Literature Synthesis

Using Measures of Educator Effectiveness to Strengthen Educator Preparation: Improving Outcomes for Students With and Without Disabilities

George H. Noell
Louisiana State University

Mary T. Brownell
University of Florida

Heather M. Buzick
Educational Testing Service
Princeton, New Jersey

Nathan D. Jones
Boston University

June 2013
CEEDAR Document No. RS-1

Disclaimer:
This content was produced under U.S. Department of Education, Office of Special Education Programs, Award No. H325A120003. Bonnie Jones and David Guardino serve as the project officers. The views expressed herein do not necessarily represent the positions or polices of the U.S. Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this website is intended or should be inferred.

Recommended Citation:

Noell, G.H., Brownell, M.T., Buzick, H.M., Jones, N.D., (2013) Using Educator Effectiveness to Strengthen Educator Preparation: Improving Outcomes for Students With and Without Disabilities (CEEDAR Doc. No. RS-1). Gainesville: University of Florida, Collaboration for Effective Educator Development, Accountability, and Reform.

The CEEDAR Center

University of Florida
360 Norman Hall
PO Box 117050
Gainesville, FL 32611

www.ceedar.org
ceedar@coe.ufl.edu
352.273.4259

Table of Contents

Improving Outcomes for Students With and Without Disabilities 4

Student learning outcomes 7

Value-added modeling 7

Does existing research provide validity evidence for using VAM? 8

Student learning targets 10

Classroom Practice Assessments 12

CLASS—Classroom Assessment Scoring System 13

Does existing research provide evidence of validity and reliability for using CLASS? 14

FFT—The Framework for Teaching 17

Does existing research support the validity and reliability of FFT? 18

Using CLASS and FFT to assess initial preparation for teaching students with disabilities 19

Emerging practices in classroom assessment for teachers working with students with disabilities 21

Perception Surveys 24

Using perception surveys to evaluate preparation programs for teachers working with students with disabilities 25

Summary and Recommendations 26

References 29

Improving Outcomes for Students With and Without Disabilities

With research suggesting that effective teachers are the most important influence on student educational attainment in schools, public policy discourse has increasingly emphasized how to improve educator effectiveness (Aaronson, Barrow, & Sander, 2003; Clotfelter, Ladd, & Vigdor, 2007; Rivkin, Hanushek, & Kain, 1998). Policy work and debate has focused principally on content standards, professional development, teacher evaluation, standards for licensure, and teacher preparation. Although these elements of the educational context and labor market are leverage points for increasing teacher effectiveness, teacher preparation as a proactive intervention that is the source of many new educators each year has unique appeal. Approximately 150,000 new teachers are hired annually, representing approximately 4.6% of the teacher workforce (Feistritzer, 2011).

Federal interest in teacher preparation is evident in the U. S. Department of Education's Race to the Top initiative and the negotiated rule making regarding teacher preparation as an element in the reauthorization of the Higher Education Act. In both instances, the Department has emphasized strengthening preparation and developing accountability for teacher preparation programs based on student learning outcomes. This policy focus on preparation has aligned with findings from an emerging literature suggesting that preparation programs vary in the extent to which recent graduates contribute to student achievement gains (Boyd, Grossman, Lankford, Loeb, & Wycoff, 2009; Gansle, Noell, & Burns, 2012; Goldhaber & Liddle, 2012); see Koedel, Parsons, Podgursky, and Ehlert (2012) for a contrary finding. However, the contributions and interactions of program features, such as recruitment, admissions, field experiences, candidate assessment, pedagogical knowledge, and content knowledge, are less understood.

Although it remains unclear which combinations of features lead to the most successfully prepared new teachers, research in special education has supported the hypothesis that preparation matters. In an analysis of a statewide database, special education teachers who had a degree in special education, a certificate in special education, or 30 hours of course work in special education were found to produce larger student gains in reading than special education teachers who lacked such preparation (Feng & Sass, 2010). In mathematics, special education teachers with advanced degrees in their area secured larger student gains than teachers without such degrees. These findings supported previous research based on classroom observations which suggested that teachers trained in formal preparation programs were more effective than teachers receiving minimal preparation (Nougaret, Scruggs, & Mastropieri, 2005; Sindelar, Daunic, & Rennells, 2004).

