**STATE OF NEW MEXICO**
**COUNTY OF SANTA FE**
**FIRST JUDICIAL DISTRICT COURT**

**STATE OF NEW MEXICO EX REL.**
**THE HONORABLE MIMI STEWART,**
**THE HONORALBE SHERYL WILLIAMS STAPLETON,**
**THE HONORABLE HOWIE C. MORALES,**
**THE HONORALBE LINDA M. LOPEZ,**
**THE HONORABLE WILLIAM P. SOULES,**
**AMERICAN FEDERATION OF TEACHERS—**
**NEW MEXICO, ALBUQUERQUE FEDERATION OF**
**TEACHERS, JOLENE BEGAY, DANA ALLEN,**
**NAOMI DANIEL, RON LAVANDOSKI, TRACEY**
**BRUMLIK, CRYSTAL HERRERA, and**
**ALLISON HAWKS,**

**Plaintiffs,**

**v.** **No. D-101-CV-2015-00409**

**NEW MEXICO PUBLIC EDUCATION**
**DEPARTMENT and SECRETARY-DESIGNEE**
**HANNA SKANDERA in her official capacity,**

**Defendants.**

**AFFIDAVIT OF AUDREY AMREIN-BEARDSLEY**

STATE OF ARIZONA )
) ss
COUNTY OF MARICOPA )

I, Audrey Amrein-Beardsley, being first duly sworn, depose and say as follows:

I am a professor of Educational Policy at Arizona State University. A copy of my CV is attached as *Exhibit A.* Among other things, I have been recognized as an expert in testing, assessment, and growth and value-added modeling. I have published over 60 articles and presented approximately 150 scholarly papers on these topics, and I have published two books about growth and value-added modeling: "Rethinking Value-Added Models in Education Critical Perspectives on Tests and Assessment-Based Accountability" and "Student Growth

Measures in Policy and Practice: Intended and Unintended Consequences of High-Stakes Teacher Evaluations."

Within this statement I express my professional opinion about the teacher evaluation model used in the state of New Mexico and provide evidence that the system is both arbitrary and capricious, and that: (1) plaintiffs will continue to prevail on the merits after trial, (2) plaintiffs will suffer irreparable harm, professionally, if the preliminary injunction is not upheld until evidence of validity, reliability, and a lack of bias is established by the defendant (i.e., the New Mexico Public Education Department [NMPED]), (3) the threatened injury to plaintiffs outweighs the injury that upholding the injunction until such evidence is provided, and (4) the injunction will continue to *not* be adverse to the public interest.

In order to support this professional opinion I, with the support of my doctoral student – Tray J. Geiger, PhD candidate, see a copy of his CV attached as *Exhibit B* – examined the data shared by the defendant and analyzed for purposes of this affidavit. These data files were titled as follows (see also which tabs from within each data file that were used):

- DEF 010546 SY1314_RFP_1_5.xlsx - Contained PerformancebyTeacher (used), DemographicsByTeacherCourseGrp (unused)
- DEF 010547 SY1415_RFP_1_5.xlsx - Contained SY1415Performance (used), StudentDemogsbyTeacherCourseGrp (unused)
- DEF 010548 SY1516_RFP_1_5.xlsx - Contained SY1516Performance (used), StudDemograhibyTeacherCG (unused)

In sum, these data were used to conduct the analyses pertinent to this case and, more specifically, necessary to assess (1) the defendant's teacher evaluation system (current at the time that the preliminary injunction was filed) and (2) the system's teacher-level value-added model

(VAM) scores as a critical component of said system on primarily three key educational and psychological measurement indicators: reliability, validity, and bias (or a lack thereof).

## Policy Landscape

The first part of the twenty-first century brought several iterations of educational accountability policies, fixed upon educational reform by increasingly holding students, teachers, schools, districts, and states responsible for their measureable impacts on student achievement over time. Subsequently, and as per the federal government's expanded role in public education at the time (e.g., President Obama's Race to the Top Competition, 2011; No Child Left Behind [NCLB], 2001; NCLB waivers excusing states from not meeting NCLB's prior 100% student proficiency by 2014 goals should they adopt stronger accountability measures for teachers, U.S. Department of Education, 2010, 2014; see also Duncan, 2011), compliant states and districts spent much effort (and much expense) to adopt and implement such policy-mandated metrics (e.g., as largely based on growth models and VAMs).

In the simplest of terms, statisticians use growth models and VAMs (hereafter referred to more generally as VAMs, and as specific to the state of New Mexico teachers' value-added scores [VASs]) to measure the predicted then actual "value" a teacher "adds" to (or detracts from) student growth in achievement from one year to the next. Modelers typically do this by measuring student growth over time on large-scale standardized tests (e.g., as mandated by NCLB, 2001), and aggregating this growth at the teacher-level, while statistically controlling for confounding variables such as students' prior test scores and other student-level (e.g., free-and-reduced lunch [FRL] eligibility, English language learner [ELL] status, special education [SPED] classification) and school-level variables (e.g., class size, school resources), although control variables vary by model.

However, much debate continues to exist about the actual, intended successes such policies have had on improving educational outcomes, and the extent to which pernicious, unintended consequences have resulted instead (see, for example, American Educational Research Association [AERA] Council; American Statistical Association [ASA], 2014; Amrein-Beardsley, 2014; Ballou & Springer, 2015; Blazar, Litke, & Barmore, 2016; Holloway-Libell, 2015; Kane, 2017; National Association of Secondary School Principals [NASSP], n.d.; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Polikoff & Porter, 2014; Rothstein, 2009, 2010, 2014; Rothstein & Mathis, 2013; Schochet & Chiang, 2013; Yeh, 2013).

Because of these and many other reasons (e.g., students spending too much instructional time testing), the federal government reduced such centralized policy mandates with its Every Student Succeeds Act (ESSA, 2016), allowing for greater state/local control over such policies. In April of 2016, for example, all 50 states submitted their ESSA plans to the federal government, with most states diminishing the accountability plans they had recently put into place. Fewer than half still intend to measure student achievement as based on proficiency, the majority significantly lessened or removed the consequences attached to schools', teachers', or students' test-based performance, and the majority decreased or withdrew prior efforts to hold teachers accountable for their effectiveness using their students' achievement as measured by test scores (e.g., via VAMs; Excel*in*Ed, 2017).

In the state of New Mexico, via its ESSA plan, teachers' VAS estimates were to carry less weight than prior, with state leaders adjusting the weight of teachers' VASs from 50% to 35%. Inversely, the weight teachers' observation-based estimates (see also forthcoming) were to carry was increased by the same proportion (i.e., from 25% to 40%). The state also made other changes, for example, allowing teachers to be absent six versus three times per academic year,

which is twice as many days absent as was written into the state's prior teacher evaluation plans, and which was also to now count at 5%, alongside student survey data at 5%, and other observational variables taken from the state's observational system's Domains 1 and 4 (i.e., Planning, Preparation, and Professionalism [PPP]) at 15%. All of these latter variables (i.e., not including teachers VASs and observational scores from Domains 2 and 3) were to total the other 25% of a New Mexico teacher's teacher evaluation score[1] (NMPED, n.d.).

It is important to note here that while these are the weights proposed, these are also the weights currently being used throughout the state. These weights are also different than those being used across the state when this lawsuit was initially heard in court (see forthcoming). Hence, the data analyzed herein do not take any weights into consideration (as moving as well as still arguably arbitrary; see also Kane & Staiger, 2012). Accordingly, only the educational metrics themselves (i.e. teachers' VASs, observational scores from Domains 2 and 3, observational scores from Domains 1 and 4, and student survey scores) were used to examine the actual and unweighted data as shared by the defendants for purposes of these analyses and this lawsuit.

## Purpose

When in December of 2015 state District Judge David K. Thomson, who continues to preside over this lawsuit in New Mexico, granted a preliminary injunction preventing consequences from being attached to the state's teacher evaluation data, Judge Thomson ruled that the state could proceed with "developing" and "improving" its teacher evaluation system, but the state was *not* to make any consequential decisions about New Mexico's teachers using

[1] If a teacher has no VAS data, then observation scores are to count 50%, PPP is to count 40%, and the other two variables above are to remain at 5% each. Because researchers only included teachers in the analyses herein if they had VAS data, researchers will not comment further about the percentages for teachers without such scores.

the data the state collected until the state (or others external to the state) could evidence to the court that the system was reliable, valid, unbiased, fair, uniform, and the like. Hence, within this affidavit are the findings as per subsequent analyses of the aforementioned state data regarding the state's teacher evaluation and VAS-based system's levels of reliability, validity, and bias (or lack thereof), with issues of fairness and uniformity also duly noted.

Accordingly, the purpose of this examination was to assess system-level reliability, validity, and bias using (1) teacher-level VASs as calculated by the state, classroom observation scores derived via a modified version of Charlotte Danielson's Framework for Teaching (Danielson Group, n.d.), with (2) scores from modified Domains 2 and 3 weighted differentially than (3) scores from modified Domains 1 and 4 (i.e., PPP), and (4) student survey scores. These four measures were used to determine to what extent the above-mentioned measures, with primary emphases on teachers' VASs, were reliable, valid, and unbiased, with these three issues also aptly defined and framed using the *Standards for Educational and Psychological Testing,* hereafter referred to as the *Standards* (AERA, American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

**Reliability**

As per the *Standards* (AERA et al., 2014), reliability is defined as the degree to which test-based scores "are consistent over repeated applications of a measurement procedure [e.g., a VAM] and hence and inferred to be dependable and consistent" (p. 222-223) for the individuals (e.g., teachers) to whom the scores pertain. In terms of VAMs, reliability is quantifiable and numerically observed when VAM estimates of teacher effectiveness are consistent from one year to the next, regardless of the types of students and subject areas teachers teach. This is typically captured using "standard errors, reliability coefficients per se, generalizability coefficients,

error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency" (AERA et al., 2014, p. 33) that help to situate and make explicit VAM estimates and their (sometimes sizeable) errors. This is also done to make transparent the errors that come along with VAM estimates, to help situate and inform the validity of the inferences that result, which is also important given that reliability is a precondition to validity. Without adequate reliability, in other words, never can validity be obtained (AERA et al., 2014; Brennan, 2006, 2013; Crocker & Algina, 1986; Kane, 2006, 2013; Lord & Novick, 1968; Messick, 1975, 1980, 1995).

**Validity**

As per the *Standards* (AERA et al., 2014), validity "refers to the degree to which evidence and theory support the interpretations of test scores for [the] proposed uses of tests" (p. 11). Likewise, "[v]alidity is a unitary concept" and is measured by "the degree to which all the accumulated evidence supports the intended interpretation of [the test-based] scores for [their] proposed use[s]" (p. 14). Accordingly, when establishing evidence of validity, one must be able to support with quantitative and/or qualitative evidence that accurate inferences can be drawn from the data derived (Brennan, 2006, 2013; Kane, 2006, 2013; Messick, 1980, 1989).

