibilities Standard 1** | 3.31 | 0.22 | 2.93 | 3.84 | 0.28 | 0.13 |
| **Professional Responsibilities Standard 2** | 3.51 | 0.20 | 3.00 | 4.00 | 0.12 | 0.36 |
| **Professional Responsibilities Standard 3** | 3.14 | 0.30 | 2.63 | 4.00 | 0.72 | 1.20 |
| **Professional Responsibilities Standard 4** | 3.30 | 0.19 | 2.95 | 3.83 | 0.79 | 1.04 |
| **Student Outcomes (SLG score)** | 2.98 | 0.69 | 1.24 | 4.00 | -0.86 | 1.08 |

**Figure 14. Distribution of School-Level NEPF Administrator Instructional Leadership Scores (All Years)**

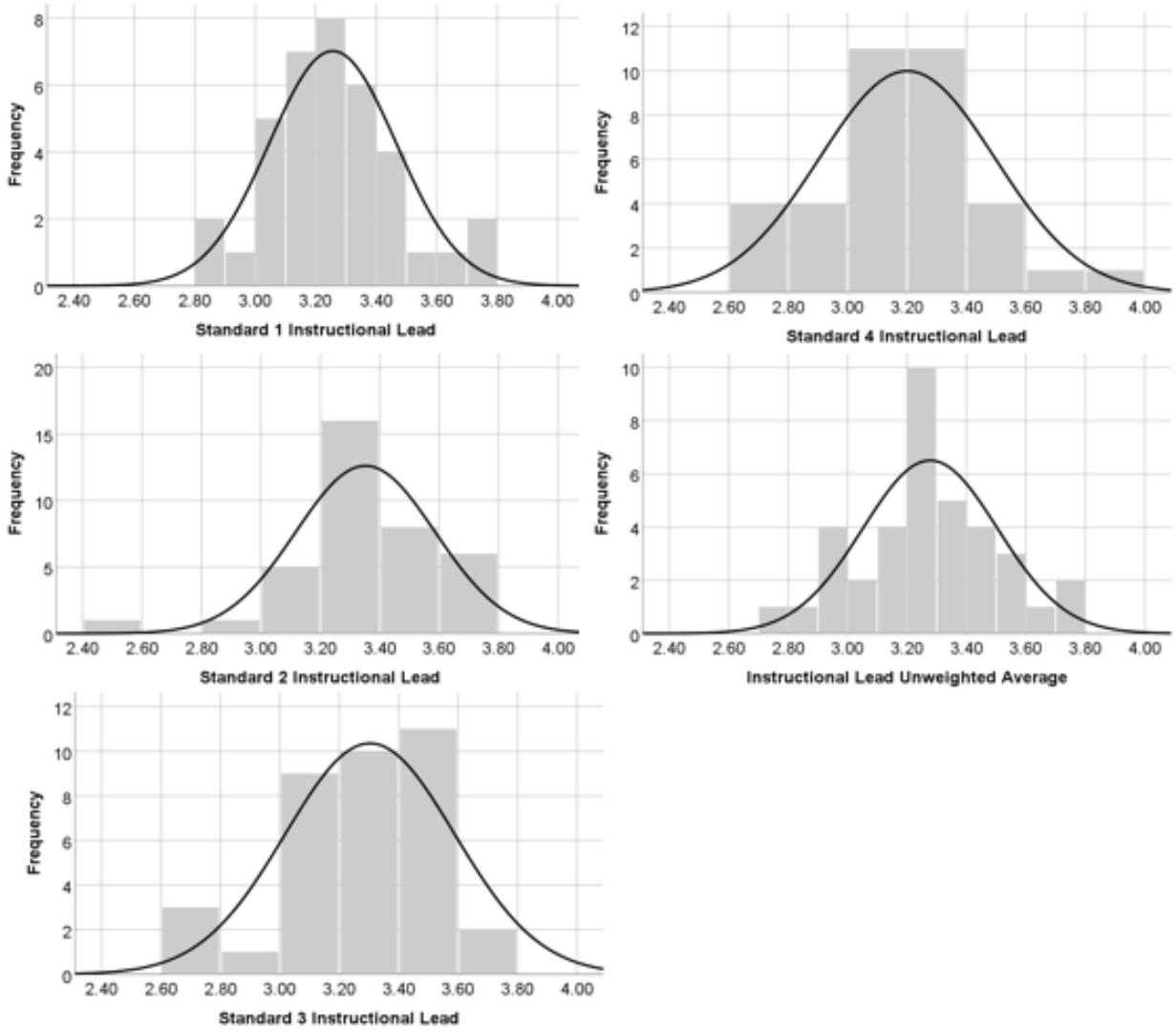
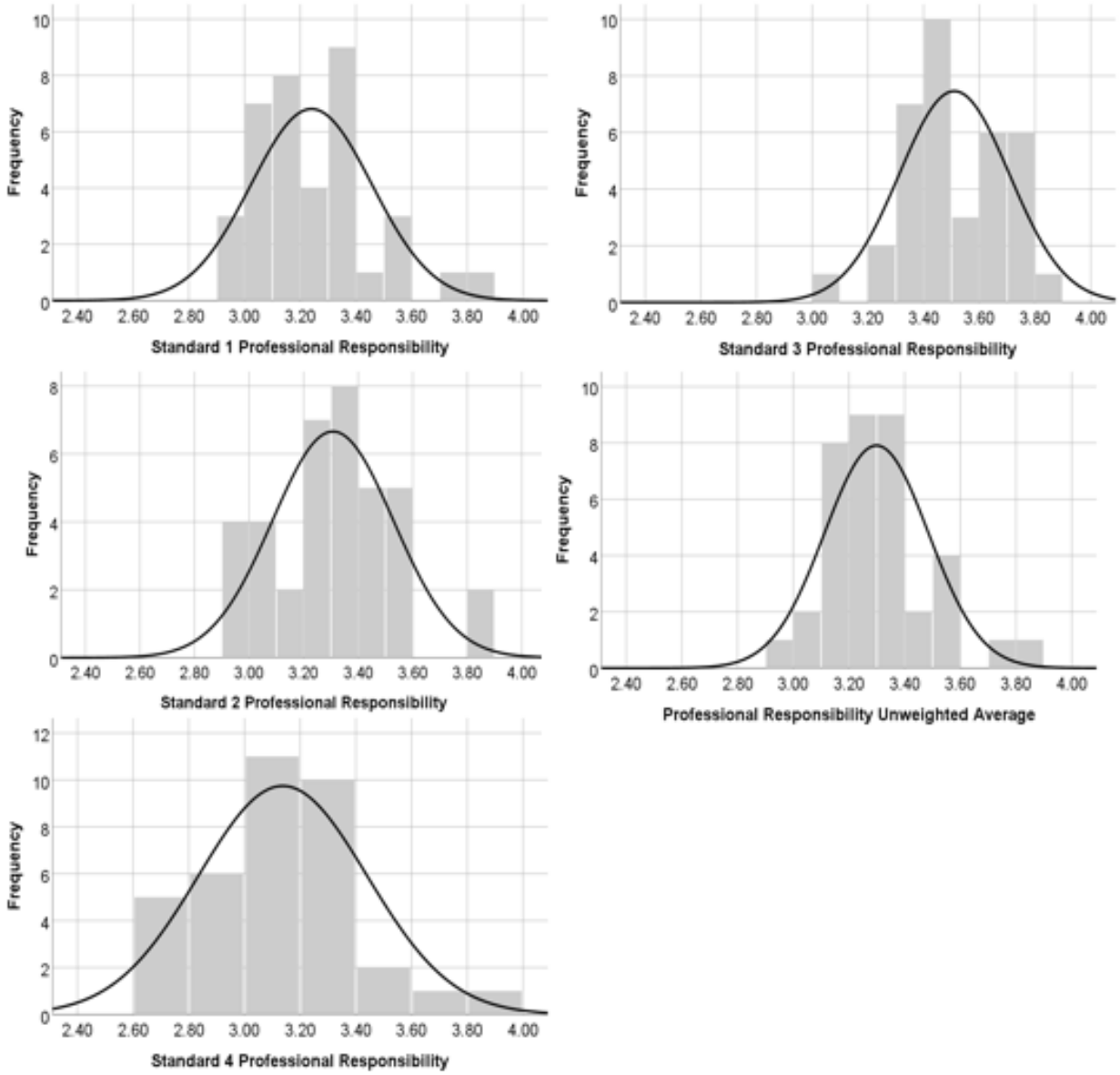


**Figure 15. Distribution of School-Level NEPF Administrator Professional Responsibilities Scores (All Years)**



Let's now look at the distribution of final scores. Figure 16 shows the distribution school-level NEPF teacher final scores (all years) with no weighting applied. The black vertical lines show the lower and upper bounds of the cut score for a teacher to receive a final rating of Effective. It's important to note here, that without any weighting applied, no schools maintain an average that could be classified as Ineffective and very few maintain an average of Developing. Because these are school-averages, the data are certainly masking the true number of Developing or Ineffective teachers, especially if these types of teachers do not cluster by school. Table 20 shows that the school-average teacher NEPF score with no weighting is 3.28, which sits in the middle of the Effective range. The skew of this distribution is 0.73, which indicates a small right-skew, and the kurtosis is 0.34, which means the distribution is slightly steeper than a typical normal distribution.

**Figure 16. Distribution of School-Level NEPF Teacher Final Scores (Unweighted, All Years)**

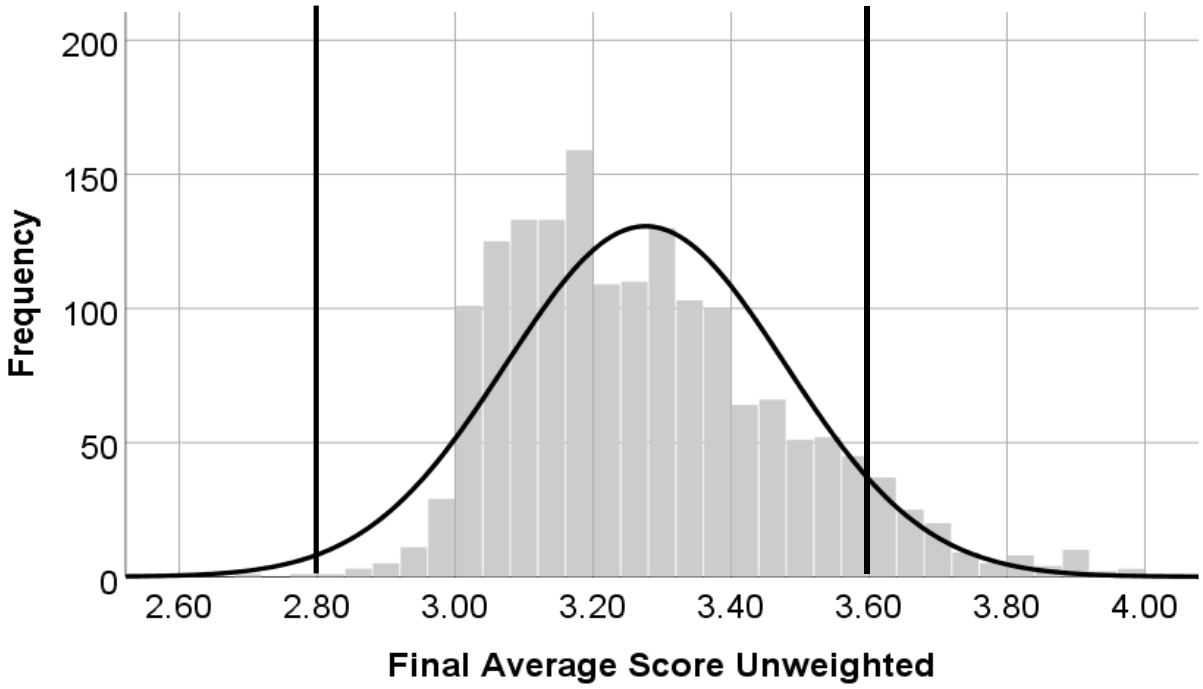


**Table 20. Summary Statistics for School-Level NEPF Teacher Final Score (All Years)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 3.28 | 0.20 | 2.70 | 4.00 | 0.73 | 0.34 |
| **Final Avg. Score (2019-20 weights)** | 3.27 | 0.20 | 2.73 | 3.99 | 0.60 | 0.14 |
| **Final Avg. Score (2018-19 weights)** | 3.28 | 0.23 | 2.26 | 3.99 | 0.53 | -0.03 |
| **Final Avg. Score (2017-18 weights)** | 3.27 | 0.21 | 2.68 | 3.99 | 0.58 | 0.09 |

**Table 21. Percentage of School-Level NEPF Teacher Final Scores Classified by Effectiveness Level (All Years)**

| | Ineffective | Developing | Effective | Highly Effective |
|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 0 | 0.10 | 92.20 | 7.70 |
| **Final Avg. Score (2019-20 weights)** | 0 | 0.40 | 92.60 | 7.00 |
| **Final Avg. Score (2018-19 weights)** | 0 | 0.70 | 87.30 | 12.00 |
| **Final Avg. Score (2017-18 weights)** | 0 | 0.40 | 91.50 | 8.10 |

Table 21 shows the percentage of school-level NEPF teacher final scores classified by effectiveness level. When using the final average score with no weighting applied, we find that 0% of schools have an average final rating in the Ineffective range and less than one percent have an average final rating in the Developing range. Instead, 92% of schools have an average rating of Effective and 8% have an average rating of Highly Effective.

What happens to the distribution when we start to apply the weighting? As a reminder, the TLC recommends the weighting of NEPF domains in an educators' final evaluation score and the final weights are established in statute (and have changed over time as mentioned above). The weighting is applied after an educator receives their final score in each of the NEPF domains. An example of how this works is provided below in Table 22.

Because the domain weights have changed over time, we recalculated the school-average final teacher scores from the School-Aggregate Teacher NEPF Scores file and the district-average final administrator scores from the District-Aggregate Administrator NEPF Scores file in each year using different weighting schemes. We used the 2019-20 weights (as shown in Table 22), the 2018-19 weights, and the 2017-18 weights (which were identical to the 2016-17 weights). This allows for a comparable look at growth in NEPF scores over time under the different weight sets.