Collectively, however, this research has provided little information about the features of special education preparation that contribute to teacher effectiveness. Additionally, the field has developed few rigorous measures for evaluating the quality of teacher preparation programs for special educators and general educators who will serve students with disabilities.

Responding to the interest in stronger preparation for teachers of students with disabilities, policy makers and teacher educators have faced challenging questions, including: (a) how to prepare these teachers, (b) how to measure teacher effectiveness, and (c) how to establish the reliability and validity of the proposed measures. Currently state education agencies and teacher educators rely on process measures without evidence to understand the relationships of each measure to preparation programs or teacher effectiveness (e.g., review of curricular design, surveys of graduates, certification examination pass rates, and portfolio assessments) to make important decisions about programs and graduates (Wineburg, 2006). As policy makers have demanded a higher level of accountability for teacher preparation programs, the existing measures are inadequate and have failed to provide information about how the programs can be improved. The context is made more challenging by the reality that state education agencies and teacher educators have to continue to make decisions now, rather than defer decisions until better measures become available.

As policy makers and teacher educators have considered how to create a system assessing preparation programs for teachers serving students with disabilities, the growing body of research on the evaluation of in-service teachers has offered some direction. Studies of K-12 teacher evaluation have suggested that value-added scores based on student achievement measures, specific tools for observing classroom practice, and evaluations of teachers by principals can be used to identify effective general education teachers (e.g., Bill & Melinda Gates Foundation, 2010, 2012, 2013).

Yet, before applying this research to preparation programs for teachers serving students with disabilities, there are two important considerations:

- The goals and contexts for evaluating teacher preparation programs and evaluating individual teachers are different. As such, promising measures used in the teacher evaluation literature must be reevaluated to examine how well these fit this context and these goals.
- The context for evaluating preparation programs for teachers of students with disabilities presents challenges beyond those inherent in evaluating teacher preparation generally.

These challenges will be discussed in greater detail in the following sections.

States need data on individual preparation programs to determine the degree to which these programs are preparing effective teachers for students with disabilities. Teacher value-added data, scores on valid observation protocols, and valid supervisor rating tools have provided summative evaluation data that can be used to make decisions about programs. Teacher educators, by comparison, have to make both summative and formative evaluation decisions. Teacher educators need:

- summative data to decide if individual teacher candidates can be recommended for licensure
- evaluation data to use in formative ways to identify when program revisions are needed

- evaluation data to develop hypotheses about how aspects of their programs can be strengthened.

The differentiated needs of students with disabilities present additional challenges for assessing the preparation of their teachers. The teacher evaluation literature has offered little guidance on how measures researched for general education teachers (e.g., value-added models, classroom performance assessment, and administrator surveys) perform when applied to evaluations of special education teachers. Most evaluations of classroom instruction have been based on conceptions of effective teaching in general education. Thus far, no research has examined whether existing observation systems are valid and reliable for use with special educators. Students with disabilities have diverse learning needs that are not well represented in existing measures. Academic achievement, social competence, independent living skills, and other outcomes may be equally important. From a measurement perspective, there are substantial limitations in applying what we know from the teacher evaluation literature to the evaluation of teacher preparation programs. For example, we know from the literature on conducting classroom observations that raters frequently make errors when scoring teachers' lessons (Bill & Melinda Gates Foundation, 2013). Although we may expect raters to make similar errors when observing pre-service teachers, these errors are likely to be reduced across multiple teachers. Therefore, as we draw on findings from the teacher evaluation literature, we are also careful to articulate ways in which this evidence may be less useful when evaluating teacher preparation programs.