Following suit, VAM researchers have delved into searching for evidence of many sub-area evidences of validity, including but not limited to: (1) content-related evidence of validity in that "what is to be validated is not the test…but the inferences derived from [the] test scores" (Messick, 1989, p. 5); (2) concurrent-related evidence of validity or "the degree of relationship between the test scores and [other] criterion scores" taken at the same time (Messick, 1989, p. 7; see also Messick, 1980); and (3) consequence-related evidence of validity, which is"[t]he only form of validity evidence [typically] bypassed or neglected in these traditional formulations," but

that "bears on the [intended and unintended] social consequences of test interpretation and use" (Messick, 1980, p. 8; see also Kane, 2013). It should also be noted that while all of these evidences of validity help to support construct-related evidence of validity, in this area of research most researchers rely on gathering concurrent-related evidence of validity, as was done herein.

When examining concurrent-related evidence of validity it is necessary to assess, for example, whether teachers who post large and small value-added gains over time are those deemed effective, and inversely, using other independent quantitative and qualitative measures of teacher effectiveness (e.g., supervisors' observational scores, student survey results). If all measures line up (i.e., correlate) and essentially validate one another, confidence in them as independent measures of the same construct subsequently increases (Messick, 1980, 1989; see also Chin & Goldhaber, 2015; Hill, Kapitula, & Umlan, 2011). If all indicators do not correlate or correspond, however, something may be wrong with either, both, or all indicators being used for the purpose of measuring or capturing the construct defined, which in this case is teacher effectiveness.

**Bias**

As per the *Standards* (AERA et al., 2014), bias pertains to validity or, rather, the validity of the inferences to be drawn from such test-based scores. Bias is defined as the "construct underrepresentation of construct-irrelevant [variant (CIV)] components of test scores that differentially affect the performance of different groups of test takers and consequently the…validity of interpretations and uses of their test scores" (p. 216). Also known as systematic error given "[t]he systematic over- or under-prediction of criterion performance" (p. 222), biased estimates in the cases of VAMs are observed when performance varies for "people [i.e.,

teachers] belonging to [or in this case teaching]…groups [of students] differentiated by characteristics not relevant to the criterion performance" (p. 222) of measurement (e.g., impoverished, gifted, SPED, ELL students). In the cases of VAMs, it is important to note that most VAM modelers argue that the sophisticated statistical controls build into their systems block such bias, so as to "level the playing field" to invoke a familiar idiom; hence, should estimates (dis)favor certain groups of teachers in this particular case, this serves as an indicator that such controls are not working as intended, or rather not working well enough to adequately block such bias (Ballou, Sanders, Wright, 2004; Koedel, Mihaly, & Rockoff, 2015; Lavery, Holloway, Amrein-Beardsley, Pivovarova, & Hahs-Vaughn, in progress; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Michelmore, & Dynarski, 2016; Newton et al., 2010; Rothstein, 2009, 2010, 2014; Sanders, Wright, & Rivers, 2009; Tekwe et al., 2004; Wright, White, Sanders, & Rivers, 2010).

Accordingly, using teacher effectiveness data provided by NMPED, investigated herein were issues specific to NMPED's NMTEACH teacher evaluation system, more specifically described as follows: (1) reliability, assessed by examining teachers' VASs, observation scores (i.e., Domains 2 and 3), PPP scores (i.e., Domains 1 and 4), and student survey scores, over time and given the extent to which each indicator fluctuated from year-to-year; (2) validity (i.e., concurrent-related evidence of validity) assessed by investigating the statistical (cor)relationships between teachers' VASs, observation scores, PPP scores, and student survey scores; and (3) bias, defined by the extent to which teachers' VASs, observation scores, PPP scores, and student survey scores might be influenced by the types of students non-randomly assigned into teachers' classrooms, also given the statistical controls typically built into VAMs to block the bias

typically caused by typical biasing factors (e.g., student race, income level, English fluency; teacher gender, teacher years of experience, subject taught [ELL, SPED, gifted]).

## Methods

### Data

The following data were used for all analyses: (1) VASs – the total number of points earned, per teacher, per year; (2) Observation scores – the proportion of points earned divided by possible points, per teacher, per year, on Domains 2 and 3; (3) PPP scores – the proportion of points earned divided by possible points, per teacher, per year, on Domains 1 and 4; and (4) Student survey scores – the proportion of points earned divided by possible points, per teacher, per year. Also used were a variety of teacher-level (e.g., gender, years of experience) and school-level (e.g., proportion of low-income students, proportion of non-white students) demographics to help compare groups of teachers (e.g., for analyses of bias; see Appendix A for a full list of the demographics used).

### Population and Subsample

The initial data files provided by the NMPED included 26,966 unique teachers across three academic years: 2013-2014 (Year 1), 2014-2015 (Year 2), and 2015-2016 (Year 3). Of the 26,966 educators, 97.0% (n = 26,160) were certified teachers. For purposes of this analysis, this sample was restricted to include certified teachers who had VAS and observation data for all three academic years, as is standard and recommended practice (Brophy, 1973; Cody, McFarland, Moore, & Preston, 2010; Glazerman & Potamites, 2011; Goldschmidt, Choi, & Beaudoin, 2012; Harris, 2011; Ishii & Rivkin, 2009; Sanders as cited in Gabriel & Lester, 2013), also taking into consideration that these two measures carry the majority of weight in the

NMTEACH system (i.e., these two measures are of most evaluative value as proportionally weighted even though, again, weights were not used in these analyses).

This resulted in the final sample including 7,777 teachers, which was 28.8% of the full dataset or 29.7% of all certified teachers. That the final sample included approximately 30% of the teachers included in the main data files is also important to note as directly related to issues of fairness (and uniformity), which were also of concern to the court in this particular case.

As per the *Standards* (AERA et al., 2014), fairness is defined as the impartiality of "test score interpretations for intended use(s) for individuals from *all* [emphasis added] relevant subgroups" (p. 219). Accordingly, issues of fairness (and uniformity) arise when a test or test-based inference or use impacts some more than others in unfair or prejudiced, and often consequential ways.

The main issue here is that, currently, VAM-based estimates can be produced for approximately 30-40% of all teachers across the nation (Baker, Oluwole, & Green, 2013; Gabriel & Lester, 2013; Harris, 2011, Harris, & Herrington, 2015; Jiang, Sporte, & Luppescu, 2015; Papay, 2012). Related, the statistic reported above (i.e., 30%) indicates that, relatively, New Mexico is not as "fair" or "uniform" as states using similar types of teacher evaluation systems. Rather, there seems to be 70% or so of New Mexico teachers who are accountability ineligible, and this proportion held constant across the three years of data analyzed herein. This subsequently makes this 70% or so of New Mexico teachers immune from the state's VAS-based systems (see also Footnote 1 above regarding teachers without VASs) which is important, again, knowing that when these data are used to make consequential decisions, issues with fairness and uniformity become even more important, given accountability-eligible teachers are also those more likely to realize the negative or reap the positive consequences attached to VAS-based

estimates. Notwithstanding, this approximately 70% of teachers were also those who were justifiably excluded from these analyses.

**Analyses**

First, in terms of reliability, the distribution of VAS estimates per teacher over time was investigated, as well as the correlations among scores over the same period of time. For comparative purposes, also calculated were the correlations among teachers' observation scores, PPP scores, and student survey scores over the same period of time. Chi-square ($\chi^2$) tests were conducted to determine if teachers' overall summative ratings along with the four variables' score distributions and degrees of score variation significantly differed from year-to-year. To do this, quintiles of each measure's scores were calculated and then used to determine what percentages of teachers moved across and among quintiles from one year to the next. Also examined was whether there were any teacher- or school-level differences in stability over time.

Second, in terms of concurrent-related evidence of validity, the (cor)relationships between teachers' VAS estimates, observation scores, PPP scores, and student survey scores were investigated. The correlations among all four variables via the calculation of Pearson's *r* correlation coefficient between each pair of variables for each year were analyzed. To determine if the differences between bivariate correlations from year-to-year were significant, a Fisher's *Z* test[2] was used (Dunn & Clark, 1969, 1971).

Third, the issue of bias (or the lack thereof) was investigated by comparing the scores of all four measures per multiple teacher- and school-level subgroups, as also described above (see also Appendix A for a full list of the demographics used). More specifically, descriptive statistics

[2] SPSS 23.0 (IBM Corp., 2015) does not contain a mechanism to conduct Fisher's *z* test. Therefore, researchers calculated the appropriate *z* values and corresponding *p* values by inputting data into the Simple Interactive Statistical Analysis (SISA) calculator, found at http://www.quantitativeskills.com/sisa/statistics/correl.htm

were calculated for teachers' scores by teacher- and school-level subgroups, with statistically significant differences analyzed using *t* tests or fixed effects analyses of variance (ANOVA).

## Results

### Descriptive Statistics

Overall, teachers' summative effectiveness ratings (i.e., teachers' final evaluation scores ranging between one as "Ineffective" and five as "Exemplary") were slightly below the scale and sample median of 3.0 or "Effective". Teachers' mean scores were 2.86, 2.88, and 2.94 for Year 1, Year 2, and Year 3, respectively. The distributions of these summative scores were normal or, rather, yielded a bell curve (see Figure 1). This is something about which the state apparently takes pride (see, for example, NMPED, n.d.), although illustrations of such normal distributions do not stand without their own set of controversies (e.g., a general assumption that such data yield "truth" given they are mathematically derived, sans systemic or human error, bias, policy/political decisions, manipulation, and the like; see also Amrein-Beardsley & Geiger, under review; VAMboozled, 2017).

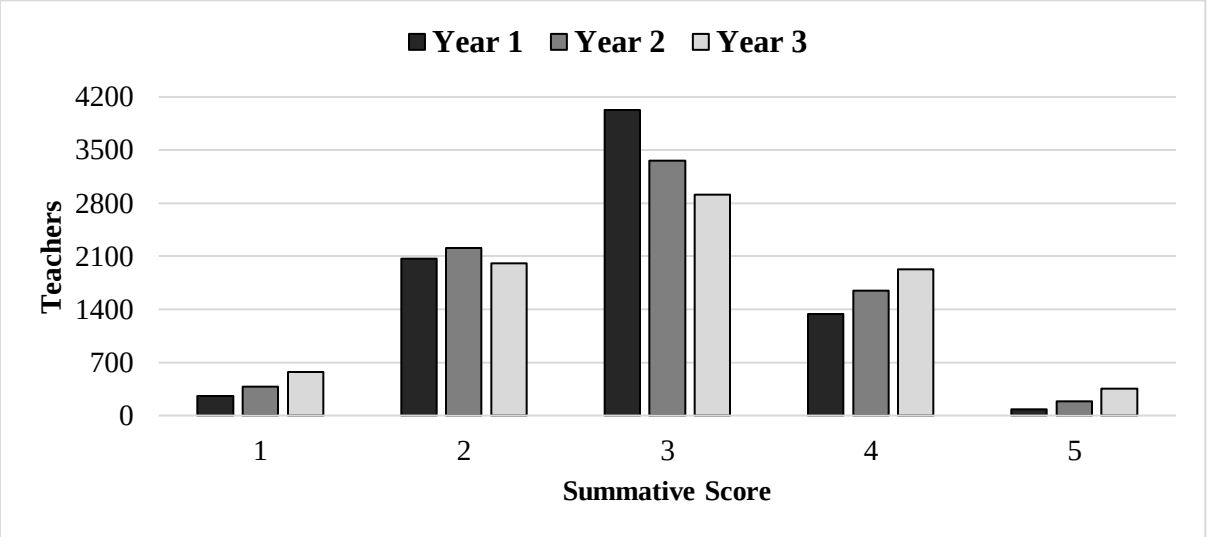



*Figure 1*. Distributions of teachers' overall summative ratings.