**Table 22. Example of Final NEPF Score Calculation (2019-20)**

| Domain | Weight | X | Score | = | Weighted Score |
|---|---|---|---|---|---|
| Instructional Practice | .65 | X | 4 | = | 2.6 |
| Professional Responsibilities | .20 | X | 4 | = | .80 |
| Student Outcomes | .15 | X | 3 | = | .45 |
| | | | Final Score | = | 3.85 |

Cut Scores for Final Performance Rating: Highly Effective (3.6-4.0); Effective (2.8-3.59) Developing (1.91-2.79); Ineffective (1.0-1.9). For teachers or administrators to receive a rating of Effective, they must have scored a 2, 3, or 4 on the Student Outcomes domain For them to receive a rating of Highly Effective, they must have received a 3 or 4 on the Student Outcomes domain.

Once an educator's weighted final score is calculated, they are assigned a final rating category based on the established cut adopted by the State Board of Education. The example in Table 22 shows that this educator, with a final weighted score of 3.85 would have received a final rating of Highly Effective on their evaluation. If all domains were weighted equally and just took the simple average of their NEPF score, they would receive a 3.6 and would still receive a Highly Effective rating.

Now consider a case where the Instructional Practice score remains at 4 but the Professional Responsibilities and Student Outcomes score are both at 2 (the lowest possible Student Outcomes score to still receive an effective rating). An educator, in this case would receive a final weighted score of 2.6 + .4 + .30 = 3.3, and would receive an Effective rating. If taking just the simple average of this educator's domain scores (weighting all domains equally), they would receive a 2.7 and would be rated as Developing. In short, the weighting of the domains could change the distribution of educator effectiveness in meaningful ways.

Figures 17, 18, and 19 show the school-level NEPF teacher final score distributions with the 2019-20, 2018-19, and 2017-18 weighting schemes applied. Tables 20 and 21 show the summary statistics for the final scores and the percentage of schools with average scores fitting each of the effective ratings by each of the different weighting schemes.

We find that the different NEPF weighting schemes do not appreciably shift the distribution of school-average educator performance. The distributions in Figures 17, 18, and 19 look very similar to the unweighted distributions and Table 20 confirms this. The mean scores and standard deviations obtained under the different weighting schemes look very similar to the unweighted version of the NEPF. Table 21 shows that percentage of schools scoring in the different teacher rating categories does not change meaningfully based on the different weighting schemes, with perhaps one exception.

More schools are rated as Highly Effective when using the 2018-19 weighting, that gives more weight to the Student Outcomes Domain. In 2018-19, the Student Outcomes domain received a 40% weight versus a 20% weight in 2016-17 and 2017-18 and a 15% weight in 2019-20.

**Figure 17. Distribution of School-Level NEPF Teacher Final Scores (2019-20 Weights, All Years)**

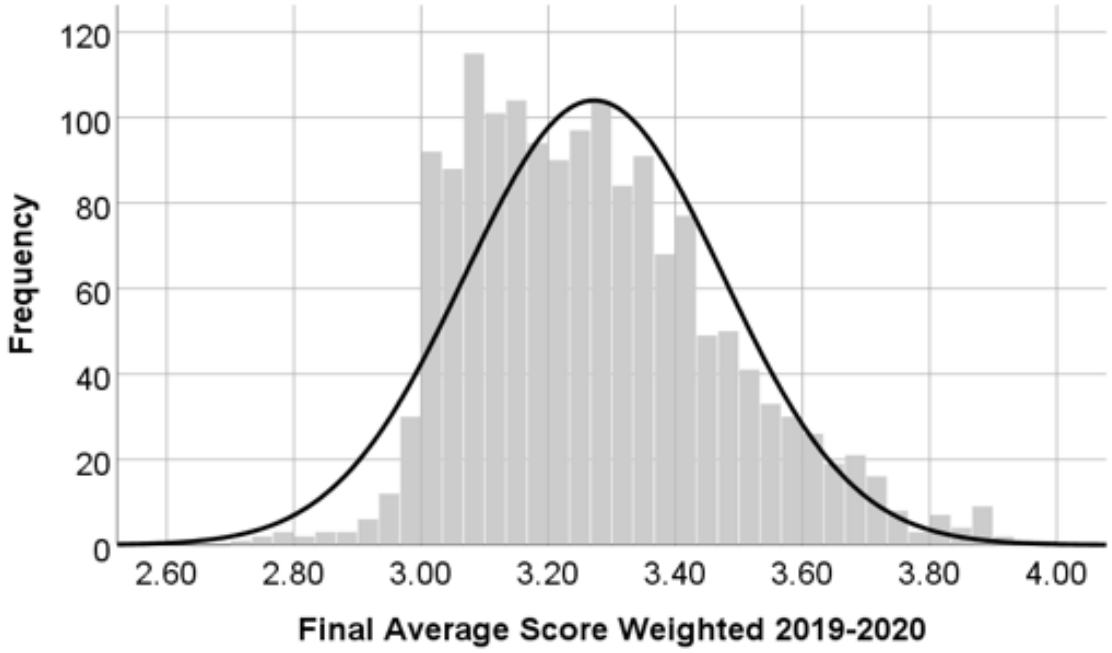


**Figure 18. Distribution of School-Level NEPF Teacher Final Scores (2018-19 Weights, All Years)**

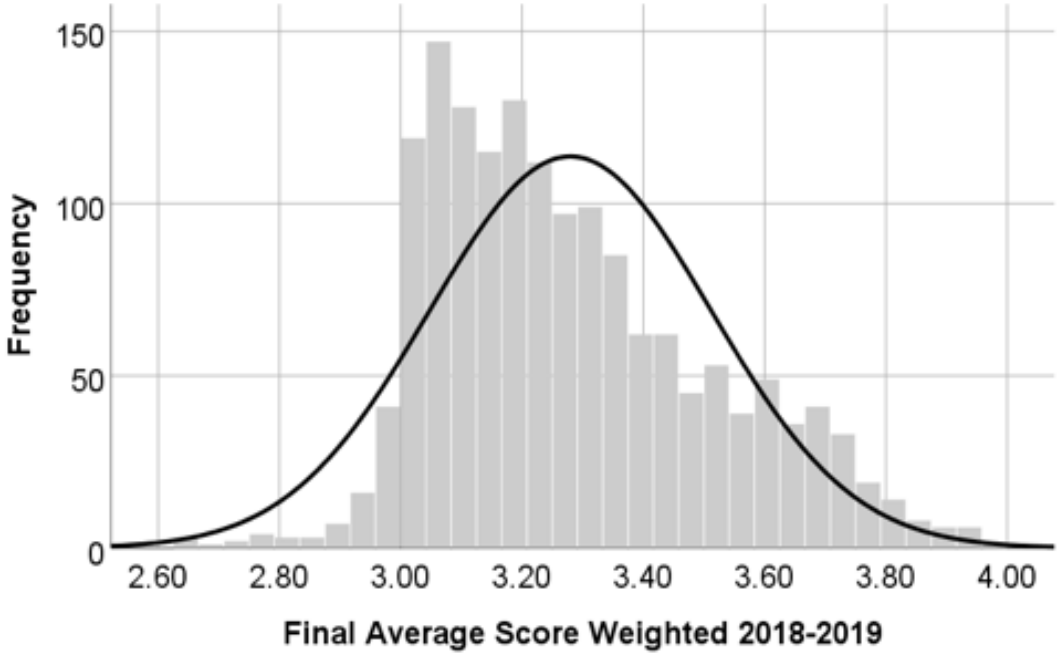
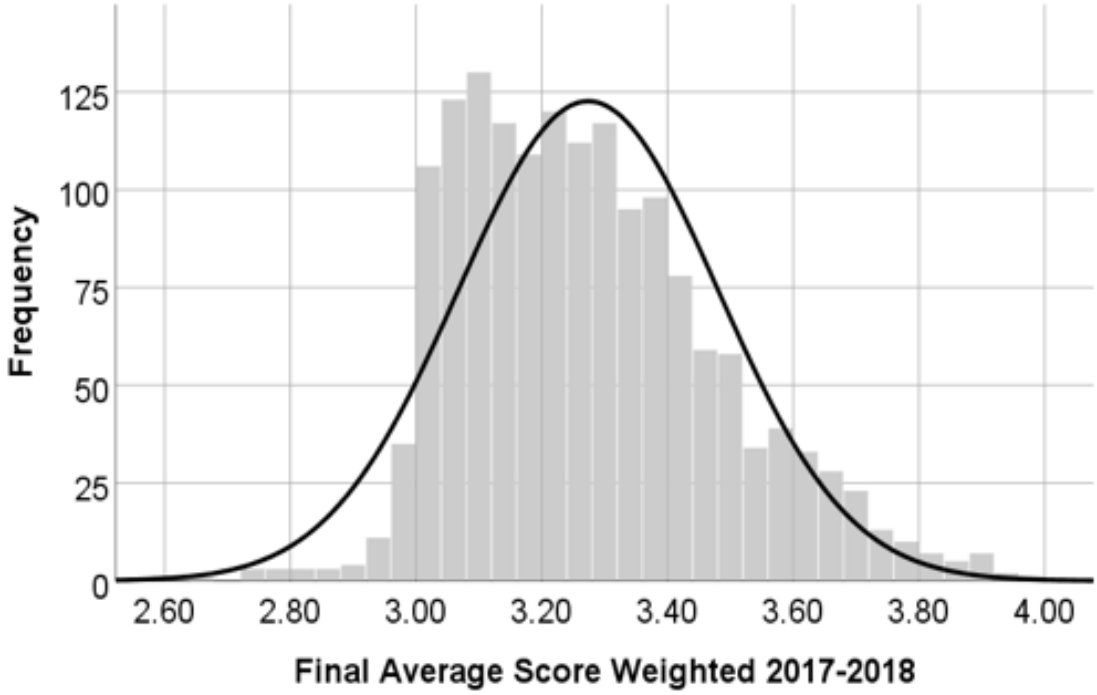



**Figure 19. Distribution of School-Level NEPF Teacher Final Scores (2017-18 Weights, All Years)**



We ran the same analyses for administrators (at the district-level). Figure 20 shows the distribution of district-level NEPF administrator final scores (all years) with no weighting applied, Figure 21 shows the distribution with the 2019-20 weighting, Figure 22 shows the distribution with the 2018-19 weighting, and Figure 23 shows the distribution with the 2017-18 weighting. Table 23 provides the summary statistics for the district-level NEPF administrator final scores based on each of the weighting schemes and Table 24 shows the percentage of districts with average scores fitting each of the effective ratings by each of the different weighting schemes. Altogether, we find very similar results to the school-level teacher NEPF analysis. In short, weighting does not significantly change the distribution of district-aggregate final administrator NEPF scores. The 2018-19 weighting, which privileges more of the Student Outcomes domain allows slightly more districts to score in the Highly Effective rating category as shown in Table 24.

