Toward Addressing the Issues of Site Selection in District Effectiveness Research: A 2-Level Hierarchical Linear Growth Model[1]

Alex J. Bowers[2,3]
Teachers College, Columbia University

ABSTRACT

Purpose
District effectiveness research (DER) is an emerging field concerned with identifying the organizational structures, administration, and leadership practices at the school district level that help districts find success with all of their students across the schools within the system. This work has mirrored much of the early school effectiveness research (SER). However, to date across the DER literature, site selection for in-depth studies of districts deemed "effective" has been haphazard and nonsystematic. This is problematic given the long history of critiques centered on site selection in SER. The purpose of this study is to address and adapt the critiques from SER to a method of site selection for DER and test the method using a large multi-year dataset to identify districts that are significantly unusual and effective.

Research Methods
A 2-level hierarchical linear growth model which nests multiple time points per district (level 1) within districts (level 2) was used to predict gains in district achievement for all school districts in the state of Ohio during a seven-year period, 2001-02 through 2007-08.

Findings
Districts that statistically significantly outperformed their predicted gains in achievement, controlling for background and demographic variables over the period are identified as possible sites for in-depth qualitative studies for DER in comparison to districts performing at the norm.

Implications for Research and Practice
This study proposes and tests a method for district identification in DER that addresses the critiques from SER through controlling for achievement covariates, modeling district gains over time, and examining the population of districts within an entire state.

Keywords: School Districts, District Effectiveness Research, School Effectiveness Research, Hierarchical Linear Modeling, Growth Modeling

INTRODUCTION

District effectiveness research (DER) has recently come to the fore as researchers work to identify and understand the roles, practices, and leadership models of effective school districts that find success with the vast majority of their students (Bowers, 2008; Hightower, Knapp, Marsh, & McLaughlin, 2002; Honig & Coburn, 2008; McLaughlin & Talbert, 2002; Murphy & Hallinger, 1988; Opfer, Henry, & Mashburn, 2008; Rorrer, Skrla, & Scheurich, 2008). This has been in response to the longstanding debate (Purkey & Smith, 1985) about the roles, policies and functions of districts given the findings from school effectiveness research (SER). In a search to identify how effective districts achieve and maintain high performance, DER researchers have mirrored the early SER literature by conducting both surveys and deep qualitative studies of school districts to understand and describe the leadership and management practices of administrators in an effort to identify practices that could be beneficial to other districts (Cuban, 1984; Dailey et al., 2005; Opfer et al., 2008; Rorrer et al., 2008). These DER studies have identified many factors that appear to be consistent across these district studies including consistent and community-wide support for district initiatives, a focus on continuous improvement and coherent sustained professional development, the goal to improve instruction through a core focus on instructional leadership, and an ability to bring together district resources in service to improving instruction, among others (Bowers, 2008; Cuban & Usdan, 2003; Elmore, 2003; Elmore & Burney, 1999; Firestone, Mangin, Martinez, & Polovsky, 2005; Hannaway & Stanislawski, 2005; Hightower & McLaughlin, 2005; Murphy & Hallinger, 1988; Opfer et al., 2008; Rorrer et al., 2008; Stein & D'Amico, 2002; Supovitz, 2006). However, as with the early SER literature, the question of site selection for district effectiveness studies has gone largely unaddressed.

The vast majority of research around school districts has historically detailed the broad roles of districts in education. As recently detailed in a thorough review of the district literature (Rorrer et al., 2008), district research has focused on many issues. These issues have included, but are not limited to, the role of districts in school reform and instructional improvement, the creation and implementation of policy initiatives, descriptions of district organizational and finance functions, as well as the role of district central offices and individual's roles such as the superintendent and school board members, among many others (Rorrer et al., 2008). District effectiveness research (DER) is a subset of this more general school district research and, over the

---

[1] This document is a preprint of an article published in 2010 in *Educational Administration Quarterly*. Recommended Citation: Bowers, A.J. (2010) Toward Addressing the Issues of Site Selection in District Effectiveness Research: A 2-Level Hierarchical Linear Growth Model. *Educational Administration Quarterly*,46(3), 395-425.doi:10.1177/0013161X10375271.

[2] Teachers College, Columbia University; Bowers@tc.edu; 525 W. 120th Street, New York, New York 10027.
ORCID: 0000-0002-5140-6428 ResearcherID: C-1557-2013
[3] Formerly at The University of Texas at San Antonio at the time of publication.

Note: This document last updated on November 15, 2013

past 25 years, DER has emerged from the school effectiveness research (SER) movement.

School effectiveness research gained considerable attention in the late 1970s and early 1980s, as championed by Edmonds (Edmonds, 1979), as a means to demonstrate the processes present in unusually effective schools. As researchers focused on the school as the unit of analysis in effectiveness, others began to ask what roles the school district could play in individual school success (Clark, Lotto, & Astuto, 1984; Cuban, 1984; Purkey & Smith, 1985) or lack thereof (Floden et al., 1988; Hill, 1994). However, this research kept the focus on the school as the unit of analysis. Murphy and Hallinger (1988) noted this lack of attention to the district as the unit of analysis in this earlier school effectiveness literature and proposed that districts themselves could be instructionally effective, identifying twelve instructionally effective districts, one of the earliest DER studies. As detailed by these authors, effective districts research mirrors SER in which researchers examine some set of school districts, conclude by some criterion that the districts are effective system-wide or not, then study those districts with the aim to find and disseminate the aspects of the effective districts that appear to have led to their success in the hope of providing guidance to other districts looking to improve (Murphy & Hallinger, 1988; Murphy, Hallinger, Peterson, & Lotto, 1987; Murphy, Peterson, & Hallinger, 1986; Peterson, Murphy, & Hallinger, 1987). Their findings demonstrated that for the identified districts, the district administration attended to aligning instruction, curriculum, professional development, and student assessments across the district in coherent and mutually reinforcing ways. Subsequent work has expanded from the original focus on instruction to include an understanding of the interconnectedness of the system-wide organization of effective districts, and how all district operations, from instruction to administration, finance, human resources, and professional development can work to positively influence individual student success (Bowers, 2008; Elmore & Burney, 1999; Opfer et al., 2008; Rorrer et al., 2008). To be clear, while much of the literature that includes school districts has focused on describing what districts provide, their role in reform and policy, and the actions of the administrators (Rorrer et al., 2008), only studies that claim a district as exemplary should be considered as district effectiveness research. However, when considering district effectiveness research, as stated by Murphy, Peterson and Hallinger (1986) "previous criticisms which have been applied to studies of instructionally effective schools also apply to research on effective school districts" (p.152).

As noted by Opfer, Henry and Mashburn (2008) in reviewing the research on district effects, they state that "case-based evidence has shown that school districts can impact teaching and learning; however, it is as yet unclear whether the districts studied are anecdotes or instances where real district effects that are broadly feasible and replicable have been observed" (p.303). This quote highlights three of the main critiques from the parallel SER literature that have to date received little attention in the DER literature, two of which are the focus of the current study and will be detailed below. First is the issue of site selection, the second is the justification that the organizations studied are indeed effective, both of which are necessary to address before the research can turn to the final issue which is if any observed district effect is in fact district caused.

To date, DER studies have selected districts as sites for in-depth qualitative studies and surveys based on a variety of methods. As just a few examples, these selection strategies have ranged from recruiting districts interested in consulting or action research interventions (Firestone et al., 2005; Ikemoto & Marsh, 2007; Marsh et al., 2005; Supovitz, 2006; Wayman, Cho, & Johnston, 2007) to selecting districts based on raw achievement score rankings or gains (Hentschke, Nayfack, & Wohlstetter, 2009; Petersen, 1999; Togneri & Anderson, 2003), to ranking districts by standardized test scores within any one year across a sample using linear regression and controlling for background variables such as socioeconomic status (Bowers, 2008; Murphy & Hallinger, 1988; Peterson et al., 1987). This lack of agreement upon a method for identification of effective districts prior to the initiation of a qualitative study is problematic. As with SER, the goal is to study organizations that are unusually effective (Bryk & Raudenbush, 1988; Edmonds, 1979; Klitgaard & Hall, 1975; Luyten, Visscher, & Witziers, 2005). The difficulty stems from the differences in attempting to warrant the argument that any one school organization is both unusual and effective (Rowan, Bossert, & Dwyer, 1983). Nevertheless, the question remains that given the vast population of hundreds of school districts available on average to study within any one state, how should researchers and policymakers interested in district effectiveness select individual districts and justify the claim of effectiveness prior to investing the time and resources needed to perform the labor-intensive work of deep qualitative studies of these organizations? The focus of this study is to address this question of district site selection through addressing the long history of critiques encountered in the highly similar school effectiveness research and then demonstrate one method adapted from SER that researchers can use to select districts for DER studies that addresses the critiques.