Across the four measures of teacher effectiveness used in the NMTEACH system, the distributions of teachers' scores varied considerably (see Appendix B for normal, positively, and negatively skewed figures), which also suggests that when these measures were averaged, or combined together (e.g., using weights), the measures yielded a normal distribution such as that illustrated in Figure 1 above (i.e., a result of taking averages of averages skewed and un-skewed). For perhaps a better illustration of how this normal distribution in Figure 1 above might have occurred (or been constructed) in the case of New Mexico, for example, see Figure 2.

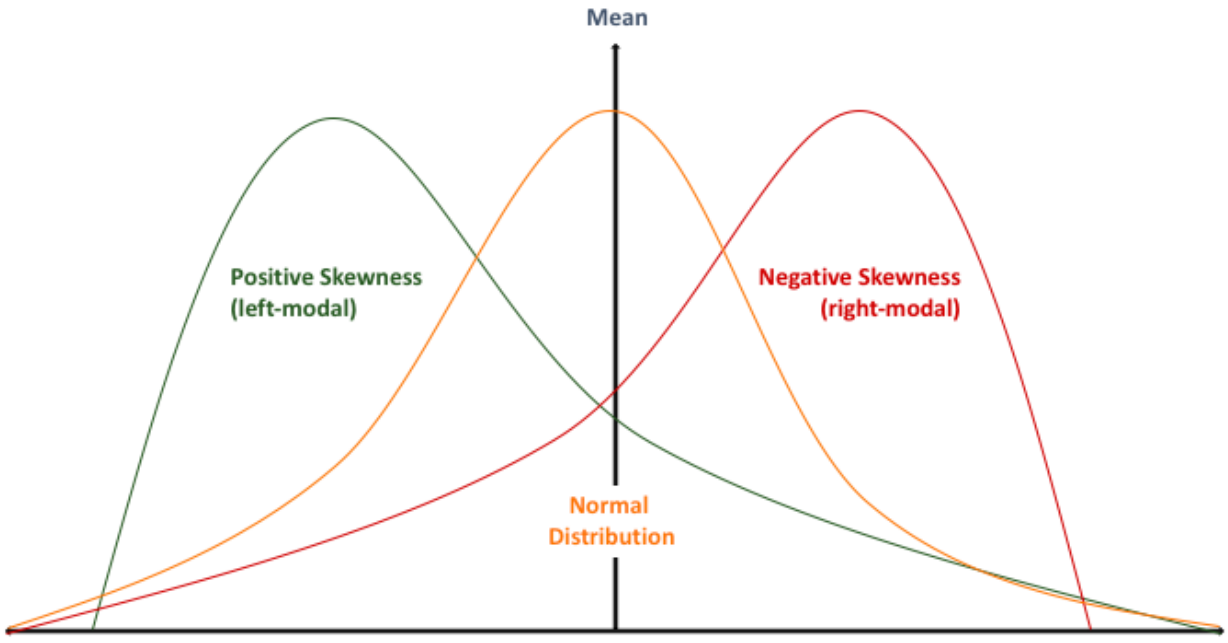



*Figure 2*. Illustration of how when averaging or combining (e.g., using weights) positively and negatively skewed data can yield a normal, bell curve distribution

More specifically, the distributions of VASs were positively skewed across the board (i.e., Years 1, 2, and 3), though distributions became less skewed over time. Inversely, the distributions of teachers' observation and PPP scores were negatively skewed, as were distributions of teachers' student survey scores across all three years. When putting these altogether, also with weights, a normal distribution as illustrated in Figures 1 and 2 above is

likely to result, not because it represents a "true" distribution of "actual" or "real" teacher effects that were accurately or effectively captured, but a "true" distribution of multiple evaluative indicators that were perhaps arbitrarily and otherwise weighted for "best fit" purposes (see also the discussion about New Mexico's ESSA plans and state changes in measurement weights above). Nonetheless, this ultimately makes claims like:

- "[New Mexico] has rapidly moving [*sic*] away from what Weisburg, Sexton, Mulhern and Keeling termed the 'widget effect' in their report issued almost a decade ago," in which Weisberg et al. criticized negatively skewed teacher effectiveness data across states given, as they argued, these data should be more evenly or normally distributed (Weisburg et al., 2009, 121-122), and
- "[New Mexico] teacher eval[uation]s are toughest in the nation" given data like those illustrated in Figure 1 above evidence New Mexico's "commitment to putting students first" *not* the "painting [of] a picture that we know is not accurate by rating nearly all teachers effective or better" (Burgess, 2017),

suspect in terms of the extent to which such distributions are "real" or illustrate "real" levels of teachers' effectiveness levels. This is akin decades of debates pertaining to what have been called the *Bell Curve Debates* (see, for example, Jacoby, Glauberman, & Herrnstein, 1995).

Notwithstanding, teachers demonstrated a significant increase from year-to-year via all teacher effectiveness measures for all years, except on the student survey from Year 2 to Year 3 (see Table 1) when teachers scored significantly lower in Year 3 than in Year 2. It is also uncertain, however, to what extent these increases in these scores were also caused by "real" or "true" increases in New Mexico teachers' actual effectiveness or other statistical and policy, political, etc. changes, as is also possible and sometimes common with almost any type of test-

based system such as this (i.e., artificial inflation; see also AERA et al., 2014; Amrein & Berliner, 2002; Haladyna, Nolen, & Haas, 1991; Nichols & Berliner, 2007; Shepard, 1990).

Table 1

*Means and Standard Deviations for Measures of Teacher Effectiveness*

| Year | VAM | | Observation | | PPP | | Student Survey | |
|---|---|---|---|---|---|---|---|---|
| | M | sd | M | sd | M | sd | M | sd |
| 2013-2014 | 26.35 | 17.93 | 0.67 | 0.847 | 0.69 | 0.096 | 0.78 | 0.119 |
| 2014-2015 | 38.01 | 22.46 | 0.70 | 0.095 | 0.74 | 0.106 | 0.82 | 0.102 |
| 2015-2016 | 46.16 | 24.44 | 0.73 | 0.100 | 0.75 | 0.115 | 0.81 | 0.102 |

*Note*: All comparisons within measures across years resulted in *p* values of .000.

**Reliability**

Overall, the distributions of teachers' overall summative scores per year all significantly differed from each other (Year 1 vs. Year 2: $\chi^2 = 371.31$, $p < .001$; Year 1 vs. Year 3: $\chi^2 = 1814.03$, $p < .001$; Year 2 vs. Year 3: $\chi^2 = 375.36$, $p < .001$) (see Appendix C for distributions of the overall summative scores, as well as distributions of the four measures of teacher effectiveness). Additionally, while a plurality of teachers' VAS-based quintile rankings were either identical from year-to-year (i.e., 31.6% [n = 2,455/7,771] from Year 1 to Year 2, and 31.6% [n = 2,454/7,766] from Year 2 to Year 3) or differed by one quintile (i.e., 40.6% [n = 3,157/7,771] from Year 1 to Year 2 and 39.4% [n = 3,060/7,766] from Year 2 to Year 3), many teachers also received dissimilar VAS-based quintile rankings over the same period of time, with over one quarter of all teachers having scores from year-to-year that differed by two or more quintiles (i.e., 27.8% [n = 2,159/7,771] from Year 1 to Year 2; 29.0% [n = 2,252/7,766] from Year 2 to Year 3) (see Appendix D, Table A6 for quintile deviations in scores over time).

These results make sense when situated in the current literature, whereas teachers classified as "effective" one year typically have a 25%-59% chance of being classified as "ineffective" the next, or vice versa, with other permutations also possible (Chiang, McCullough, Lipscomb, & Gill, 2016; Martinez, Schweig, & Goldschmidt, 2016; Schochet & Chiang, 2013;

Shaw & Bovaird, 2011; Yeh, 2013). In New Mexico approximately 40% of teachers differed by one quintile and approximately 28% of teachers differed, from year-to-year, by two or more quintiles in terms of their VAS-derived effectiveness ratings. What this also means is that across VAMs, not only just in New Mexico, reliability is a hindrance, thwarting the usability of teacher-level VAM estimates, again, especially when unreliable measures are to be used for consequential decision-making purposes.

Overall, teachers' observation scores appeared to be the most stable or reliable over time, when compared to each of the other three measures of teacher effectiveness. However, this apparent stability could also be due to the fact that teachers' observation scores in New Mexico and elsewhere (Weisberg et al., 2009) are typically negatively skewed and thus, have less variation than other "more objective" measures including VAMs (and VASs in the case of New Mexico; see Appendix B).

Otherwise, teachers' VAS-based quintile scores fluctuated significantly more than the scores of all other measures, except for that of the survey scores for just one year. Teachers' survey scores yielded the most fluctuation from Year 1 to Year 2, but not Year 2 to Year 3, which could have also been caused by measurement issues with survey research (e.g., inadequate response rates; Nunnally, 1978). Without adequate student responses (e.g., 70% of the population sampled), reliability *and* validity issues such as those observed herein can arise.

**Validity**

Overall, correlations among all measures were weak to very weak[3] across all years, with the exception of the correlations between teachers' observation and PPP scores, which were

[3] Interpreting *r*: $0.8 \leq r \leq 1.0$ = a very strong correlation; $0.6 \leq r \leq 0.8$ = a strong correlation; $0.4 \leq r \leq 0.6$ = a moderate correlation; $0.2 \leq r \leq 0.4$ = a weak correlation; and $0.0 \leq r \leq 0.2$ = a very weak correlation, if any at all (Merrigan & Huston, 2008).

strong (see Table 2). The strong correlations observed, however, make sense (i.e., as potentially valid as per concurrent-related evidence of validity) given teachers' observation and PPP scores came from within the same rubric/protocol (i.e., as modified from Charlotte Danielson's Framework for Teaching [Danielson Group, n.d.]). In fact, if anything, such strong correlations might, rather, suggest that separating out teachers' observation and PPP scores is *not* defensible or warranted, as is also current practice in New Mexico, given such high or strong correlation coefficients often suggest (1) a universal but not divisible or detachable (e.g., Domains 2 and 3 versus Domains 1 and 4) factor structure and that (2) the factor structure pragmatically posited may not empirically hold (Sloat, 2015; see also Polat & Cepik, 2015; Sloat, Amrein-Beardsley, & Sabo, 2017).

