The topics of school and teacher accountability have become permanent fixtures in the ongoing discussion of educational reform in the United States. The federal government has long pushed for high-quality teachers in high-quality schools, and federal initiatives continue to demand that school districts provide evidence of student growth. While student growth may be defined and measured in a variety of ways, the primary evidence base for growth has been—and will continue to be in the foreseeable future—scores on standardized tests. Students in the K-12 public school system take annual high-stakes tests, which determine how much students have increased their knowledge and skills during the academic year.

Student achievement outcomes are collected and analyzed for the purpose of measuring student growth, and researchers and practitioners alike continually seek to identify which factors most contribute to an increase in student achievement. In theory, if particular inputs to the educational system can be identified and separated, the ones most impacting student performance can then be targeted for improvement. One such input of interest is teacher quality, which research has demonstrated to be one of the most important contributing factors to student achievement (e.g., Rivkin, Hanushek, & Kain, 2005). Researchers continue to explore which particular aspects make an individual teacher highly effective and to what degree highly effective or ineffective teachers affect student performance on standardized tests.

Although teacher quality is important to the educational process, teacher evaluation systems have been criticized for their lack of reliability, validity, and objectivity. In response to criticisms and the Obama administration's Race to the Top initiative, many states are adopting value-added models (VAMs) to evaluate and predict teacher effectiveness. VAMs are growth models designed to capture student performance on standardized tests over several years and match that data to teachers and schools. In simplest terms, VAMs compute individual teacher effectiveness by measuring student achievement data and controlling for other variables that may impact student performance.

Although there is much debate over whether VAMs are a valid and reliable measure of teacher quality, many school districts are still relying upon these growth models to evaluate general education teachers.[1] There are considerable impediments to including standardized test scores of students with disabilities in VAMs, which makes measuring the quality of special education teachers very difficult. School districts should correctly evaluate, promote, and retain high-quality special education teachers, but further research is needed to

[1] General education teachers, sometimes referred to as regular education teachers, are those licensed/certified to teach specific grade levels and/or specific subject areas, not including special education.

determine the extent to which VAMs can assist in this process. This literature review begins by critically examining the development and use of VAMs as an evaluative measure for general education teachers in public schools. There are methodological and practical concerns associated with VAMs in general, and these concerns serve as a necessary conceptual foundation for discussing the particular application of VAMs to special education teachers. The review continues with an in-depth analysis of the issues pertaining specifically to the use of VAMs for measuring special education teacher quality.

An Overview of the Literature on VAMs

Educational Production Function

Educational production function analysis is often traced back to the Coleman Report (Coleman, 1966), which is one of the best known and most controversial inventories of school resources (Hanushek, 1986).[2] Mandated by the Civil Rights Act of 1964, the Coleman Report surveyed more than half a million students in 3,000 schools across the United States. The study sought to examine various factors that influence the education process, including school resources, family background, student characteristics, teacher quality, and school quality. The Coleman Report is often cited as the beginning of extensive examinations into which inputs produce the greatest gains in student achievement.

To consider our nation's educational system as a production function, one must accept that the output of the educational process—individual student achievement—is directly related to a series of inputs (Hanushek, 1986). Not all people, however, believe that the outputs of the educational process can be so easily defined and measured. Test scores are in large measure relied upon as evidence of student achievement, although many believe that standardized test scores are a narrow reflection of individual student progress (Koretz, 2002). According to Hanushek (1986), "some people, including many school practitioners, simply reject this line of research entirely because they believe that educational outcomes are not or cannot be quantified" (p. 1150).

While it can be argued that a student's growth goes well beyond scores on purely academic measures, many educators continue to believe that test scores have a place in measures of education. The advantage of standardized test scores

---

[2] Used in economics, a production function is an equation that expresses the relationship between inputs (e.g., capital and labor) and outputs (e.g., goods and services). When applied to education, a production function considers how various inputs (e.g., teachers, peers, and family resources) impact outputs of interest (e.g., student test scores, graduation rates, and college attendance).

is that they provide an observable and measurable outcome variable, whereas other forms of student growth and achievement (e.g., social-emotional and behavioral) remain much more difficult to operationally define and objectively measure. Educational production function studies rely upon standardized test scores as an outcome variable, and educational systems currently use test scores to provide at least some indication of student growth, teacher quality, and school quality.

Teacher Quality

It is generally acknowledged that teacher quality is one of the most important factors in promoting student achievement; however, there is no consensus on which factors define, signal, and enhance teacher quality (Harris & Sass, 2011). One line of research has examined observable teacher qualifications such as undergraduate degrees, graduate degrees, and years of experience, as possible factors in differentiating effective from ineffective teachers. Results of these studies have been mixed (see the review section in Harris & Sass, 2011), and the inconsistent findings have caused researchers and practitioners to question whether teacher qualifications have any impact at all on student achievement.