Critiques from School Effectiveness Research

In school effectiveness research, researchers have historically chosen a few schools that are deemed highly effective for in-depth qualitative study in comparison to schools that are deemed not as effective. These studies have generally shown that the effective schools demonstrate strong administrative leadership, high expectations for student learning, an orderly school atmosphere, a high priority on teaching basic skills, and a willingness to prioritize school resources in service to these goals (Edmonds, 1979; Muijs, Harris, Chapman, Stoll, & Russ, 2004; Reynolds, Teddlie, Creemers, Scheerens, & Townsend, 2000). However, there have been many critiques of SER, with the majority of these critiques centered on four major issues with the methodology; the definition of effectiveness, effective school selection, external validity, and appropriate comparisons (Clark et al., 1984; Coe & Taylor-Fitz-Gibbon, 1998; Gibson & Asthana, 1998; Goldstein & Woodhouse, 2000; Luyten et al., 2005; Purkey & Marshall, 1983; Rowan et al., 1983; Stringfield, 1994; Teddlie, Reynolds, & Sammons, 2000; Thrupp, 2001).

The first in the list of critiques of SER methodology is the definition of effectiveness. Most of the early SER studies relied on single standardized assessment scores in one or two subjects (usually mathematics or English) at a single grade level to gauge effectiveness, assuming a high degree of reliability and validity of the assessments across multiple schools and contexts. However, the research community has never settled upon a definition of the term "effective", and single standardized test scores at single grade levels are arguably only one dimension of many for measuring

highly successful learning for all students (Coe & Taylor-Fitz-Gibbon, 1998; Luyten et al., 2005; Reynolds et al., 2000; Rowan et al., 1983). While test scores themselves can be questioned as to their reflection of actual instructional processes within schools, the SER literature has appeared to come to some consensus that as a summative measure of student knowledge within any one year, standardized test scores can help researchers determine which schools are more or less effective in obtaining those scores. That these scores are a reflection of student knowledge, and when examined over time and across subjects and grade levels, are reflective of student gains in knowledge, is generally accepted (Luyten et al., 2005; Teddlie, 1994; Teddlie et al., 2000). The question of actual instructional processes then must be one of the central questions within subsequent qualitative studies of the selected schools. Thus, rather than base site selection on a single test score, the critiques have argued for the inclusion of multiple tests, subjects, grade levels and years rather than rely on one or two single time-points, to examine achievement across subjects, grade levels and time.

In the second critique, how a school is selected for in-depth qualitative SER study has been heavily criticized, and this was especially relevant to the early SER studies (Luyten et al., 2005; Purkey & Marshall, 1983; Rowan et al., 1983; Teddlie et al., 2000). Early SER school selection methods were based on samples of convenience, local knowledge, or anecdotal evidence from administrators, parents or state policymakers. However, the focus on site selection quickly shifted to attempting to sift through the large sample of schools available in any one region to statistically identify schools that out performed their demographics (Dyer, Linn, & Patton, 1969; Klitgaard & Hall, 1975; Marco, Murphy, & Quirk, 1976; Teddlie et al., 2000). In response to the critiques, methods have ranged from gauging effectiveness based on overall test score levels, to using multiple linear regression models to control for demographic predictors of student success, such as socioeconomic status, to more recent innovations using hierarchical linear modeling to more accurately account for the nested nature of school-level data (Aitkin & Longford, 1986; Creemers & Kyriakides, 2006; Harker & Nash, 1996; Raudenbush & Bryk, 2002; Teddlie et al., 2000; Willms & Raudenbush, 1989). However, much of the foundational work in SER that set the stage for future studies employed multiple linear regression, and it is with this method that many of the critiques have issues. Using ordinary least squares (OLS) multiple linear regression, researchers first controlled for the effects of context and demographics on school performance, such as the socioeconomic status (SES) of the enrolled students, in an attempt to compare school test scores on a more "equal" basis, holding constant confounding variables such as SES. Researchers would then rank schools based on how far the school outperformed or underperformed the regression predicted test scores based on the school's demographics, stating that the schools that ranked highly were far outperforming their peers who had similar student demographics for that year and sample. These "outliers" were then selected for further in-depth qualitative analysis (Dyer et al., 1969; Klitgaard & Hall, 1975; Rowan et al., 1983; Teddlie et al., 2000).

The criticisms of this technique have mainly focused on the problems of outlier studies that focus on individual time-points and small intact samples (Stringfield, 1994). The critiques point out that these single year regression methods usually have low correlations year-to-year, identifying different schools as effective from one year to the next (Purkey & Marshall, 1983; Reynolds et al., 2000; Rowan et al., 1983; Thrupp, 2001). This "snapshot research" overly focuses on single points in time rather than on a "moving picture" of the organization over time (Luyten et al., 2005; Reynolds, Hopkins, & Stoll, 1993). In addition, the argument has been that the year-to-year fluctuations in ranking indicates that any one school identified within one year's data is most likely an outlier only due to chance alone, and the following year will score more closely to the mean of the normal distribution due to multiple random effects and confounding variables. Thus, the critique is that an outlier school that appears to be far outperforming its demographics in any one year has a high chance of having randomly scored as an outlier for that year, and so choosing that school for an in-depth qualitative analysis of effectiveness will give unreliable results. The main response to these critiques was the application of Hierarchical Linear Modeling (HLM) to control for these issues between schools and across time (Aitkin & Longford, 1986; Bryk & Raudenbush, 1988; Raudenbush & Bryk, 2002; Raudenbush & Chan, 1993; Raudenbush & Willms, 1995; Singer & Willett, 2003). Instead of warranting the argument that a school was outperforming its demographics within any one year, the argument from the HLM standpoint is that schools are nested within time. Thus, the critiques of the year-to-year variance issues are controlled for within longitudinal HLM, as schools are compared to how far they outperform or underperform their predicted rate of growth over time, rather than on overall scores for any one year, controlling for covariates in their student population.

Due to these issues of measurement and selection, the third major critique of SER is that the qualitative studies of these effective schools have little external validity. This lack of generalizability comes from the point that even if the school was not randomly successful that year, but actually successful, because the measures of effectiveness were so narrowly defined within a single year and the number of schools selected for the study sample so low, that whatever the findings, those findings could only be generalized to other schools for those specific test scores with similar students (Goldstein & Woodhouse, 2000; Luyten et al., 2005; Purkey & Marshall, 1983; Rowan et al., 1983; Teddlie et al., 2000; Thrupp, 2001). Addressing these critiques has involved vastly increasing the number of tested subjects and grade levels included in the initial site selection as well as initially comparing many more schools for site selection, allowing for generalizations to a much broader population of schools.

The fourth major critique has centered on the comparisons used in many SER studies. The question raised by the critiques has been: effective in comparison to whom? Effectiveness is a relative term, and so if a study is to argue that one school is effective, it must inherently answer this question of the comparison school that is deemed less effective. The critique has focused on this issue mainly because much of the SER literature has focused exclusively on detailing effective schools with few to no schools in the qualitative study sample that were deemed not as effective, or for studies that have examined less effective schools, those schools were chosen from the bottom of the ranking and compared to the top (Stringfield, 1994). The argument of the critics has been that if one wishes to understand what one school is doing differently over another that may make them effective, in an effort by the researchers to translate that research into specific recommendations for all or the majority of schools (an issue that

relates back to the point above about external validity), then studies should not compare a few top ranked schools to the bottom. Rather, to justify the term "unusual", a study should compare multiple schools from the top to the norm, or from the bottom to the norm, understanding that the differences between an effective school and the norm may be very different than those between an effective school and a non-effective school, and that samples much larger than two or three are desirable to increase both internal and external validity (Klitgaard & Hall, 1975; Purkey & Marshall, 1983; Teddlie et al., 2000).