Table 2

*Correlations Among Measures, per Year*

| Measure | 2013-2014 | 2014-2015 | 2015-2016 |
|---|---|---|---|
| VAS & Observation | 0.153*** | 0.187*** | 0.210*** |
| VAS & PPP | 0.128*** | 0.154*** | 0.189*** |
| VAS & Student Survey | 0.031* | 0.063** | 0.135*** |
| Observation & PPP | 0.771*** | 0.774*** | 0.788*** |
| Observation & Student Survey | 0.211*** | 0.219*** | 0.235*** |
| PPP & Student Survey | 0.196*** | 0.175*** | 0.202*** |

* $p < .05$ ** $p < .01$ *** $p < .001$

As illustrated, correlations were the weakest between teachers' VAS estimates and student survey scores across all years, ranging between $r = 0.031$ and 0.135. Conversely, other than the strongest relationships observed between teachers' observation and PPP scores (i.e., as taken from within the same instrument), correlations were the strongest between teachers' observation and student survey scores across all years, ranging from 0.211 to 0.235.

Perhaps more importantly, as also situated within the current literature, the correlations between teachers' VASs and observational scores ranged from $r = 0.153$ to $r = 0.210$. The

correlations between teachers' VASs and scores taken from the lesser weighted PPP indicators ranged from $r = 0.128$ to $r = 0.189$. Not only are these correlations very weak,[4] they are also very weak as appropriately situated within the literature, via which it is evidenced that correlations between multiple VAMs and observational scores typically range from $0.30 \leq r \leq 0.50$ (see, for example, Grossman, Cohen, Ronfeldt, & Brown, 2014; Hill et al., 2011; Kane & Staiger, 2012; Polikoff & Porter, 2014; Wallace, Kelcey, & Ruzek, 2016). New Mexico's correlations, with regards to evidence of concurrent-related validity, is very weak and also very weak in comparison to other models, including other models respected as some of the best models in the nation (see, for example, Grossman et al., 2014; Kane & Staiger, 2012; Polikoff & Porter, 2014).

**Bias**

As previously noted and explained, biased estimates of teacher effectiveness are observed when performance varies for different subgroups of teachers, even despite the sophistication of statistical controls put into place to control for or block such bias (see also Amrein-Beardsley, 2014; Amrein-Beardsley & Geiger, under review; Baker et al. 2010; Collins, 2014; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Green, Baker, & Oluwole, 2012; Kappler Hewitt, 2015; Koedel et al., 2015; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Michelmore, & Dynarski, 2016; Newton et al., 2010; Rothstein & Mathis, 2013). In this case, bias was evidenced at both the teacher level (e.g., teacher gender, race/ethnicity) and the school level (e.g., proportion of ELL students, proportion of minority students) across multiple measures of teacher effectiveness.

[4] Ibid.

**Teacher-level differences.** Teacher-level differences were observed across all four measures of teacher effectiveness, and included differences based on all of the five teacher-level subgroups analyzed (see Appendix E, Tables A7-A10).

*Gender.* Compared to female teachers, male teachers had significantly higher VASs in Year 1 ($t = 3.471$, $p = .001$), but significantly lower VASs in Year 3 ($t = 7.150$, $p < .001$). Female teachers had significantly greater observation scores, PPP scores, and survey scores compared to male teachers for each of the three years; hence, if anything, these three measures (i.e., not including VAS estimates across years) might be biased in favor of female teachers (see, for example, Bailey, Bocala, Shakman, & Zweig, 2016; Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014 for recent research on bias in observational systems), although some might argue that female teachers were or were perceived to be better.

*Race/Ethnicity.* While there were no significant differences between Caucasian and non-Caucasian teachers' VASs (with non-Caucasian teachers defined as Asian, African American, Hispanic, and Native American as per the data), there were significant differences between the two groups of teachers on the other three measures. Specifically, Caucasian teachers had significantly higher observation and PPP scores than non-Caucasian teachers for each of the three years. Hence, it could be concluded, again, that Caucasian teachers may be perceived as better teachers than non-Caucasians given these instruments and/or the scorers observing teachers in practice may be biased against some versus other teacher types (i.e., in this case defined by race; see, for example, Bailey et al., 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014). Given the racial demographics of teachers' observers, this might also suggest that at least some bias by race exists. Interestingly, non-Caucasian teachers had higher survey scores than Caucasian teachers for all three years, with Year 2 ($t = 4.258$, $p < .001$) and Year 3 ($t =$

5.607, $p < .001$) being significantly different. Again, this could also be due to bias (i.e., that teachers' students' have differing perspectives of their teachers' qualities, by race) or standard issues with surveys (e.g., low response rates that might distort validity; Nunnally, 1978).

*Years of experience.* Overall, teachers with fewer years of experience had VAS estimates that were significantly lower than teachers with more years of experience. Similar patterns were observed for teachers' observation scores and PPP scores, with teachers with the least amount of experience routinely earning scores that were significantly lower than their more experienced counterparts across each of the three years. This could mean, as also in line with common sense as well as the research (see, for example, Darling-Hammond, 2010), that teachers with more experience are typically better teachers. Given this finding can also be situated in the literature, these findings might support the validity of teachers' VASs and observational scores in this regard. Survey scores did not follow this pattern, however, which might be due to issues with survey research also noted prior.

*Grades taught.* In Year 1, elementary school teachers had significantly lower VASs than middle school teachers and high school teachers ($F = 149.465$, $p < .001$), and high school teachers had significantly higher VAS estimates than both elementary teachers and middle school teachers in Year 2 ($F = 6.789$, $p = .001$). This might mean, in the simplest of terms, that elementary school teachers might be worse than teachers in high school, or that VAS estimates and the ways they are calculated (e.g., using different tests at different levels), might be biased against teachers of younger students. There were no other significant differences observed across observation and PPP scores based on grades taught, although significant differences did exist for survey scores, again, and inversely, in that high school teachers had the lowest scores as based

on their students' survey responses, and elementary school teachers had the highest scores in Years 1 ($F = 60.929$, $p < .001$) and Years 2 ($F = 178.239$, $p < .001$).

*Subject taught.* Overall, teachers who taught ELL or SPED students had lower VASs across all three years than those who did not teach such students. Those that were statistically significant were for ELL teachers who had significantly lower VASs than non-ELL teachers in Year 2 ($t = 2.001$, $p = .046$), and SPED teachers who had significantly lower VAS estimates than non-SPED teachers in Year 2 ($t = 2.248$, $p = .025$) and Year 3 ($t = 7.354$, $p < .001$).

Contrariwise, teachers who taught gifted students had significantly higher VASs than non-gifted teachers in Year 1 ($t = 4.724$, $p < .001$) and Year 2 ($t = 3.147$, $p = .002$). This runs counter to the current research evidencing that teachers' gifted students oft-thwart or prevent them from demonstrating growth given ceiling effects (see, for example, Cole, Haimson, Perez-Johnson, & May, 2011; Kelly & Monczunski, 2007; Koedel & Betts, 2007, 2010; Linn & Haug, 2002; Wright, Horn, & Sanders, 1997). This does, or rather did not seem to be the case in New Mexico given this dataset.

Otherwise, patterns similar to those mentioned above were also observed for ELL, SPED, and gifted teachers on their observation and PPP scores. Consistently across all years, non-ELL, non-SPED, and gifted teachers had significantly better observation and PPP scores than their ELL, SPED, and non-gifted counterparts (see also Bailey et al., 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014). The one exception to this pattern was for ELL teachers in Year 3, as while their PPP scores were higher than non-ELL teachers' PPP scores, the difference was not significant. Regarding survey scores, ELL teachers had significantly higher scores than non-ELL teachers for each of the three years, and SPED teachers had significantly higher scores than non-

SPED teachers in Year 3. There were no significant differences between gifted and non-gifted teachers' survey scores.

**School-Level Differences.** School-level differences were observed across all four measures of teacher effectiveness, and included differences based on all six of the school-level subgroups analyzed (see Appendix E, Tables A11-A14).

*Total enrollment.* Teachers in schools with low enrollments (i.e., enrollment less than the sample median; hereafter referred to as "low enrollment schools") as compared to teachers in schools with high enrollment (i.e., enrollment greater than the sample median; hereafter referred to as "high enrollment schools") had significantly higher VASs in Year 1 ($t = 2.428$, $p = .015$) and Year 2 ($t = 5.017$, $p < .001$) than teachers in high enrollment schools, although this was reversed in Year 3 where teachers in high enrollment schools had significantly greater VASs ($t = 4.300$, $p < .001$). Observation scores only significantly differed in Year 3, where teachers in low enrollment schools also scored significantly higher than teachers in high enrollment schools ($t = 2.888$, $p = .004$). Teachers in low enrollment schools had significantly lower PPP scores in Year 1 ($t = 5.386$, $p < .001$), but significantly higher PPP scores in Year 3 ($t = 2.807$, $p = .005$) compared to teachers in high enrollment schools. Lastly, teachers' survey scores had the most established pattern: teachers in low enrollment schools had significantly higher survey scores than teachers in high enrollment schools for each of the three years (recall concerns about these measures, also as per their contradictory patterns noted above).

*SPED student population.* Teachers in low SPED schools consistently had significantly greater VASs, observation scores, and PPP scores than teachers in high SPED schools across each of the three years. This would suggest that teachers in low SPED teachers are as a group better, and/or that VAS estimates might be biased against teachers teaching in such schools,

preventing them from demonstrating comparable growth. This pattern was reversed for teachers' survey scores, however, as teachers in low SPED schools also had lower survey scores than teachers in high SPED schools across the board, although the only significant difference was in Year 3 ($t = 3.003$, $p = .003$).

*ELL student population.* Similar to teachers in low SPED schools, teachers in low ELL schools consistently had significantly greater VASs, observation scores, and PPP scores than teachers at high ELL schools for each of the three years. Teachers in low ELL schools also had lower survey scores than teachers in high ELL schools for each of the three years, though the only significant difference was, again, in Year 3 ($t = 9.131$, $p < .001$). This would suggest that teachers in low ELL schools are as a group better, and/or that VAS estimates might be biased against teachers teaching in such schools, preventing them from demonstrating comparable growth.

*FRL student population.* Again, similar to teachers in low ELL and low SPED schools, teachers in low FRL schools consistently had significantly greater VAS estimates, observation scores, and PPP scores than teachers at high FRL schools for each of the three years. Teachers in low FRL schools also had significantly lower survey scores than teachers in high FRL schools for each of the three years. Again, this would suggest that teachers in low FRL schools are as a group better, and/or that VAS estimates might be biased against teachers teaching in such schools, preventing them from demonstrating comparable growth.

*Gifted student population.* Related to the discussion about gifted teachers prior, teachers at high gifted schools also had significantly greater VASs as a whole than teachers at low gifted schools in Year 1 ($t = 2.980$, $p = .003$) and Year 2 ($t = 6.075$, $p < .001$). This pattern was the same for observation and PPP scores; hence, this would, again, suggest that teachers in high

gifted schools are as a group better, and/or that VAS estimates might be biased against teachers teaching in schools with fewer gifted students, preventing them from demonstrating comparable growth. Teachers in low gifted schools, however and perhaps not surprisingly at this point, had significantly greater survey scores than teachers in high gifted schools for each of the three years.