To explore this issue further, Harris and Sass (2011) sought to identify specific and observable teacher qualifications that contribute to teacher productivity using student-level achievement test data for both math and reading as an outcome measure. The study analyzed the impact of teacher experience, post-baccalaureate degrees, in-service professional development, and pre-service undergraduate education. The study also distinguished between forms of training, types of coursework at the undergraduate level, and the quality of undergraduate training, controlling for the innate ability of future teachers as measured by college entrance exam scores. The results of the study indicate that experience increases teacher productivity at the elementary and middle school levels, but formal training acquired while teaching does not. The attainment of advanced degrees does not impact student achievement, in-service professional development has little or no effect on teacher productivity, and specific undergraduate coursework in education also has no effect. In sum, observable teacher qualifications appear to be only weakly related to student achievement (Harris & Sass, 2011; Koedel & Betts, 2007).

Although observable teacher qualifications do not directly produce student achievement gains, the federal government does demand that teachers be highly qualified. According to No Child Left Behind (NCLB, 2002), a highly-qualified teacher is one who possesses a bachelor's degree, has full state certification or licensure, and demonstrates competency in the subject area to be taught. A teacher must be deemed "qualified" in order to provide instruction, but, as Harris and Sass (2011) demonstrate, these qualifications do not necessarily make a

teacher effective. A high-quality or effective teacher is generally thought of as one who delivers instruction that helps students learn, and student learning is generally demonstrated through increased scores on assessments. Teachers in the United States are asked to produce, and their production is measured by how well their students perform on academic achievement tests. In keeping with the research studies reviewed in this paper, a quality or effective teacher is defined as one who contributes to students' academic growth as measured by student achievement test scores.

The federal government continues to focus on the idea of teacher quality because there is strong evidence that teacher quality is one of the most important contributors to student achievement (e.g., Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Wright, Horn, & Sanders, 1997). Nye et al. (2004) found that which teacher a student happens to be assigned to within a school matters more than which school a student attends. In this experimental study, which randomly assigned students to classrooms within 79 elementary schools across 42 school districts in Tennessee, the authors found substantial differences among teachers in their ability to produce student achievement gains, and those teacher effects were larger than school effects. In low-socioeconomic status (SES) schools, there was a much larger distribution of teacher effectiveness than in high-SES schools; therefore, for students in low-SES schools, teacher assignment matters more than for students in high-SES schools. The authors state that "naturally occurring teacher effects are typically larger than naturally occurring school effects" (p. 247), which suggests that policies focusing on school choice are less promising than those that focus on improving individual teacher quality.

Teacher Evaluation Systems

Even though teacher quality has been demonstrated to be important to student achievement, existing educator evaluation systems have been criticized for failing to properly identify and subsequently remove ineffective teachers, which are those that fail to deliver "instruction that helps students learn and succeed" (Weisberg et al., 2009, p. 5). In a survey of 15,000 teachers and 1,300 administrators in 12 districts across four states, Weisberg et al. (2009) found that 99 percent of teachers receive a rating of satisfactory in districts that use binary evaluation ratings (i.e., 'satisfactory' or 'unsatisfactory'). In districts that use more rating options, less than one percent of teachers are rated as unsatisfactory.

Additionally, how teachers are evaluated is of considerable concern. Weisberg et al. (2009) note that teacher evaluations are often based upon classroom observations that are typically short (often one class period or less than 60 minutes) and infrequent (two or fewer classroom observations per academic year). As Weisberg et al. (2009) suggest, if equal classroom effectiveness is

assumed and teacher efficacy is not reliably and validly measured, then "[e]xcellent teachers cannot be recognized or rewarded, chronically low-performing teachers languish, and the wide majority of teachers performing at moderate levels do not get the differentiated support and development they need to improve as professionals" (p. 4). Teacher evaluation, then, should do much more than deem a teacher satisfactory or unsatisfactory; strong evaluation systems should ideally recognize effective teachers, identify and support teachers who need to improve, and ultimately serve as a basis for removing ineffective teachers.

Race to the Top

The Obama administration has made an effort to improve upon teacher evaluation systems through a competitive grant program entitled Race to the Top. The program offers large monetary incentives for states that demonstrate success in raising student achievement, as measured by standardized test scores. States must create a plan to accelerate their reform, and a major component of the plan must be specific steps to improve the evaluation of teacher and principal effectiveness based on performance. According to the U.S. Department of Education (2009), states are eligible for funding if they establish an approach to measuring student growth, design fair evaluation systems that take student growth into account as a significant factor, and provide teachers with data on the growth of their students. Schools must also use the data on student growth to inform decisions regarding compensating, retaining, and removing tenured and untenured teachers. In essence, Race to the Top places emphasis on measuring student achievement and linking student growth directly to teachers and principals. School districts are required to create systems whereby teachers receive recognition for their direct contribution to student growth or are removed should their instruction fail to result in increases in student achievement.