Thus, overall, these four main critiques can be summed up as: 1) Limited number of subjects and grade levels tested to warrant the term effective. 2) Snapshot research which is overly focused on single year point estimates using regression estimates, and in many cases will identify organizations that are randomly high in outperforming their demographics in comparison to the other organization considered. 3) Limited samples rather than analyze the entire population, such as an entire state. Initial limited samples hamper future generalizability. 4) Limited ability to compare to the norm. Many studies identify only the effective organizations, or more rarely only the non-effective. However, the recommendations from the SER literature indicate that more useful comparisons would be to compare the high to the mean or the low to the mean.

To date, these issues with systematic site selection have not been a part of the conversation around district effectiveness research. However, these issues raised in SER apply directly to DER site selection. The danger is that if the issues from SER are not addressed, then DER will be opened to many of the same critiques discussed above that have plagued the findings from SER for more than 30 years. Since a growing number of researchers have become interested in if there is such a thing as a "district effect" and what it might entail, raising these issues from the SER literature is timely, as more researchers look to find effective districts and invest the vast number of resources needed to perform in-depth qualitative studies within these organizations. In this study, I outline one potential method that may be useful for identifying districts for DER, addressing and adapting the methods and critiques from SER to the district level using a 2-level hierarchical linear growth model. The method compares districts not on single year regression estimates, but on their growth in achievement over a sustained period, controlling for known covariates. I then test the method using the entire population of districts from the state of Ohio over a seven-year time span, 2001-02 through 2007-08. The method identifies multiple districts that outperform the HLM predicted growth estimates, and provides a means to visualize and recommend districts for DER sites.

METHOD

Seven years of publicly available district level achievement and demographic data for the school years 2001-02 through 2007-08 from Ohio were analyzed for all 608 school districts in the state. Publically available data were obtained from the Ohio Department of Data Services (ODE, 2008). The state of Ohio was attractive as an initial state with which to test the district identification method for three main reasons. First, the data records for Ohio are consistent across all seven years for the majority of variables analyzed in which the state both collected and reported the data in a consistent manner. The state of Ohio does report data from earlier academic years, however individual variables were not consistently collected nor reported thus seven years of data was the maximum available at the time of the study. Second, coincidentally, the span of time covered by the Ohio data begins the year during implementation of the No Child Left Behind act of 2001 ("NCLB," 2002), academic year 2001-02, and continues until the most recent academic year available. Third, the primary outcome measure analyzed was the Ohio Performance Index Score (PIS) for all seven years that aggregates district level state standardized test performance into a single indicator and is used by the state to calculate district Adequate Yearly Progress (AYP). The state of Ohio defines the PIS as:

> The performance index score is calculated by multiplying the percentage of students that are untested, below basic, basic, proficient, accelerated or advanced by weights ranging from 0 for untested to 1.2 for advanced students. The products are summed across all tested subjects in grades 3, 4, 5, 7, 8, and 10 to compute the performance index score for the school or district. The PI score (PIS) is on a scale of 0 to 120 (ODE, 2008).

Tested subjects in Ohio include mathematics, reading, writing, science, and social studies at the six grade levels indicated above. A PIS below 69 is usually grounds for the state of Ohio to designate the district as not meeting AYP and thus designating it as "Academic Emergency", which then triggers specific policy interventions under NCLB (ODE, 2008). Thus, the PIS is a single outcome measure that represents multiple grade-level and subject indicators of district standardized test performance.

A 2-level hierarchical linear growth model was used to estimate each district's growth or decline in PIS over the seven years. The use of a multilevel model nested in time is preferred given the critiques from the SER literature discussed above, as well as the superiority of multilevel growth models over OLS point regression estimates (Hox, 2002; Raudenbush & Bryk, 2002; Raudenbush & Chan, 1993; Singer & Willett, 2003). This stems from the increased precision gained when using a single HLM time-nested model over multiple OLS regression estimates, as well as the multilevel model providing weighted estimates of growth trajectories (Singer & Willett, 2003). Rather than estimating each district's yearly PIS for each of the seven years, controlling for covarying demographic variables, a hierarchical linear growth model allows for the estimation and comparison of each district's change in PIS through time, controlling for district performance covariates. This allows for the examination of the variance within and between districts' slopes through time, while controlling for district background time-varying covariates. Thus, a 2-level hierarchical linear growth model using district PIS addresses many of the critiques discussed above through estimating district growth in achievement through time, controlling for demographic covarying variables, and using PIS as a dependent variable as a weighted measure of district standardized test performance across multiple different subjects and grade levels.

A 2-level hierarchical linear growth model was estimated for the entire population of school districts in the state of Ohio over the seven academic years, 2001-02 through 2007-08, following the recommendations of the multilevel models for change literature for estimating change over time in a 2-level nested model (Hox, 2002; Raudenbush & Bryk, 2002; Singer & Willett, 2003). For this study, the model nests multiple time points per district (level 1) within districts (level 2). The dependent variable was district PIS within

any one year. Year was coded as 0 through 6. All variables were grand mean centered except for year. The level 1 model includes the fixed effects for the intercept, year, and the time-varying covariates for each district, described below. The level 2 model allows the intercepts and the slope for year to vary randomly, holding all other slopes for the covariates as fixed effects. No level 2 covariates were included, since all of the covariates were time-varying, and thus must be included in the level 1 model (Singer & Willett, 2003). Following the nomenclature recommended from the hierarchical linear modeling literature (Raudenbush & Bryk, 2002), the general form of the hierarchical linear growth model can be represented by the following equations:

Level1: $PIS_{ij} = \pi_{0j} + \pi_{1j} YEAR_{ij} + \pi_{2j} X_{ij} ... e_{ij}$

Level 2: $\pi_{0j} = \gamma_{00} + r_{oj}$

$\pi_{1j} = \gamma_{10} + r_{1j}$

$\pi_{2j} = \gamma_{20}$

:

:

In which:

PIS_{ij} = District performance index score for time i.

$YEAR_{ij}$ = Year for each district's data.

X_{ij} = Time varying covariates for each district in each year.

π_{0j} = The slope of the intercepts varying randomly across districts; district j's estimated PIS score in 2001-2002.

π_{1j} = The slope of time varying randomly across districts; the annual rate at which district j's PIS scores grew between 2001-02 and 2007-08.

π_{2j} = The slope of a level 1 predictor across districts.

Due to missing data for certain districts within any one year, eight districts were deleted from the dataset, leaving 600 districts at level 2, with 4113 total records at level 1. HLM 6.04 was used to estimate the model parameters and residuals (Raudenbush, Bryk, Cheong, Congdon, & duToit, 2007). To identify districts that outperformed or underperformed their 2-level HLM predicted growth over the six time periods represented (six periods over seven years), the model predicted slope for each district was calculated by summing the fitted slope for year and the empirical Bayes residual for each district. These empirical Bayes coefficients were then multiplied by six to represent the model predicted growth for each district over the timespan. Actual growth in PIS for each district was calculated by subtracting each previous year's PIS from the following year, and summing the results for each district. Each district's predicted PIS gains were then subtracted from the district's actual gains, and districts falling outside of the 95% confidence interval are deemed to have statistically significantly outperformed or underperformed their demographically controlled predicted seven-year growth in PIS and thus can be considered unusual in comparison to the majority of districts in the state. To aid in visualization, district actual PIS growth was then regressed on district predicted growth and plotted for all 600 districts.

RESULTS

A Model for the Identification of Districts for District Effectiveness Research

To date, as discussed above for district effectiveness research, the question of the selection of districts for in-depth qualitative analysis has gone mostly unexamined. However, given the long history of critiques from the highly relevant school effectiveness literature presented above, the question of how to warrant the argument that any one school district is "unusually effective" has remained, despite the growing interest in district research. Adapting the critiques from SER, the main issues in district identification stem from the need to define both "unusual" and "effective" prior to the selection of a school district for an in-depth qualitative study. These critiques reviewed above from SER can be summed up through four main points. First, while defining organizational effectiveness prior to a qualitative study of educational organizations must inherently rest upon the use of standardized test scores, since few other comparable measures exist across a large number of organizations, the test scores used must span as many different subjects and grade levels as possible. Relying on math or English at one or two grade levels is insufficient. Second, overreliance on what has come to be called "snapshot research" is problematic, given the inherent random variability year to year with single-year regression estimates or raw test scores, and the more longitudinal nature of educational organizations. Third, limited or intact samples from which sites are selected for in-depth qualitative research are problematic given the general purpose of effectiveness research to identify practices, norms or procedures that may be generalizable to a much larger population. Fourth, selection procedures must include a means to compare not only organizations deemed effective to each other, but to select and compare those organizations to the norm (Luyten et al., 2005; Reynolds et al., 1993; Reynolds et al., 2000; Rowan et al., 1983; Teddlie et al., 2000; Thrupp, 2001). While originally articulated in the terms of school effectiveness research, together, these issues directly apply to district effectiveness research and thus must be addressed by any district selection procedure.

