

---

## RAYMA HARCHAR OUTSTANDING RESEARCH PAPER AWARD

### **The Significance of the Imbalance of Accountability Mandates**

*Lakesha N. Reese-Penn and Adam C. Elder  
Southeastern Louisiana University*

#### **Abstract**

Student achievement on standardized testing is a core component of accountability measures for teachers, schools, and districts nationwide, but extant research on this practice is inconclusive about its validity and differs depending on various state-level policies. This study examines how accountability outcomes vary for teachers and school districts in Louisiana using publicly available data. The findings showed that there is an imbalance in accountability outcomes for various stakeholders that warrants discussion in accountability policies and research.

*Keywords: accountability, teacher evaluation, value-added*

#### **Introduction**

There has been a conscientious effort to foster significant improvement in student achievement nationwide in the past two decades. This has been commissioned through federal accountability mandates such as No Child Left Behind (NCLB) and the Every Student Succeeds Act (ESSA), as well as incentivized in the Race to the Top (RTT) grant program. NCLB was implemented in 2002, and it focused on accountability systems and required standardized assessment of student performance 3rd through 8th grade (No Child Left Behind Act, 2001). RTT defined an overhaul of teacher evaluation to include performance measures and targets based primarily on student performance (Race to the Top Executive Summary, 2009). ESSA was passed in 2010 and replaced NCLB by shifting power from the federal level to the states, but it continued the premise of tracking student achievement across subgroups and measuring the success of schools and districts on the performance of teachers and students (Every Child Succeeds Act, 2015). In addition to implementing more stringent accountability expectations, these initiatives placed an increased focus on “at risk” student populations, which include students who are economically disadvantaged, minorities, special education, and English as a second language students.

These federal mandates have provided a framework of policies that state legislators and departments of education have enacted to overhaul and improve teacher evaluation and student performance. The state of Louisiana followed suit with a litany of educational reform legislation. The most prominent is Act 54, which was passed in 2010 and was enacted statewide in the 2011-2012 school year. This legislation required teacher evaluation to encompass both teacher performance and student performance, a uniformed evaluation tool, teachers’ professional ratings, and a high-stake policy that would be determined from teacher evaluation scores and student assessment results for teachers where both components were both weighted equally (La. Act 54, 2010).

The reformed model of Louisiana’s teacher evaluation system, called COMPASS, was highly focused on the components of student growth and professional practice. This mandated that teachers receive an evaluation score and rating that is calculated such that half of the score is determined by their classroom observation and the other half their value-added model (VAM) or student learning target (SLT) score. The first component of the COMPASS evaluation requires teachers be evaluated twice per year both formally and informally using the COMPASS framework for teaching rubric or another state-approved rubric that is then

converted to align with the COMPASS evaluation system. The second component of COMPASS requires teachers to submit two SLTs, which are academic goals focused on student achievement. The targets set on SLTs are rigorous and evaluative, and they are set based on student performance data from previous standardized state assessments or district benchmark assessments that mirror state assessments. Although all teachers were required to complete SLTs, the score was only used initially for teachers that were not evaluated using VAMs. VAMs are predictive models that forecast students' test scores based on a variety of factors that are used to evaluate a teacher's effectiveness. Since its implementation, the COMPASS system has undergone many revisions with new adjustments being implemented from modifications based on the implementation of ESSA. The various components of the evaluation are then compiled to assign a teacher one of four ratings that classify the educator's effectiveness as *Ineffective*, *Effective Emerging*, *Effective Proficient*, or *Highly Effective*; these ratings become a permanent part of a teachers' professional record.

Accountability structures in Louisiana are multifaceted. Schools and districts receive letter grade ratings based on student performance on state assessments in 3<sup>rd</sup> through 12<sup>th</sup> grade, as well as other relevant metrics such as student retention and graduation. Teacher ratings and letter grading have collectively created disparity between teachers who work at schools whose population consists of a higher percentage of minority students compared to a lower percentage of minority students. The rise in accountability mandates has placed increased pressure on principals, teachers, and students and has jeopardized teacher and student efficacy. The common practice has become hinging decisions about program effectiveness, student learning, student growth, teacher effectiveness, administrator effectiveness, school climate, and a host of other identifiers about public schooling on scores set from the performance of students on standardized assessments (Warlop, 2016). Act 54 was enacted with the premise that it would improve and properly align teacher instruction, adequately measure student performance, provide teachers with meaningful feedback to foster continuous improvement, and supply data to determine professional development for teacher support.