In the Race to the Top executive summary, the U.S. Department of Education (2009) defines effective teachers as those "whose students achieve acceptable rates (*e.g.*, at least one grade level in an academic year) of student growth" (p. 12) and highly effective teachers as those "whose students achieve high rates (*e.g.*, one and one-half grade levels in an academic year) of student growth" (p. 12). Race to the Top will only consider funding states that evaluate teachers "in significant part" (p. 12) by student growth; consequently, many states have already adopted or are in the process of developing teacher assessment systems that use student achievement data as the primary outcome variable of interest (Winters & Cowen, 2013). According to Winters and Cowen (2013), 19 states have developed policies that dismiss teachers for ineffective teaching, and 13 of those states use student achievement data as the primary determinant of ineffective teaching.

Value-Added Models

In an effort to improve teacher evaluation systems, and in response to the Race to the Top initiative, many states and school districts have adopted value-added models to measure teacher quality based on student growth. One of the best known and most widely used models, the Educational Value-Added Assessment System (EVAAS) or Tennessee Value-Added Assessment System (TVAAS), was first developed by William Sanders in 1993 for use in Tennessee (Sanders & Horn, 1998). The model then, and as it is often applied now, involves student achievement data, as measured by standardized test scores from two or more years, matched to teacher and/or school-level data (Buzick & Laitusis, 2010). VAMs are essentially statistical models that take into account student prior achievement on standardized tests to estimate a teacher-specific effect on achievement (Holdheide, Browder, Warren, Buzick, & Jones, 2012). VAMs also attempt to control for other variables, including student and school characteristics such as race, peer influence, and percentage of students receiving free and reduced price meals.

According to Holdheide et al. (2012), VAMs are an improvement upon other systems of teacher evaluation because they provide a standardized, common metric; are based on large-scale standardized assessments with more desirable psychometric properties; can be evaluated for validity; and do not require students to meet set proficiency levels.[3] VAMs take into account students' prior achievement so as to measure growth, rather than focusing on a uniform achievement target across all populations. VAMs are also intended to make causal inferences about a teacher's direct influence on student achievement (Holdheide et al., 2012). In other words, VAMs are designed to isolate and quantify a teacher's direct impact on student learning, and a teacher's quality score can be ranked relative to the scores of other teachers in the same school or district.

Assumptions

Most value-added modeling, which purports to measure teacher quality, focuses exclusively on standardized test scores as an outcome of interest (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). Since VAMs rely solely on standardized test scores, the use of VAMs then requires a set of

[3] Under NCLB, states are mandated to set proficiency targets on standardized tests. Committees in each state determine a "cut score" (e.g., 90 out of 100 correct answers on a given test) that students must reach to demonstrate proficiency in an academic area.

assumptions about those scores (Hill, 2009). The assumptions, as outlined by Hill (2009), are presented in the following discussion.

Scores represent quality. If VAMs are to measure teacher quality using standardized test scores, then it must be assumed that value-added scores converge with other measures of quality. In essence, different student achievement assessments should produce the same results and scores should not be affected by varying district, school, and classroom inputs (Hill, 2009). However, Lockwood et al. (2007) found that different outcome measures (i.e., different student assessments) do affect teacher ranks, and in a follow-up to that study, Papay (2009) found that "using different achievement tests produces substantially different estimates of individual teacher effectiveness" (p. 166). Thus, a highly ranked teacher according to one type of student assessment may be deemed an ineffective teacher when using a different type of student achievement assessment in a value-added model.

The assumption that student achievement scores accurately represent teacher quality is also violated if VAMs do not control for the multiple influences on student learning. While teacher quality may affect student performance, there are a variety of factors that also influence student gains. Family resources, school resources, student health, family mobility, and the influence of peers are but a few factors that may influence how well students are able to perform on standardized tests (Baker et al., 2010; Braun, 2005; Hill, 2009; Koretz, 2002; McCaffrey, 2012). In addition, it is difficult to disentangle the effects of multiple teachers over time (Baker et al., 2010).

A major violation of the assumption that scores represent quality frequently discussed in the literature is the non-random sorting of students into classrooms and of teachers into schools and classrooms (e.g., Baker et al., 2010).[4] It is widely acknowledged that some parents choose to live in particular areas for the purpose of enrolling their children in particular schools. Parents can also influence which teachers their children are assigned, and can request that a child be moved from one teacher's classroom to another. Districts may also assign particular teachers to particular schools, and then, schools can make non-random

[4] Randomization is necessary when making causal inferences because the statistical model must ensure that the teacher effects alone produced changes in student achievement scores and not some other variable. For example, if students are not randomly assigned to classrooms, it could be that one teacher is assigned students who all receive more academic support at home. The assistance from parents could produce higher achievement gains, and the teacher may have very little influence over the students' performance. Without random assignment, teacher effects cannot be isolated from other possible influences on student learning.

assignments of teachers to grade levels and classrooms. Some principals elect to give the highest achieving students to the strongest or most experienced teachers, leaving newer and weaker teachers with lower achieving and more challenging populations of students. When students and teachers are not randomly assigned to classrooms, confounding factors can obscure the true effects of teachers on student achievement.

