Leadership Where It Matters: Principal Effectiveness and Equity in Wisconsin

Jeffery Dean
Department of Education Reform
College of Education and Health Professions
University of Arkansas

August 2012

Abstract

The importance of school leaders in the improvement of student achievement has long been discussed but little understood. A great deal of qualitative and conceptual research has been supported by too little solid empirical work. Particularly, researchers have had difficulty separating principal effects from overall school effectiveness, which may be influenced by factors beyond a principal's control. In recent years, research into principal effectiveness has improved through the examination of changes in school achievement under different principals.

I examine school achievement under principals in Wisconsin over a 13-year period (1998-99 to 2010-11). Three hypotheses are tested, building upon the existing research literature. I first examine the influence of principals on student achievement as measured by a model which incorporates changes in school leadership into typical value-added measures of school effectiveness. I find that a one standard deviation improvement in principal effectiveness is associated with a student achievement increase of 0.27 standard deviations in reading and 0.34 SDs in math. Second, I examine whether the importance of principals in the production of student achievement varies according to student population characteristics. Third, I examine whether highly effective principals are more likely to leave high-needs schools than are average and ineffective principals. I find that, while pay varies little between highly effective and ineffective principals, highly effective ones are more likely to seek out and obtain jobs in schools with wealthier and whiter student populations. Equity and policy implications are discussed.

This paper complements recent studies of principal effectiveness in a few ways. It draws upon a large sample of principals (nearly 1,500) over a longer period of time (eleven years), so its findings may have greater external validity. Further, it controls for the impact of teacher turnover and characteristics on student achievement. In terms of findings it largely aligns with prior studies.

Background

Over the last decade, the pressures of school accountability have changed how schools behave, as well as how both teachers and school leaders view their profession. For example, Li (2011) shows how the pressures of accountability help to discourage highly effective principals from staying in low-performing schools. While the merits and flaws of such accountability are endlessly debated, there is little question that the debate has shone a light on school and teacher effectiveness. Not only have researchers and policymakers become more interested in measuring effectiveness, but also in understanding what drives teachers and students to succeed or fail. Most of this effort has focused on teachers.

Teachers matter; this claim has been quantified, tested, and of course confirmed. But while we expect principals to likewise impact student achievement, albeit in different ways, they have received less attention. Despite extensive research in the last decade on the measurement and importance of teacher value-added measures (henceforth teacher VAMs), only since 2010 has much progress been made in estimating and analyzing principal effectiveness. The reasons for this may be various. At first glance, teachers stand to more directly influence student achievement than school leaders; teachers, after all, deliver instruction, maintain classroom order, and provide enrichment and remediation. The principal, insofar as he affects achievement, likely does so indirectly, by supporting teachers and providing the necessary conditions within which learning can occur. In addition to the perception that principal influences on students are indirect, methodological challenges have also hindered research into principal effects.

Researchers have proceeded with efforts to reliably measure teacher effectiveness, with much of the research focusing not only on the determinants of teacher effectiveness, but also on the impact of having a highly effective teacher. How well does teacher pay align with their value added? Where are these teachers found, and how might their distribution affect achievement gaps? Where do high (or low) value-added teachers have the greatest impact? What would the recruitment of high value-added teachers, or the removal of low ones, do to student achievement? Can they be distributed in a way that is Pareto improving, i.e. rearranging them so that some students benefit and none lose?

While this research has focused on the value of teachers, there has been relatively little research on environmental factors influencing teacher value added. Some of the most prominent research (Kane 2010) has focused on how well observable characteristics of individuals predict their effectiveness as instructors. Though this stands to be a helpful line of inquiry for the training and hiring of teachers, the question nonetheless ignores environmental factors influencing effectiveness. A surface-level knowledge of the literature may tempt one to think that teacher value added is strictly exogenous: teachers' effectiveness is self-determined, and their distribution need not affect their quality. Yet teachers not only influence students, but are themselves influenced by their environment. This statement may seem like a truism, but it has been under-examined in much of the value-added debates. Teachers fit better with some co-workers than others, whether socially, culturally, or pedagogically. Their experience with one group of students may not carry over to another group. They are variously influenced by their peers and superiors. They flourish or flounder depending on the support given them by their leadership and fellow teachers.

This last concern feeds into the subject of the current paper, the effect of principals on student learning. While students are ultimately taught by teachers, principals affect student achievement indirectly in countless ways, most of which go through the teacher (Hallinger & Heck 1998). The principal has the responsibility to create a safe and stable working environment. He also has the job of hiring teachers who fit well together, often one of the prime concerns of principals. He must facilitate cooperation within and across curricula, as well as professional development. He must see that each teacher has all the tools necessary to teach in the best way possible. Each of these responsibilities is distinct, yet collectively they may form the most important environmental factor influencing teacher effectiveness. Thus while principals themselves are rarely in a given classroom, they conceivably have the potential to greatly influence student achievement in subtler ways for more children: where a teacher directly influences achievement in a single classroom, a principal is likely to have an indirect influence across many classrooms.