**Figure 20. Distribution of District-Level NEPF Administrator Final Scores (Unweighted, All Years)**

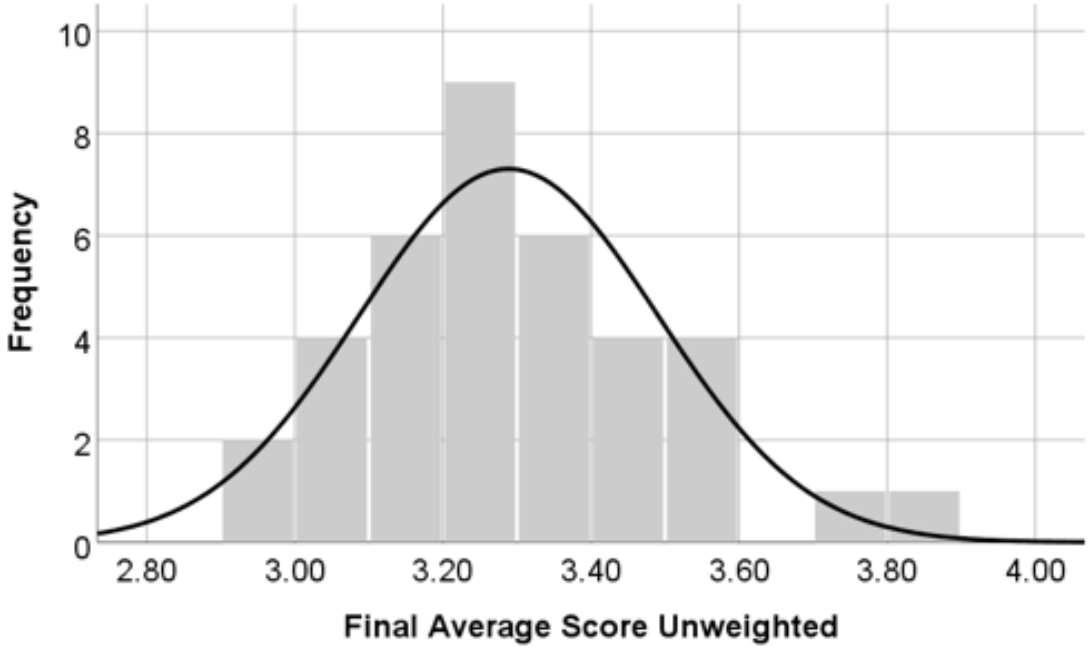


**Figure 21. Distribution of District-Level NEPF Administrator Final Scores (2019-20 Weights, All Years)**

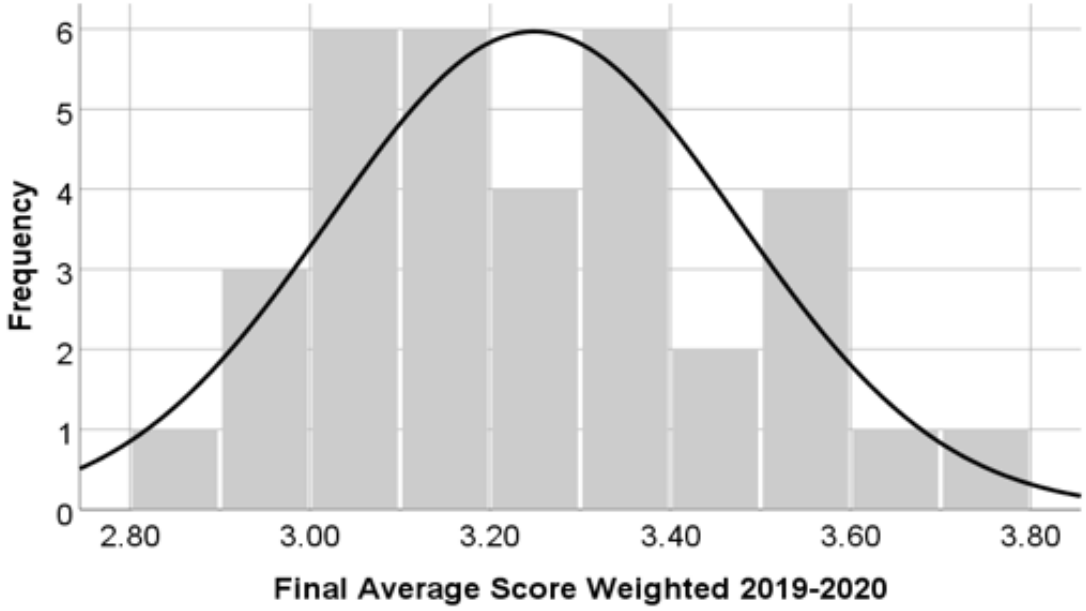



**Figure 22. Distribution of District-Level NEPF Administrator Final Scores (2018-19 Weights, All Years)**

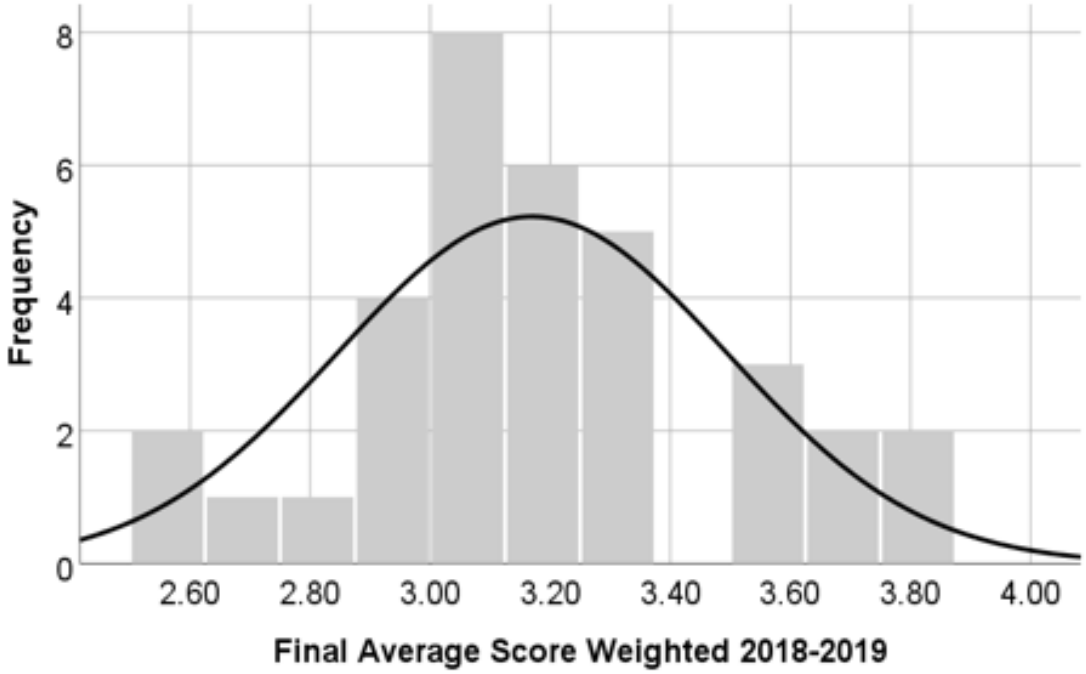


**Figure 23. Distribution of District-Level NEPF Administrator Final Scores (2017-18 Weights, All Years)**

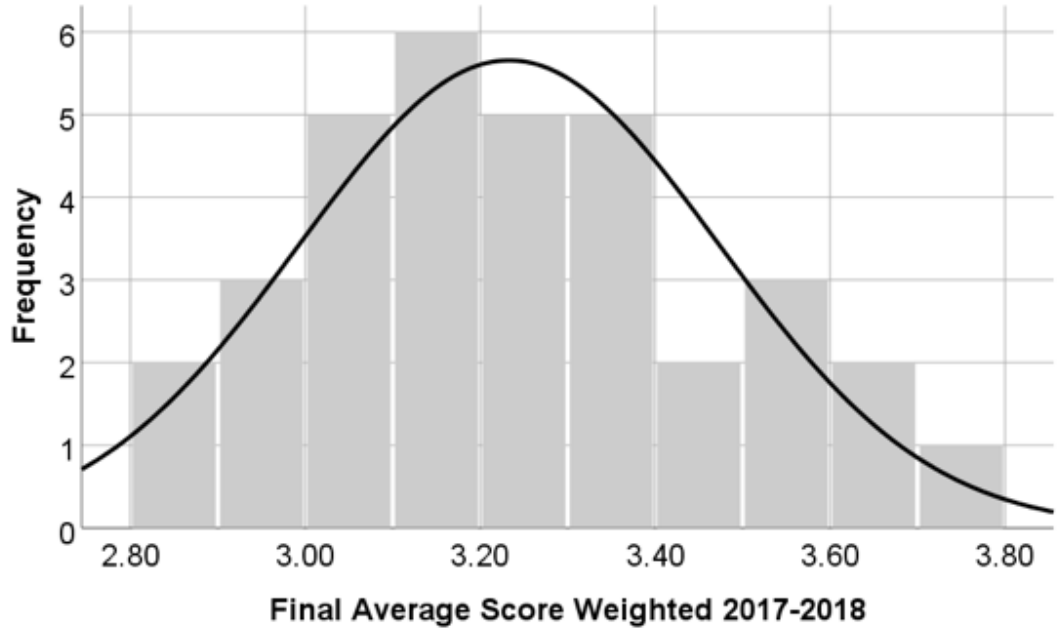



**Table 23. Summary Statistics for District-Level NEPF Administrator Final Scores (All Years)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 3.29 | 0.20 | 2.91 | 3.81 | 0.48 | 0.38 |
| **Final Avg. Score (2019-20 weights)** | 3.25 | 0.23 | 2.88 | 3.78 | 0.40 | -0.40 |
| **Final Avg. Score (2018-19 weights)** | 3.17 | 0.32 | 2.53 | 3.85 | 0.16 | -0.09 |
| **Final Avg. Score (2017-18 weights)** | 3.23 | 0.24 | 2.83 | 3.79 | 0.46 | -0.32 |

**Table 24. District-Level NEPF Administrator Final Scores Classified by Effectiveness Level (All Years)**

| | Ineffective | Developing | Effective | Highly Effective |
|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 0 | 0 | 94.60 | 5.40 |
| **Final Avg. Score (2019-20 weights)** | 0 | 0 | 94.10 | 5.90 |
| **Final Avg. Score (2018-19 weights)** | 0 | 0 | 88.20 | 11.80 |
| **Final Avg. Score (2017-18 weights)** | 0 | 0 | 91.20 | 8.80 |

What happens if we explore this information at the individual level? We presented the school-aggregate and district-aggregate information first because we have more years of data and data by standards, domains, and overall scores. However, looking at the individual-level will help unpack just how many educators are scoring in the bottom and upper ends of the performance distribution (which may be masked by the school-level and district-level averages mentioned above).

Figure 24 shows the distribution of teacher NEPF final scores for the 2018-19 school year. The distribution somewhat mirrors the school-aggregate scores shown in Figures 17-19 above, where most teachers score in the Effective range (between 2.80 and 3.60) with very few teachers scoring in the Ineffective/Developing ranges and a moderate number of teachers reaching Highly Effective. Interestingly, there tends to be a group of teachers clustering around 3.00 (likely because they are receiving 3.00 on all domains). Table 25 reveals that the average final NEPF score for teachers in 2018-19 was 3.28 with a standard deviation of 0.29 and a minimum and maximum of 1.00 and 4.00. Table 26 shows that only a tenth of a percent of teachers were classified as Ineffective and only 1.7% as Developing. The vast majority (81.90%) are classified as Effective with another 16% as Highly Effective. Again, weighting does not appear to change these percentages much—except the 2018-19 weights again appear to increase the percentage of teachers in the Highly Effective category. Importantly, it appears the school-level averages presented above tend to deflate the percentage in the Highly Effective category, but not the percentage in the Ineffective or Developing categories.

**Figure 24. Distribution of Teacher NEPF Final Scores (2018-19, Unweighted)**

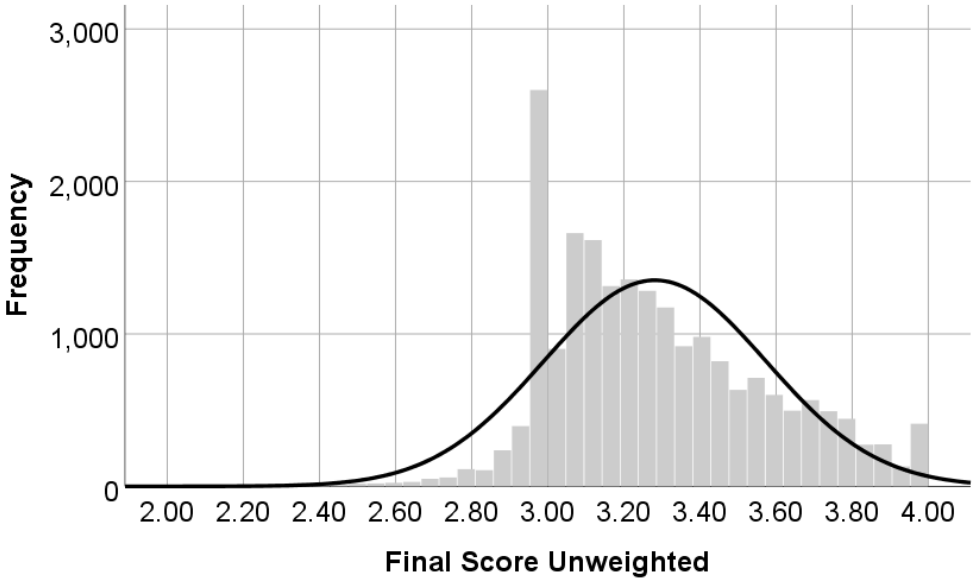



**Table 25. Summary Statistics for Teacher NEPF Final Scores (2018-19)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 3.28 | 0.29 | 1.00 | 4.00 | 0.18 | 1.24 |
| **Final Avg. Score (2019-20 weights)** | 3.31 | 0.29 | 1.00 | 4.00 | 0.11 | 1.10 |
| **Final Avg. Score (2018-19 weights)** | 3.33 | 0.32 | 1.00 | 4.00 | 0.08 | 0.34 |
| **Final Avg. Score (2017-18 weights)** | 3.31 | 0.29 | 1.00 | 4.00 | 0.11 | 0.96 |

**Table 26. Teacher NEPF Final Scores Classified by Effectiveness Levels (2018-19)**

| | Ineffective | Developing | Effective | Highly Effective |
|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 0.10 | 1.70 | 81.90 | 16.30 |
| **Final Avg. Score (2019-20 weights)** | 0.10 | 1.50 | 79.90 | 18.50 |
| **Final Avg. Score (2018-19 weights)** | 0.10 | 1.50 | 75.50 | 22.90 |
| **Final Avg. Score (2017-18 weights)** | 0.10 | 1.50 | 79.30 | 19.10 |

Figure 25 shows the distribution of administrator NEPF final scores for the 2018-19 school year. The distribution somewhat mirrors the district-aggregate scores shown in Figures 20-23 above. Table 27 shows that the average administrator final score in 2018-19 was 3.36. Not a single administrator had a final score below 2.00. Table 28 shows that no administrator received a rating of Ineffective in 2018-19 and only 1 percent received a rating of Developing. Most administrators (79%) receive ratings of Effective with another 21% rated as Highly Effective. Again, weighting does not appear to change these percentages much—except the 2018-19 weights again appears to increase the percentage of administrators in the Highly Effective category.

**Figure 25. Distribution of Administrator NEPF Final Scores (2018-19, Unweighted)**

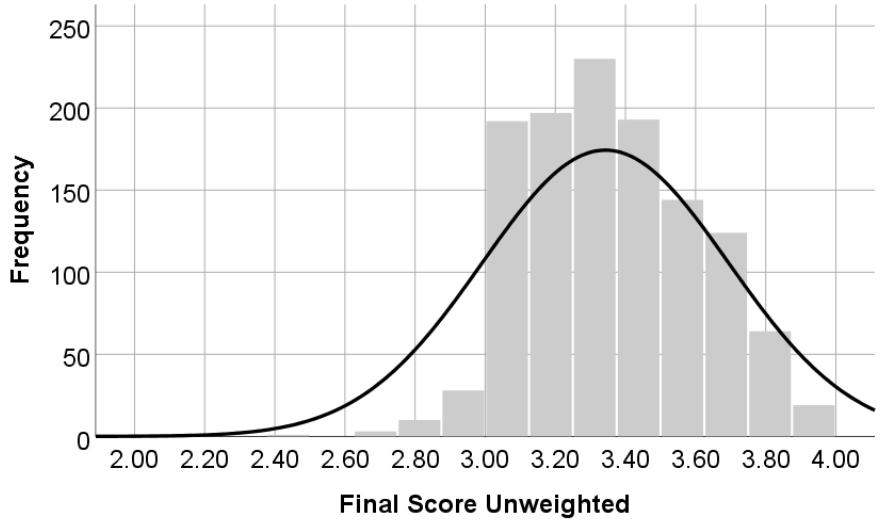



**Table 27. Summary Statistics for Administrator Final Scores (2018-19)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 3.36 | 0.26 | 2.02 | 4.00 | 0.11 | 0.06 |
| **Final Avg. Score (2019-20 weights)** | 3.37 | 0.27 | 2.32 | 4.00 | 0.05 | -0.17 |
| **Final Avg. Score (2018-19 weights)** | 3.37 | 0.32 | 1.78 | 4.00 | -0.37 | 1.61 |
| **Final Avg. Score (2017-18 weights)** | 3.37 | 0.27 | 2.38 | 4.00 | -0.01 | -0.02 |

**Table 28. NEPF Administrator Final Scores Classified by Effectiveness Level (2018-19)**

| | Ineffective | Developing | Effective | Highly Effective |
|---|---|---|---|---|
| **Final Avg. Score (Unweighted)** | 0 | 1.20 | 78.90 | 19.80 |
| **Final Avg. Score (2019-20 weights)** | 0 | 1.50 | 76.00 | 22.50 |
| **Final Avg. Score (2018-19 weights)** | 0 | 1.00 | 71.60 | 26.40 |
| **Final Avg. Score (2017-18 weights)** | 0 | 1.80 | 76.00 | 22.20 |

## THE NEPF HAS MODERATE PREDICTIVE VALIDITY

Table 29 returns to the school-level analysis of NEPF teacher scores and showcases the bivariate correlations between school-average NEPF scores and the percentage of students at the school-level Emerging, Approaching, Meets, Exceeds, and Proficient (Meets/Exceeds) on the SBAC Math and Reading exams. The correlations are presented for the unweighted final NEPF scores as well as the scores calculated using the 2019-20, 2018-19, and 2017-18 weights. High (above 0.7) or moderate (between 0.3 and 0.7) correlations between school-average NEPF scores and school proficiency scores suggests a certain degree of predictive validity for the NEPF (i.e. the NEPF is related to measures that we think it should be related to, like student achievement).