Scores are reliable. A second assumption within VAMs is that standardized test scores are reliable and unbiased. Teacher ranking should also be relatively stable over time (Hill, 2009). The testing of this assumption has produced mixed results in the research literature. Kane and Staiger (2008) found that "standard teacher value-added models are able to generate unbiased and reasonably accurate predictions of the causal *short-term* impact of a teacher on student test scores" (p. 33); however, the authors also found significant fade-out effects, suggesting that VAMs may not yield consistent predictions about a teacher's long-term impact.

A study conducted by McCaffrey, Sass, Lockwood, and Mihaly (2009) found that VAMs do not produce consistent results. In five large urban districts, VAMs were used to measure teacher quality for two consecutive years. The authors found that teachers were ranked with only moderate stability from year to year: "roughly one-third of top-quintile teachers remain in the top quintile the next year, while approximately one in ten falls to the bottom quintile of the teacher effectiveness distribution" (p. 599). Thus, only one third of highly ranked teachers in one year will continue to be highly ranked the following year; furthermore, one in 10 highly ranked teachers will fall to the bottom of teacher rankings in a subsequent year. The inconsistent results appear to violate the assumption that VAMs are able to produce stable data on teacher quality over time.

Scores are free from manipulation. One of the most controversial assumptions about VAM scores is that they are not vulnerable to corruption (Koretz, 2002). A longstanding concern regarding the dependence on standardized test scores as an outcome measure is that teachers will knowingly or unknowingly "game the test" (Hill, 2009). When the stakes are high, teachers and schools feel enormous pressure to produce student achievement gains on standardized tests. Schools will often narrow the curriculum (Baker et al., 2010; Koretz, 2002) such that it aligns as closely as possible with the content covered on high-stakes tests (Koretz, 2005). Teachers will also adjust what is taught and when it is taught so that students are as prepared as possible for the test questions. Some teachers "coach" students on test content (Koretz, 2005), which may inflate test scores, but the most egregious offense occurs when teachers actually supply answers to students during test administration. Although cheating is difficult to

prove, Koretz (2002) notes anecdotal evidence of teachers giving answers directly to students, and Baker et al. (2010) report numerous cheating scandals.

Another problem with VAMs is simply that they are prone to measurement error (Hill, 2009; Koretz, 2002; McCaffrey et al., 2004). Student test score data is gathered from relatively small classes, and often with missing values (McCaffrey et al., 2004). As with any statistical model, there is always some amount of measurement error; in the case of evaluating teachers using VAMs, error presents a challenge to obtaining precise estimates of teacher effects (McCaffrey et al., 2004). As so aptly stated by Hill (2009):

> Rather than assuming a value-added score is an indicator of teacher quality or effectiveness, as is often done in current debates, we must more accurately characterize these scores as representing not only teacher quality but also bias due to student selection, the effect of other resources on student achievement, and a generous amount of measurement error. (p. 706)

VAMs may potentially offer some estimate of teacher quality, but researchers argue that the estimate is neither unbiased nor error free. Careful consideration of the aforementioned assumptions is necessary when interpreting value-added scores and applying those scores toward teacher evaluations and rankings.

Further Criticisms

Although many school districts have adopted VAMs for teacher evaluation, there is widespread concern among researchers over the aggressive pursuit of VAMs. Of practical concern are the extensive computing resources required to perform high-quality longitudinal data analysis. Many districts simply do not have the equipment, personnel, or expertise to perform the necessary computing for VAMs (McCaffrey et al., 2004). For states and school districts that do possess the knowledge and equipment, there is increased concern over model and policy designs. According to Goldhaber, Goldschmidt, and Tseng (2013), estimates of teacher effectiveness are model specification dependent, meaning that "the choice of how to control for student ability is a nontrivial one" (p. 229). In a study conducted by Winters and Cowen (2013), a larger number of teachers would be removed under a dismissal policy that averaged teacher scores over two years than under a policy based on consecutive years of below-standard performance, even when both policies used the same percentile cutoff. How districts design their policies, set performance criteria for teachers, and choose the specific value-added model to employ will all affect teacher rankings.