Prior Research on the Determinants of Principal Effectiveness

As long as principals have autonomy, whether to succeed or fail, this effectiveness is likely significant, if for no other reason than mere plausibility: a major part of what teachers need to succeed is provided through a principal's leadership.

Hallinger and Heck (1998) summarized research on principal effects since 1980 into empirical, conceptual, and methodological factors. Empirical factors concern what data are available, especially prior to the advent of widespread accountability testing in the late 1990s and early 2000s. Conceptual questions include *how* principals may influence student achievement: whether through staffing decisions, motivation, discipline, or other factors over which a principal may have influence. Methodological factors concern how to isolate and estimate principal effects from non-principal factors influencing student achievement. Most importantly, it is difficult to attribute school effects to principals; many school factors influencing principals are either beyond the control of principals (parental involvement, perhaps teacher retirement or in some cases teacher quality), or are inherited by the principal upon arrival in the school (e.g. prior teacher quality and experience). Summarizing the research on principal effectiveness up to 1998, Hallinger and Heck found principals' influence to be indirect, affecting achievement through principals' effects on teachers. They found consistent evidence that principals likely did not influence achievement apart from teachers.

Before Hallinger and Heck, Pitner (1988) attempted to conceptualize how principals may affect student achievement. Pitner distinguished effects which were antecedent to principal effects from intervening effects, through which principal leadership may ultimately influence student achievement. He proposed a variety of non-experimental models through which principal effects could be estimated as well as understood: models in which antecedent effects (such as teacher and student demographics) may affect both principal leadership and student achievement, models in which antecedent effects are ignored in favor of examining intervening effects, and models which mix both antecedent and intervening effects. Lastly, Pitner examined whether changes in student achievement may independently influence principal leadership—a concern which, if true, complicates estimation of principal effects. Pitner's conceptual work pointed forward to much of the research done in recent years regarding principal effects as measured through value-added models.

Dhuey and Smith (2011) examined principal value added in British Columbia, finding the effect size of principal effectiveness to be about 0.20 standard deviations in both reading and math. This implies that a one-standard deviation improvement in principal effectiveness (relative to all other principals) increases school-wide student achievement by one-fifth of a standard deviation, where variation is measured against student-level achievement rather than school-level averages. Dhuey's finding of highly significant principal effects in British Columbia harmonizes with other recent findings by Branch and Hanushek (2012) in Texas, who found a principal effect size of 0.11 in math.

Li (2011) likewise found significant principal effects in North Carolina. Li's analysis of principal effects found that, subsequent to the introduction of academic sanctions as a result of No Child Left Behind, highly effective principals in schools serving disadvantaged students were induced by the policy to move to higher-achieving (and typically wealthier) schools. Key to this problem was the author's finding that districts had little flexibility to increase principal compensation as a means of offsetting this policy inducement. The suggestion was that principals, on average, behaved rationally, seeking out the best available combination of pay and working conditions; if job pressure disproportionately increased at schools at under greater threat of sanction, without some compensating increase in pay, they would seek out better conditions with roughly equal pay elsewhere.

Where effectiveness matters most also likely varies depending on school characteristics. Branch and Hanushek (2011) find that principal effectiveness varies more in low-income and low-achieving schools, consistent with findings on teacher effectiveness. A third question on principal effectiveness is its relationship to school fit. This question is not one merely of selection, where principals tend to sort themselves based upon their own preferences. Rather, "fit" concerns how principal effectiveness itself changes depending on the characteristics of the school he or she leads. For example, some principals may be effective in schools with few minorities but ineffective in a high-minority population, independently of how demographics independently influence achievement. Dhuey (2011) uses "match" or "spell" effects—indicator variables for principal-school combinations—to investigate this relationship, finding that roughly half of a principal's effectiveness is due to the match between principal and school. These findings suggest that principal effects are not fixed over time, nor across schools in the case of switchers. Here as elsewhere, an individual's effectiveness is ecological, both a cause and an effect. These connected findings show that as the subtlety of value added models grow, the researcher increasingly realizes how student achievement depends upon a web of influences and effects. In other words, no individual is exogenous—even if he's a principal.