Table 29 shows that school-level NEPF teacher scores are negatively related to the percentage of students scoring Emerging in math (-0.35) and reading (-0.32) and the percentage of students scoring Approaching in math (-0.04) and reading (-0.05). Alternatively, we find that school-level NEPF teacher scores are positively associated with the percentage of students scoring Meets (0.29), Exceeds (0.30), and Proficient (0.33) in math and the percentage of students scoring Meets (0.25), Exceeds (0.23), and Proficient (0.28) in reading. The weighting does not meaningfully alter these correlations.

Table 30 shows the same analysis at the district-level for administrators. We find that district-level NEPF administrator scores are negatively related to the percentage of students scoring Emerging in math (-0.34) and reading (-0.22), positively related to the percentage of students scoring Approaching in math (0.16), but negatively associated with the percentage of students scoring Approaching in reading (-0.36). Alternatively, we find that district-level NEPF administrator scores are positively associated with the percentage of students scoring Meets (0.19), Exceeds (0.33), and Proficient (0.28) in math and the percentage of students scoring Meets (0.41), Exceeds (0.15), and Proficient (0.33) in reading. The weighting does not considerably change these correlations.

**Table 29. Bivariate Correlations Between Final School-Level NEPF Teacher Scores and Student Proficiency**

| | Final Avg. Unweighted | Final Avg. (2019-20 Weights) | Final Avg. (2018-19 Weights) | Final Avg. (2017-18 Weights) |
|---|---|---|---|---|
| Final Avg. Unweighted | 1.00 | | | |
| Final Avg. (2019-20 Weights) | 0.96 | 1.00 | | |
| Final Avg. (2018-19 Weights) | 0.83 | 0.94 | 1.00 | |
| Final Avg. (2017-18 Weights) | 0.94 | 0.99 | 0.96 | 1.00 |
| Math % Emerging | -0.35 | -0.38 | -0.39 | -0.38 |
| Math % Approaching | -0.04 | -0.06 | -0.07 | -0.08 |
| Math % Meets | 0.29 | 0.31 | 0.29 | 0.31 |
| Math % Exceeds | 0.30 | 0.35 | 0.36 | 0.35 |
| Math % Proficient (Meets/Exceeds) | 0.33 | 0.37 | 0.37 | 0.37 |
| Reading % Emerging | -0.32 | -0.37 | -0.37 | -0.37 |
| Reading % Approaching | -0.05 | -0.08 | -0.08 | -0.10 |
| Reading % Meets | 0.25 | 0.29 | 0.29 | 0.29 |
| Reading % Exceeds | 0.23 | 0.27 | 0.28 | 0.28 |
| Reading % Proficient (Meets/Exceeds) | 0.28 | 0.33 | 0.34 | 0.34 |

All correlations are significant at the 0.01 level (2-tailed). N=1,269

It's important to note that these correlations do not mean that that the NEPF causes higher student performance. There could be several school- and district- factors that explain these results that have little to do with the NEPF. For example, it may be that teachers or administrators who perform higher on the NEPF self-select into schools with more proficient and fewer emerging/developing students. This would lead to positive correlations between teacher/administrator NEPF scores and the percentage of students scoring at the upper end of the achievement distribution and would lead to negative correlations between teacher/administrator NEPF scores and the percentage of students scoring at the bottom end of the achievement distribution. We try to disentangle some of these factors further in the impact analysis presented below.

**Table 30. Bivariate Correlations Between Final District-Level NEPF Administrator Score and Student Proficiency**

| | Final Avg. Unweighted | Final Avg. (2019-20 Weights) | Final Avg. (2018-19 Weights) | Final Avg. (2017-18 Weights) |
|---|---|---|---|---|
| Final Avg. Unweighted | 1.00 | | | |
| Final Avg. (2019-20 Weights) | 0.88 | 1.00 | | |
| Final Avg. (2018-19 Weights) | 0.54 | 0.87 | 1.00 | |
| Final Avg. (2017-18 Weights) | 0.81 | 0.99 | 0.93 | 1.00 |
| Math % Emerging | -0.34 | -0.52 | -0.54 | -0.53 |
| Math % Approaching | 0.16 | 0.15 | 0.10 | 0.14 |
| Math % Meets | 0.19 | 0.41 | 0.51 | 0.45 |
| Math % Exceeds | 0.33 | 0.44 | 0.42 | 0.44 |
| Math % Proficient (Meets/Exceeds) | 0.28 | 0.46 | 0.50 | 0.48 |
| Reading % Emerging | -0.22 | -0.42 | -0.48 | -0.44 |
| Reading % Approaching | -0.36 | -0.39 | -0.32 | -0.38 |
| Reading % Meets | 0.41 | 0.49 | 0.44 | 0.49 |
| Reading % Exceeds | 0.15 | 0.36 | 0.45 | 0.39 |
| Reading % Proficient (Meets/Exceeds) | 0.33 | 0.51 | 0.54 | 0.53 |

All Correlations above .35 are significant at the 0.05 level, above .41 significant at the .01 level (2-tailed). N=34

## ADMINISTRATOR AND TEACHER NEPF SCORES ARE MODERATELY CORRELATED

Here we assess the correlation between district-average NEPF domain scores for teachers and administrators using a bivariate correlation matrix (Table 31). High correlations entail that school districts where teachers are performing higher also have higher administrator NEPF performance. The most instructive correlations to look at are between district-average administrator Instructional Leadership and teacher Instructional Practice scores, between district-average administrator Professional Responsibilities and teacher Professional Responsibilities scores, between district-average administrator Student Outcomes and teacher Student Outcomes scores, and between district-average administrator Final Score and teacher Final Score.

We find a positive moderate correlation between district-average administrator Instructional Leadership and teacher Instructional Practice scores (0.64), between district-average administrator Professional Responsibilities and teacher Professional responsibilities scores (0.48), between district-average administrator Student Outcomes and teacher Student Outcomes scores (0.28), and between district-average administrator Final Scores and teacher Final Scores (0.62). These moderate positive

correlations tend to indicate, again, that districts with higher performing teachers on the NEPF also have higher performing administrators.

**Table 31. Bivariate Correlations Between District-Level Teacher and Administrator NEPF Scores (All Years)**

| | | Administrator | | | |
|---|---|---|---|---|---|
| | | Instructional Leadership | Professional Responsibilities | Student Outcomes | Final Avg. Score (unweight.) |
| Teacher | Instructional Practice | **0.64** | 0.61 | -0.08 | 0.64 |
| | Professional Responsibilities | 0.52 | **0.48** | 0.29 | 0.51 |
| | Student Outcomes | 0.57 | 0.51 | **0.28** | 0.56 |
| | Final Avg. Score (unweight.) | 0.62 | 0.59 | 0.11 | **0.62** |

All correlations above .40 significant at the 0.01 level (2-tailed)
All correlations above .35 significant at the 0.05 level (2-tailed)

## THERE IS VERY LITTLE CHANGE IN SCHOOL AND DISTRICT NEPF FINAL SCORES OVER TIME

Table 32 presents the average year-to-year change in school-aggregate teacher NEPF scores (termed NEPF growth) for all years and then for 2016-17 to 2017-18 and 2017-18 to 2018-19 separately. The mean in this table captures the change in the final teacher NEPF scores from one year to the next. A positive mean entails that on average, schools tend to improve their NEPF scores over time. A negative mean entails that on average, schools tend to regress in their NEPF scores over time. Table 32 makes clear that we see very little year-to-year change in school average teacher NEPF final scores. The mean change between years is 0.03 points with a standard deviation of 0.13 points. The average school NEPF growth from 2016-17 to 2017-18 is essentially zero and the average school NEPF growth from 2017-18 to 2018-19 is 0.06 points. We plot the distribution of change in Figure 26. Notice that the distribution centers around 0, with a few schools experiencing a year-to-year decrease as high as 0.73 points and a year-to-year improvement as high as 0.66 points.

**Table 32. Summary Statistics for School-Level NEPF Teacher Final Score (All Years)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **NEPF Growth (all years)** | 0.03 | 0.13 | -0.73 | 0.66 | -0.19 | 4.79 |
| **NEPF Growth (2016-17 to 2017-18)** | 0.00 | 0.13 | -0.73 | 0.66 | -0.45 | 5.30 |
| **NEPF Growth (2017-18 to 2018-19)** | 0.06 | 0.12 | -0.69 | 0.59 | 0.18 | 4.52 |

**Figure 26. Distribution of Year-to-Year Change in School-Level NEPF Teacher Final Scores**

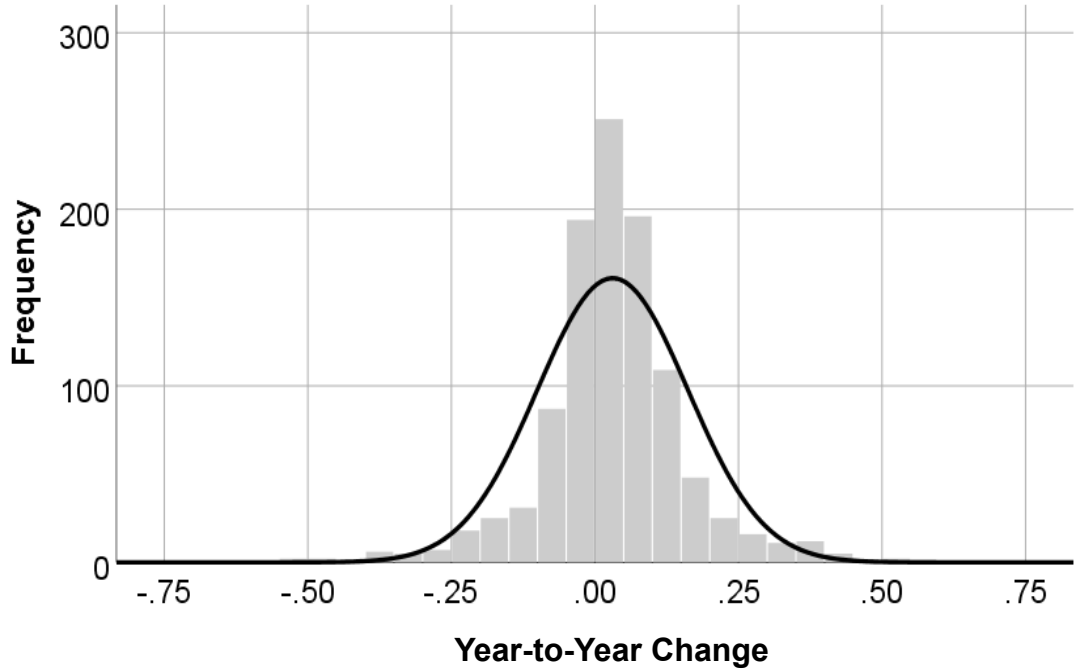
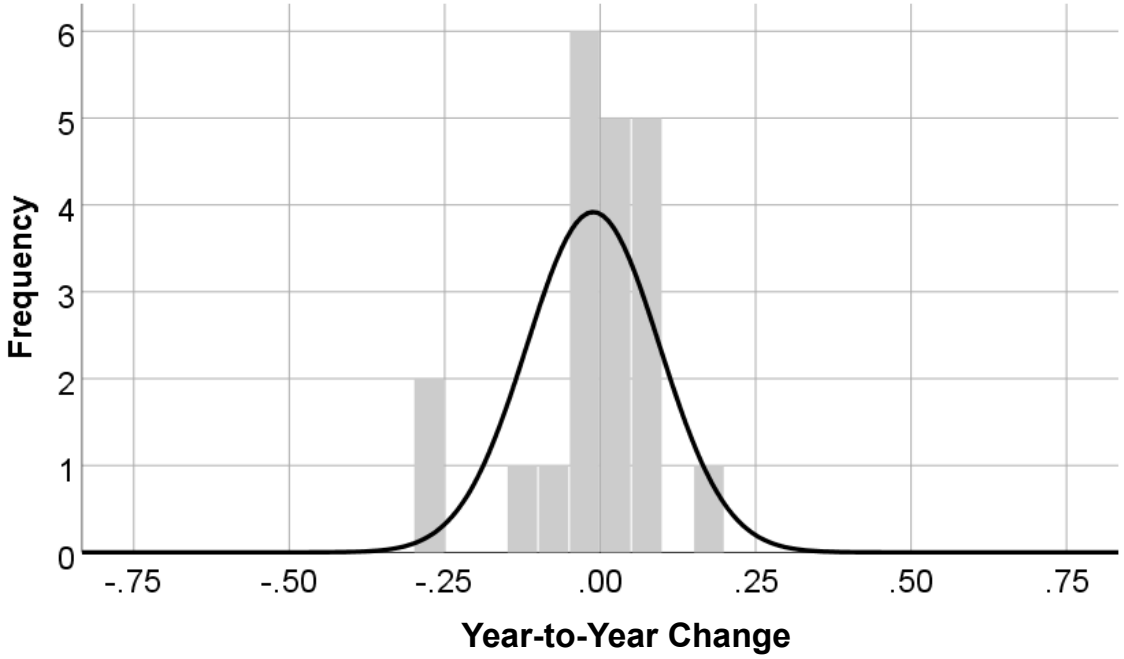



Table 33 and Figure 27 presents this same information for administrators and the findings are the same. We find that, on average, school districts do not make much improvement year-to-year in their final administrator NEPF scores. On average, the year-to-year change for a school district is -0.01 points with a standard deviation of 0.11 points. School districts have experienced a decline in their average administrator NEPF score by as large as -0.29 points (the min) and an improvement by as much as 0.15 points (the max). The results are similar when looking at the year-to-year change from 2016-17 to 2017-18 and the year-to-year change from 2017-18 to 2018-19 separately. Figure 27 makes clear that any changes are centered around a mean change of zero.