There are several other practical concerns that states and districts should consider before incorporating VAMs into teacher assessment systems. Currently, there is a lack of appropriate tests for all grade levels and subjects, which affects a

district's ability to use student achievement data as an outcome measure. This is especially problematic in middle and high schools when students are exposed to several subjects, often taught by several different teachers. In lower grades, students are typically assigned to one teacher for the duration of the year, and high-stakes tests are administered at one time point during the academic year. In middle school, and even more so in high school, high-stakes tests do not cover every subject that is taught, and it is difficult to control for the effects of exposure to various teachers throughout the school day.

What is most disconcerting are the implications that the rampant use of VAMs will create disincentives for teachers to work with the neediest and most challenging student populations and that teachers will be less likely to work cooperatively with other teachers as the field becomes increasingly competitive (Baker et al., 2010). It is possible that teachers will limit sharing of instructional techniques and materials with colleagues and that individual performance will take precedence over the collaborative efforts necessary to improve schools as a whole.

Still a Better Evaluation System

Despite the criticism and concerns, some still argue that teacher evaluation systems based at least in part on VAMs are better indicators of teacher quality and will yield improvements in teacher effectiveness over time (Goldhaber & Hansen, 2010; Glazerman et al., 2010). Several researchers argue that VAMs should not be used alone when making decisions about teacher effectiveness, but including VAM scores can help meaningfully differentiate teacher performance (Glazerman et al., 2010). Glazerman et al. (2010) aptly compare the use of VAMs to the nation's dependence on SAT scores as a predictor of success in college. While SAT scores are not a perfect predictor, they are thus far the best available and are still a heavily weighted component of a student's college application. As Glazerman et al. (2010) argue, VAMs are not a perfect estimate of teacher quality, but they are able to capture at least a portion of a teacher's impact on student performance, and should not be so easily dismissed because of methodological or practical concerns. An imperfect classification system—as are VAMs—is better than district evaluation systems that rate nearly every teacher as satisfactory while student achievement rates continue to lag.

Measuring Special Education Teachers

Improving accountability measures for teacher effectiveness has become increasingly important for students with disabilities (SWDs), who are often educated within general education classrooms but generally exhibit lower

performance on standardized tests (e.g., Thurlow, Bremer, & Albus, 2011). In the United States, approximately 10% of students in K-12 public schools receive special education services (U.S. Department of Education, 2011). The majority of the students are educated in mainstream classrooms, often co-taught by general and special education teachers, especially since the Individuals with Disabilities Education Act (IDEA) mandated that special education students be educated in the least restrictive environment.[5]

Although students with disabilities are included in general education classrooms, they are often left out of the student data sets used in VAMs. This may occur for a variety of reasons, one of which being that SWDs often take alternative or modified state assessments if determined by an Individualized Education Plan (IEP) team.[6] Scores from general, modified, and alternate assessments are on different scales and it may be impossible to combine them in some longitudinal models (Buzick & Laitusis, 2010). In a survey of 15 states using growth models for teacher assessment, 13 of those states did not include students who took alternate assessments in their growth model outcomes (Ahearn, 2009). Feng and Sass (2010), conducted a study using VAMs to measure teacher quality, but SWDs instructed in co-taught classrooms were eliminated from the model. The exclusion of SWDs from VAMs is problematic if teacher assessment systems are ever to be fairly applied to all teachers, including special educators.

There is a paucity of research on the use of VAMs to measure the quality of special education teachers because special education programs, classrooms, and students present unique challenges—including those aforementioned—that complicate straightforward growth models. These challenges violate the previously discussed assumptions of VAMs, and render estimates of special education teacher quality rather difficult. Four major areas of concern, discussed in more detail below, include the testing instruments, the use of testing accommodations and modifications, multiple influences on student learning, and nonrandom assignment.

---

[5] IDEA mandates that children with disabilities be educated with children who are not disabled to the maximum extent possible. Removal from the regular educational environment must only occur when the nature or severity of a child's disability warrants additional support.

[6] An IEP team must include, at a minimum, the following members: the child's parent(s), at least one regular education teacher of the child, at least one special education teacher of the child, a representative of the public agency, other individuals who have expertise in areas of related services (when appropriate), and the child with the disability (when appropriate).

Testing Instruments

As previously mentioned, SWDs often exhibit lower performance on state assessments than their general education peers (Thurlow, Bremer, & Albus, 2011). SWDs only qualify for special education if a suspected area of disability is adversely impacting educational performance; therefore, by definition, SWDs will generally score lower on standardized tests. The purpose of special education instruction and support is to assist SWDs in achieving growth, but the amount of growth on purely academic measures may not occur at the same rate or to the same degree as typically developing peers. A dependence on standardized test scores alone as an outcome measure for special education teacher quality may provide an inaccurate and incomplete picture of the extent to which the special education teacher is contributing to student growth.