What does prior research say on the relationship between principal value added and their observables? Here again, much of the findings are consistent with those on teacher value added, though still tentative. According to Dhuey (2011), principal experience matters in the early years, but principals' effectiveness plateaus between 3-5 years on the job. Independently of experience, Dhuey also finds that how a long a principal has been in a school also shows little relationship to effectiveness; a principal hired from within is as likely to be effective as one hired from outside the school. Earlier research by Rockoff et al. (2009) in New York accords with Dhuey's research, finding little or no impact of highest degree (assuming at least a four-year bachelor's degree), total experience, or total pay on principal effectiveness. They find that prior experience as a principal matters in the first years of one's tenure, but afterwards levels off. In this light, findings from Clotfelter & Ladd (2006) that principals in lower-achieving schools often have lower qualifications appear less problematic.

Challenges

As Lipscomb (2010) points out, much of the early research on principal value added assumed that principal effects and school effects were one and the same. If a school was effective, then so was the school's principal, and vice versa. Many of these early models, such as those used in the Dallas Independent School District and in Tennessee, simply attempted to control for observable characteristics of each school to account for non-principal factors influencing student achievement. Yet anyone who has spent time working in or visiting schools can attest that observable characteristics often capture only a small part of what makes a school successful. Consider two schools with equal percentages of minorities and low-income students. One may have a very strong Parent-Teacher Organization while the other has none. One may by chance have a dedicated core of teachers, and the other an inexperienced or fractious staff. One may have a tradition of high expectations, and the other a habit of excusing failure. While all these factors are certainly related to demographics, the relationship is not one-to-one. Thus controlling for demographics doesn't account for every non-principal factor related to a school's effectiveness, and it becomes necessary to control for school unobservables through the use of dummy variables for each school.

In determining whether to estimate effectiveness using either levels or changes in test scores, one must consider the nature of the counterfactual: what would be happening in a school without the principal it has? Most likely, an effective principal will yield a short-term increase in student test scores, after which her school will remain at the same high level—higher than it would be without the principal, but not showing a relative increase year-to-year. If this is the nature of the counterfactual, then expecting effective principals to produce unusual gains in test scores year after year is unrealistic. I argue that the counterfactual is better captured by estimating the difference in each year between the level of a school's achievement under a given principal and the level a school would be achieving without that given principal.

Using levels of achievement rather than growth also recommends itself in other ways. One is that, if I have correctly described the counterfactual above, principal estimates using levels will be less noisy than estimates using growth. Second, using z-scores for growth introduces the threat of reversion to the mean, in which low-scoring schools are likely to see higher gains and high-scoring schools are likely to see drops in achievement, as a statistical matter. Third, growth estimates using state accountability tests require two years of data. This requirement significantly reduces the frequency of connections, and therefore the size of principal networks which are necessary for estimating principal effects.

Method

Keeping all of the above factors in mind, I adopt a model of achievement which controls for student demographics, school staff characteristics, test idiosyncrasies, and school unobservables to estimate principal effectiveness. The model is thus:

$$Y_{gspt} = \beta X_{gst} + \gamma S_s + \delta P_p + \varepsilon_{gspt}$$

where Y_{gspt} is the current average test score for grade *g* in school *s* with principal *p* at year *t*, X_{gst} is a vector of school demographics, S_s is a vector of school dummy variables intended to capture school unobservables, P_p is a vector of principal dummy variables intended to capture principal effects, and ε_{gspt} is remaining random error in test scores. The model also controls for teacher experience, turnover, and characteristics. To account for irregularities in testing across grades and unexpected sample-wide changes in test scores across years, dummy variables are added for grades and years:

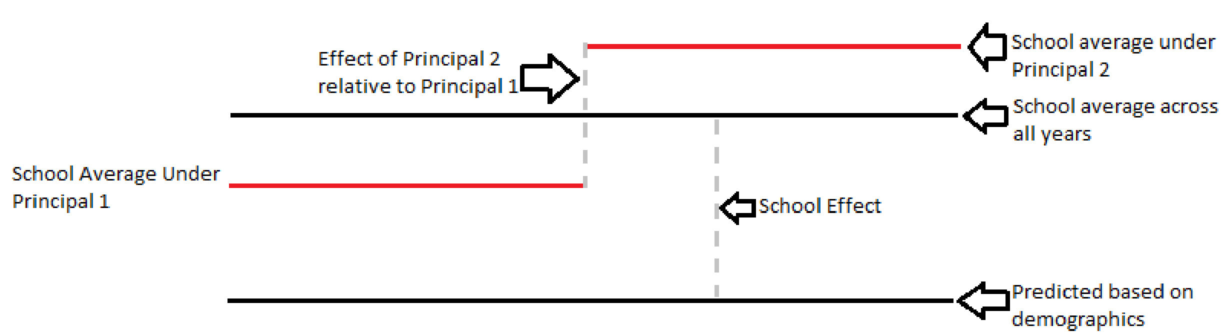
$$Y_{it} = \alpha Y_{i,p-1} + \beta X_{it} + \gamma S_{it} + \delta P_i + \mu T_t + \rho G_i + \varepsilon_{it}$$

where T_t is a vector of year dummies and G_i is a vector of grade dummies. The last two vectors are included exclusively for the purpose of controlling for irregularities and shocks.