*Minority student population.* Lastly, and in line with teachers in low SPED, ELL, and FRL schools, teachers in low minority schools consistently had significantly higher VAS estimates than teachers in high minority schools for each of the three years. Again, this was also the case for teachers' observation and PPP scores, again suggesting that teachers in low minority schools are as a group better, and/or that VAS estimates might be biased against teachers teaching in schools with more minority students, preventing them from demonstrating comparable growth. Survey scores were slightly more varied based on minority student population, with the only significant difference between scores in Year 2, where teachers in low minority schools had significantly higher survey scores than teachers in high minority schools ($t = 2.902$, $p = .004$).

## Conclusions

The key conclusion as pertinent to reliability, validity, and bias (or the lack thereof) follow. In terms of reliability, evidenced was that New Mexico teachers' VAS-derived effectiveness ratings were unstable (i.e., unreliable) over time, as over one quarter (i.e., 25%) of teachers in the sample received a VAS that differed by at least 40% (i.e., two quintiles) from one year to the next.

In terms of validity, substantiated was weak to very weak evidence of concurrent-related validity, as assessed by the (cor)relationships among the four measures of teacher effectiveness and especially in comparison to other well-known and respected models. More specifically, as

also previously noted, other than the strong (cor)relationships observed between teachers' observation scores and PPP scores, all other evidence of concurrent-related validity was quite poor given the weak to very weak correlations observed among all measures. Even the strongest relationship, which was observed between teachers' observation and survey scores, was also weak at best, with the strongest correlation across the three years being an $r = 0.235$. Perhaps more important, also as per the current research, was the strength of the correlations observed between teachers' VASs and observation scores and VASs and PPP scores. These correlation coefficients were poor, especially in light of these other models, with the strongest correlation between New Mexico's VASs and observation scores being $r = 0.210$ and between VASs and PPP scores being $r = 0.189$, which is substantively lower than the correlations observed between multiple (and well-respected) VAMs and observational scores (i.e., $0.30 \leq r \leq 0.50$; Grossman et al., 2014; Hill et al., 2011; Kane & Staiger, 2012; Polikoff & Porter, 2014; Wallace et al., 2016).

In terms of bias, evidenced was that bias was evident at the teacher level, specific to teachers' gender, race/ethnicity, years of experience, grades taught, and subject taught, and it was evident at the school level, as specific to the proportions of students that teachers taught, including SPED, ELL, FRL, gifted, and minority students. The sub-results that can be taken from the additional analyses conducted on bias were as follows.

At the teacher-level, evidenced herein was that: (1) female teachers performed significantly better on the three non-VAS measures of effectiveness (i.e., observations, PPP, and student surveys) compared to male teachers; (2) Caucasian teachers performed significantly better than non-Caucasian teachers on the observation and PPP measures, while non-Caucasian teachers performed significantly better than Caucasian teachers on the student survey; (3) across the board, teachers with the least amount of experience performed worse than teachers with three

or more years of experience, and this difference was consistently significant for VASs and routinely significant for both observation and PPP scores; (4) elementary school teachers tended to receive significantly lower VASs than high school teachers, yet tended to receive significantly higher observation and survey scores; and (5) overall, teachers who were assigned to specific subgroups of students, such as ELL or SPED students, had lower VASs across all years compared to teachers who did not teach such groups of students.

At the school-level, evidenced herein was that: (1) teachers in schools with higher relative proportions of SPED students consistently had significantly lower VAS, observation, and PPP scores than teachers in low SPED schools; (2) teachers in schools with higher relative proportions of ELL students consistently had significantly lower VAS, observation, and PPP scores than teachers in low ELL schools; (3) teachers in schools with higher relative proportions of FRL students consistently had significantly lower VAS, observation, and PPP scores than teachers in low FRL schools; (4) teachers in schools with higher relative proportions of gifted students, overall, had significantly better VAS scores than teachers at low gifted schools; and (5) teachers in schools with higher relative proportions of minority students had significantly lower VAS scores than teachers at low gifted schools.

Whether, indeed, all of these teachers were (or are) in fact better or worse than other teachers teaching different proportions or populations of students, though, is certainly something of interest, and also of methodological and pragmatic concern. While some might argue that the by-school or by-teacher-type findings above simply suggest that better teachers are resident within certain types of schools, others might argue that the types of students who "better" teachers teach are simply more likely to demonstrate growth, compared to the other types or proportions of students that "worse" teachers teach (i.e., as non-randomly assigned into their

classrooms). Hence, the real questions as pertinent to the findings above are (1) whether students, regardless of the type of school, have equal opportunities to demonstrate growth, and consequently (2) whether students' teachers, regardless of the type of school, have equal opportunities (i.e., a "level the playing field") to demonstrate growth once their students' scores are aggregated.

Again, should estimates (dis)favor certain groups of teachers, this serves as an indicator that the statistical controls included within this or any other VAM are not working as intended, or rather not working well enough to adequately block such bias. If such controls were to work as intended, and work well, indeed, all students *and* teachers would have equal opportunities to demonstrate growth and be able to be evaluated and judged appropriately and accordingly. Hence, the findings above would consequently evidence that at least some level of bias still exists, regardless of the sophistication of the statistics used to control for it.

Finally of note is that more or less across all analyses was that the scores derived via student surveys yielded scores most often "opposite" of what was observed across the other measures included within this particular NMTEACH system. Again, this could very well be due to the measurement issues pertaining to survey research noted above, as well as in general (e.g., reliability, validity, and bias, as often related to inadequate response rates) when evaluating teachers in public schools.

## References


American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Educational Research Association (AERA) Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Educational Researcher, *X*(Y), 1-5. doi:10.3102/0013189X15618385

American Statistical Association (ASA). (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). doi:10.14507/epaa.v10n18.2002 Retrieved from http://epaa.asu.edu/epaa/v10n18/

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability.* New York, NY: Routledge

Amrein-Beardsley, A., & Geiger, T. J. (under review). Potential sources of invalidity when using teacher value-added and principal observational estimates: Artificial inflation, deflation, and conflation. *Educational Policy.*

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). *Teacher demographics and evaluation: A descriptive study in a large urban district.* Washington DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2017189.pdf

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

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

Ballou, D., Sanders, W. L. & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44(*2), 77-86. doi: 10.3102/0013189X15574904

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a ‘‘high’’ or ‘‘low’’ value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal, 53*(2), 324–359. doi: 10.3102/0002831216630407

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

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

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal, 10*(3), 245–252. doi:10.2307/1161888

Burgess, K. (2017). Expert: NM teacher evals are toughest in the nation. *Albuquerque Journal*. Retrieved from https://www.abqjournal.com/1029370/expert-nm-teacher-evals-toughest-in-us.html

Chiang, H., McCullough, M., Lipscomb, S., & Gill, B. (2016). Can student test scores provide useful measures of school principals' performance? Washington DC: U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/2016002/pdf/2016002.pdf

Chin, M., & Goldhaber, D. (2015). *Exploring explanations for the "weak" relationship between value added and observation-based measures of teacher performance*. Cambridge, MA: Center for Education Policy Research (CEPR), Harvard University. Retrieved from http://cepr.harvard.edu/files/cepr/files/sree2015_simulation_working_paper.pdf

Cody, C. A., McFarland, J., Moore, J. E., & Preston, J. (2010, August). *The evolution of growth models.* Public Schools of North Carolina. Raleigh, NC. Retrieved from http://www.dpi.state.nc.us/docs/intern-research/reports/growth.pdf

Cole, R., Haimson, J., Perez-Johnson, I., & May, H. (2011, September). *Variability in pretest-posttest correlation coefficients by student achievement level*. Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from ies.ed.gov/ncee/pubs/20114033/pdf/20114033.pdf

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future.* New York, NY: Teachers College Press.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93(6), 8-15. Retrieved from http://www.kappanmagazine.org/content/93/6/8.full.pdf+html

Duncan, A. (2011). *Winning the future with education: Responsibility, reform and results. Testimony given to the U.S. Congress.* Washington, DC: Retrieved from http://www.ed.gov/news/speeches/winning-future-education-responsibility-reform-and-results

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64*(325), 366-377. doi: 10.2307/2283746

Dunn, O. J., & Clark, V. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association, 66*(336), 904-908. doi: 10.2307/2284525

Every Student Succeeds Act (ESSA) of 2016, Pub. L. No. 114-95, § 129 Stat. 1802. (2016). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Excel*in*Ed. (2017). ESSA state plans: 50-state landscape analysis. Tallahasee, FL. Retrieved from https://www.excelined.org/wp-content/uploads/2017/12/ExcelinEd.Quality.ESSA_.50StateAnalysis.Dec072017.pdf

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

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

Goldschmidt, P., Choi, K., & Beaudoin, J. B. (2012, February). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Technical Issues in Large-Scale Assessment State Collaborative on Assessment and Student Standards. Council of Chief State School Officers.

Green, P. C., Baker, B. D., & Oluwole, J. (2012). Legal and policy implications of value-added teacher assessment policies. *The Brigham Young University Education and Law Journal,* 2012, 1.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of

assessment. *Educational Researcher, 43*(6), 293–303. doi:10.3102/0013189X14544542

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

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

Harris, D. N., & Herrington, C. D. (2015). Editors' introduction: The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher, 44*(2), 71-76. doi:10.3102/0013189X15576142

Hill, H. C., Kapitula, L, & Umlan, K. (2011, June). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. doi:10.3102/0002831210387916

Holloway-Libell, J. (2015). Evidence of grade and subject-level bias in value-added measures. *Teachers College Record, 117*.

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

Jacoby, R., Glauberman, N., & Herrnstein, R. J. (1995). *The bell curve debate: History, documents, opinions*. New York, NY: Times Books.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher, 44*(2), 105-116. doi:10.3102/0013189X15575517 Retrieved from http://edr.sagepub.com/content/44/2/105.full.pdf+html?ijkey=4xULkEGAZXa.s&keytype=ref&siteid=spedr

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

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

Kane, M. T. (2017). *Measurement error and bias in value-added models.* Princeton, NJ: Educational Testing Services (ETS RR–17-25). doi:10.1002/ets2.12153 Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12153/full

Kane, T. J., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda

Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

Kappler Hewitt, K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: Undermined intentions and exacerbated inequities. *Education Policy Analysis Archives, 23*(76), 1-49. Retrieved from http://epaa.asu.edu/ojs/article/view/1968

Kelly, S. & Monczunski, L. (2007). Overcoming the volatility in school-level gain scores: A new approach to identifying value added with cross-sectional data. *Educational Researcher, 36*(5), 279-287. doi:10.3102/0013189X07306557

Koedel, C., & Betts, J. R. (2007, April). *Re-examining the role of teacher quality in the educational production function.* Working Paper No. 2007-03. Nashville, TN: National Center on Performance Initiatives.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47,* 180–195. doi: 10.1016/j.econedurev.2015.01.006

Lavery, M. R., Holloway, J., Amrein-Beardsley, A., Pivovarova, M., & Hahs-Vaughn, D. L. (in progress). Evaluating the validity evidence surrounding the use of student standardized test scores to evaluate teachers: A centennial, systematic mega-review.