Therefore, the question is, from the vast constellation of school districts within any one state, how are researchers to first find some set of districts for effectiveness studies and then to warrant the argument that the selected districts are both unusual and effective before devoting the large amounts of time and resources needed for qualitative studies on district effectiveness? To address these issues from SER and adapt them to DER, this study proposes and tests the use of a 2-level hierarchical linear growth model to predict the seven-year growth in achievement for the entire population of school districts from the state of Ohio across multiple subjects and grade levels, controlling for background and demographic variables. Predicted rates of growth in performance are then compared to actual district performance gains, and districts that statistically significantly outperform the control variables are considered both significantly unusual and effective (*see methods*).

This method addresses the issues reviewed from SER in the following ways. First, in addressing the issue of defining effectiveness as performance across multiple tests, subjects and grade levels, rather than on a few selected tests at one or two grade levels, the outcome variable, the Ohio Performance Index Score (PIS), is a district-level aggregate of standardized test performance. The PIS is a weighted average indicator of a district's performance in reading, writing, mathematics, English, science, and social

TABLE 1: Ohio District Descriptive Variables for Seven Years, 2001-02 through 2007-08, included in the 2-Level Hierarchical Linear Growth Model

	Grand Mean	*Standard Deviation*
Performance Index Score	91.84	8.248
Enrollment	2,819.46	4,539.092
Student/Teacher ratio	16.45	1.977
Years teacher experience	14.43	2.875
Teacher salary ($)	46,247.99	6,616.121
% Teacher attendance	94.38	10.185
% Student attendance	95.12	0.982
% Students high mobility	34.80	10.082
% Asian students	0.66	1.354
% African American students	5.33	13.885
% Hispanic students	1.30	2.939
% Economically disadvantaged students	25.95	16.645
Total number of records per district level-1 (*N*)	4113	
Total number of districts level-2 (*n*)	600	

studies at multiple grade levels across the organization (*see methods*). In addition, given the privileged status by district and school administrators of standardized tests that are linked to overall state policy sanctions (Guskey, 2007), the PIS is an interesting variable to consider as a measure of district effectiveness due to its use by the state to enact sanctions against districts that do not meet the criterion for AYP under NCLB.

Second, the use of HLM modeling growth over a sustained period also addresses the issue of defining effectiveness and addresses the problems associated with the "snapshot research" term. Rather than rely on single year regression estimates or raw performance scores, examining district growth in achievement over time is desirable since single year estimates of achievement on individual tests are known to vary significantly year to year. In addition, through the use of nesting districts in time using HLM, change over time controlling for covarying demographic and district community variables provides a means to more precisely model and control for the effects of covariates on district performance and estimate a district's achievement trajectory in relation to all of the other districts in the state. This point also relates to the third issue above of limited or intact samples. Rather than select a sample and estimate the means of the population, a main caveat of inferential statistics, with the use of the entire population of all school districts within a large state a researcher is able to calculate the means directly since the sample is the entire population. This both increases the precision of the model and decreases bias. In addition, a hierarchical linear growth model using such a dataset helps to address the fourth main issue from SER, of selecting unusual sites in comparison to the norm. Through including the entire population of districts within the analysis, predicting each district's growth in PIS over the seven years controlling for a district's background then comparing the predicted PIS growth to actual growth provides a means to compare each district to each of the other districts in the state. For districts that significantly outperform or underperform their demographics, if those districts are selected for in-depth qualitative study, districts that performed at the norm can be selected as interesting comparisons representing school districts that perform near the average, which inherently is the majority of all districts in the state and the districts one eventually wishes to generalize to. In these ways, the proposed method for district selection for DER addresses the main issues from SER. I now turn to testing the method using seven years of data and the entire population of school districts in Ohio.

A 2-Level Hierarchical Linear Growth Model Predicting District Achievement

Table 1 presents descriptive statistics detailing the grand means and standard deviations for variables included in the HLM. Variables were chosen for inclusion in the model based on two factors. First was if past evidence across the literature indicated that the variable could affect average student achievement and was to some extent outside the control of the district. These included enrollment (Rorrer et al., 2008), student/teacher ratio (Rumberger & Palardy, 2005), years of teacher experience (Wayne & Youngs, 2003), teacher salary (Rumberger & Palardy, 2005), teacher attendance (Taylor & Bogotch, 1994), student attendance (Dailey et al., 2005), student mobility (Rumberger & Palardy, 2005), student ethnicity and socioeconomic status (Sirin, 2005). Second, variables were included that were consistently recorded by the state of Ohio across the seven years studied. Variable names followed the nomenclature reported for each variable from the Ohio Department of Education (ODE, 2008). Variables included represented three main categories aggregated to the district level. First, district performance and enrollment were reported as the district PIS and overall district student enrollment. Second, teacher variables were included, including student-teacher ratio, average years of teacher experience, average teacher salary, and teacher attendance percentage. Third, average student variables were included, including percent student attendance, percent of students attending a district between one and two years (classified here as students high mobility), percent Asian students, percent African American students, percent Hispanic students, and percent of economically disadvantaged students. In 2001-02 the state of Ohio did not report for any district the percent of economically disadvantaged students enrolled, and in 2002-03 the state of Ohio

TABLE 2: Predicted State-Wide Seven-Year District Performance 2001-02 through 2007-08: A Two-Level Hierarchical Linear Growth Model Controlling for Teacher Parameters and Student Demographics

Dependent Variable *Yearly Performance Index Score (PIS)*	*Coefficient*	*Standardized Coefficient*	*Standard Error*
Fixed effects			
Intercept	86.430***	---	0.265
Year	1.798***	0.436	0.052
Enrollment (in thousands)[a]	0.564*	0.054	0.223
Student/Teacher ratio	-0.221***	-0.053	0.037
Years of teacher experience	0.063*	0.022	0.027
Teacher salary (in thousands)	0.153***	0.122	0.023
% Teacher attendance	-0.002	-0.002	0.005
% Student attendance	1.327***	0.158	0.098
% Students high mobility[b]	0.618***	0.060	0.083
% Asian students[a]	1.279***	0.079	0.226
% African American students[a]	-1.841***	-0.264	0.135
% Hispanic students[a]	-0.471**	-0.039	0.168
% Economically disadvantaged students[b]	-0.924***	-0.184	0.076
Hierarchical linear modeling reliability			
Intercepts	0.905		
Slopes of district improvement through time	0.632		
Within-District variance explained (%)	75.6		
Between-District variance explained (%)	31.7		

NOTE: All variables are grand mean centered, except for Year.
a. Transformed variable (natural log)
b. Transformed variable (square root)
*p<0.05, **p<0.01, ***p<0.001

did not report the percent of students enrolled for only one or two years (high mobility). To increase the total number of years available to model using HLM from five to seven, due to the need to have near complete datasets for each variable for each year, and to increase the reliability of the subsequent HLM growth model (Raudenbush & Chan, 1993), these two single data points for these two years were imputed using linear interpolation (SPSS, 2006) and the six other years of data available.

Ohio school district yearly PIS from 2001-02 through 2007-08 was estimated as the dependent variable using a 2-level hierarchical linear growth model (*Table 2*). In the 2-level model, multiple time points are nested within districts. Using all districts from the state of Ohio for all years with available data, the goal of the model is to estimate growth in district PIS through time controlling for known district achievement covariates (*see methods*). The model contains the intercept, year and time-varying covariates at level 1, with the intercepts and slopes for year varying randomly at level 2 with no predictors. In this way, each district's predicted gain in PIS is modeled through the seven years, while controlling for district demographics and background. Table 2 details the results of the 2-level HLM. Variables that were either natural log or square root transformed are indicated in Table 2. As indications of model fit and as a result of the use of many years of data (Raudenbush & Chan, 1993) for the entire population of districts, the overall reliability measures for the model are high, 0.905 for the intercepts and 0.632 for the slopes for year, as are the within district and between district variance explained, 75.6% and 31.7% respectively (Table 2, lower section, first column). However, there is some disagreement in the literature over if the between-district variance explained can be interpreted for multilevel growth models (Hox, 2002; Singer & Willett, 2003), so caution is recommended when assessing the between-district variance.