Assessments have created a standard evaluation of students, teachers, principals, and schools, whereas school letter grades are determined by student performance scores. The current system attaches a letter grade to determine the success of the school based on a formula and rating defined by the Louisiana Department of Education (LDOE). This practice implemented by the LDOE has based the success or the failure of schools and teachers largely on students' performance on the state assessments. The majority of the schools that have been labeled "struggling schools" (D and F schools) have an enrollment that consists of a higher percentage of minority students, socioeconomically disadvantaged students, and students with disabilities, relative to schools that have been labeled as "exemplar schools" (A and B schools) with a higher percentage of White students and lower percentages of socioeconomically disadvantaged students and students with disabilities.

The purpose of this study was to investigate if there is an imbalance in the accountability outcomes in Louisiana public schools based on demographic factors outside of student performance on standardized assessments. Specifically, the study was interested in examining whether there were disparities in elementary teachers' professional evaluation scores and in district performance scores based on school performance scores and various demographic factors. The research questions for this study were the following:

1. Is there a difference in the percent of teachers with good evaluation ratings at high-performing public elementary schools compared to low-performing public elementary schools in Louisiana?
2. Does student enrollment and the percentage of minority students in Louisiana's public school districts make a difference in district performance scores?

## Literature Review

Teachers differ considerably in their effectiveness to promote their students' academic achievement, and this variability in teacher effectiveness can be large (Nye, Konstanopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2003). Education policy over the past decade has shifted from emphasizing school performance to focusing on individual teachers (Goldhaber, 2015). Additionally, the present shift in accountability has placed most of the responsibility of the school performance measures on

teacher performance and their students' performance. Teachers working to close achievement gaps have varied contextual factors and school factors such as student demographics, faculty characteristics, class sizes, and available resources; therefore, it is of critical importance that researchers and policy makers not rely solely on value-added results (Franco & Seidel, 2014). Moreover, these factors consistently have been shown to have an impact on standardized test scores. Therefore, evaluation criteria that is confined to students' performance on standardized tests independent of "social, cultural, and economic context and [of] policies, practices, and resources of schools is unfair to teachers, administrators, students, and others because it holds them fully accountable for outcomes that they have limited power to produce" (Murray & Howe, 2017, p. 11). Murray and Howe (2017) further emphasized that restricting evaluation criteria exclusively to student academic performance hinders leaders' ability to ascertain how various policies and practices interplay with student outcomes—knowledge that is necessary to ultimately improve the school. A significant amount of research exists that explores teacher efficacy and its relationship with student achievement. The literature also highlights the ineffectiveness, inconsistency, and/or inequity of the use of high-stakes teacher evaluation measures including VAM and teacher observations.

VAM models have been utilized as a tool that identifies teacher performance or effectiveness with the ultimate objective of rewarding or penalizing teachers (Konstantopoulos, 2014). Research shows using teacher evaluation to inform high-stakes decisions is controversial and lacks consensus among the research community for the use of evaluation and decision-making (Goldhaber, 2014). The disagreement in part is the use of the statistical properties, the validity of a measurement of teacher performance, and the variations of valued-added measures throughout the grade and academic content (Goldhaber & Hausen, 2013; Kane, McCaffrey, Miller, & Staiger, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009) and has led to questions of the reliability for its use in high-stakes purposes (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). In addition to VAM, research has also shown that instruments used in classroom observations can be biased; therefore, these also pose issues with the validity of the rankings they assign teachers (Bell et al., 2012; Elder, 2016). The lingering questions in the literature related to the measures used to evaluate teachers and the documented disparities that present themselves with standardized testing demonstrates a need for research into possible inequities in state evaluation mandates.

## **Data and Analytic Approach**

The data for this study were collected from the publicly available LDOE data center website. This study focused exclusively on accountability data from the 2015-2016 school year because this was the last year of the hiatus for VAM that was implemented by the LDOE due to a change in the statewide standardized assessments. Furthermore, only elementary schools were considered for the first question in this study since comparing accountability outcomes across school levels could confound the results due to the differing methods of school evaluation by level in Louisiana. The data were pulled from various spreadsheets and merged into one dataset using the unique identifiers assigned to each school and district by the LDOE. Since the data that are made available to the public are only presented in aggregate form at either the school or district level, the units of analysis for this study were schools and districts for each research question, respectively.

### **Differences Between Low- and High-Rated Schools**

An independent samples *t*-test was conducted to analyze the data for the first research question. The outcome of interest for the analysis was the percentage of teachers at the school that scored a *Proficient* or *Highly Effective* evaluation rating on the COMPASS instrument. The independent variable was a dichotomous variable that indicates whether the school was classified as a high or low performing school. Schools were classified into the high performing category if they received an A or B rating on their annual report card for the 2015-2016 school year, and schools that received a D or F grade were classified as low performing. Schools that earned a C rating were excluded from the analysis because the focus for this question was to determine if there was a difference between high and low performing schools, not to examine schools with an "average" grade. In

total, 504 public elementary schools in Louisiana were included in the analysis. There were 335 schools that were classified as high performing and 169 that were classified as low performing.