**Table 33. Summary Statistics for School-Level NEPF Administrator Final Scores (All Years)**

| | Mean | SD | Min | Max | Skew | Kurt |
|---|---|---|---|---|---|---|
| **NEPF Growth (all years)** | -0.01 | 0.11 | -0.29 | 0.15 | -1.42 | 2.51 |
| **NEPF Growth (2016-17 to 2017-18)** | -0.01 | 0.06 | -0.10 | 0.09 | 0.33 | -0.14 |
| **NEPF Growth (2017-18 to 2018-19)** | -0.01 | 0.14 | -0.29 | 0.15 | -1.33 | 1.01 |

**Figure 27. Distribution of Year-to-Year Change in District-Level NEPF Administrator Final Scores**



# IMPACT OF THE NEPF

## RESEARCH QUESTIONS 4 AND 5

### GROWTH ON THE TEACHER NEPF HAS NO IMPACT ON SCHOOL ACHIEVEMENT GROWTH

Here we examine the relationship between growth in school-aggregate teachers' NEPF scores and school achievement (see Table 34). As reminder, these models investigating growth use a fixed effect approach that helps us account for several fixed differences (i.e., time differences between schools or districts that the models presented above do not. Each school or district serves as their own control group, allowing us to assess impacts as deviations from within school (or within district) averages over time. We can still not be totally sure that we have isolated a cause and effect relationship (i.e. that changes in NEPF scores causes changes in student achievement) because there still may be important time-varying differences between schools and districts that we have not accounted for, though we do our best by using available control variables.

Table 34 shows that we do not find a statistically significant relationship between changes in the percentage of teachers rated Effective or Highly Effective and change in school reading or math achievement scores. Similarly, we do not find a statistically significant relationship between changes in school-average NEPF scores and school achievement.

**Table 34. Relationship Between Growth in School-Aggregate Teacher NEPF Scores and Growth in School Achievement**

| | Reading | | Math | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| % Teachers Rated Effective or Highly Effective | 0.00<br>(0.00) | | 0.00<br>(0.00) | |
| NEPF Final Scoring (2019-20 weighting) | | -0.04<br>(0.05) | | 0.04<br>(0.06) |
| Demographic Controls | X | X | X | X |
| School Fixed Effect | X | X | X | X |
| Year Fixed Effect | X | X | X | X |

Standard errors clustered at the school level in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Standardized scores are derived from uncoarsening total school performance levels by subject and year

## GROWTH ON THE ADMINISTRATOR NEPF HAS NO IMPACT ON DISTRICT ACHIEVEMENT GROWTH

The results are similar when we examine the relationship between growth in the percentage of Administrators rated Effective or Highly Effective or district-aggregate administrators' NEPF scores and student achievement (see Table 35). We again find no statistically significant relationships. In short, we do not find convincing evidence that growth in administrators' NEPF scores is associated with growth in student achievement. Note that the estimates on NEPF Final Score (0.08 and 0.21) are positive and relatively large. However, given the small number of district-by-year observations, we do not have enough power at the district level to confidently detect any statistically significant relationships.

**Table 35. Relationship Between Growth in District-Aggregate Administrator NEPF Scores and Growth in District Achievement**

| | Reading | | Math | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| % Administrators Rated Effective or Highly Effective | 0.00<br>(0.00) | | 0.00<br>(0.00) | |
| NEPF Final Scoring (2019-20 weighting) | | 0.08<br>(0.18) | | 0.21<br>(0.34) |
| Demographic Controls | X | X | X | X |
| District Fixed Effect | X | X | X | X |
| Year Fixed Effect | X | X | X | X |

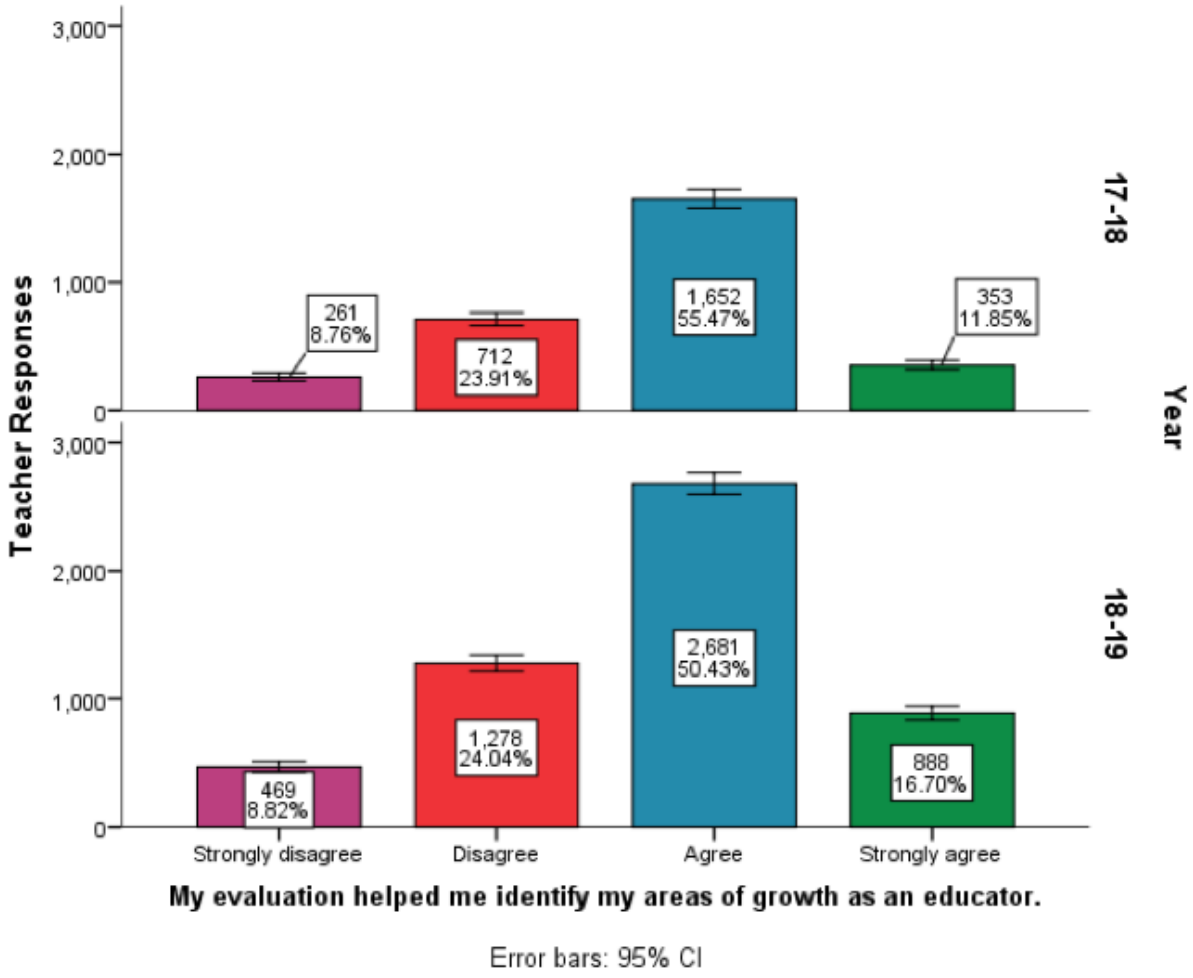
Standard errors clustered at the district level in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Standardized scores are derived from uncoarsening total school performance levels by subject and year

## MOST EDUCATORS AGREE THAT THE NEPF HELPS THEM IDENTIFY AREAS FOR GROWTH

While we did not detect any statistical impacts of the NEPF, we further explore whether educators perceive any impact of the NEPF on their practice. Our growth analysis above is limited by the small amount of detectable growth on the NEPF. In other words, because most teachers and administrators regularly score Effective or Highly Effective, it is really difficult to assess the impact of growth over time. Consequently, if we expect the NEPF to have an impact on educator practice and student outcomes, the result will not be identified in growth in NEPF scores, but instead in the quality of the feedback they receive from their NEPF evaluation cycles. While we cannot assess the statistical relationship between feedback quality and student achievement, we can explore whether educators believe that the NEPF yields quality feedback that helps them change their practice in positive ways.

First, we use items from the NDE Annual NEPF survey to assess whether teachers believe that NEPF feedback helps them identify areas for growth (Figure 28). Among the 2,978 teacher responses in 2017-18, we find that 55% of teachers in (2017-18) agree and 12% strongly agree that the NEPF helps them identify areas of growth as an educator. Thirty-three percent of teachers disagreed or strongly disagreed. The results are very similar in 2018-19.

**Figure 28. Teachers' Response to Survey Item: My Evaluation Helped Me Identify My Areas of Growth as an Educator**




We now turn to our CREA survey which sought to further establish the extent to which educators are receiving helpful feedback through the NEPF in ways that might drive improvements to practice and student outcomes. We first asked teachers how much feedback they typically receive from their supervisor as a result of their evaluation cycle. We asked this question overall and for each domain. The results are presented in Table 36.

We find that the average teacher believes that they receive “some” feedback on the NEPF (mean= 3.19, standard deviation= 0.76). Nearly 40% of teachers believe they receive a large amount of feedback from their supervisor throughout their evaluation cycle. However, 14% of teachers receive minimal feedback and 2% receive no feedback. In short, most teachers are receiving at least some feedback (83%), but 16% are still receiving minimal or no feedback. We further find that these amounts do not vary with large magnitude across the individual Instructional Practice and Professional Responsibilities standards. This finding draws into question the extent to which supervisors and teachers distinguish between the individual standards in their evaluations (or whether teachers tend to receive the same score across each of the standards and more generalized feedback). Recall that the EFA results in the Validity section supported the idea that the NEPF had high internal consistency but low dimensionality (meaning the Instructional Practice and Professional Responsibilities domains were measuring a similar construct).

**Table 36. Teacher Response to Survey Item: How Much Feedback Do You Typically Receive...?**

| | Mean | SD | None | Minimal | Some | A large amount |
|---|---|---|---|---|---|---|
| Overall | 3.19 | 0.76 | 2.2% | 14.4% | 45.0% | 38.4% |
| Instructional Practice Standard 1 | 3.06 | 0.82 | 4.9% | 16.4% | 46.8% | 31.9% |
| Instructional Practice Standard 2 | 3.08 | 0.83 | 4.7% | 16.0% | 45.3% | 33.9% |
| Instructional Practice Standard 3 | 3.16 | 0.81 | 4.0% | 14.3% | 43.4% | 38.3% |
| Instructional Practice Standard 4 | 3.08 | 0.84 | 5.2% | 16.2% | 44.0% | 34.6% |
| Instructional Practice Standard 5 | 3.07 | 0.83 | 5.0% | 16.3% | 45.0% | 33.6% |
| Professional Responsibilities Standard 1 | 2.94 | 0.91 | 8.7% | 18.7% | 42.8% | 29.7% |
| Professional Responsibilities Standard 2 | 3.02 | 0.89 | 7.0% | 17.6% | 41.5% | 33.9% |
| Professional Responsibilities Standard 3 | 2.98 | 0.90 | 8.0% | 17.4% | 43.3% | 31.3% |
| Professional Responsibilities Standard 4 | 2.93 | 0.92 | 9.2% | 18.6% | 42.3% | 29.9% |
| Professional Responsibilities Standard 5 | 2.95 | 0.92 | 8.8% | 18.0% | 42.1% | 31.1% |

We subsequently asked teachers whether they agreed that the feedback they received is helping them achieve progress on each of the Instructional Practice and Professional

Responsibilities standards (and overall in each domain). The results are shown in Table 37.

We find that, on average, teachers agree that the feedback they receive helps them achieve progress on the Instructional Practice and Professional Responsibilities standards. We find similar amounts of agreement across all of the standards. For example, on Instructional Practice standard 1, 41% of teachers agree and 27% of teachers strongly agree that the feedback they receive helps them achieve progress on the standard. Approximately, 6 percent disagree and 5 percent strongly disagree. The last two rows in Table 37 show whether teachers believe the feedback they receive helps them achieve growth overall in Instructional Practice and Professional Responsibility. Again, the percentages are very similar to what we find with the individual standards.

Table 38 explores teachers' perceptions of the SLG. Teachers were asked whether their SLG is based on their students' needs, whether the chosen assessment(s) to judge progress was appropriate, and whether the SLGs were set through a collaborative process with their supervisor. We find that most teachers agree (43%) or strongly agree (36%) that their SLG is based on student needs. Only 11% disagree or strongly disagree. We observe similar percentages for whether teachers believe the assessments used to judge progress are appropriate and whether the SLG is set through a collaborative process with their supervisor.

We now turn to the administrator responses on the CREA survey. Table 39 shows administrator perceptions on the amount of feedback they receive from their supervisor. We find that the average administrator believes that they receive "some" feedback on the NEPF (mean= 3.17, standard deviation 0.79). Identical to the teachers, 39% of administrators believe they receive a large amount of feedback from their supervisor throughout their evaluation cycle. Approximately 17% of administrators receive minimal feedback and 2% receive no feedback. Most administrators are receiving at least some feedback (81%).

We do find some variation in the amount of feedback administrators receive on each standard. Administrators more commonly believe they receive minimal or no feedback on the Professional Responsibilities standards versus the Instructional Leadership standards. For example, 19% of administrators believe they receive minimal or no feedback on Instructional Leadership Standard 1 but 26% believe they receive minimal or no feedback on Professional Responsibilities Standard 1. This is the highest for Professional Responsibilities Standard 5 where 30% of administrators believe they receive minimal or no feedback.