Variables were transformed by either natural log or square root transformations as needed to correct for skewness in the data. To aid in interpretation, all variables were grand mean centered (*see Table 1*) except for year, which was coded 0 through 6 for each of the academic years 2001-02 through 2007-08. Thus, the intercept represents the mean PIS for the average school district in Ohio with average attributes on all variables during the first year of data, 2001-02, 86.430. The coefficient for year, 1.798, represents the average district growth per year in PIS, controlling for district background variables state-wide. Of interest to note is that the coefficient for year is positive, indicating a general upward trend for achievement through time for the average district in the state of Ohio. This represents approximately a one-fifth of a standard deviation increase per year in PIS for the average district.

For the remaining variables, many are significant in the model and separate into either positive or negative coefficients (Table 2). State-wide for this period, the significant positive coefficients appear to contribute to district PIS, including increasing enrollment, years of teacher experience, increasing student attendance, high mobility students and percent Asian students. Among these, increasing enrollment and high mobility students are interesting in that positive coefficients are somewhat unexpected

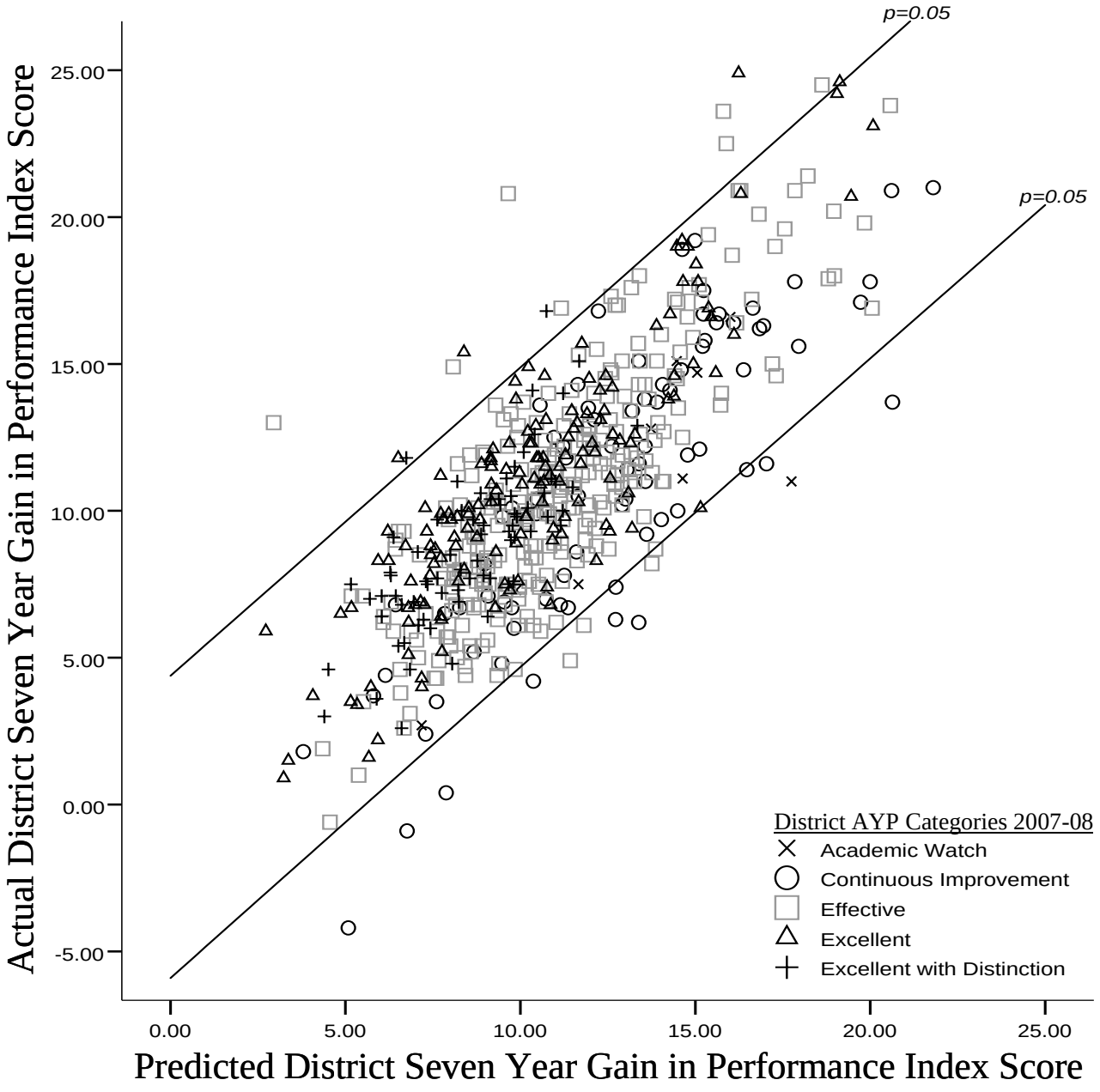

FIGURE 1: Comparison of Actual versus HLM Predicted State-Wide Gains in Performance Index Score, for every District in Ohio 2001-02 through 2007-08. A comparison of seven-year hierarchical linear growth model predicted gains in the Performance Index Score controlling for school district demographics for each of the school districts in Ohio to actual gains over the same time period indicates that the rate of growth in achievement for the majority of school districts is predicted by the HLM growth model controlling for demographics. However, multiple school districts significantly either outperform or underperform their predicted rate of growth controlling for background and demographic variables in the model (districts above or below p=0.05). Ohio AYP district performance categories for the final academic year included (2007-08) are indicated in the legend.

given past research (Kerbow, 1996; Odden, Goertz, & Picus, 2008; Rumberger & Larson, 1998). However, these may be positive due to controlling for the other demographic and enrollment trends in the school districts simultaneously, especially percent economically disadvantaged. For the significant negative coefficients, as the variable increases the model suggests that district PIS over time decreases. These coefficients are not unexpected given past research on class size (Finn & Achilles, 1999; Hanushek, 1999; Odden et al., 2008) and on issues of ethnicity and SES in school achievement, in particular in Ohio (Coleman et al., 1966; Ogbu, 2003; Rothstein, 2004). Table 2 also lists the standardized coefficients that indicate the overall contribution of each of the variables to the model (Table 2, standardized coefficient column). By far, the largest contribution to the model is year, indicating that district achievement is rising over time on average, state-wide, and accounts for the majority of the variance in the model. This is followed by percent African American students, percent economically disadvantaged students, student attendance and teacher salary. For the remaining coefficients, while most are statistically significant, they contribute only minimally to the model.

Therefore, overall, the 2-level hierarchical linear growth model of district PIS over the years 2001-02 to 2007-08 predicts district achievement controlling for the covariates included in the model. This model was used in the next section to compare predicted district gains in PIS to their actual gains.

Districts that Outperform their HLM Growth Model Predicted Gains

To examine which districts may outperform their demographics and background variables, predicted gains in PIS over the seven years controlling for the variables included in the HLM growth model were calculated for each district in Ohio (*see methods*). These predicted gains were compared to district actual gains, and districts that fall outside of the 95% confidence interval are considered to statistically significantly either outperform or underperform their background and demographic variables included in the model. A plot of the predicted gains in PIS for each district in Ohio by the actual gains provides a visual comparison to display the population of districts and visualize which districts significantly outperform the HLM predicted gains (Figure 1). Figure 1 demonstrates that while the vast majority of school districts gained in PIS over the seven years, a number of school districts significantly outperformed their demographic and background variables (Figure 1, above the upper p=0.05 line) or underperformed (Figure 1, below the lower p=0.05 line).

To address an issue raised in the SER literature in which school comparison studies mainly explore only relational norm-referenced regression differences (Luyten et al., 2005), Figure 1 includes AYP information for each district. Each district is labeled as to its state designated AYP category from the final year included in the dataset, 2007-08. This issue stems from the problem of a normed measure of district performance, based on a statistical model, in comparison to a criterion reference of performance outside of the model, to help judge both the model and district selection. As an example, a school identified in an SER study might outperform its peers, but if all of the schools in the sample hypothetically failed to teach mathematics to their students, as defined by state AYP criteria, then outperforming that sample is not very informative (Luyten et al., 2005). Thus, displaying final year AYP categories for districts in Figure 1 helps to address this issue for DER by providing a means to visualize each district by its difference in gains from the model and on the state's own AYP criterion, which is itself based on the dependent variable in the model, PIS.