Whether to control for lagged achievement in every year remains a major question in the econometric and applied research literature on fixed effects estimation. Angrist and Pischke (2009) warn that using both a fixed effect and a lag is likely to introduce bias into one's model. They advise using lags and fixed effects separately to set bounds on model estimates. However, in addition to the threat of bias, concerns over interpretation bear upon the decision of whether and how to use lagged scores. In particular, the inclusion of lagged scores implies a different interpretation on coefficients in the model. Rather than estimating the effect of covariates on the *level* of the dependent variable, one is estimating effects on *changes* in the dependent variable, because its lag is now on the righthand side of the equation. If the dependent variable is something that should be expected to change over time, the inclusion of lags may be desirable, with relatively straightforward interpretation of coefficients. However, the standardization of test scores within grades and years in the model used here means that the mean of the dependent variable (zero by definition) remains stable over time. As a result, the inclusion of lags, especially over many periods, is more likely to tell us something about regression to the mean than it is to shed light on persistent principal effectiveness.

Consistent with Li (2010) and Chiang (2012), the intentional exclusion of lagged scores means that the interpretation of principal effects is cumulative, rather than sequential. This interpretation is reasonable if one considers the nature of the counterfactual. Suppose a principal enters a school, at which time the school sees a sizable jump in achievement relative to similar schools. In subsequent years, this advantage is maintained but does not grow. Should one then conclude that the principal is no longer effective? Hardly; the counterfactual remains that the school would likely be performing lower without that principal, even if it does not continue to "distance" itself from its peer schools. Figure 1 crudely models the estimation of principal effects without lags, controlling for school observables as well as school fixed effects. The model assumes that principal effectiveness is time-invariant. This assumption is questionable, though for the sake of simplicity and illustration it is useful.[1] Principal effectiveness is considered to be the difference at school s between test scores under principal 1 and principal 2, after controlling for time-varying demographics and a school fixed effect.

Figure 1. Illustration of Principal Effects Estimation



A school fixed effect *S* is employed to account for time-invariant school unobservables. Demographics are observable, so unobservables in the current model are everything except demographics. There are a broad range of factors which may influence a school's effectiveness, from a tradition of achievement or a strong Parent-Teacher Organization to a highly committed staff. The fixed effect absorbs any of these factors which do not change over time. The model may thus be insufficient if there are non-demographic factors systematically related to school achievement which are time-varying. Though some of these factors may be impossible to control for, future extensions of the model may introduce controls for such observables as staff turnover.

The parameter of interest in the original equation is δ, which here is represented by the distance between the two red lines representing the effectiveness of two principals. Each individual P_i gets his or her own coefficient in δ. If effective controls are made elsewhere for prior test scores, demographics, school unobservables, and systemic irregularities, then δ for each principal should estimate that principal's effectiveness. In the value added literature, there have been two ways so far of estimating individual effectiveness, whether of teachers or principals. The first way is to regress on test scores everything unrelated to the teacher or principal in question, and then use the residual as a measure of effectiveness. The second way is employed here: rather than use residuals as measures of effectiveness, include dummy variables for each individual and rank them based on the coefficients on each dummy in the estimated model. The main advantage of this method is that it allows inference on the coefficients. Statistical significance can be established, whereas error terms by definition are not subject to inference.

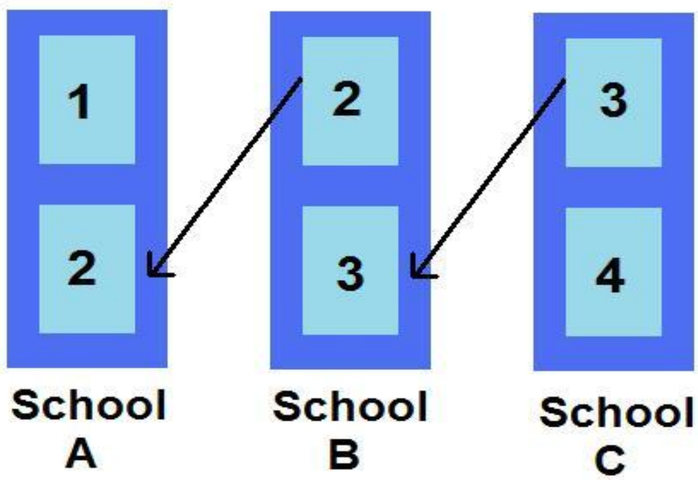
For Whom Are Estimates Possible?