**Table 37. Teacher Survey Response to Item: Do You Agree the Feedback You Receive Helps You Achieve Progress On...?**

| | Mean | SD | Strongly Disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| Instructional Practice Standard 1 | 3.79 | 1.07 | 5.0% | 6.2% | 20.9% | 40.6% | 27.3% |
| Instructional Practice Standard 2 | 3.78 | 1.06 | 5.0% | 6.4% | 20.5% | 41.4% | 26.7% |
| Instructional Practice Standard 3 | 3.83 | 1.06 | 4.9% | 5.9% | 19.2% | 41.4% | 28.7% |
| Instructional Practice Standard 4 | 3.78 | 1.07 | 5.1% | 6.6% | 20.9% | 40.4% | 27.1% |
| Instructional Practice Standard 5 | 3.80 | 1.06 | 4.9% | 6.1% | 20.8% | 40.9% | 27.3% |
| Professional Responsibilities Standard 1 | 3.65 | 1.08 | 5.8% | 7.4% | 25.3% | 39.3% | 22.2% |
| Professional Responsibilities Standard 2 | 3.73 | 1.09 | 5.6% | 6.9% | 21.9% | 40.2% | 25.5% |
| Professional Responsibilities Standard 3 | 3.69 | 1.08 | 5.7% | 6.9% | 23.6% | 40.0% | 23.8% |
| Professional Responsibilities Standard 4 | 3.64 | 1.09 | 6.0% | 7.5% | 25.2% | 38.7% | 22.5% |
| Professional Responsibilities Standard 5 | 3.67 | 1.09 | 6.0% | 7.0% | 24.7% | 38.8% | 23.5% |
| Growth due to feedback from Instructional Practice | 3.71 | 1.13 | 6.5% | 8.4% | 19.1% | 40.0% | 26.1% |
| Growth due to feedback from Professional Responsibility | 3.64 | 1.14 | 6.7% | 9.0% | 21.4% | 39.0% | 23.9% |

**Table 38. Teacher Response to Survey Item: To What Extent Do You Agree With the Following Regarding Your SLG...?**

| | Mean | SD | Strongly Disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| SLG based on students' needs | 4.02 | 1.04 | 4.3% | 5.8% | 10.2% | 43.4% | 36.3% |
| Assessment to judge progress Is appropriate | 3.87 | 1.09 | 5.0% | 7.9% | 12.8% | 43.9% | 30.4% |
| SLG is set through a collaborate process with supervisor | 3.72 | 1.15 | 6.5% | 9.8% | 16.2% | 40.2% | 27.3% |

**Table 39. Administrator Response to Survey Item: How Much Feedback Do You Typically Receive...?**

| | Mean | SD | None | Minimal | Some | A large amount |
|---|---|---|---|---|---|---|
| Overall | 3.17 | 0.79 | 2.2% | 17.3% | 41.5% | 39.0% |
| Instructional Leadership Standard 1 | 3.06 | 0.85 | 5.6% | 16.5% | 44.1% | 33.8% |
| Instructional Leadership Standard 2 | 3.10 | 0.88 | 5.9% | 16.2% | 40.3% | 37.6% |
| Instructional Leadership Standard 3 | 3.04 | 0.88 | 5.9% | 18.5% | 40.9% | 34.7% |
| Instructional Leadership Standard 4 | 3.07 | 0.85 | 5.3% | 17.4% | 42.6% | 34.7% |
| Professional Responsibilities Standard 1 | 2.93 | 0.93 | 8.2% | 22.4% | 37.4% | 32.1% |
| Professional Responsibilities Standard 2 | 2.97 | 0.91 | 7.1% | 21.5% | 39.1% | 32.4% |
| Professional Responsibilities Standard 3 | 2.95 | 0.89 | 7.1% | 21.2% | 41.8% | 30.0% |
| Professional Responsibilities Standard 4 | 2.81 | 0.92 | 10.3% | 22.6% | 42.9% | 24.1% |

Table 40 shows the percentage of administrators that strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree with the statement that the feedback they receive throughout their evaluation cycle helps them achieve progress on the various NEPF standards. Here we find similar levels of agreement regardless of the standard. For example, 41% of administrators agree and 28% strongly agree that the

feedback they receive helps them achieve progress on Instructional Leadership Standard 1. On Professional Responsibilities Standard 1, 36% agree and 26% strongly agree. Overall, for the Instructional Leadership domain, 67% of administrators agree or strongly agree and 13% disagree or strongly disagree that the feedback they receive helps them achieve progress on Instructional Leadership. For the Professional Responsibilities domain, 65% of administrators agree or strongly agree and 13% disagree or strongly disagree that the feedback they receive helps them achieve progress on Professional Responsibilities

**Table 40. Administrator Survey Response to Item: Do You Agree the Feedback You Receive Helps You Achieve Progress On...?**

| | Mean | SD | Strongly Disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| Instructional Leadership Standard 1 | 3.80 | 1.08 | 5.2% | 6.4% | 19.9% | 40.7% | 27.8% |
| Instructional Leadership Standard 2 | 3.83 | 1.09 | 5.5% | 5.8% | 19.0% | 39.9% | 29.8% |
| Instructional Leadership Standard 3 | 3.76 | 1.13 | 6.4% | 5.8% | 22.1% | 36.5% | 29.1% |
| Instructional Leadership Standard 4 | 3.76 | 1.10 | 5.8% | 6.4% | 20.9% | 39.3% | 27.6% |
| Professional Responsibilities Standard 1 | 3.69 | 1.12 | 5.9% | 7.8% | 24.2% | 36.0% | 26.1% |
| Professional Responsibilities Standard 2 | 3.72 | 1.10 | 5.6% | 6.5% | 25.2% | 36.0% | 26.7% |
| Professional Responsibilities Standard 3 | 3.70 | 1.10 | 5.9% | 6.5% | 24.5% | 37.6% | 25.5% |
| Professional Responsibilities Standard 4 | 3.60 | 1.11 | 6.2% | 8.1% | 28.0% | 35.4% | 22.4% |
| Growth due to feedback from Instructional Leadership | 3.80 | 1.12 | 5.5% | 7.1% | 20.2% | 36.2% | 31.0% |
| Growth due to feedback from Professional Responsibility | 3.72 | 1.10 | 5.9% | 6.8% | 22.7% | 38.5% | 26.1% |

We also assessed administrator perceptions on their SLG (see Table 41). Most administrators agree or strongly agree (77%) that their SLG is based on students' needs. Similarly, a smaller percentage (67%) agree or strongly agree that the assessment used judge progress towards their SLG is appropriate, and an even smaller percentage agree or strongly agree that the SLG is set through a collaborative process with their supervisor.

**Table 41. Administrator Response to Survey Item: To What Extent Do You Agree With the Following Regarding Your SLG...?**

| | Mean | SD | Strongly Disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| SLG based on students' needs | 3.98 | 0.91 | 3.3% | 1.2% | 18.4% | 48.5% | 28.6% |
| Assessment to judge progress Is appropriate | 3.70 | 0.98 | 3.9% | 7.5% | 21.7% | 48.8% | 18.1% |
| SLG is set through a collaborate process with supervisor | 3.64 | 1.10 | 5.4% | 10.5% | 20.2% | 42.2% | 21.7% |

## ADMINISTRATORS ARE CONFIDENT IN THEIR ABILITY TO PROVIDE QUALITY FEEDBACK ON THE NEPF STANDARDS

We further asked administrators about their confidence in providing quality feedback to teachers on each of the Instructional Practice and Professional Responsibilities standards (see Table 42). Administrators' overwhelmingly agreed that they felt confident in their ability to provide quality feedback and their agreement was fairly consistent across all standards. For example, 91% of administrators agree or strongly agree that they feel confident in their ability to provide quality feedback on Instructional Practice Standard 1. Only 1% disagree or strongly disagree. Additionally, 92% of administrators agree or strongly agree that they feel confident in their ability to provide quality feedback on Professional Responsibilities Standard 1. Only 2% disagree or strongly disagree.

When asked whether they have seen teachers grow in Instructional Practice as a result of the feedback they have given, 88% agreed or strongly agreed, while only 1% disagreed or strongly disagreed. We found only slightly lower agreement percentages for Professional Responsibilities—82% agree or strongly agree.

**Table 42. Administrator Response to Survey Item: I Feel Confident as an Evaluator in My Ability to Provide Quality Feedback on…**

| | Mean | SD | Strongly Disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| Instructional Practice Standard 1 | 4.32 | 0.67 | 0.7% | 0.7% | 5.6% | 52.0% | 41.1% |
| Instructional Practice Standard 2 | 4.31 | 0.66 | 0.3% | 0.7% | 7.0% | 52.0% | 40.1% |
| Instructional Practice Standard 3 | 4.37 | 0.64 | 0.3% | 0.3% | 6.0% | 48.3% | 45.0% |
| Instructional Practice Standard 4 | 4.21 | 0.75 | 0.7% | 2.3% | 8.6% | 52.3% | 36.1% |
| Instructional Practice Standard 5 | 4.33 | 0.67 | 0.3% | 0.7% | 7.3% | 48.7% | 43.0% |
| Professional Responsibilities Standard 1 | 4.25 | 0.69 | 1.0% | 0.7% | 6.3% | 56.8% | 35.2% |
| Professional Responsibilities Standard 2 | 4.26 | 0.66 | 0.3% | 0.7% | 8.0% | 54.5% | 36.5% |
| Professional Responsibilities Standard 3 | 4.26 | 0.66 | 0.7% | 0.3% | 7.3% | 56.1% | 35.5% |
| Professional Responsibilities Standard 4 | 4.18 | 0.71 | 0.7% | 1.3% | 9.6% | 56.5% | 31.9% |
| Professional Responsibilities Standard 5 | 4.19 | 0.75 | 1.0% | 1.7% | 9.6% | 53.2% | 34.6% |
| Seen teacher growth due to feedback given for Instructional Practice | 4.11 | 0.63 | 0.3% | 0.7% | 10.6% | 64.2% | 24.2% |
| Seen teacher growth due to feedback given for Professional Responsibility | 4.01 | 0.71 | 0.7% | 1.7% | 15.9% | 59.8% | 21.9% |

We also asked administrators how confident they felt in collaborating with teachers on their SLG (Table 43). Again, administrators overwhelmingly responded with confidence. Eighty-seven percent agree or strongly agree that they feel confident in their ability to collaborate with teachers to set their SLG, and 82% agree or strongly agree that they feel confident in their ability to collaborate with teachers on the assessments in determining SLG progress.

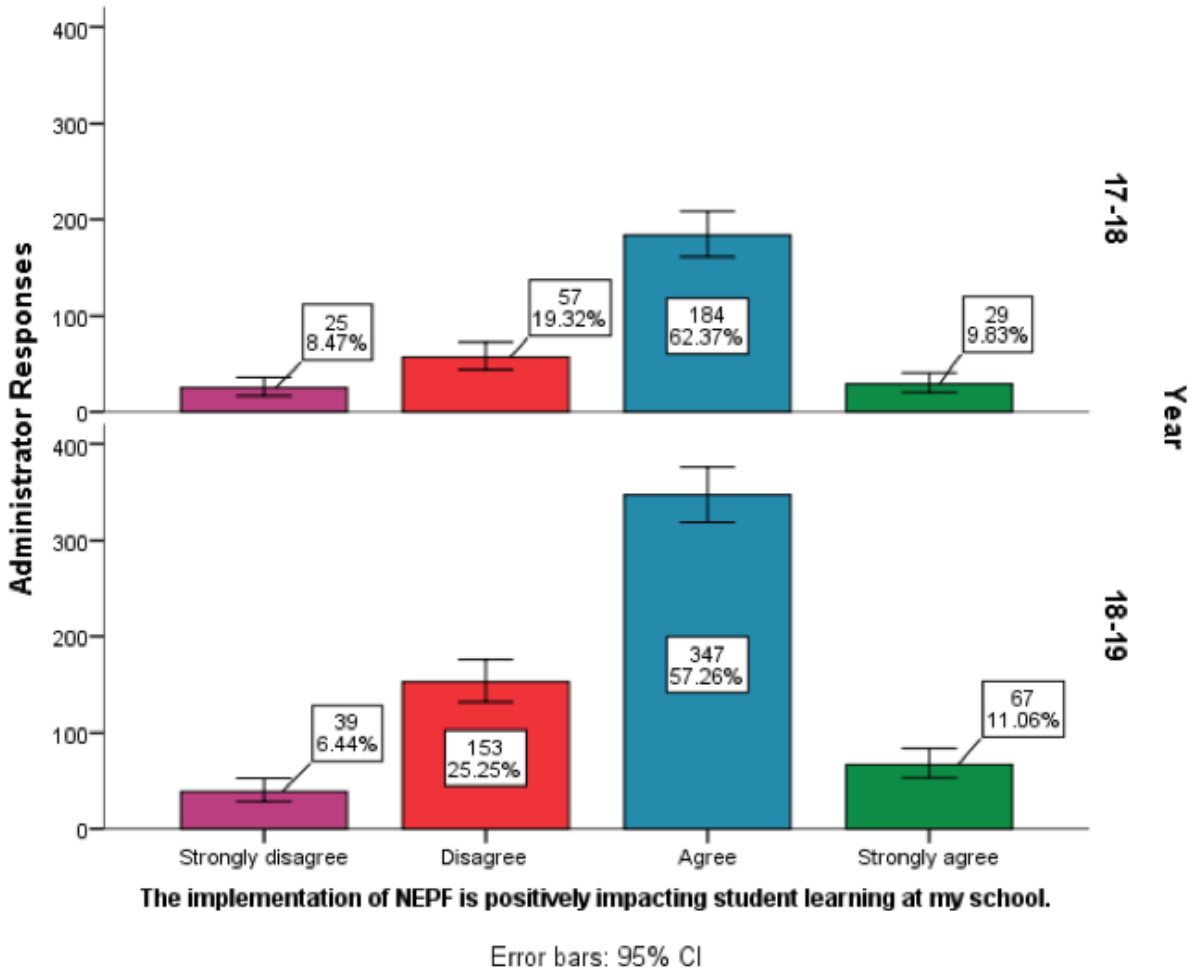
**Table 43. Administrator Response to Survey Item: I Feel Confident in My Ability to Do the Following in Relation to the Student Learning Goal…**

| | Mean | SD | Strongly Disagree | Disagree | Neither agree nor disagree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| Collaborate with teachers to set their SLG | 4.15 | 0.76 | 0.7% | 2.7% | 10.7% | 53.3% | 32.7% |
| Collaborate with teachers on assessments in determining SLG progress | 4.07 | 0.84 | 1.3% | 3.7% | 13.3% | 50.3% | 31.3% |

## MOST ADMINISTRATORS AGREE THAT THE NEPF IS POSITIVELY IMPACTING STUDENTS

Drawing again on the NDE Annual Survey, we show the percent of administrators that believe the NEPF is positively impacting student learning at their school (Figure 29). In 2017-18, 62% of administrators agree and 10% strongly agree that the NEPF is assisting student learning, whereas 19% disagree and 8% strongly disagree. The percentage of administrators that agreed to the same question in 2018-19 is slightly lower—57% agreed and 11% strongly agreed whereas 25% disagreed and 6% strongly disagreed.