When SWDs do take state assessments, they have additional options that distinguish their assessments from those of their general education peers. If agreed upon by an IEP team, a special education student may take the general assessment, the general assessment with accommodations, the general assessment with modifications, a state-adopted modified assessment, or a state-adopted alternate assessment. In addition, SWDs have the options in combination, meaning that a student could take the general assessment with accommodations for English Language Arts, but take the modified assessment for Mathematics. Special education students often move between the general, modified, and alternative assessments in different years and for different subjects (Buzick & Laitusis, 2010). The lack of consistency in test-taking patterns and the psychometric properties of alternate and modified assessments make it nearly impossible to use the standardized test scores of SWDs as a reliable outcome measure in VAMs.

Testing Accommodations and Modifications

Testing accommodations and modifications are special education supports that are available for SWDs and are designed to improve accessibility. Testing *accommodations* are those that do not alter the construct being measured; for example, the use of large print on tests that do not measure vision, or extended time on tests that do not measure speed (Buzick & Laitusis, 2010). According to Koretz and Hamilton (2006), "The psychometric function of accommodations is to increase the validity of inferences about students with [disabilities] by offsetting specific disability-related, construct-irrelevant impediments to performance" (p. 562). Testing *modifications* are changes that do alter the construct being measured; for example, providing the use of a calculator for test

questions designed to measure multiplication. When SWDs make use of testing modifications, measurements of growth cannot be directly interpreted.

Testing accommodations and modifications are added and/or removed from a student's IEP annually, at a minimum. IEP teams convene to re-evaluate which accommodations and modifications are necessary for individual students. Once accommodations and/or modifications are written into a student's IEP, it simply means that those supports are available to the student, but it is the student's choice as to whether or not to avail him or herself of those supports. The annual review and amendment of IEP accommodations/modifications, along with students' freedom to refuse the accommodations/modifications, result in variability in the number and type of testing accommodations/modifications used from year to year (Fuchs & Fuchs, 2001). Variability is the result of the changing needs of SWDs over time or the changing preferences of the student, but variation across years can also be due to external factors such as changes to state policy (Christensen, Lazarus, Crone, & Thurlow, 2008).

The problem with the inconsistent application of testing accommodations is that standardized test scores can inflate or deflate depending on the addition or removal of supports (Jones, Buzick, & Turkan, 2013). Research indicates that the use of testing accommodations results in differential score changes for SWDs (Sireci, Scarpeti, & Li, 2005). Especially when testing accommodations are added, a performance boost may occur in one particular year for a student with a disability. According to Buzick and Laitusis (2010), "The implication is that the change in test scores from year to year may be related to inconsistency in the use of accommodations and modifications rather than true changes in knowledge, skills, and abilities over time" (p. 540). It is difficult then to isolate true special education teacher effects from the effects of testing supports in the growth of SWDs' scores.

Multiple Influences on Student Learning

As previously mentioned, the assumption that VAMs actually measure teacher quality is violated when individual teacher effects cannot be isolated from other influences on student growth (Holdheide et al., 2012; Jones et al., 2013). For special education teachers, this assumption is often violated because of the nature of special education programs. SWDs, especially those with learning disabilities or emotional and behavioral disorders, often move into or out of special education or experience changes to the type and frequency of services. Some special education students receive services for part of the school day and may be pulled out of the general education classroom for specific services. For example, a student who receives speech and language services will typically be pulled out of the classroom for individual or group therapy on a daily or weekly basis.

In elementary schools, SWDs may receive reading or math instruction in a separate classroom or program depending on individual student need. In middle and high schools, it is common to see SWDs enrolled in a special education class for one subject area, but be integrated with general education peers for a different subject area. For example, a student with a learning disability who experiences difficulty with math calculation may receive additional special education support in a mathematics class but does not require special education services for other subject areas. This variability in instruction, frequency of services, and service providers makes it much more challenging to make causal inferences about an individual teacher's effects on student growth.

Another challenge to estimating the impact of special education teachers on the educational performance of SWDs is the influence of peers. Part of the mandate to educate SWDs in the least restrictive environment is the explicit assumption that SWDs benefit from interaction with their non-disabled peers (Feng & Sass, 2010). Many SWDs have goals for social-emotional, behavioral, or social skills written into their IEPs for the express purpose of improving academic outcomes by way of targeting other impeding factors. Special education teachers often address these goals through examples and interactions with typically developing peers. As students learn to engage and interact appropriately with classmates, they are able to be more successful in general within a classroom environment. A special education student's academic growth, then, could be partially attributable to the influence of his or her general education peers.

In addition to peer influence, individuals other than the teacher of record may impact growth for SWDs. In co-taught or collaborative classrooms, general education and special education teachers team up to teach both general and special education students in a mainstream classroom. In these cases, it is impossible to isolate individual teacher effects on student performance since both teachers are responsible for instruction. Classroom support for SWDs may also be offered in the form of assistance from an instructional aide. Whether assigned to individual students or to the class as a whole, the instructional aide typically assists with learning and ideally contributes to academic growth. In sum, the extra resources provided to SWDs obscure the effects of an individual teacher on achievement gains.