Linn, R L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24,* 29-36. doi:10.3102/01623737024001029

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND Corporation. Retrieved from www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics,* 29(1), 67-101. Retrieved from www.rand.org/pubs/reprints/2005/RAND_RP1165.pdf

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis, 38*(4), 738–756. doi: 10.3102/0162373716666166 Retrieved from http://journals.sagepub.com/doi/pdf/10.3102/0162373716666166

Merrigan, G., & Huston, C. L. (2008). *Communication research methods*. New York, NY: Oxford University Press.

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

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

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

Michelmore, K., & Dynarski, S. (2016). The gap within the gap: Using longitudinal data to understand income differences in student achievement. Cambridge, MA: National Bureau of Economic Research (NBER). Retrieved from http://www.nber.org/papers/w22474

National Association of Secondary School Principals (NASSP). (n.d.). *Value-added measures in teacher evaluation: Position statement.* Reston, VA: NASSP Board of Directors

New Mexico Public Education Department (NMPED). (n.d.) *New Mexico rising: New Mexico's state plan for the Every Student Succeeds Act.* Santa Fe, NM. Retrieved from https://www2.ed.gov/admins/lead/account/stateplan17/nmconsolidatedstateplan.pdf

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23). doi: 10.14507/epaa.v18n23.2010

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

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425. (2002).

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Papay, J. P. (2012). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal.* doi: 10.3102/0002831210362589

Polat, N., & Cepik, S. (2015). An exploratory factor analysis of the Sheltered Instruction Observation Protocol as an evaluation tool to measure teaching effectiveness. *TESOL Quarterly, 50(*4), 817–843. doi: 10.1002/tesq.248

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399-416. doi: 10.3102/0162373714531851

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

Rothstein, J. (2009). *Student sorting and bias in value-added estimation: Selection on observables and unobservables.* Cambridge, MA: National Bureau of Economic Research.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175-214. doi: 10.1162/qjec.2010.125.1.175

Rothstein, J. (2014). *Revisiting the impacts of teachers* (Working paper). Berkeley, CA: University of California, Berkeley.

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET Project*. Boulder, CO: National Education Policy Center.

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009, November). *A response to criticisms of SAS EVAAS*. Cary, NC: SAS Institute Inc. Retrieved from http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*(2), 142-171. doi: 10.3102/1076998611432174

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

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice, 9*, 15-22. doi:10.1111/j.1745-3992.1990.tb00374.x

Sloat, E. F. (2015). *Examining the validity of a state policy-directed framework for evaluating teacher instructional quality: Informing policy, impacting practice* (Unpublished doctoral dissertation). Arizona State University, Tempe, AZ.

Sloat, E., Amrein-Beardsley, A., & Sabo, K. E. (2017). Examining the factor structure underlying the TAP System for Teacher and Student Advancement. *AERA Open, 3*(4), 1–18. doi:10.1177/2332858417735526 Retrieved from http://journals.sagepub.com/doi/full/10.1177/2332858417735526

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293-317. doi:10.3102/0162373715616249 Retrieved from

http://static.politico.com/58/5f/f14b2b144846a9b3365b8f2b0897/study-of-classroom-observations-of-teachers.pdf

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., …Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29* (1), 11-36. doi:10.3102/10769986029001011

U.S. Department of Education. (2010). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act*. Retrieved from http://www2.ed.gov/policy/elsec/leg/blueprint/index.html

U.S. Department of Education. (2014). *States granted waivers from No Child Left Behind allowed to reapply for renewal for 2014 and 2015 school years.* Washington D.C. Retrieved from http://www.ed.gov/news/press-releases/states-granted-waivers-no-child-left-behind-allowed-reapply-renewal-2014-and-2015-school-years

VAMboozled. (2017). The "Widget Effect" report revisited. Retrieved from http://vamboozled.com/the-widget-effect-report-revisited/

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, 53(6), 1834–1868. doiI:10.3102/0002831216671864

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: New Teacher Project. Retrieved from https://tntp.org/publications/scroll/evaluation-and-development

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts.* Washington, DC: Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2016/06/Evaluating-Teachers-with-Classroom-Observations.pdf

Wright, P., Horn, S., & Sanders, W. L. (1997). Teachers and classroom heterogeneity: Their effects on educational outcomes. *Journal of Personnel Evaluation in Education, 11*(1), 57-67.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010, March 25). SAS® EVAAS® statistical models. Retrieved from www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf

Yeh, S. S. (2013). A re-analysis of the effects of teacher replacement using value-added modeling. *Teachers College Record, 115*(12).

## Appendix A

Demographics Used in Analyses

| | |
|---|---|
| Teacher Gender | Male<br>Female |
| Teacher Ethnicity | Asian<br>African American<br>Caucasian<br>Hispanic<br>Native American |
| Teacher Ethnicity | White<br>Non-White |
| Teacher Years of Experience* | 0-2 Years<br>3-8 Years<br>9-15 Years<br>16+ Years |
| Grade Level** | Elementary School<br>Middle School<br>High School |
| ELL Teacher | Yes<br>No |
| SPED Teacher | Yes<br>No |
| Gifted Teacher | Yes<br>No |
| Student Enrollment | Low (4 – 502 students)<br>High (503 – 2389 students) |
| Student SPED School Population | Low (0.00% – 13.99% of all students)<br>High (14.00% – 100.00% of all students) |
| Student ELL School Population*** | Low (0.00% – 10.69% of all students)<br>High (10.70% – 85.20% of all students) |
| Student FRL School Population | Low (0.00% – 74.49% of all students)<br>High (74.50% – 100.00% of all students) |
| Student Gifted School Population*** | Low (0.00% – 3.49% of all students)<br>High (3.50% – 44.44% of all students) |
| Student Minority School Population*** | Low (5.60% – 78.29% of all students)<br>High (78.30% – 100.00% of all students) |

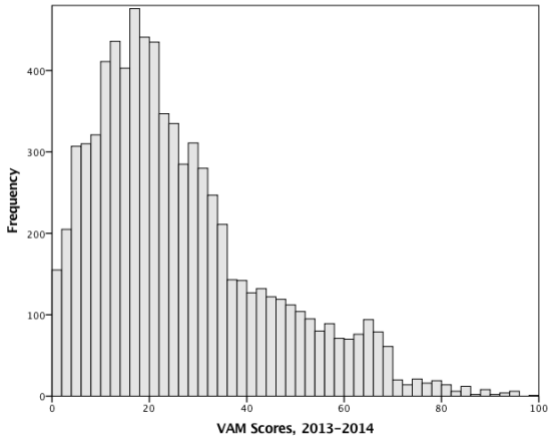
* These categories were calculated by determining quartiles as the Year of Experience variable in the initial dataset provided by the NMPED was continuous.

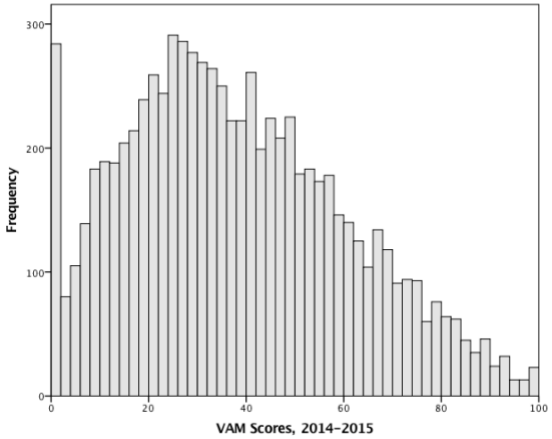
** This data was only available for 2013-2014 and 2014-2015; thus, grade level analyses were not conducted for the 2015-2016 school year.

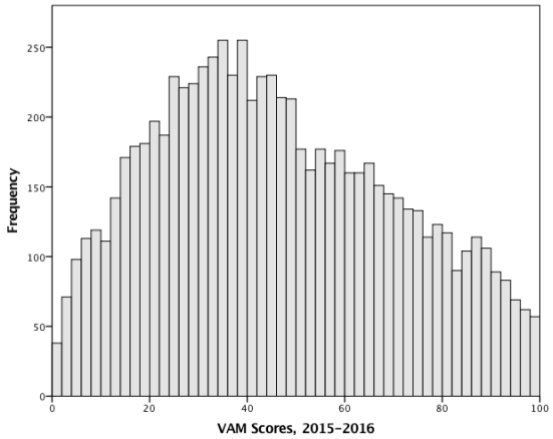
*** The low end of the "Low" range and the high end of the "High" range represents the lowest and highest proportions in the data set, respectively; as such, not all low ends of the range are 0.00% and not all high ends of the range are 100.00%.

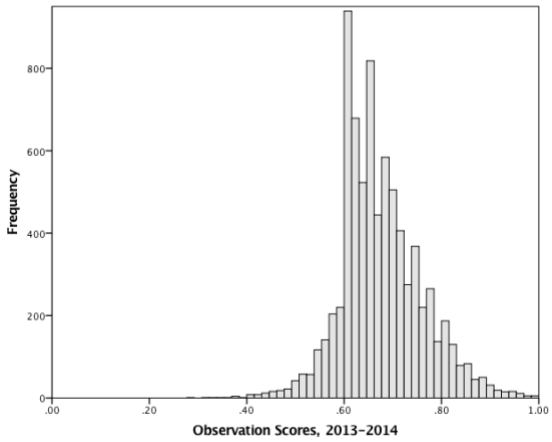
## Appendix B

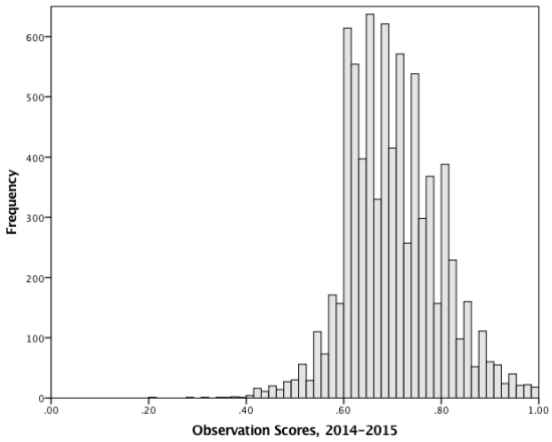
Distributions of Teacher Effectiveness Measures, per Year