For district effectiveness research, Figure 1 suggests that the majority of districts in Ohio did not significantly outperform or underperform their HLM model predicted gains in PIS. For those districts that did fall outside of the 95% confidence interval, 12 outperformed the HLM growth model predicted gains in PIS, while 15 school districts underperformed the model (Figure 1). To aid in reading Figure 1, district names are not attached to each point. For the outperforming school districts, working from the outer most points inwards towards the center, starting from the upper left center, the Ohio school districts that significantly outperformed their background and demographic variables were: Dawson-Bryant, Steubenville, Gorham Fayette, Bloom-Vernon, Orange, Williamsburg, Norwalk, Lake, Southeast, Coshocton, Vanlue, South Range[1]. Thus, in addressing the critiques of SER for DER site selection by using the entire population of school districts from a single state over a sustained period of seven years, districts that resemble the majority of school districts in the state but yet outperform their background and demographic variables can be identified using the HLM growth model comparison method proposed and tested here. However, identifying what these districts are doing, why they do it, and how they go about achieving this level of performance is the difficult work of subsequent studies. I now turn to a discussion of the issues and assumptions of the method for DER site selection.

DISCUSSION

This study reviewed the main critiques of site selection for organizational effectiveness studies in education from the school effectiveness research literature, applied, and adapted those critiques to a proposed method for site selection for district effectiveness research. Using a 2-level hierarchical linear growth model to first model and control for background and demographic characteristics of districts in predicting overall district gains in achievement over an extended period, predicted district gains in achievement were modeled using the entire population of school districts in the state of Ohio with data over the seven years studied, from 2001-02 through 2007-08. These predicted gains were then compared to the actual district gains over the period, and a set of districts were identified as significantly outperforming or underperforming the background and demographic variables included in the model. Through proposing a method that addresses the main critiques from SER, then testing the method for DER, the purpose of this study is to identify a useful method for future DER site selection and subsequent in-depth qualitative analysis of effective school districts. However, as in the SER literature, many issues with this identification method remain.

The first issue is with the identification of districts based on how far they outperform or underperform the background and demographic model predicted gains in achievement scores. As is well stated in SER, while it can be assumed that a school identified using this type of method as outperforming its background and demographic predictors is a school effect, this does not necessarily mean that the school effect is school caused (Coe & Taylor-Fitz-Gibbon, 1998; Goldstein, 1997; Luyten et al., 2005; Thrupp, 2001). This point also applies to DER, in that while the method identifies a "district effect", certain districts appear to perform

differently from their predicted performance, it does not necessarily lead to the assumption that the effect is "district caused". The method for DER site selection detailed here can only identify the magnitude of the effect, not the cause. Only surveys or qualitative studies can begin to explore if the effect identified is actually caused by the district organization or central office. Stated another way, the underlying assumption of the identification method is that one wishes to control for and remove the confounding covariates that occlude the district effect, such as district background and student demographic variables. However, if the HLM growth model does not include a major confounding covariate, then that variable may be an explanation for the district effect that would not be district caused. Nevertheless, subsequent qualitative studies exploring and comparing different districts with the norm would conceivably discover such an omission, and future revisions of the identification method would need to control for such a variable. For this study, although many potential district covariates were included in the 2-level HLM, variable selection was limited to only those variables that the state of Ohio consistently recorded and reported by the state over the seven years. For future work using Ohio data, as well as replicating the method in other states, this issue with consistent long-term reporting of variables by state agencies may lessen as states continue to refine and work to report education data under federal policy mandates.

A second critique of the method is that one could argue that a 2-level time nested model is insufficient to model the complexities of district performance and instead a more complex model is needed that takes into account other levels within the system. This issue is well articulated in SER, as researchers work to understand the effects of schooling on individual students, and thus use 3-levels or more in their models of students in classrooms in schools over time (Bryk & Raudenbush, 1988; Teddlie et al., 2000). Such a hierarchical model in DER could conceivably include up to five levels, with students in classrooms in schools in districts over time. This is a difficult prospect conceptually, methodologically, computationally and for interpretation. Nevertheless, this critique cannot be discounted, and so future research will focus on testing if the inclusion of more complex nested data structures of districts aids in district selection. One recommendation of the multilevel modeling literature (Raudenbush & Bryk, 2002) is to include student and school-level data within the growth model, increasing model fit, and the precision of coefficient estimates and standard errors, and decreasing aggregation bias. However, student-level data may not be available in most cases, so future research will concentrate first on replicating and expanding the method presented here to a 3-level model. Conceptually, from a policy and administrative perspective, a 3-level model, in which schools are nested in time within districts would help to address one of the remaining issues with the method presented here. That issue is that the district effect modeled with 2-levels assumes that the effect is constant across schools within each district school system. This is problematic given that previous district research has shown that district efforts can be focused unevenly across the system, most often at the elementary level since researchers have indicated that many districts see more opportunities for improvement at the primary grades rather than at the secondary level (Bowers, 2008; Cuban, 1984, 2003; Cuban & Usdan, 2003; Elmore, 2003; Purkey & Smith, 1985). A 3-level model conceivably could allow schools to vary in time within districts, modeling the variance across schools within districts. However, this would necessitate the use of a different outcome variable, one at the school level rather than district level, which would change the focus of the model and the identification method. The purpose of this study was to propose an initial district-level identification method and test it using a long-term state-wide dataset to provide an initial means for researchers to select sites for DER as researchers continue to work in this domain. Thus, while outside the scope of this study, future work will focus on comparing different methods and investigating if the inclusion of the school level as a third level in the model is beneficial or not.

Another critique of the method also stems from SER, in which researchers argue that prior student achievement should be included when modeling school effectiveness. While it is generally agreed that change in achievement over time is more reflective of the effect of the organization, rather than on the specific level of student achievement (Luyten et al., 2005), for SER, controlling for prior student achievement helps to account for past student experiences when examining the school that a student is enrolled in within any one year. However, the argument here is that this is where SER and DER diverge. In SER, one of the main questions is to examine the effect of a specific school on the learning of a specific student. For example, would the same student have gained 0.2 standard deviations in achievement in school X versus having attended school Y, controlling for past experiences and school background and demographics? In many ways, SER is concerned with the classic input/output production function question (Bryk & Raudenbush, 1988; Hanushek, 1997, 2003; Todd & Wolpin, 2003). However, the question of interest and the unit of analysis are different for DER given that it is at a different stage than SER and at this point has a different focus. One of the central questions still for DER is if there is such a thing as a positive district effect that is also district caused. In other words, the argument here is that a central question for DER is in identifying districts that are different in effectiveness, controlling for background and demographic covariates in an effort to identify which districts to study with in-depth qualitative studies to determine the extent to which district effects are district caused. Thus, the unit of analysis is the district-level, not the student-level. While determining if one student would have done better or worse in district A in comparison to district B is of interest, at this stage in the DER literature, it is not the focus of DER. Rather, one of the central concerns of DER is in the long-term performance of the district, controlling for student background and district-level covariates that are outside the control of the district. The question is, are differences in the long-term gain in achievement at the district level, district caused and are these differences due in any way to organizational or leadership properties that are not the effect of the happenstance of which community the district resides in, and thus may be extended to other districts looking for guidance on how to improve? This question inherently lies at the district-level, and this is why this study focuses on a 2-level district-level model, examining gains in the district performance index score, an indicator that the state rates districts on for AYP purposes. Hence, to control for previous student achievement before students enter the district would change the focus of the study from examining the significant differences in district achievement gains in an effort to identify outperforming districts, to the production function question. While of interest, this type of shift in focus is outside the scope of this study.