With these considerations in mind, the purpose of this paper is: (a) to examine the research base on three assessments currently collected in some states and by some institutions of higher education and (b) to consider the degree to which these assessments can be used to evaluate the effectiveness of programs preparing teachers to work with students with disabilities. As such, we present a synthesis of available research about teacher preparation and effectiveness along three domains:

- student learning outcomes
- measures of classroom practice
- supervisors' ratings of educator effectiveness

We discuss the current state of research evidence around measurement in each of these domains, along with some critical considerations for use in evaluating teacher preparation programs. Specifically, we discuss the current status of research examining the indicators' technical adequacy and potential uses in both the formative and summative evaluation of teacher preparation for working with students with disabilities. We conclude each section by describing specific issues relevant to assessing preparation for teaching students with disabilities.

Student Learning Outcomes

One research focus on a potentially informative element of evaluating and improving teacher preparation programs has been to tie preparation to the academic learning of students taught by program graduates once they enter the teaching profession. Such an approach offers face validity in that student achievement is a valued outcome for policy makers and the general public alike. Also, emerging evidence has suggested that student outcomes can predict adult benefits for students taught by teachers deemed as effective based on student outcomes (Chetty, Friedman, & Rockoff, 2011). To use student outcomes in program evaluation has required an analytic model that links teacher preparation programs to student achievement outcomes for recent program completers; for example, see Gansle et al. (2012) and Goldhaber and Liddle (2012). It may also be possible to use some student outcome-based indicators for decisions about the progressions of individual candidates through their training programs and program completions. However, the influence of teachers on student test scores is conflated with other factors, both in and out of school. There are also technical challenges associated with linking student performance and teacher effectiveness. In the following sections, we describe the promise and challenges in using student learning outcomes in program evaluation, focusing our discussion on two categories of indicators based on student outcomes—value-added modeling (VAM) and student learning targets (SLTs).

Value-Added Modeling

Using VAM, researchers attempt to isolate the contribution of programs, individual teachers, or interventions to student learning, typically drawing on large-scale, annual state assessments. Although the bulk of the research in this area has focused on individual teachers, the use of VAM to evaluate educational programs has also been common in research and practice. VAM creates aggregate scores for the unit assessed—teacher preparation programs in our case—derived from the difference between the actual score and predicted score of each student taught by program graduates who have been placed in schools. The predicted score for each student can be computed in a number of ways; but in each approach VAM attempts to control for factors influencing student performance that are outside of the control of the teacher preparation program. Prior-year student test scores are used to do this in all varieties of models. Some models also include student factors (e.g., demographic characteristics); classroom factors (e.g., percentage of students receiving special education services); and school factors (e.g., the percentage of students receiving free or reduced-priced lunch). Each student's predicted score comes from a model that includes all students currently in the same grade taking the same subject-area assessment in a district (or state). See Braun (2005) for a nontechnical description of VAM as well as McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) for a technical introduction.

The advantages of VAM scores in teacher evaluation have also been apparent when used for program evaluation. VAM scores can provide a standardized, objective measure of student learning that is comparable across students in different schools and districts within a state. The technical properties

of assessments used to estimate VAM scores are known or can be routinely obtained—allowing for quantitative analysis and validation. Because many students with disabilities participate in general standardized state assessments, some special education teachers and many general education teachers who teach students with disabilities have scores on effectiveness indicators that reflect the learning outcomes of many students receiving special education services.

The technical properties of VAM scores are strongest when aggregating over many teachers. The study by Feng and Sass (2010) is an example of using VAM in the aggregate. The authors estimated the impact of teacher special education training on the achievement of students with disabilities using a very large sample of teachers in Florida. Teachers whose students took the same assessment can be aggregated—that is, those who are practicing in the same state or, after 2014, those who practice in states in the same consortium (i.e., Smarter Balanced, www.smarterbalanced.org or PARCC, www. parcconline.org). By enlarging the sample of comparable students with disabilities, common assessments can improve estimated VAM scores and make it possible to evaluate specific specializations within preparation programs (e.g., special education).