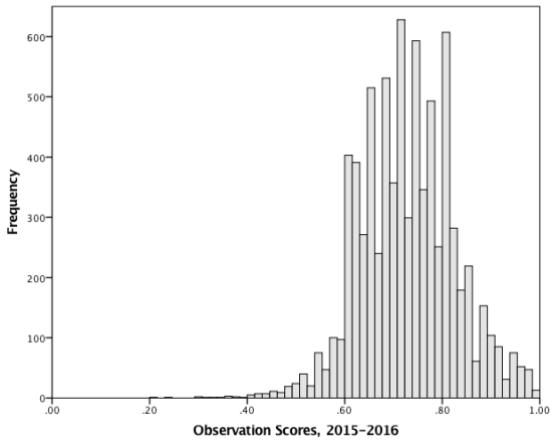


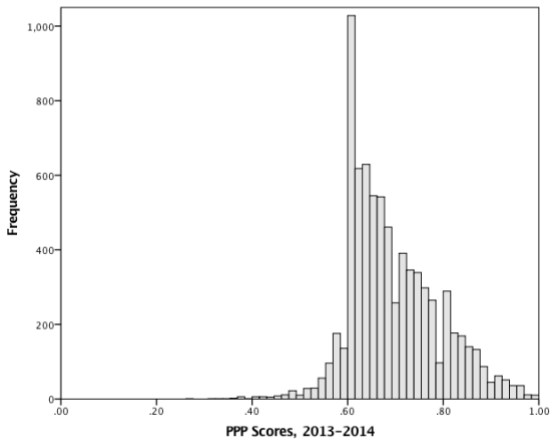



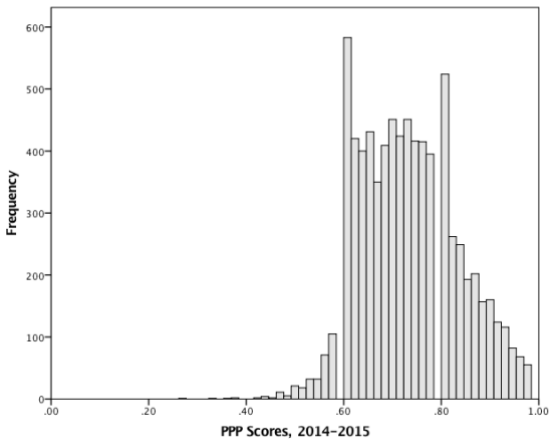



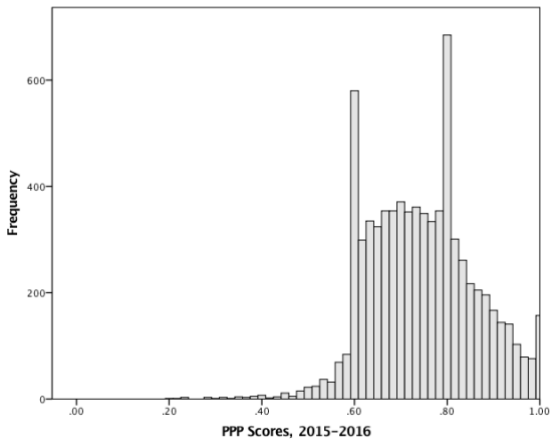



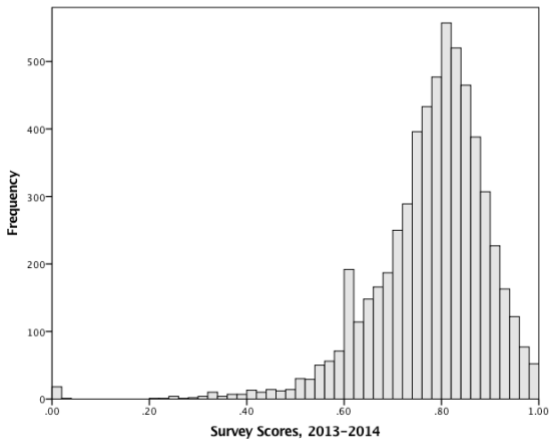



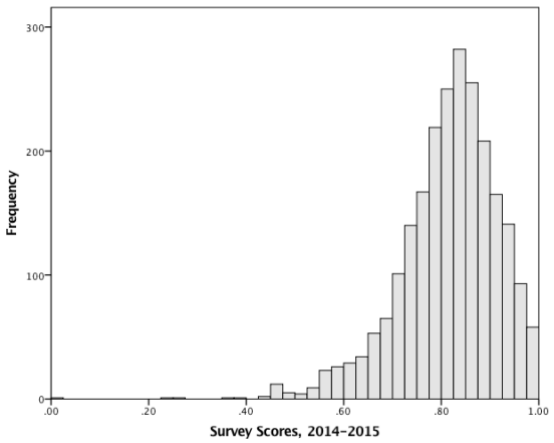



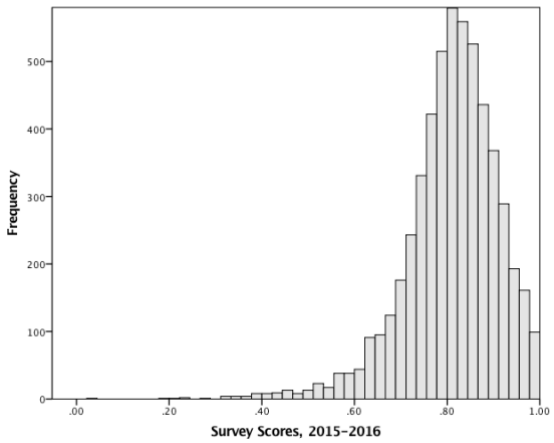

Frequency
1,000
800
600
400
200
0
.00
.20
.40
.60
.80
1.00
PPP Scores, 2013-2014


Frequency
600
500
400
300
200
100
0
.00
.20
.40
.60
.80
1.00
PPP Scores, 2014-2015


Frequency
600
400
200
0
.00
.20
.40
.60
.80
1.00
PPP Scores, 2015-2016


Frequency
500
400
300
200
100
0
.00
.20
.40
.60
.80
1.00
Survey Scores, 2013-2014


Frequency
300
200
100
0
.00
.20
.40
.60
.80
1.00
Survey Scores, 2014-2015


Frequency
500
400
300
200
100
0
.00
.20
.40
.60
.80
1.00
Survey Scores, 2015-2016

## Appendix C

Distributions of Summative Scores and Teacher Effectiveness Measure Scores

Table A1

*Distributions of Summative Scores Over Time*

| | Score | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Total |
| Year 1 Scores | 3.3% (n=255/7777) | 26.6% (n=2072/7777) | 51.8% (n=4030/7777) | 17.2% (n=1338/7777) | 1.1% (n=82/7777) | 100% (n=7777/7777) |
| Year 2 Scores | 4.9% (n=382/7777) | 28.4% (n=2206/7777) | 43.2% (n=3357/7777) | 21.2% (n=1646/7777) | 2.4% (n=186/7777) | 100% (n=7777/7777) |
| Year 3 Scores | 7.3% (n=571/7777) | 25.8% (n=2003/7777) | 37.5% (n=2915/7777) | 24.8% (n=1932/7777) | 4.6% (n=356/7777) | 100% (n=7777/7777) |

Table A2

*Distributions of VAM Scores Over Time*

| | Score | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Total |
| Year 1 Scores | 19.5% (n=1518/7777) | 19.4% (n=1512/7777) | 20.0% (n=1554/7777) | 20.3% (n=1581/7777) | 20.7% (n=1612/7777) | 100% (n=7777/7777) |
| Year 2 Scores | 2.9% (n=229/7771) | 18.0% (n=1396/7771) | 24.5% (n=1903/7771) | 25.3% (n=1965/7771) | 29.3% (n=2278/7771) | 100% (n=7771/7771) |
| Year 3 Scores | 13.7% (n=1067/7772) | 15.6% (n=1213/7772) | 17.9% (n=1392/7772) | 23.1% (n=1794/7772) | 29.7% (n=2306/7772) | 100% (n=7772/7772) |

Table A3

*Distributions of Observation Scores Over Time*

| | Score | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Total |
| Year 1 Scores | 12.0% (n=933/7777) | 23.0% (n=1790/7777) | 20.8% (n=1617/7777) | 24.7% (n=1921/7777) | 19.5% (n=1516/7777) | 100% (n=7777/7777) |
| Year 2 Scores | 14.5% (n=1129/7777) | 18.7% (n=1451/7777) | 17.1% (n=1331/7777) | 26.8% (n=2081/7777) | 23.0% (n=1785/7777) | 100% (n=7777/7777) |
| Year 3 Scores | 14.3% (n=1112/7777) | 18.9% (n=1467/7777) | 19.0% (n=1481/7777) | 22.3% (n=1732/7777) | 25.5% (n=1985/7777) | 100% (n=7777/7777) |

Table A4

*Distributions of PPP Scores Over Time*

| | Score | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Total |
| Year 1 Scores | 6.8% (n=475/6956) | 28.9% (n=2010/6956) | 19.8% (n=1379/6956) | 21.0% (n=1463/6956) | 23.4% (n=1629/6956) | 100% (n=6956/6956) |
| Year 2 Scores | 19.0% (n=1310/6893) | 17.1% (n=1182/6893) | 18.6% (n=1284/6893) | 12.1% (n=832/6893) | 33.1% (n=2285/6893) | 100% (n=6893/6893) |
| Year 3 Scores | 14.2% (n=917/6475) | 15.6% (n=1013/6475) | 16.6% (n=1077/6475) | 26.6% (n=1722/6475) | 27.0% (n=1746/6475) | 100% (n=6475/6475) |

Table A5

*Distributions of Student Survey Scores Over Time*

| | Score | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Total |
| Year 1 Scores | 16.4% (n=1162/5877) | 14.4% (n=1203/5877) | 35.2% (n=1277/5877) | 12.9% (n=1275/5877) | 21.1% (n=960/5877) | 100% (n=5877/5877) |
| Year 2 Scores | 12.9% (n=433/2337) | 9.9% (n=477/2337) | 33.3% (n=525/2337) | 13.6% (n=496/2337) | 30.2% (n=406/2337) | 100% (n=2337/2337) |
| Year 3 Scores | 17.3% (n=1022/5389) | 11.8% (n=1136/5389) | 25.6% (n=1125/5389) | 10.5% (n=1116/5389) | 34.8% (n=990/5389) | 100% (n=5389/5389) |

## Appendix D

Deviations in Scores Over Time

Table A6

*Teachers' Variations in Score Quintiles Over Time*

| | VAM Scores | | | | Observation Scores | | | |
|---|---|---|---|---|---|---|---|---|
| | Year 1 to Year 2 | | Year 2 to Year 3 | | Year 1 to Year 2 | | Year 2 to Year 3 | |
| | % | n | % | n | % | n | % | n |
| No Variation | 31.6% | 2455 | 31.6% | 2454 | 39.0% | 3031 | 41.3% | 3210 |
| One Quintile Variation | 40.6% | 3157 | 39.4% | 3060 | 39.7% | 3087 | 39.7% | 3084 |
| Two Quintile Variation | 20.1% | 1564 | 19.9% | 1547 | 15.8% | 1230 | 13.6% | 1061 |
| Three Quintile Variation | 6.0% | 465 | 7.3% | 570 | 4.7% | 366 | 4.4% | 343 |
| Four Quintile Variation | 1.7% | 130 | 1.7% | 135 | 0.8% | 63 | 1.0% | 79 |
| Total | 100.0% | 7771 | 100.0% | 7766 | 100.0% | 7777 | 100.0% | 7777 |
| | PPP Scores | | | | Student Survey Score | | | |
| | Year 1 to Year 2 | | Year 2 to Year 3 | | Year 1 to Year 2 | | Year 2 to Year 3 | |
| | % | n | % | n | % | n | % | n |
| No Variation | 37.2% | 2298 | 41.2% | 2369 | 33.5% | 778 | 41.5% | 894 |
| One Quintile Variation | 39.9% | 2468 | 36.1% | 2079 | 38.3% | 889 | 39.1% | 844 |
| Two Quintile Variation | 15.6% | 964 | 15.4% | 889 | 16.2% | 375 | 13.8% | 297 |
| Three Quintile Variation | 6.3% | 389 | 6.0% | 343 | 9.0% | 208 | 5.0% | 108 |
| Four Quintile Variation | 1.0% | 63 | 1.3% | 76 | 3.1% | 71 | 0.7% | 14 |
| Total | 100.0% | 6182 | 100.0% | 5756 | 100.0% | 2321 | 100.0% | 2157 |