In addition, controlling for prior district achievement, such as controlling for year 2000-01 district PIS in the model, the year prior to the first year included in the HLM growth model, would overly focus the district selection method on gains only during the seven years included. If this was the research question, such as identifying school districts that have significantly changed since the introduction of NCLB, one would then want to control for prior achievement in this way. However, the question of interest for this study is to propose and test a method to determine which districts may be outperforming their peers. Thus, controlling for prior district-level performance is problematic. It may be that certain districts have instituted organizational changes at the district level beginning many years before the first year included, and that it is these changes that have led to their continued gains with their students. Controlling for prior district achievement would penalize these types of districts in the model, requiring districts to have made their changes only during the period included within the model. Thus, since this study is concerned with identifying school districts that have outperformed or underperformed their background and demographic covariates, even if the changes were prior to the first year included, prior performance was not included in the HLM growth model.

Overall, this study demonstrates that school districts that are unusually effective can be identified from state-wide datasets. Additionally, the study is also significant in five other aspects. First, by addressing the critiques from the SER literature, the proposed method adapts the recommendations from the past literature on SER site selection to district effectiveness research. Second, by using a hierarchical linear growth model to model the gains in overall district performance across grades and subjects, the method compares districts not on individual yearly raw scores or gains, but on multiyear gains controlling for significant district background and demographic variables. Third, the model was tested using seven years of data. Rarely have studies examined district gains in performance over a sustained amount of time that extends from the beginning of a national policy, NCLB, to the most recent data available. Fourth, the sample used to test the model was the entire population of districts with consistent data for the state of Ohio. Rather than estimating population means from a sample, the overall model was improved since the sample was the entire population. Fifth, while not the focus of this study, the variables within the HLM growth model itself are of interest for future research, since rarely have gains in district performance been modeled in such a way using state-wide data over a sustained period to examine differences at the district level. Although much has been done around school level research of this type, the point here is that districts have often gone unexamined. As discussed above, these types of models using the entire dataset within a policy region are of interest to examine organizational responses to state and national level policies. Overall, the HLM growth model and the overall method for district identification appear to work well. Future work will continue to incorporate the ongoing work in the school effectiveness domain into a model and the method for district effectiveness identification. However, while acknowledging that the method should continue to be tested, replicated and refined, the argument here is that this method of district selection is an improvement over the past range of methods in warranting a study's assertion that a district is effective or not.

In the end, while the identification method tested here appears to identify school districts that have significantly outperformed or underperformed their background and demographics, identifying districts that appear to have an effect on achievement, determining what the districts are doing differently to cause the effect is a question that remains. As discussed above, getting at the question of if the district effect is actually district caused, and exactly if and how district operations, organization, administration and leadership may play a role in the district effect, are the main types of questions that can only be addressed through in-depth qualitative studies and surveys. These future studies should aim to compare the outperforming districts with districts at the norm, or underperforming districts with districts at the norm. As reviewed above, much is known about district operations and administration from the existing district effectiveness research. However, what is not known from the past studies is if the studied districts were at the norm, significantly outperforming, or significantly underperforming their peers. Until otherwise shown, one assumption might be that past districts deemed as effective in DER may be closer to the norm, given the critiques presented here and in the past SER literature.

However, for future research studying districts identified using the method presented in this study, the question remains as to what might help to explain a district's high performance. Is it in fact district caused and not just an effect of district location or student demographics? If the performance gains are district caused, the subsequent qualitative studies may reveal a few different possible explanations. One is that a district may be cheating or gaming the accountability system in some way. A statistically significant difference, as demonstrated here, could be taken as evidence of this. However, this is a fairly pessimistic view. Alternatively, the literature on high performing schools does provide some insight into how an organization may outperform its predicted achievement, and how this is an administrative issue. Much of the SER literature indicates that effective schools maintain strong leadership, high expectations, orderly schools, and savvy resource management (Luyten et al., 2005; Reynolds et al., 1993; Reynolds et al., 2000; Teddlie, 1994). In addition, the DER studies to date in many ways mirror these findings, yet often lack the comparison school districts at the norm that this study argues should be included in future DER work. Nevertheless, for school leadership and administration, and through extension district leadership, studies have shown that schools that demonstrate multiple types of leadership that both manage the system and engage and empower the members of the system are effective when compared using both overall indicators and when controlling for background covariates (Hallinger, 2003; Marks & Printy, 2003; Printy & Marks, 2006). Future research will work to explore this relationship at the district level.

ENDNOTES

[1.] Information for each district and variable is publically available online through the Ohio Department of Education (ODE, 2008).

ACKNOWLEDGEMENTS:

The author would like to thank Gloria Crisp, Anne-Marie Nunez, Bruce Barnett, Daniel Sass and the EAQ reviewers for their in-depth and timely suggestions and critiques of this study.

RECOMMENDED CITATION

Bowers, A.J. (2010) Toward Addressing the Issues of Site Selection in District Effectiveness Research: A 2-Level Hierarchical Linear Growth Model. *Educational Administration Quarterly*,46(3), 395-425.doi:10.1177/0013161X10375271.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General), 149*(1), 1-43.

Bowers, A.J. (2008). Promoting excellence: Good to Great, NYC's District 2, and the case of a high performing school district. *Leadership and Policy in Schools*, 7(2), 154-177. doi:10.1080/15700760701681108

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65-108.

Clark, D. L., Lotto, L. S., & Astuto, T. A. (1984). Effective schools and school improvement: A comparative analysis of two lines of inquiry. *Educational Administration Quarterly, 20*(3), 41-68.

Coe, R., & Taylor-Fitz-Gibbon, C. (1998). School effectiveness research: criticisms and recommendations. *Oxford Review of Education, 24*(4), 421-438.

Coleman, J., Campbell, E., Hobsen, C., McPartland, J., Mood, A., Weinfeld, F., et al. (1966). *Equality of educational opportunity survey*. Washington, D.C.: U.S. Government Printing Office.

Creemers, B. P. M., & Kyriakides, L. (2006). Critical analysis of the current approaches to modeling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*(3), 347-366.

Cuban, L. (1984). Transforming the frog into a prince: Effective schools research, policy and practice at the district level. *Harvard Educational Review, 54*(2), 129-151.

Cuban, L. (2003). *Why is it so hard to get good schools?* New York: Teachers College Press.

Cuban, L., & Usdan, M. (2003). Fast and top-down: Systemic reform and student achievement in San Diego city schools. In L. Cuban & M. Usdan (Eds.), *Powerful reforms with shallow roots* (pp. 77-95). New York: Teachers College Press.

Dailey, D., Fleischman, S., Gil, L., Holtzman, D., O'Day, J. A., & Vosmer, C. (2005). *Toward more effective school districts: A review of the knowledge base*. Washington, DC: American Institutes for Research.

Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal, 6*(4), 591-605.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 31*(1), 15-24.

Elmore, R. F. (2003). Accountability and capacity. In M. Carnoy, R. Elmore & L. S. Sisken (Eds.), *The new accountability: High schools and high stakes testing* (pp. 195-209). New York: RoutledgeFalmer.

Elmore, R. F., & Burney, D. (1999). Investing in teacher learning: Staff development and instructional improvement. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession : Handbook of policy and practice* (pp. 263-291). San Francisco: Jossey-Bass.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*(2), 97-109.

Firestone, W. A., Mangin, M. M., Martinez, M. C., & Polovsky, T. (2005). Leading coherent professional development: A comparison of three districts. *Educational Administration Quarterly, 41*(3), 413-448.

Floden, R. E., Porter, A. C., Alford, L. E., Freeman, D. J., Irwin, S., & Schmidt, W. H. (1988). Instructional leadership at the district level: A closer look at autonomy and control. *Educational Administration Quarterly, 24*(2), 96-124.

Gibson, A., & Asthana, S. (1998). School performance, school effectiveness and the 1997 white paper. *Oxford Review of Education, 24*(2), 195-210.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8*(4), 369-395.

Goldstein, H., & Woodhouse, G. (2000). School effectiveness research and educational policy. *Oxford Review of Education, 26*(3/4), 353-363.

Guskey, T. R. (2007). Multiple scources of evidence: An analysis of stakeholder perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice, 26*(1), 19-27.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education, 33*(3), 329-352.

Hannaway, J., & Stanislawski, M. (2005). Flip-flops in school reform: An evolutionary theory of decentralization. In F. M. Hess (Ed.), *Urban school reform: Lessons from San Diego* (pp. 53-70). Cambridge, Mass.: Harvard Education Press.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164.

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis, 21*(2), 143-163.

Hanushek, E. A. (2003). The failure of input-based school policies. *The Economic Journal, 113*(485), F64-F98.