Nonrandom Assignment

Nonrandom assignment of students to teachers or teachers to specific classrooms threatens the validity of causal inferences in growth models. This is of particular concern for special education teachers because they often do not have a choice in their assignment. Some principals opt to assign highly effective and/or experienced teachers to classrooms of SWDs with lower test scores on average, which results in an underestimation of the true effectiveness of the

teachers because the students may generally exhibit more difficulty with performance on academic tests (Feng & Sass, 2010). Other principals choose instead to assign ineffective or inexperienced teachers to special education classrooms (Feng & Sass, 2010; Holdheide et al., 2010). Since newer teachers often do not have a choice in their classroom assignment, many principals will assign them to more difficult student populations that are not already taught by veteran teachers. In practice, special education teachers are rarely randomly assigned to their classrooms, which violates the assumption that VAMs provide an accurate estimate of true teacher quality.

Conclusions

Researchers and practitioners alike acknowledge the important influence teachers have on improving academic outcomes for students. There is also general agreement that most teacher evaluation systems used across school districts fail to differentiate between highly effective and highly ineffective teachers. Of utmost concern is that inadequate evaluation systems are considerably lacking in their ability to identify ineffective teachers before those teachers receive tenure, after which time they are nearly impossible to remove from a school district. Students who happen to be assigned to the classroom of an ineffective teacher run the risk of academically languishing at best and regressing at worst.

VAMs offer an opportunity to improve upon teacher evaluation systems by providing objective outcome measures, rather than relying upon traditional teacher evaluations based largely on a small number of classroom observations performed by administrators. By collecting student test data from multiple years, and controlling for as many potentially confounding variables as possible, VAMs can produce estimates of teacher effects on student growth. There are several methodological and practical concerns with the use of VAMs, but growth models have at least some potential to reform and improve teacher evaluation.

While there are challenges associated with measuring general education teachers using VAMs, measuring special education teachers is even more problematic. Special education teachers often work collaboratively with general education teachers, service providers, and support staff to ensure growth for students with disabilities. The field of special education is inherently collaborative, as families, schools, and communities pool their resources to maximally support students with special needs. Special education teachers are not encouraged to work in isolation, which makes isolating their individual effects on student achievement very difficult.

A major concern is the dependence on special education students' standardized test scores as outcome measures in VAMs. Special education

students should demonstrate improved academic gains, but there is a great deal more involved with supporting a special education student's overall growth and success in school. Special education teachers serve as case carriers for SWDs, and they are charged with identifying areas that may be impeding academic performance; those may be social, emotional, behavioral, or other areas that adversely affect a student's ability to make progress in the general education classroom. A child with Autism Spectrum Disorder (ASD) may exhibit social deficits, which impact his or her ability to participate in academic activities within the classroom. A student with an emotional disorder may withdraw and have difficulty fully engaging with a teacher's instruction. A student with a behavioral disorder may need assistance with behaviors that are impacting his or her ability to appropriately manage academic tasks. These are merely a few examples of areas that a special education teacher will target and teach to in order to offer a student with a disability the best chance of achieving his or her potential in an academic environment.

Special education teachers need to be held accountable for student growth, and SWDs should exhibit at least some gains on academic measures. Special education students are entitled to be educated by highly-qualified teachers, and evaluation systems should reveal as accurately as possible an individual teacher's effectiveness. VAMs offer the potential to objectively estimate teacher quality, but a great deal more research is needed to address methodological concerns. Future research should examine the impact of modified assessments and testing accommodations, especially as states prepare to assess special education students on Common Core Standards.[7] Future research should also explore ways in which VAMs can account for variations in special education programs, including co-taught classrooms, services delivered outside of the classroom setting, and multiple teachers for multiple subject areas. Standardized test scores have a place in measures of special education students' achievement, and VAMs could very well become a predominant component of special education teacher evaluations. Research is needed to determine to what degree and reliability VAMs can include scores of SWDs in the measure of special education teacher quality, and to what degree VAMs truly capture a special educator's effect on student growth.

---

[7] The Common Core Standards are a single set of clear educational standards designed to ensure that students finish their K-12 public education fully prepared for two and four year college programs or the workforce. The Common Core Standards have been voluntarily adopted by forty-five states and the District of Columbia. For more information, see www.corestandards.org

References

Ahearn, E. (2009). *Growth models and students with disabilities: Report of state interviews*. Alexandria, VA: National Association of State Directors of Special Education. Retrieved from http://www.projectforum.org

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. L., Linn, R. L., …Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: The Economic Policy Institute.