### Differences Between Districts

A factorial ANOVA was utilized to analyze the data for the second research question. The outcome of interest was the district performance score (DPS), which is a numerical score assigned to districts based on a variety of measures including student performance on standardized assessments as well as retention and graduation rates. The first independent variable was the district size, which was grouped comparatively by size into three equal groups of districts that were classified as having either large (more than 8,400), medium (3,201 to 8,400), or small (less than 3,200) student enrollment. The second independent variable was the ethnic composition of the district. Specifically, districts were classified as having a low (0-33%), moderate (34-66%), or high (67-100%) percentage of minority students (defined as all non-White students). All 69 traditional school districts in the state were included in this analysis. Districts such as the Recovery School District and charter networks were excluded due to their nontraditional structure and governance.

### Findings

The results of the independent samples *t*-test showed that there was a statistically significant difference in the percentage of teachers receiving high evaluation ratings between high and low performing schools,  $t(190.53) = 8.95, p < .001$ . The variances between the two groups were heterogeneous ( $F = 164.10, p < .001$ ), so the degrees of freedom were adjusted using the Welch-Satterthwaite method. The effect size was large,  $d = 0.94$ , indicating schools identified as high performing ( $M = 97.12, SD = 5.27$ ) had a significantly larger proportion of teachers rated as effective compared to low performing schools ( $M = 86.77, SD = 14.57$ ).

The results of the factorial ANOVA showed there was not a significant interaction between the amount of minority students and the district size,  $F(4,60) = 0.23, p = .92$ . However, there were significant differences between categories for both the amount of minority students in the district,  $F(2,60) = 23.52, p < .001$ , as well as the district enrollment size,  $F(2,60) = 3.67, p = .03$ . A Tukey post hoc test was conducted for each of the main effects to determine which categories were significantly different. The means and standard deviations for each group are presented in Table 1. The Tukey post hoc tests indicated that there are significant differences in district performance scores when the amount of minority students is taken into account. Specifically, each of the classifications was statistically significantly different from the others ( $p < .01$ ). The results showed that districts with more minority students had lower district performance scores. There was not a statistically significant difference between districts with medium and large levels of enrollment ( $p = .64$ ), but small districts had statistically significantly lower district performance scores than medium ( $p = .01$ ) and large ( $p < .01$ ) districts.

Table 1. Means and Standard Deviations of District Performance Scores by Percentage of Minority Students and Enrollment Size

District Size	Percentage of Minority Students			
	Low	Moderate	High	Total
Small	97.6 (5.2)	83.8 (12.1)	70.5 (10.3)	81.6 (14.4)
Medium	98.8 (10.7)	91.4 (11.4)	75.8 (14.0)	90.7 (12.8)
Large	105.9 (5.0)	94.8 (10.0)	77.2 (4.9)	93.4 (12.7)
Total	100.9 (7.6)	90.8 (11.6)	73.4 (9.7)	88.5 (14.1)

## Discussion

This study found that disparate accountability ratings are being assigned in Louisiana's public schools. The results of the independent *t*-test showed that there is a significant difference in teachers' professional evaluation ratings between high and low performing elementary schools. The factorial ANOVA showed that significant differences exist in district performance scores across districts with varying amounts of minority students and enrollment counts. These differences were prevalent across all classifications of percentage of minority students, but significant differences between enrollment sizes were only found between the smallest one-third of districts and the largest two-thirds of districts in the state. These results have important implications for policy and practice when it comes to measuring accountability for teachers, schools, and districts in Louisiana.

According to Hanushek and Raymond (2005), accountability has tended to help White achievement more than Black achievement, and observed movement toward higher minority concentrations in schools has a detrimental effect on Black achievement, which pushes toward a wider distribution of achievement. Most schools that have highly rated teachers have more resources and are usually not located in urban areas. Human capital or high turnover is usually not an issue; therefore, it provides a greater effect on change and improvement of student academic performance. Principals are able to effectively support teachers more effortlessly, continuously, and frequently. For this reason, principal support is an important variable that should be included when discussing achievement in addition to more commonly measured attributes such as demographics, student ethnic composition, and school resources.

One of the overarching reasons for the implementation of COMPASS and Act 54 was to have teacher evaluation data be more closely reflective of student performance data. This performance data is gathered for elementary schools using student performance scores and SLTs for third through fifth grade and SLTs only for Pre-K through second grade. The difference in teacher evaluation ratings between schools furthers the idea that the COMPASS observation tool needs to be revisited and revised. In its current state, it is not inclusive of all teachers. Letter grading systems fail to validly measure and represent school quality, and they typically fail to drive or promote school improvement (Murray & Howe, 2017). The results of this study found that teachers who teach in schools that receive poor ratings using the current system are more likely to receive poor evaluation ratings.