Harker, R., & Nash, R. (1996). Academic outcomes and school effectiveness: Type "A" and Type "B" effects. *New Zealand Journal of Educational Studies, 32*(2), 143-170.

Hentschke, G. C., Nayfack, M. B., & Wohlstetter, P. (2009). Exploring superintendent leadership smaller urban districts: Does district size influence superintendent behavior. *Education and Urban Society, 41*(3), 317-337.

Hightower, A. M., Knapp, M. S., Marsh, J. A., & McLaughlin, M. W. (2002). The district role in instructional renewal: Setting the stage for dialogue. In A. M. Hightower, M. S. Knapp, J. A. Marsh & M. W. McLaughlin (Eds.), *School districts and instructional renewal* (pp. 1-6). New York: Teachers College Press.

Hightower, A. M., & McLaughlin, M. W. (2005). Building and sustaining an infrastructure for learning. In F. M. Hess (Ed.), *Urban school reform: Lessons from San Diego* (pp. 71-92). Cambridge, Mass.: Harvard Education Press.

Hill, P. T. (1994). *Reinventing public education*. Santa Monica, CA: RAND.

Honig, M. I., & Coburn, C. E. (2008). Evidence-based decision making in school district central offices. *Educational Policy, 22*(4), 578-608.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the "data-driven" mantra: Different conceptions of data-driven decision making. In P. A. Moss (Ed.), *Evidence and decision making: The 106th yearbook of the National Society for the Study of Education, Part 1* (pp. 105-131). Malden, Mass: Blackwell Publishing.

Kerbow, D. (1996). Patterns of urban student mobility and local school reform. *Journal of Education for Students Placed at Risk, 1*(2), 147-170.

Klitgaard, R. E., & Hall, G. R. (1975). Are there unusually effective schools? *The Journal of Human Resources, 10*(1), 90-106.

Luyten, H., Visscher, A., & Witziers, B. (2005). School effectiveness research: From a review of the critism to recommendations for further development. *School Effectiveness and School Improvement, 16*(3), 249-279.

Marco, G. L., Murphy, R. T., & Quirk, T. J. (1976). A classification of methods using student data to assess school effectiveness. *Journal of Educational Measurement, 13*(4), 243-252.

Marks, H. M., & Printy, S. M. (2003). Principal Leadership and School Performance: An Integration of Transformational and Instructional Leadership. *Educational Administration Quarterly, 39*(3), 370-397.

Marsh, J. A., Kerr, K. A., Ikemoto, G. S., Darilek, H., Suttorp, M., Zimmer, R. W., et al. (2005). *The role of districts in fostering instructional improvement: Lessons from three urban districts partnered with the institute for learning*. Santa Monica, CA: RAND Corporation.

McLaughlin, M. W., & Talbert, J. E. (2002). Reforming districts. In A. M. Hightower, M. S. Knapp, J. A. Marsh & M. W. McLaughlin (Eds.), *School districts and instructional renewal* (pp. 193-202). New York: Teachers College Press.

Muijs, D., Harris, A., Chapman, C., Stoll, L., & Russ, J. (2004). Improving schools in socioeconomically disadvantaged areas - A review of research evidence. *School Effectiveness and School Improvement, 15*(2), 149-175.

Murphy, J., & Hallinger, P. (1988). Characteristics of instructionally effective school districts. *Journal of Educational Research, 81*(3), 175-181.

Murphy, J., Hallinger, P., Peterson, K. D., & Lotto, L. S. (1987). The administrative control of principals in effective school districts. *Journal of Educational Administration, 25*(2), 161-192.

Murphy, J., Peterson, K. D., & Hallinger, P. (1986). The administrative control of principals in effective school districts: The supervision and evaluation functions. *The Urban Review, 18*(3), 149-175.

No Child Left Behind Act of 2001, (2002).

Odden, A. R., Goertz, M. E., & Picus, L. O. (2008). Using available evidence to estimate the cost of educational adequacy. *Education Finance and Policy, 3*(3), 374-397.

ODE. (2008). *Office of data services*. Retrieved Aug 2008 from http://www.ode.state.oh.us.

Ogbu, J. U. (2003). *Black American students in an affluent suburb: A study of academic disengagement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Opfer, V. D., Henry, G. T., & Mashburn, A. J. (2008). The district effect: Systemic responses to high stakes accountability policies in six southern states. *American Journal of Education, 114*(2), 299-332.

Petersen, G. J. (1999). Demonstrated actions of instructional leaders: An examination of five California superintendents. *Education Policy Analysis Archives, 7*(18).

Peterson, K. D., Murphy, J., & Hallinger, P. (1987). Superintendents' perceptions of the control and coordination of the technical core in effective school districts. *Educational Administration Quarterly, 23*(1), 79-95.

Printy, S. M., & Marks, H. M. (2006). Shared leadership for teacher and student learning. *Theory Into Practice, 45*(2), 125-132.

Purkey, S. C., & Marshall, S. S. (1983). Effective schools: A review. *The Elementary School Journal, 83*(4), 426-452.

Purkey, S. C., & Smith, M. S. (1985). School reform: The district policy implications of the effective schools literature. *The Elementary School Journal, 85*(3), 352-389.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & duToit, M. (2007). *HLM 6: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International Inc.

Raudenbush, S. W., & Chan, W.-S. (1993). Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. *Journal of Consulting and Clinical Psychology, 61*(6), 941-951.

Raudenbush, S. W., & Willms, D. J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307-335.

Reynolds, D., Hopkins, D., & Stoll, L. (1993). Linking school effectiveness knowledge and school improvement practice: Towards a synergy. *School Effectiveness and School Improvement, 4*, 37-58.

Reynolds, D., Teddlie, C., Creemers, B. P. M., Scheerens, J., & Townsend, T. (2000). An introduction to school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 3-25). New York: Falmer Press.

Rorrer, A. K., Skrla, L., & Scheurich, J. J. (2008). Districts as institutional actors in educational reform. *Educational Administration Quarterly, 44*(3), 307-357.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap*. Washington, DC: Economic Policy Institute.

Rowan, B., Bossert, S. T., & Dwyer, D. C. (1983). Research on effective schools: A cautionary note. *Educational Researcher, 12*(4), 24-31.

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education, 107*(1), 1-35.

Rumberger, R. W., & Palardy, G. J. (2005). Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American Educational Research Journal, 42*(1), 3-42.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

SPSS. (2006). SPSS Inc. (Version 15.0): SPSS Inc.

Stein, M. K., & D'Amico, L. (2002). Inquiry at the crossroads of policy and learning: A study of a district-wide literacy initiative. *Teachers College Record, 104*(7), 1313-1344.

Stringfield, S. (1994). Outlier studies of school effectiveness. In D. Reynolds, B. P. M. Creemers, P. S. Nesselrodt, E. C. Schaffer, S. Stringfield & C. Teddlie (Eds.), *Advances in school effectiveness research and practice* (pp. 73-84). Tarrytown NY: Elsevier Science.

Supovitz, J. A. (2006). *The case for district-based reform: Leading, building, and sustaining school improvement*. Cambridge, Massachusetts: Harvard Education Press.

Taylor, D. L., & Bogotch, I. E. (1994). School-level effects of teachers' participation in decision making. *Educational Evaluation and Policy Analysis, 16*(3), 302-319.

Teddlie, C. (1994). The integration of classroom and school process data in school effectiveness research. In D. Reynolds, B. P. M. Creemers, P. S. Nesselrodt, E. C. Schaffer, S. Stringfield & C. Teddlie (Eds.), *Advances in school effectiveness research and practice* (pp. 111-132). Tarrytown NY: Elsevier Science.

Teddlie, C., Reynolds, D., & Sammons, P. (2000). The methodology and scientific properties of school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 55-134). New York, NY: Falmer Press.

Thrupp, M. (2001). Recent school effectiveness counter-critiques: Problems and possibilities. *British Educational Research Journal, 27*(4), 443-457.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113*(485), F3-F33.

Togneri, W., & Anderson, S. E. (2003). *Beyond islands of excellence: What districts can do to improve instruction and achievement in all schools*. Washington, DC: The Learning First Alliance and The Association for Supervision and Curriculum Development.

Wayman, J. C., Cho, V., & Johnston, M. T. (2007). *The data-informed district: A district-wide evaluation of data use in the Natrona county school district*. Austin: The University of Texas at Austin.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89-122.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209-232.