## Appendix E

Means of Teacher Effectiveness Measures, per Teacher and School Subgroups

Table A7

*VAS Means, per Teacher Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Gender | | | |
| Male | 27.69 | 37.42 | 42.67 |
| Female | 25.94 | 38.24 | 47.29 |
| Race/Ethnicity | | | |
| Asian | 27.25 | 36.18 | 45.53 |
| African American | 23.71 | 34.96 | 42.24 |
| Caucasian | 26.64 | 38.30 | 46.56 |
| Hispanic | 26.18 | 38.12 | 45.64 |
| Native American | 22.06 | 31.63 | 46.75 |
| Non-Caucasian | 25.94 | 37.65 | 45.60 |
| Years of Experience | | | |
| 0-2 Years | 22.82 | 31.37 | 40.52 |
| 3-8 Years | 24.24 | 35.03 | 45.51 |
| 9-15 Years | 27.42 | 40.39 | 47.56 |
| 16+ Years | 28.02 | 46.26 | 47.27 |
| Grade Level | | | |
| Elementary School | 21.91 | 37.27 | N/A |
| Middle School | 29.54 | 37.75 | N/A |
| High School | 30.87 | 39.48 | N/A |
| ELL Teachers | | | |
| Yes | 24.67 | 35.48 | 44.88 |
| No | 26.41 | 38.10 | 46.20 |
| SPED Teachers | | | |
| Yes | 25.86 | 36.20 | 40.62 |
| No | 26.40 | 38.19 | 46.73 |
| Gifted Teachers | | | |
| Yes | 32.66 | 43.06 | 43.63 |
| No | 26.21 | 37.89 | 46.24 |

*Note*: Teacher grade level assignment was not provided for Year 3, so disaggregating scores by grades taught was not possible.

Table A8

*Observation Score Means, per Teacher Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Gender | | | |
| Male | 0.663 | 0.682 | 0.709 |
| Female | 0.679 | 0.708 | 0.738 |
| Race/Ethnicity | | | |
| Asian | 0.640 | 0.665 | 0.691 |
| African American | 0.658 | 0.686 | 0.715 |
| Caucasian | 0.680 | 0.707 | 0.734 |
| Hispanic | 0.670 | 0.697 | 0.729 |
| Native American | 0.634 | 0.651 | 0.389 |
| Non-Caucasian | 0.667 | 0.693 | 0.726 |
| Years of Experience | | | |
| 0-2 Years | 0.662 | 0.682 | 0.712 |
| 3-8 Years | 0.669 | 0.700 | 0.730 |
| 9-15 Years | 0.678 | 0.704 | 0.735 |
| 16+ Years | 0.682 | 0.707 | 0.733 |
| Grade Level | | | |
| Elementary School | 0.677 | 0.705 | N/A |
| Middle School | 0.671 | 0.697 | N/A |
| High School | 0.676 | 0.701 | N/A |
| ELL Teachers | | | |
| Yes | 0.657 | 0.680 | 0.710 |
| No | 0.676 | 0.702 | 0.732 |
| SPED Teachers | | | |
| Yes | 0.658 | 0.680 | 0.709 |
| No | 0.677 | 0.704 | 0.733 |
| Gifted Teachers | | | |
| Yes | 0.708 | 0.751 | 0.764 |
| No | 0.674 | 0.700 | 0.730 |

*Note*: Teacher grade level assignment was not provided for Year 3, so disaggregating scores by grades taught was not possible.

Table A9

*PPP Score Means, per Teacher Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Gender | | | |
| Male | 0.676 | 0.708 | 0.713 |
| Female | 0.700 | 0.744 | 0.757 |
| Race/Ethnicity | | | |
| Asian | 0.673 | 0.714 | 0.704 |
| African American | 0.673 | 0.719 | 0.726 |
| Caucasian | 0.702 | 0.743 | 0.753 |
| Hispanic | 0.685 | 0.726 | 0.740 |
| Native American | 0.647 | 0.682 | 0.689 |
| Non-Caucasian | 0.682 | 0.723 | 0.736 |
| Years of Experience | | | |
| 0-2 Years | 0.677 | 0.707 | 0.727 |
| 3-8 Years | 0.689 | 0.732 | 0.745 |
| 9-15 Years | 0.697 | 0.738 | 0.751 |
| 16+ Years | 0.701 | 0.745 | 0.748 |
| Grade Level | | | |
| Elementary School | 0.693 | 0.734 | N/A |
| Middle School | 0.692 | 0.738 | N/A |
| High School | 0.698 | 0.734 | N/A |
| ELL Teachers | | | |
| Yes | 0.667 | 0.712 | 0.737 |
| No | 0.695 | 0.736 | 0.746 |
| SPED Teachers | | | |
| Yes | 0.679 | 0.715 | 0.726 |
| No | 0.696 | 0.737 | 0.748 |
| Gifted Teachers | | | |
| Yes | 0.728 | 0.777 | 0.786 |
| No | 0.693 | 0.734 | 0.745 |

*Note*: Teacher grade level assignment was not provided for Year 3, so disaggregating scores by grades taught was not possible.

Table A10

*Survey Score Means, per Teacher Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Gender | | | |
| Male | 0.763 | 0.793 | 0.792 |
| Female | 0.784 | 0.824 | 0.816 |
| Race/Ethnicity | | | |
| Asian | 0.732 | 0.747 | 0.766 |
| African American | 0.760 | 0.790 | 0.798 |
| Caucasian | 0.777 | 0.807 | 0.803 |
| Hispanic | 0.786 | 0.831 | 0.822 |
| Native American | 0.767 | 0.802 | 0.805 |
| Non-Caucasian | 0.782 | 0.825 | 0.819 |
| Years of Experience | | | |
| 0-2 Years | 0.770 | 0.815 | 0.803 |
| 3-8 Years | 0.788 | 0.828 | 0.819 |
| 9-15 Years | 0.780 | 0.820 | 0.809 |
| 16+ Years | 0.776 | 0.805 | 0.806 |
| Grade Level | | | |
| Elementary School | 0.803 | 0.863 | N/A |
| Middle School | 0.768 | 0.795 | N/A |
| High School | 0.756 | 0.781 | N/A |
| ELL Teachers | | | |
| Yes | 0.803 | 0.860 | 0.863 |
| No | 0.778 | 0.813 | 0.808 |
| SPED Teachers | | | |
| Yes | 0.772 | 0.825 | 0.829 |
| No | 0.780 | 0.815 | 0.808 |
| Gifted Teachers | | | |
| Yes | 0.773 | 0.789 | 0.795 |
| No | 0.779 | 0.816 | 0.810 |

*Note*: Teacher grade level assignment was not provided for Year 3, so disaggregating scores by grades taught was not possible.

Table A11

*VAS Means, per School Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Total Enrollment | | | |
| High | 26.81 | 39.30 | 44.96 |
| Low | 25.82 | 36.73 | 47.36 |
| Student SPED Population | | | |
| High | 25.04 | 37.06 | 45.38 |
| Low | 27.57 | 38.99 | 46.89 |
| Student ELL Population | | | |
| High | 24.30 | 36.77 | 44.29 |
| Low | 28.19 | 39.22 | 47.85 |
| Student FRL Population | | | |
| High | 25.10 | 36.03 | 43.68 |
| Low | 27.47 | 39.93 | 48.45 |
| Student Gifted Population | | | |
| High | 26.90 | 39.50 | 46.05 |
| Low | 25.67 | 36.39 | 46.25 |
| Student Minority Population | | | |
| High | 25.09 | 36.18 | 42.33 |
| Low | 27.49 | 39.79 | 49.73 |

Table A12

*Observation Score Means, per School Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Total Enrollment | | | |
| High | 0.677 | 0.701 | 0.727 |
| Low | 0.673 | 0.702 | 0.734 |
| Student SPED Population | | | |
| High | 0.671 | 0.697 | 0.728 |
| Low | 0.678 | 0.706 | 0.734 |
| Student ELL Population | | | |
| High | 0.663 | 0.690 | 0.721 |
| Low | 0.686 | 0.712 | 0.740 |
| Student FRL Population | | | |
| High | 0.661 | 0.690 | 0.722 |
| Low | 0.688 | 0.713 | 0.739 |
| Student Gifted Population | | | |
| High | 0.679 | 0.706 | 0.732 |
| Low | 0.670 | 0.697 | 0.729 |
| Student Minority Population | | | |
| High | 0.660 | 0.686 | 0.718 |
| Low | 0.689 | 0.717 | 0.743 |

Table A13

*PPP Score Means, per School Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Total Enrollment | | | |
| High | 0.700 | 0.733 | 0.742 |
| Low | 0.688 | 0.738 | 0.750 |
| Student SPED Population | | | |
| High | 0.691 | 0.729 | 0.742 |
| Low | 0.697 | 0.742 | 0.749 |
| Student ELL Population | | | |
| High | 0.680 | 0.719 | 0.735 |
| Low | 0.707 | 0.750 | 0.756 |
| Student FRL Population | | | |
| High | 0.675 | 0.721 | 0.735 |
| Low | 0.712 | 0.749 | 0.755 |
| Student Gifted Population | | | |
| High | 0.702 | 0.738 | 0.747 |
| Low | 0.685 | 0.732 | 0.744 |
| Student Minority Population | | | |
| High | 0.678 | 0.718 | 0.729 |
| Low | 0.709 | 0.752 | 0.761 |

Table A14

*Survey Score Means, per School Subgroup*

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| Total Enrollment | | | |
| High | 0.770 | 0.806 | 0.794 |
| Low | 0.790 | 0.824 | 0.825 |
| Student SPED Population | | | |
| High | 0.781 | 0.820 | 0.814 |
| Low | 0.779 | 0.813 | 0.806 |
| Student ELL Population | | | |
| High | 0.780 | 0.819 | 0.823 |
| Low | 0.779 | 0.812 | 0.798 |
| Student FRL Population | | | |
| High | 0.786 | 0.817 | 0.820 |
| Low | 0.774 | 0.815 | 0.801 |
| Student Gifted Population | | | |
| High | 0.775 | 0.793 | 0.791 |
| Low | 0.784 | 0.830 | 0.829 |
| Student Minority Population | | | |
| High | 0.777 | 0.810 | 0.812 |
| Low | 0.782 | 0.822 | 0.807 |