This result is unsurprising when the fact that both school letter grades and teacher evaluations are largely tied to standardized test performance. However, the research has demonstrated the fallibility of this method, especially considering teachers can account for as little as one percent of a student's performance on standardized assessments (American Statistical Association, 2014). This is especially problematic when the stakes that are attached to these evaluation ratings are taken into consideration. Teachers are ultimately viewed, judged, and characterized by their evaluations, which are entered and stored into COMPASS Information System (CIS), a web-based system that allows the transparency of other school districts the ability to review teachers' evaluations for any employment opportunities. The results of this study showed that without the background context of school performance and student performance trends, as well as myriad contextual factors, this practice becomes maligned and more punitive toward the teacher, which could in turn affect their efficacy and contribute to a culture of attrition and stifled growth for potential teacher leaders. This demonstrates one way in which the extensive accountability mandates and measures in Louisiana have distorted and negatively highlighted the performance of disadvantaged schools, likely leading to decreased teacher and student efficacy, environments of stress and anxiety, weakened pedagogy, and routinized teaching.

These findings also suggest that school districts should be measured comparatively as analyzed in this study. District size and ethnic composition are just some of the factors that should be considered when reporting data publicly and using it to draw comparative conclusions about district performance. The Louisiana Department of Education can seek to improve the quality of student achievement and school improvement by revising, readjusting, and restructuring the current letter grade formula by tiering districts into levels to properly analyze students' performance. This would make reporting school and district performance more equitable.

## Limitations and Future Research

There are limitations of this study that are worth noting, and these limitations lend themselves to future research opportunities. The school-level analysis only looked at elementary schools in the state. Expanding to other levels in future studies could provide additional meaningful information about equity issues surrounding evaluation ratings in Louisiana. Additionally, since the results supported extant literature that says accountability outcomes differ across various demographic factors, future research should incorporate these factors. This study did not consider these factors at the school-level and only examined size and ethnic composition at the district level. Additional controls would better isolate the impacts of any factors that are of interest when examining teacher, school, and district evaluation.

## References

- American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA: Author.
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, 17(2-3), 62-87.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012, February 29). Evaluating teacher evaluation. *Education Next*. Retrieved from <http://www.edweek.org/ew/articles>
- Elder, A. C. (2016). Examining the equity of a teacher observation rubric: A mixed methods approach. *Research Issues in Contemporary Education*, 1(1), 33-42.
- Every Student Succeeds Act, S.1177, 114th Cong. (2015). Retrieved from [http://edworkforce.house.gov/uploadedfiles/every\\_student\\_succeeds\\_act\\_-\\_conference\\_report.pdf](http://edworkforce.house.gov/uploadedfiles/every_student_succeeds_act_-_conference_report.pdf) 2.
- Franco, M. & K. Seidel (2014) Evidence for the need to more closely examine school effects in value-added modeling and related accountability policies. *Education and Urban Society*, 46(1) 30-58. doi: 10.1177/0013124511432306.
- Goldhaber, D. (2015) Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, 44(2), 87-95 doi: 10.3102/0013189X15574905
- Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long term stability of estimated teacher performance. *Economica*, 80(319), 589-612.
- Hanushek, E., & Raymond, M. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297-327.
- Hanushek E., & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers and Proceedings*, 100, 267-271.
- Kane, T. J., McCaffrey, D., Miller, T., & Staiger, D. (2013, January). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (MET Project Research Paper) Seattle, WA: Bill and Melinda Gates Foundation.
- Konstantopoulos, S. (2014). Teacher effects, value added models, and accountability. *Teachers College Record*, 116, 1-21.
- La. Act 54, H. B. 1033 (2010).
- Lockwood, J., McCaffrey, D., Hamilton, L., Stecher, B., Le, V., & Martinez, J. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47-67.
- Murray, K., & Howe, K. R. (2017). Neglecting democracy in education policy: A-F school report card accountability systems. *Education Policy Analysis Archives*, 25(109), 1-31. <http://dx.doi.org/10.14507/epaa.25.3017>
- No Child Left Behind Act of 2001, Public Law 107-110. (2002). Retrieved from <http://www2.ed.gov/policy/elsec/leg/esea02/index.html>

- Rivkin, S., Hanshushek, E., & Kain J. (2005). Teachers, schools and academic achievement. *Econometrica*, 73, 417-458.
- Warlop, D.M. (2016). Threats to validity in accountability structures for public education. *Curriculum and Teaching Dialogue*, 18(1&2), 41-55.