The most significant limitation of the model used here and elsewhere for principal effectiveness is that not all principals can be compared to each other. In order to disentangle school and principal effects, a change in leadership must be observed in a school. This change provides the best quasi-experimental counterfactual available if principal estimates are to be meaningful. The implication of this necessity is that effectiveness for principals observed in a school across all years in the study period cannot be estimated. Principals for whom an estimation are available fall into two categories: principals who switch schools in the study period ("switchers"), and principals observed in one school for a portion of the study period, with a change in leadership prior or subsequent to his tenure. To have an estimate of effectiveness, a principal must work in a school that sees a change in leadership at some point during the study period.

To accomplish this task, this paper borrows an estimation routine from Mihaly et al. (2010). This routine, designed originally to estimate the effectiveness of workers who switch employers, has been adapted previously by Dhuey and Smith (2011) for estimating principal effectiveness.

Because estimates are available only for principals observed in a school with a leadership change, estimates of effectiveness are relative to other principals. Specifically, a principal can be compared on effectiveness only if she is observed in the same network of school switchers, whether she is a switcher or is only observed in a single school in which a switcher also has worked.

Figure 2. Principal Switching



This relationship is shown above in Figure 2, which illustrates a network of principals for whom relative effectiveness can be estimated. Not only can principals 1 and 4 be estimated relative to principals 2 and 3 even though they are not switchers themselves, but perhaps more importantly, principals who never work in the same school can be compared. For example, the relationship between principals 1 and 3 can be thought of in terms of the transitive property. If principal 1 is more effective than principal 2, and 2 is more effective than 3, then principal 1 must be more effective than principal 3. In this way, estimates for many principals can be meaningfully obtained if large switching networks are observed in the study.

Once estimated, principals' effectiveness can be ranked within networks, examined for influence on scores across their range, and jointly estimated for how well they explain variation in test scores. Principals not in the same network cannot be compared, but there may be similar patterns distinguishing effective principals in different networks. Rather than try to directly compare all principals, one can instead ask questions about the characteristics of effective principals across networks without knowing which networks are more effective on average.

Data

Data for this paper are taken from rich panel data on staffing and test scores in Wisconsin. Wisconsin staffing data are available from the 1994-95 school year until 2010-11. These data include not only individual identifiers (name, year of birth, sex, school), but also rich data on factors which may influence effectiveness, such as experience, degree levels, pay, and prior positions held.

Wisconsin achievement data are available at the grade level from the Wisconsin Knowledge and Concepts Examination (WKCE). The WKCE has been given annually since 1992, and has periodically undergone substantial revisions. Changes in the test must be addressed in the selection and modeling of achievement data for the present study. If test data are insufficient to allow at least a school-level comparison across years, then they must be excluded from the study. For this reason, only test scores from 1998-99 onward are used here. Because of this limitation, the merged data set of achievement, staffing, and demographic data covers the years 1998-99 to 2008-09, with the latest year determined by the availability of demographic data from the US Common Core of Data.

Within the period for which comparable scores are available, the test still underwent re-scaling and revision, as well as expanding from testing in grades 4, 8, and 10, to fully include grades 3-8 as well as 10. Adjustments must be made for this as well.[2] This problem is approached in two ways. First, all achievement measures (math and reading) are converted to a standardized z-score, for which the mean is 0 and the standard deviation is 1. The distribution over which test scores are standardized is the student-level distribution for the entire state of Wisconsin, available yearly by grade and subject. Therefore a change of 1 standard deviation for a school is not relative to other schools, but to students individually. This is important for interpreting results. The distribution of student scores for

[2] One threat posed by this is that because the units of analysis are grades within schools, later years (from 2002-03 onward) have more observations per principal-year and thus principal tenure in these years gets greater weight than in the first four years of the study. This shortcoming remains unaddressed in the current draft of the paper.

the whole state is broader than the distribution of school averages across the state. This is simply a statistical property of randomness: greater sample sizes lead to smaller sampling variances, all else equal. Thus a change of 0.1 for a school may seem small, but compared to other schools this change may be large. Converting school test score averages to z-scores nonetheless helps account for the re-scaling of the test. School averages across years are always relative to the state mean, whether that mean is represented by a scale score of 250 or 700.

The second way to account for changes in the test is to include dummies for each year and grade in the estimation. This is done simply as a check upon how well z-scores account for changes in the test, as well as changes in true achievement, over time. This is specified in the model as μT_t and ρG_i. Combined with z-scores of achievement, these regressors should help prevent time or grade bias on principal effectiveness.[3]

To control for demographics, school-level data are obtained from the US Common Core of Data. These data represent the most reliable measure of school characteristics available to the public. Since achievement data are used at the grade level, demographics are also obtained by grade rather than school. Because of the limitations of achievement data in the 1990s and the waiting period for most recent demographic data from the Common Core of Data, the effective period for estimation of principal value added in Wisconsin is from the 1998-99 school year until 2008-09, representing eleven years of data on achievement, staffing, and demographics.