Braun, H. I. (2005, September). *Using student progress to evaluate teachers. A primer on value-added models.* Washington, D.C.: ETS, Policy and Information Center. Retrieved June 6, 2013, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39*(7), 537-544. doi: 10.3102/0013189X10383560

Christensen, L. L., Lazarus, S. S., Crone, M., & Thurlow, M. L. (2008). 2007 *State policies on assessment participation and accommodations for students with disabilities* (Synthesis Report 69). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Coleman, J. S. (1966). *Equality of educational opportunity* [NBER Working Paper 12828]. Washington, DC: U.S. Dept. of Health, Education, and Welfare, Office of Education.

Feng, L., & Sass, T. R. (2010, June). *What makes special education teachers special? Teacher training and achievement of students with disabilities* [Working Paper 49]. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research (CALDER). Retrieved June 6, 2013, from http://www.caldercenter.org/upload/CALDER WorkPaper_49.pdf

Fuchs, L., & Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research and Practice, 16*(3), 174-181. doi: 10.1111/0938-8982.00018

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Washington, DC: Brookings Institution.

Goldhaber, D. D., Goldschmidt, P. & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis, 35*(2), 220-236. doi: 10.3102/0162373712466938

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review, 100*(2), 250-255. doi: 10.1257/aer.100.2.250

Hanushek, E. A. (1986). The economics of schooling—Production and efficiency in public schools. *Journal of Economic Literature, 24*(3), 1141-1178.

Harris, D., & Sass, T. (2011). Teacher training, teacher quality, and student achievement. *Journal of Public Economics, 95*(7-8), 798-812. doi: 10.1016/j.jpubeco.2010.11.009

Holdheide, L., Browder, D., Warren, S., Buzick, H., & Jones, N. (2012). *Using student growth to evaluate educators of students with disabilities: Issues, challenges, and next steps.* Washington, D.C.: National Comprehensive Center for Teacher Quality. Retrieved June 6, 2013, from http://www.isbe.state.il.us/peac/pdf/using_student_growth_summary0112.pdf

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*(4), 700-709. doi: 10.1002/pam.20463

Jones, N. D., Buzick, H. M, & Turkan, S. (2013). Including students with disabilities and English learners in measures of educator effectiveness. *Educational Researcher, 42*(4), 234-241. doi: 10.3102/0013189X12468211

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* [Working Paper Series 14607]. Cambridge, MA: National Bureau of Economic Research.

Koedel, C., & Betts, J. (2007). *Re-examining the role of teacher quality in the educational production function* [Working Paper 2007-03]. Nashville, TN: Vanderbilt Peabody College, National Center on Performance Incentives.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources, 37*(4), 752-777.

Koretz, D. (2005). *Alignment, high stakes, and the inflation of test scores* [Report 655]. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE) Graduate School of Education & Information Studies, University of California.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education and Praeger.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of*

Educational Measurement, 44(1), 47-67. doi: 10.111/j.1745-3984.2007.00026.x

McCaffrey, D. F. (2012). *Do value-added methods level the playing field for teachers?* Carnegie Knowledge Network. Retrieved June 6, 2013, from http://carnegieknowledenetwork.org/briefs/value-added/level-playing-field/

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101. doi: 10.3102/10769986029001067

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606. doi: 10/1162/edfp.2009.4.4.572

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

Nye, B. Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257. doi: 10.3102/01623737026003237

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi: 10/3102/0002831210362589

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*(2), 417-458. doi: 10/1111/j.1468-0262.2005.00584.x

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student achievement* (Research Progress Report). Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256. doi: 10.1023/A:1008067210518

Sireci, S. G., Scarpeti, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490. doi: 10/3102/00346543075004457

Thurlow, M. L., Bremer, C., & Albus, D. (2011). *2008-09 publicly reported assessment results for students with disabilities and ELLs with disabilities* (Technical Report No. 59). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (2009). *Race to the top*. Retrieved from http://www2ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2011). *30th annual report to Congress on the implementation of the Individuals with Disabilities Education Act, 2008*. Washington, D.C.: Office of Special Education and Rehabilitative Services, Office of Special Education Programs. Retrieved from http://www2ed.gov/about/reports/annual/osep/index/html

Weisberg, D., Sexton, S., Mulhern, J., & Kneeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on teacher effectiveness*. New York, NY: The New Teacher Project.

Winters, M. A., & Cowen, J. M. (2013). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher, 42*(6), 330-337. doi: 10.3102/0013189X13496145

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67. doi: 10.1023/A:1007999204543

Author

Janelle Lawson is a third year joint doctoral student in special education at the University of California, Los Angeles and California State University, Los Angeles. She currently teaches courses in the Charter College of Education at Cal State LA. Her primary research interest is in the area of special education law and policy, especially how federal and state policies directly impact local school districts. She is specifically interested in special education teacher evaluation systems and mental health service provision for students with disabilities.