**Figure 29. Administrator Response to Survey Item: The Implementation of NEPF is Positively Impacting Student Learning at my School**



# POLICY RECOMMENDATIONS

Required by SB 475 (2019), the purpose of this report was to provide an overview of the NEPF for teachers and administrators, to provide an assessment on the validity of the NEPF as a mechanism for assessing teacher and administrator performance, and to provide an assessment of the impact of the NEPF on teacher and administrator practice and student performance. This evaluation includes analyses of individual teacher and administrator NEPF scores, school-aggregate and district-aggregate NEPF scores, Nevada Report Card data, NCES data, surveys of Nevada administrators and teachers, and information on other state evaluation systems. In this final section, we outline a series of policy recommendations regarding the NEPF.

## NDE SHOULD ENGAGE IN STRATEGIES TO IMPROVE DIFFERENTIATION IN SCORES BETWEEN NEPF DOMAINS

The results of the EFA indicated that the individual standards were not grouping appropriately on the intended domains. In the same way that skills like multiplication and division are related but separable in terms of procedure, the ideal outcome would be for the Practice standards to load strongly onto the Practice domain (and the Professional Responsibilities standards to load strongly onto the Professional Responsibilities domain). These domains should be related, but the rating scores should load more strongly on their respective factors to demonstrate they are being used to evaluate distinct skills associated with good teaching and leading.

A result like this is symptomatic of a larger issue—that is a lack of differentiation between the Instructional Practice and Professional Responsibilities domains by evaluators of teachers and a lack of differentiation between the Instructional Leadership and Professional Responsibilities domains by evaluators of administrators. In short, evaluators are assigning similar ratings across all standards in both domains. Sure, we might expect high-quality teachers or administrators to score highly and similarly on several domains of the NEPF—but the correlations between domains suggest that the practice of assigning uniform ratings is fairly widespread. This is a persistent problem with other teacher and administrator evaluation systems around the country (Lash, Tran, & Huang, 2016; Liu et al., 2019; Grissom, Blissett, & Mitani, 2018)

Why is this a problem? First, if the goal of the evaluation process is to generate feedback that allows educators to assess opportunities for growth and make progress in those areas, then the lack of differentiation between NEPF domains might by symptomatic of a shortage of tailored, specific feedback from evaluators for each NEPF domain. Second, the lack of differentiation between the standards during the evaluation process has implications for the weights assigned to the NEPF domains. From a measurement perspective, our results suggest that the NEPF standards are all measuring the same skill. This means that any weights applied to the Instructional Practice or Professional Responsibilities domains to increase or decrease their influence in the overall effectiveness ratings produces negligible effects in terms of

shaping an educators' final rating. For example, this can be seen in Table 26, which demonstrates the effect of weighting on proportions of teachers categorized as Effective. Only the 2018-19 weighting, which places a 40% weight on the SLG scores has a meaningful effect on final rating membership – shifting slightly more teachers into the Highly Effective category.

NDE should consider the ongoing provision of interrater reliability exercises to improve evaluator understanding of standards – particularly with regard to the various indicators and sources of evidence used to determine a rating described in the NEPF rubrics. Prior research suggests that a significant effort investment in ongoing training is needed to help prevent scoring "drift," and even then it may not be enough (Casabianca, Lockwood, & McCaffrey, 2015). Observers often engage in four different strategies when determining their ratings—reviewing scoring criteria (the recommended strategy), using internal or personal criteria, reasoning from memorable observations, or beginning with an assumed score (Bell et al., 2014). This drift increases even among experienced raters when faced with scoring many observations and time constraints (Bell et a., 2014). One strategy to improve score differentiation is to provide ongoing "think-aloud" activities to evaluators, where a brief lesson clip is provided to the rater. Raters are then asked to think aloud as they rate the clip and their scoring and thinking is analyzed by a master rater based on the scoring criteria. Engaging in these exercises could help raters gain clarity on more difficult or unclear elements of the observation protocol, help them maintain calibration of their scores with the intended ideal, and thereby improve score differentiation between the domains (Park et al., 2014). NDE could consider embedding this activity in an ongoing rater certification program to maintain score calibration over time.

## NDE SHOULD ENGAGE IN STRATEGIES TO IMPROVE THE DISTRIBUTION IN NEPF FINAL SCORES

Our examination of the underlying distributions of the NEPF standard ratings for teachers and administrators indicated that the full range of the evaluation instrument was not being utilized by evaluators. Very few teachers or administrators are categorized as Ineffective or Developing. For example, 96% of the scores for teachers fall between a score of 3 or 4, where the cutoff for Effective is a 2.8. The accumulation of scores within a narrow scoring band creates a ceiling effect that minimizes construct validity and limits the usefulness of the NEPF. Without clear definition of which teachers/administrators are Developing and which are Effective, change within a category becomes more important than change between categories. That is, a stakeholder is left to interpret what it means to move from a lower level of effectiveness (say, a score of 3) to a moderately higher level of effectiveness (say, a score of 3.2) rather than monitoring a clear transition from Developing to Effective. This makes is difficult to determine which teachers are truly growing in meaningful ways, and if the domain indicators are capturing effective instruction.

Importantly, the lack of distribution in NEPF final scores has implications for investigations of the NEPF in driving student learning growth over time. Because almost

all educators score a final rating of “Effective” there is simply not enough variation in educator growth to associate with student growth in achievement. The smaller distance between a lower and moderate effective score than the distance between an Effective and a Highly Effective score limits the ability to determine the influence of an improving educator on student learning. Furthermore, presumably when raters make greater use of the lower rating categories they provide greater feedback and incentives for lower performing teachers to improve their performance and make it more likely that Ineffective teachers exit from teaching should performance not improve (Drake et al., 2016; Grissom & Loeb, 2017). Other high stakes decision-making is also often made from the results of performance evaluation systems, including layoffs in times of economic hardship. With little variation in scoring, decisions regarding layoffs may default to alternative criteria like seniority, which may have equity implications for students (Knight & Strunk, 2016).

To be clear, the lack of a distribution in educators’ final evaluation scores is a problem that most states are still figuring out (Kraft & Gilmour, 2017). The lack of differentiation could be for a few reasons. First, there is a growing body of research that suggests administrators can get bogged down in deciphering standards and logistical aspects of the evaluation process (Darling-Hammond, 2015). In overly complex evaluation systems principals spend a large amount of time on evaluations that do not effect positive change (Marshall, 2013; Marzano & Toth, 2013). Donaldson and Woulfin (2018) find that administrators deviate in their implementation of teacher evaluation systems as they try to manage multiple and competing demands. Marsh et al., (2017) similarly found that administrators were overwhelmed with the implementation of a detailed evaluation rubric.

Second, in the implementation of the evaluation systems, some school districts require greater reporting and evidence requirements for evaluators who score educators at the bottom or top of the distribution. The enhanced paperwork burden associated with scoring educators other than Effective leads to strategic behavior and the cluster of educators at the Effective rating.

No doubt the current observation rubrics used in the NEPF are research-based, and contain important elements that capture effective teaching. This highlights a difficult tension. Rubrics need to be detailed enough to provide meaningful standards and indicators that reflect quality teaching and leading while at the same time being simple enough to be used effectively by evaluators in the face of competing time demands.

NDE could consider a few strategies. First, the ongoing provision of interrater reliability exercises, particularly the think-aloud activities could help improve the scoring distribution by helping raters understand what truly classifies as Level 1, Level 2, or Level 4 performance.

NDE could consider increasing the number of performance levels to create truly inadequate performance levels at the bottom of the scoring range that are rarely used. Although most states currently use final rating systems with only four categories, NDE could consider moving to five categories by splitting the Effective category into two

different performance levels. One could consider a system that rates educators on a five point scale like North Carolina's system as Not Demonstrated, Developing, Proficient, Accomplished, and Distinguished or Alabama's system which uses Beginning, Emerging, Applying, Integrating, and Innovating. Doing so would expand the scale, thereby helping limit the ceiling effect that presently exists in the system.

NDE should also investigate whether school districts, in their implementation of the NEPF, are requiring equal evidence requirements across the rating categories so as to remove the incentive for evaluators (especially those evaluating a high number of educators) to assign ratings of Effective.

NDE should investigate the quality of the feedback being provided to educators. In a system with little variation in scoring, the only way to drive student growth is through quality feedback that engages educators in continuous improvement. While we could not independently assess the quality of the evaluative feedback from the NEPF, our survey results indicated that about 10% to 20% of educators are receiving only minimal feedback and do not feel that any feedback they do receive drives growth. Finding ways to improve the feedback process while not overburdening administrators will be important for improving the relationship between NEPF evaluations and student outcomes. In particular, NDE might consider ensuring that feedback comes early in the school year for educators. Feedback given at the end of the school year is less likely to aid teachers in making changes in their practice (e.g., Wiggins, 2012). Additionally, NDE could consider support for an additional evaluator role for expert teachers trained in the evaluation process (Kraft & Gilmour, 2017), similar to what is done in Peer Assistance and Review Programs. These experts could help reduce the evaluation burden of administrators, especially at large schools, thereby enhancing the focus on quality feedback for improving educator practice. Additionally, NDE could rotate the standards focused on in a given year, thereby limiting the scope of the rubric a rater needs to focus on (Marsh et al., 2017). Finally, when emphasizing feedback, NDE could consider moving away from a single summative rating to focus more on the ratings on the individual standards. In this approach, the question is no longer "how effective is teacher?" but instead "how is a teacher effective?" (Kraft & Gilmour, 2017, p. 243). This approach emphasizes the specific areas where an educator is succeeding and where they might need additional assistance, rather than the single comprehensive rating.

## NDE SHOULD ENGAGE IN A MORE COMPREHENSIVE AND SYSTEMATIC DATA COLLECTION EFFORT OF INDIVIDUAL-LEVEL NEPF DATA

In this study, we primarily used aggregated school-level and district-level data to examine the validity and impact of NEPF. A consideration for future work to better understand NEPF and its effects is the availability and use of individual-level data. We discuss two primary reasons that individual-level data could support extended research on the NEPF.

First, aggregated data suffers from aggregation bias, or the idea that data that is aggregated to higher-level units can mask important information and patterns in the individual units. For instance, knowing that a school's average NEPF score is a 3.2 tells us that, on the whole, teachers in this school are Effective. However, there may be teachers that are Highly Effective and others that are Developing that, when averaged together, balance out to an effective score. To further explore this, suppose half of the teachers score a 3.8 and half of the teachers score a 2.6. In this case, an aggregated average score of 3.2 masks the fact that there are extremes in this school.

Knowing the counts and percentages of teachers scoring within each effectiveness category helps address some aggregation bias, but these patterns can exist even within a single performance category. For instance, half of the teachers in a school may score near the bottom of the Effective range at a 2.81 and the other half at the top of the Effective range at 3.59. All of the teachers have scored in the Effective category, and the overall average is still a 3.2, but, again, there is greater information that one can gather by knowing that there are individuals who are on the cusp of excelling into the Highly Effective category and others who are on the cusp of falling into the Developing category.

This aggregation bias is further compounded when considering growth, which requires differencing NEPF scores at two time points. For instance, suppose that half of the teachers in a school experience growth from a Developing score of 2.7 to an Effective score of 3.0. At the same time, the other half of teachers decrease in effectiveness from an Effective rating of 3.0 to a Developing score of 2.7. Aggregated to the school level, this averages out to zero growth, because growth among some teachers is masked by an equal loss of effectiveness among others. Further, in both the first and second time points, half of the teachers are Developing and half are Effective. At an aggregated level, one might assume that there was no growth and no loss, when in fact, there were both.

These examples are relatively simple and perhaps somewhat exaggerated, but they help illustrate the aggregation bias that school- and district-level data may induce. Further, in these examples, we have only discussed the concerns of aggregation bias as related to school-level NEPF scores for teachers. Note that this also applies to district-level administrators' NEPF scores, schools' and districts' student achievement data, and other aggregated measures.

The second concern we highlight regarding aggregated data is the reduced number of observed units, which reduces our ability to confidently identify relationships. Take, for instance, the results of our analysis examining the relationship between growth in district-aggregate administrator NEPF scores and growth in student achievement (see Table 35). We found that these two measures were positively related, and the relationship was relatively large. However, we did not have enough statistical power to confidently say that the positive relationship was statistically significant and not simply a matter of chance. The larger the number of units that we can include in the analysis, the greater the power to detect statistical significance. In this analysis, with data aggregated

at the district level, we could only observe the growth of 15 districts, but if examined at the individual level, we would be able to observe the growth of over 1,000 administrators in the state. In this case, we might have found statistically significant results in which we could be more confident in our findings.