Students moving into or out of a school just prior to testing has been and remains a major concern of principals and teachers, who commonly worry that such students might skew classroom and school achievement. Thankfully, Wisconsin provides test score data for both switching and non-switching students. To address the switching threat in this study, only students who were in the school in which they were tested from the beginning of the school year are included.

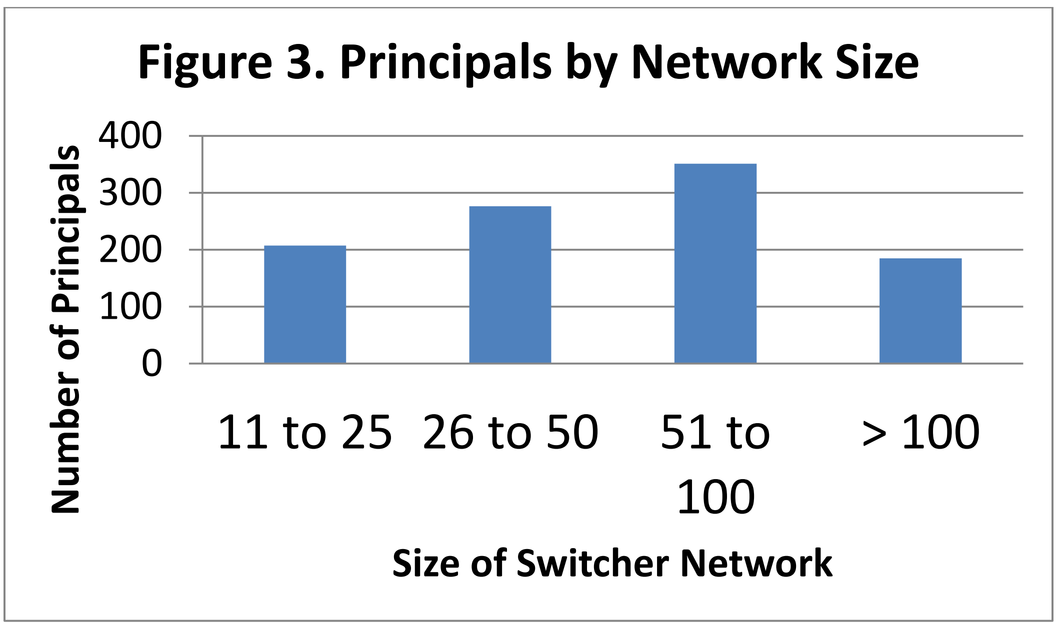
Descriptives on Wisconsin Principals

Over the study period of eleven years, 3,780 principals are observed in Wisconsin, serving in schools which have at least one tested grade. Many of these principals are observed working in the same

[3] Standardized scores would render controls for grade and year unnecessary if the sample over which standardization occurred was coextensive with the analysis sample. However, Wisconsin test scores are standardized based on statewide performance, while the analysis sample includes only schools with principals observed in switching networks, which is about one half of Wisconsin schools in tested grades.

school for the full duration, so are excluded from the study. Of all principals observed in Wisconsin schools with tested grades, 1,919 principals (50.1%) are observed in a switching network of at least two principals. Very small networks of principals may be idiosyncratic, so the analysis presented here only employs networks with more than ten principals. 1,375 principals (36%) appear in such a network, allowing more robust estimation of their relative effectiveness. This number compares favorably to prior studies. These 1,375 principals are observed in 42 distinct networks of switchers.

The size of the networks varies greatly, from the minimum of eleven to as many as 185 principals. The largest network is located in and around Milwaukee, in a network of schools which are low-performing and economically disadvantaged; the existence of such a network in these schools is unsurprising given an expectation of high leadership turnover. The distribution of principals by network size is given in Figure 3.



Results on Principal Effectiveness

Results using the model specified above are estimated on three outcomes: reading achievement, math achievement, and the average of reading and math achievement ("total achievement").

Table 1. Summary of Coefficients on Principal Effectiveness (z)

	10th percentile	25th percentile	75th percentile	90th percentile
Total Achievement	**-0.33***	-0.13	0.15	**0.29***

Reading	**-0.32***	-0.13	0.15	**0.28***
Math	**-0.35***	-0.16	0.16	**0.34***

* 95% significance

Similar across all subjects, impacts of principal effectiveness on achievement are large and significant at the margins, for the top and bottom 10% of principals as measured by their impact on reading and math achievement. Principals appear to have a slightly greater impact on math achievement than on reading. The difference between having a principal at the 10^{th} percentile and at the 90^{th} percentile is over half a standard deviation relative to student achievement statewide. This implies that switching a 10^{th} percentile principal for one at the 90^{th} percentile would have the effect of raising the achievement of every student in the school by an average of 22 percentile points (one standard deviation = 36 percentile points).[4] Even moving from a 25^{th} percentile principal to the median (50^{th} percentile) principal would increase the achievement of the average student by 5 percentile points (36 percentile points x 0.13 = 5). This effect is large by any comparison to other policy levers, though not as large as the effect of demographics on school achievement.

Roughly 1/3 of the principals in the sample had effectiveness coefficients which were significantly different from zero with 95% confidence. These principals were equally distributed at the lower and upper ends of effectiveness, so that the top 1/6 of principals were significantly effective, and the bottom 1/6 of principals were significantly ineffective.

Analysis of a Network: Milwaukee

The comparability of principals across switching networks is an open question which remains to be addressed in extensions of the current paper. Therefore for a conservative treatment of effectiveness I focus further analysis on principals in the largest switching network, containing 185 principals in and around Milwaukee during the study period. I compare principal effectiveness to both individual characteristics of principals and to characteristics of schools in which principals work.

These 185 principals were categorized three ways: significantly effective (positive), significantly ineffective (negative), and no difference from the average principal (null). It must be noted that a "null" principal is not one who has no effect on student achievement, but is simply a principal with

[4] The practicability of the suggestion is uncertain, and this statement should only be understood in a statistical sense. Probably, principal effectiveness is not itself exogenous and should be considered in light of the fit between principal and school.

average effectiveness compared to other principals. As shown in Figure 5, no measurable personal characteristic uniquely distinguished highly effective principals, but there were similar trends for both effective and ineffective principals compared to average principals. Average principals were more likely to be male and white, while both effective and ineffective ones were more likely to be female and non-white. There was no notable relationship between effectiveness and either experience or pay: mean experience across all three groups was around 18 years as either a principal or teacher, and salary averaged just below $85,000 annually in each group.

Figure 5. Principal Characteristics by Effectiveness

	Negative	Null	Positive
Male	38%	58%	23%
White	36%	59%	33%
Experience (yrs)	17.8	18.8	17.4
Average Salary	$84,332	$84,754	$84,379

Whether highly effective principals are more likely to be found in otherwise advantaged or disadvantaged schools may help shed light on the determinants of educational inequality. Figure 6 shows characteristics for schools served by highly effective, average, and ineffective principals. Because the principals observed in this network serve in an area that is low-achieving on average, it is not surprising that average math achievement for the three groups of principals is well below the state average. Yet even within this network of low schools, ineffective and highly effective principals are both more likely to be found in lower-achieving schools. This suggests that principals make a greater difference the lower is a school's achievement level, whether for better or worse. Consistent with this achievement pattern, both highly effective and ineffective principals are more likely to be found in schools with high concentrations of African-Americans, a traditionally disadvantaged student population. By contrast, Hispanic students are as likely to have an average principal as an effective or ineffective one. Perhaps most noteworthy in Fig. 6 is that highly effective principals tend to sort themselves into schools with lower poverty rates as measured by students' free and reduced lunch eligibility. This finding is discussed in more detail in the next section.

Figure 6. School Characteristics by Principal Effectiveness			
	Negative	Null	Positive
Math (pctile)	18th	29th	20th
Free Lunch (%)	76%	63%	33%
African Am. (%)	62%	50%	68%
Hispanic (%)	13%	13%	11%

Discussion of Findings

The finding of virtually no relationship between individual characteristics and principal effectiveness is consistent with prior research on principal effectiveness (Loeb et al. 2009; Branch et al. 2011). It is also consistent with the literature on teacher value added going back to Hanushek & Rivkin 2004. This consistent finding has been a source of frustration to researchers and policymakers, as the lack of relationship between supposed "inputs" on which principals are paid (experience and education) and the "output" of effectiveness implies inefficiency in how they are paid. Still, this does not imply that principals should be paid based on their value-added measures, which here and elsewhere remain very tentative and imperfect. It simply suggests that paying principals on these characteristics isn't likely to yield a better principal workforce.

Consistent also with Loeb (2009) and Branch (2011) is the finding that principals make a greater difference in high-needs schools than in average schools. This is supported by the tendency of both highly effective and ineffective principals to be found in these schools. In considering this, one need not assume that principal effectiveness is a given, and that both good and bad principals somehow sort themselves into these types of schools. On the contrary, it may be the case that a challenging environment on its own is more likely to elicit either success or failure on the part of principals, and that these principals' effectiveness may change if they instead worked in a higher-achieving suburban school.