We note that we believe the work that we have produced in this report is accurate, rigorous, and compelling given the aggregated data, and it still falls under ESSA evidence Tier 2. However, the addition of individual-level data could provide more nuanced results that may lead to more refined policy implications. Consequently, we recommend that NDE engage in a more comprehensive and systematic effort around the collection of individual-level NEPF data. Currently, NDE has individual-level data for teachers and administrators for the 2018-19 school year but the data only has district identifiers such that there is no way to link an educator's 2018-19 score with their scores in subsequent years. We recommend this data collection be continued but in a way that can identify individual educator growth over time, while also protecting the privacy of the educators. This can be done by assigning a unique state identifier for each educator that is consistent over time but is not stored with other personally identifiable information. Several other states have already engaged in this type of work.

## NDE SHOULD IMPROVE ITS CURRENT NEPF REPORTING PROCESS

Relatedly, notwithstanding the effort to collect better individual-level data, the current process of collecting NEPF information from school districts should be streamlined, made consistent across time, and made easier for both teachers and administrators. Currently, NEPF data is collected from school districts using an Excel spreadsheet. However, we uncovered several instances where Excel formulas were accidentally erased or misapplied when the data was reported to NDE. In addition, we rectified some copy-and-paste errors where data was input into incorrect rows or columns. Again, these errors were in the reporting to NDE after evaluation cycles had concluded and not in the actual calculation of educators' evaluation scores within districts.

Ultimately, in order to allow for quick and systematic data reporting, we recommend moving away from Excel spreadsheet reporting and investing in a more comprehensive data management tool—one that can handle individual-level data inputs from school districts and allow for streamlined reporting to NDE. We understand the hesitancy of school districts to adopt one more data management platform, but the accuracy of personnel performance data is important for understanding the improvement of Nevada's educators over time.

Collecting these data will help NDE engage in a more comprehensive monitoring effort of the NEPF. In particular, NDE will be able to tell whether the distribution of scoring is improving over time and whether educators are improving their NEPF scores, particularly in conjunction with the recommended improvements in training over time.

# REFERENCES


Bell, C., Qi, Y., Croft, A., Leusner, D., McCaffrey, D., Gitomer, D., & Pianta, R. (2014). Improving observational score quality: Challenges in observer thinking. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 50–97). San Francisco, CA: Jossey-Bass.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311-337.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378-387.

Darling-Hammond, L. (2015). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press.

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going “rogue”: How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531-556.

Fitzpatrick, R., & Salazar, P. (2012). Teachers and Leaders Council: Summary of anticipated final recommendations and implementation considerations. Retrieved from https://ccea-nv.org/images/stories/pdfs/TLC_Recommendations_Summary_11_14_12_pdf.pdf.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224-242.

Grissom, J. A., Blissett, R. S., & Mitani, H. (2018). Evaluating school principals: Supervisor ratings of principal practice and principal job performance. *Educational Evaluation and Policy Analysis*, *40*(3), 446-472.

Grissom, J. A., & Loeb, S. (2017). Assessing principals’ assessments: Subjective evaluations of teacher effectiveness in low-and high-stakes environments. *Education Finance and Policy*, *12*(3), 369-395.

Hoffman, R. R., Klein, G., & Militello, L., Lipshitz, R., & Schraagen, J. M. (2017). *Naturalistic decision making and macrocognition*. CRC Press.

Knight, D. S., & Strunk, K. O. (2016). Who bears the costs of district funding cuts? Reducing inequality in the distribution of teacher layoffs. *Educational Researcher*, *45*(7), 395-406.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. Educational researcher, 46(5), 234-249.

Lash, A., Tran, L., & Huang, M. (2016). Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system. REL 2016-135. *Regional Educational Laboratory West*.

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61-95.

Locke E. A., & Latham, G. P. (Eds.). (2013). *New developments in goal setting and task performance*. Routledge.

Marianno, B. D. (2015). Teachers' unions on the defensive?: How recent collective bargaining laws reformed the rights of teachers. *Journal of School Choice*, *9*(4), 551-577.

Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating teachers in the Big Easy: How organizational context shapes policy responses in New Orleans. *Educational Evaluation and Policy Analysis*, *39*(4), 539-570.

Marshall, K. (2013) *Rethinking teacher supervision and evaluation*. Jossey-Bass, San Francisco USA.

Marzano, R.J. and Toth, M.D. (2013) Teacher evaluation that makes a difference: a new model for teacher growth and student achievement. ASCD Alexandria, VA.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice*, 1*4*(4), 5-8.

Nevada Teachers and Leaders Council. (2013). Nevada Educator Performance Framework teacher and administrator evaluation and support models: White paper.

Nevada Department of Education. (2014). Teacher and administrator protocols/tools for training and validation purposes.

Nevada Department of Education. (2015). Teacher and administrator protocols/tools for 2015-2016 implementation.

Nevada Department of Education. (2016). Teacher and administrator protocols/tools for 2016-2017 implementation.

Nevada Department of Education. (2019). NEPF monitoring for continuous improvement guidance document 2018-2019.

Nevada Department of Education. (2020). 2019-2020 principal supervisor, school administrator, and teacher protocols.

Park, Y. S., Chen, J., & Holtzman, S. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 383–414). San Francisco, CA: Jossey-Bass.

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2017). Linking US School District Test Score Distributions to a Common Scale. CEPA Working Paper No. 16-09. *Stanford Center for Education Policy Analysis*.

Reardon, S., Shear, B., Castellano, K., & Ho, A. (2016). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. CEPA Working Paper No. 16-02. *Stanford Center for Education Policy Analysis.*

Shear, B. R., & Reardon, S. F. (2019). HETOP: Stata module for estimating heteroskedastic ordered probit models with ordered frequency data. Version 3.0.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*. Retrieved from https://files.eric.ed.gov/fulltext/ED515656.pdf.

WestEd. (2015). A study of the Nevada Educator Performance Framework (NEPF): Year two final report.

Wiggins, G. (2012). Seven keys to effective feedback. *Educational leadership, 70*(1), 10-16.

# APPENDICES

## APPENDIX A: CREA TEACHER SURVEY

1. The following questions ask about your perceptions of the feedback you receive throughout your own evaluation cycle (i.e. your evaluation process for a given school year) as a **teacher.** Throughout your evaluation cycle, how much feedback do you typically receive from your supervisor? Do you receive...

**Scale:** No feedback (1), A minimal amount of feedback (2), Some feedback (3), A large amount of feedback (4)

2. How much feedback do you typically receive from your supervisor on the following **NEPF Instructional Practice** standards?

**Scale:** No feedback (1), A minimal amount of feedback (2), Some feedback (3), A large amount of feedback (4)

- New learning is connected to prior learning and experience.
- Learning tasks have high cognitive demand for diverse learners.
- Students engage in meaning-making through discourse and other strategies.
- Students engage in metacognitive activity to increase understanding of and responsibility for their own learning.
- Assessment is integrated into instruction.

3. How much feedback do you typically receive from your supervisor on the following **NEPF Professional Responsibilities** standards?

**Scale:** No feedback (1), A minimal amount of feedback (2), Some feedback (3), A large amount of feedback (4)

- Commitment to the school community.
- Reflection on professional growth and practice.
- Professional obligations.
- Family engagement.
- Student perception.

4. To what extent do you agree the feedback you receive from your supervisor is helping you **achieve progress** on the following **NEPF Instructional Practice** standards?

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

- New learning is connected to prior learning and experience.
- Learning tasks have high cognitive demand for diverse learners.
- Students engage in meaning-making through discourse and other strategies.
- Students engage in metacognitive activity to increase understanding of and responsibility for their own learning.
- Assessment is integrated into instruction.

5. I have experienced growth in my **Instructional Practice** due to the feedback I have received from my supervisor.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

6. To what extent do you agree the feedback you receive from your supervisor is helping you **achieve progress** on the following **NEPF Professional Responsibilities** standards?

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

- Commitment to the school community.
- Reflection on professional growth and practice.
- Professional obligations.
- Family engagement.
- Student perception.

7. I have experienced growth in my **Professional Responsibility** due to the feedback I have received from my supervisor.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

8. The following questions ask about your perceptions of your **Student Learning Goal** during a typical evaluation cycle as a **teacher.** To what extent do you agree with the following regarding your **Student Learning Goal** during a typical evaluation cycle:

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

- My Student Learning Goal is based on my students' needs.
- The assessment(s) used to judge progress on my Student Learning Goal are appropriate.
- My Student Learning Goal is set through a collaborative process with my supervisor.

9. To what extent do you agree with the following statement: The final score obtained from my NEPF evaluation cycle is a valid measure of my performance.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

10. Is there anything else you would like to expand on or share regarding the NEPF for **teachers**? (Open-ended)

11. What is your current school district?

Carson
Churchill
Clark
Douglas
Elko
Esmeralda
Eureka
Humboldt
Lander
Lincoln
Lyon
Mineral
Nye
Pershing
Storey
Washoe
White Pine
State Public Charter School Authority
Other

12. At which type of school(s) are you currently a teacher? (Check all that apply)

Elementary School
Middle School
High School
Combined School (K-12)
Other

13. How many years of work experience do you have (including this year)?

**Scale:** This is my first year (1), 2 years (2), 3-5 years (3), 6-10 years (4), 11-15 years (5), 16-20 years (6), More than 20 years (7)

- Year(s) working as a teacher.
- Year(s) working as an educator in total.

# APPENDIX B: CREA ADMINISTRATOR SURVEY

1. The following questions ask about your perceptions of the feedback you receive throughout your own evaluation cycle (i.e. your evaluation process for a given school year) as a **building administrator.** Throughout your evaluation cycle, how much feedback do you typically receive from your supervisor? Do you receive...

**Scale:** No feedback (1), A minimal amount of feedback (2), Some feedback (3), A large amount of feedback (4)

2. How much feedback do you typically receive from your supervisor on the following **NEPF Instructional Leadership** standards?

**Scale:** No feedback (1), A minimal amount of feedback (2), Some feedback (3), A large amount of feedback (4)

- Creating and sustaining a focus on learning.
- Creating and sustaining a culture of continuous improvement.
- Creating and sustaining productive relationships (e.g. creating a caring environment, providing opportunities for productive discourse, enabling collaboration).
- Creating and sustaining structures (e.g. implementing systems to align curriculum, instruction, and assessments to standards).

3. How much feedback do you typically receive from your supervisor on the following **NEPF Professional Responsibilities** standards?

**Scale:** No feedback (1), A minimal amount of feedback (2), Some feedback (3), A large amount of feedback (4)

- Manages human capital (e.g. identify, recognize, support, and retain teachers, support the development of teachers).
- Self-reflection and professional growth.
- Professional obligations.
- Family engagement.

4. To what extent do you agree the feedback you receive from your supervisor is helping you **achieve progress** on the following **NEPF Instructional Leadership** standards?

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

- Creating and sustaining a focus on learning.

- Creating and sustaining a culture of continuous improvement.
- Creating and sustaining productive relationships (e.g. creating a caring environment, providing opportunities for productive discourse, enabling collaboration).
- Creating and sustaining structures (e.g. implementing systems to align curriculum, instruction, and assessments to standards).

5. I have experienced growth in my **Instructional Leadership** due to the feedback I have received from my supervisor.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

6. To what extent do you agree the feedback you receive from your supervisor is helping you **achieve progress** on the following **NEPF Professional Responsibilities** standards?

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

- Manages human capital (e.g. identify, recognize, support, and retain teachers, support the development of teachers).
- Self-reflection and professional growth.
- Professional obligations.
- Family engagement.

7. I have experienced growth in my **Professional Responsibility** due to the feedback I have received from my supervisor.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

8. The following questions ask about your perceptions of your **Student Learning Goal** during a typical evaluation cycle as a **building administrator.** To what extent do you agree with the following regarding your **Student Learning Goal** during a typical evaluation cycle:

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

- My Student Learning Goal is based on my students' needs.
- The assessment(s) used to judge progress on my Student Learning Goal are appropriate.
- My Student Learning Goal is set through a collaborative process with my supervisor.

9. To what extent do you agree with the following statement: The final score obtained from my NEPF evaluation cycle is a valid measure of my performance.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

10. Is there anything else you would like to expand on or share regarding the NEPF for **building administrators?** (Open-ended)

11. The following questions ask about your perceptions of the NEPF evaluation cycle for **teachers** as it pertains to your work. During a typical school year, do you evaluate **teachers** using...?

    - the Nevada Educator Performance Framework (NEPF)
    - a state-approved alternative
    - I do not evaluate teachers in my current position

12. How many educators do you currently evaluate? (Open-ended)

13. I feel confident as an evaluator in my ability to provide quality feedback to teachers on the following **NEPF Instructional Practice** standards:

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

    - New learning is connected to prior learning and experience.
    - Learning tasks have high cognitive demand for diverse learners.
    - Students engage in meaning-making through discourse and other strategies.
    - Students engage in metacognitive activity to increase understanding of and responsibility for their own learning.
    - Assessment is integrated into instruction.

14. I have seen growth in my teachers' **Instructional Practice** due to the feedback I have provided.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

15. I feel confident as an evaluator in my ability to provide quality feedback to teachers on the following **NEPF Professional Responsibilities** standards:

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

    - Commitment to the school community.

- Reflection on professional growth and practice.
- Professional obligations.
- Family engagement.
- Student perception.

16. I have seen growth in my teachers' **Professional Responsibility** due to the feedback I have provided.

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

17. I feel confident as an evaluator in my ability to do the following in relation to the **Student Learning Goal**:

**Scale:** Strongly Disagree (1), Disagree (2), Neither agree nor disagree (3), Agree (4), Strongly Agree (5)

- To collaborate with my teachers to set their Student Learning Goal.
- To collaborate with my teachers on the appropriate assessments to use in determining progress on their Student Learning Goal.

18. Is there anything else you would like to expand on or share regarding the NEPF for **teachers?** (Open-ended)

19. What is your current school district?

Carson
Churchill
Clark
Douglas
Elko
Esmeralda
Eureka
Humboldt
Lander
Lincoln
Lyon
Mineral
Nye
Pershing
Storey
Washoe
White Pine
State Public Charter School Authority
Other

20. At which type of school(s) are you currently an administrator? (Check all that apply)

Elementary School
Middle School
High School
Combined School (K-12)
Other

21. How many years of work experience do you have (including this year)?

**Scale:** This is my first year (1), 2 years (2), 3-5 years (3), 6-10 years (4), 11-15 years (5), 16-20 years (6), More than 20 years (7)

- Year(s) working as a building administrator.
- Year(s) working as an educator in total.