Uniquely in the nascent literature on principal effectiveness, highly effective principals in and around Milwaukee are much more likely to be found in lower-poverty schools than are average or ineffective principals. It is unclear from the model as currently specified whether this is due to sorting or to fit. The former would hold that good principals know they're good, so that even if they can't get a pay raise for being good (see Fig. 5), they tend to flock to schools with better "working conditions", so to speak. This is a testable hypothesis, though presently beyond the scope of this paper, and it would serve as an estimation of what in labor economics is known as a compensating differential. If this is true, then it is plausible that pay inflexibility contributes to educational inequality as good principals flee the schools where they are most needed.

Caveats

A few caveats are in order when considering the implications of this study. First, the study lacks student-level data. This poses two potential threats: students switching schools during the school year, and nonrandom sorting of students into schools and/or classrooms. Though Wisconsin data allow the exclusion of switchers from achievement figures, thereby addressing the first concern, the question of non-random sorting cannot be handled with available data. It remains simply as an asterisk upon the study, and in fact upon many studies of teacher and principal value added.

Second, there remains the possibility that estimates of principal effectiveness are biased outward slightly, with ineffective principals biased downward and effective ones biased upward. Principal effects are the sum of true effects and some random intertemporal variation across all years and grades for which a principal is observed. To correct for this, principal estimates can be reduced using a technique called empirical Bayes shrinkage, employed in prior literature by Dhuey (2011) and Branch (2011). This adjustment has yet to be included in the current analysis.

Third, it is reasonable to expect that principal effectiveness may be low in the first years of a principal's tenure, though the exact structure of such a "learning curve" is uncertain. Nonetheless, the current model does not control for principal experience or tenure in a current school, either linearly or nonparametrically through the use of dummy variables. This will be a future embellishment of the model.

Summary & Extensions

Results suggest that principal effectiveness has a large impact on student achievement. Through their effect on a great number of students, a single effective principal can on average affect student achievement more than four highly effective teachers. Their effectiveness is not simply a reflection of the schools where they work. Because observable characteristics of principals do not distinguish effective ones from ineffective ones, these findings are not useful for recruiting or promoting individuals into the principal workforce and thereby raising achievement in the aggregate. On the other hand, the simple fact that a quality distribution exists and significantly influences achievement could be useful in closing achievement gaps, whether racial, socioeconomic, or regional. This has been a major policy concern of educators over the last decade, and compared to other remedies, the recruitment of effective principals to disadvantaged schools has a promise that is statistical at the least. Finally, as a note of caution, such redistribution may come at a price if Dhuey (2011) is to be believed, namely, that a majority of a principal's effectiveness is attributable to his or her "fit" with a particular school. Thus redistributing principals to close achievement gaps may (or may not) harm effectiveness through the disruption of fit.

This paper will be extended in three ways: inclusion of controls for principal experience and tenure as well as staff turnover, extension of the analysis of determinants of effectiveness beyond the Milwaukee switching network, and analysis of the temporal switching patterns of highly effective principals. Lastly, estimates of principal effectiveness will be adjusted by empirical Bayes shrinkage to better account for random error.

References

Angrist, Joshua and Jorn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2009, 392 pp.

Branch, Gregory, Eric Hanushek and Stephen Rivkin. *Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals*. Working paper no. 66, National Center for Analysis of Longitudinal Data in Education Research (CALDER), January 2012.

Chiang, Hanley, Stephen Lipscomb and Brian Gill. *Is School Value Added Indicative of Principal Quality?* Working paper, Mathematica Policy Research, February 2012.

Clark, Damon, Jonah Rockoff and Paco Martorell. School *Principals and School Performance*. Working paper no. 38, National Center for Analysis of Longitudinal Data in Education Research (CALDER), December 2009.

Clotfelter, Charles T., Ladd, et al. *High Poverty Schools and the Distribution of Teachers and Principals*. Working paper no. 1, National Center for Analysis of Longitudinal Data in Education Research (CALDER), March 2007.

Dhuey, Elizabeth and Justin Smith. *How Important Are School Principals in the Production of Student Achievement?* Working paper, University of Toronto. November 2011.

Li, Danielle. *School Accountability and Principal Mobility: How No Child Left Behind Affects the Allocation of School Leaders*. MIT Working Paper, October 2011. Available http://economics.mit.edu/files/5437

Lipscomb, Stephen et al. *Teacher and Principal Value Added: Research Findings and Implementation Practices Final Report*. Mathematica Policy Research, Inc. September 2010.

Loeb, Susanna, Eileen Horng and Demetra Kalogrides. *Principal Preferences and the Unequal Distribution of Principals Across Schools*. Working paper no. 36, National Center for Analysis of Longitudinal Data in Education Research (CALDER), December 2009.

Mihaly, Kata et al. "Centering and Reference Groups for Estimates of Fixed Effects: Modifications to felsdvreg." Winter 2010: *The STATA Journal*, vol. 10, no. 1., pp. 82-103.