Does existing research provide validity evidence for using VAM?

In this section we review the empirical research on using student outcome-based indicators for teacher preparation programs. Given that the interest in using VAM scores to evaluate programs is relatively new, we also briefly point out relevant research evaluating VAM for estimating individual teacher effects that has spanned decades. In both bodies of literature, researchers have expressed caution in using student test scores to make high-stakes decisions about teachers and teacher preparation programs; see Floden (2012), which noted several concerns for program evaluation. Nonetheless, as we discussed previously, some useful information can be obtained from student outcome measures to evaluate teacher preparation programs; and VAM is the best available approach.

Using student data to evaluate preparation programs introduces challenges not present in the evaluation of individual teachers. The two major challenges involve selection of students into training programs—see Koedel et al. (2012)—and placement of graduates into schools—see Mihaly, McCaffrey, Sass, and Lockwood (2012). First, it is difficult, if not impossible, to separate out preparation effects from selection effects in the absence of relatively extensive data regarding candidate selection. We cannot tell whether or not a program receives a high score because the program attracted high-quality candidates or because the program transformed graduates into high-quality teachers.[1] Second, not all program graduates are placed into schools, and those who are placed are not randomly placed among various schools in the state.

It is worth noting that from a policy maker's perspective, such as a state education agency, separating

1 This will be a policy as well as a technical question when it is important to separate selection from training effects.

the effects of selection and preparation may not be a policy objective. Rather, the objective may be simply to assure that effective educators are entering the relevant workforce, whether that is as the result of selection or preparation effects notwithstanding. From this perspective, program completers who do not enter the workforce are simply not relevant to the state's policy objectives. In contrast, analyses that shed light on the relative contributions of selection and preparation would obviously be enormously valuable to teacher preparation program leaders.

The implication of these challenges is that VAM may not be able to capture variation in the quality of preparation provided to teachers if programs differ substantively on selection. Koedel et al. (2012) found few differences among programs; although others have found meaningful differences (Boyd et al., 2009; Goldhaber & Liddle, 2012; Plecki, Elfers, & Nakamura, 2012). It is unclear whether or not the mixed results were due to model specification or true characteristics of the data. Mihaly et al. (2012) showed that the rankings of preparation programs were dependent on whether or not school fixed effects were included in the model. Covariate adjustment approaches that control for individual student, peer, and school variables to address differences in school context in models estimating program effects can help mitigate the problem of nonrandom placement of program graduates into schools (Noell, Porter, Patt, & Dahir, 2008).

No research specifically addresses evaluating special education programs with VAM scores. Jones, Buzick, and Turkan (2013) point out specific challenges related to using VAM for individual teachers who educate students with disabilities. For programs that prepare teachers for students with disabilities, the inclusion of variables specific to these students in the model will likely be important when estimating VAMs for program evaluation. Previous research has demonstrated that student disability status is a statistically significant predictor of student achievement after controlling for prior achievement scores (Noell et al., 2008). Accommodations, when applied inconsistently across years, can inflate or deflate test scores (Buzick & Laitusis, 2010); for example, the read-aloud accommodation has been shown to increase student scores on reading and mathematics assessments (Buzick & Stone, 2013). Buzick and Jones (2013) showed that including variables relevant to students with disabilities (e.g., accommodation use, receiving special education services) improved the average rankings from VAM scores for teachers in classrooms in which 50% or more of the students had a documented disability.

If education programs for special education teachers are to be evaluated based on VAM scores, these kinds of decisions (e.g., adjusting for student disability status, including variables for testing accommodations) will become critical. Based on the Noell et al. (2008) findings, excluding student disability status can depress VAM results for programs whose graduates served students with disabilities. Choosing to exclude student disability status from the available predictors may create a circumstance in which programs serving these students are compared unfairly to programs whose graduates serve few students with disabilities or with milder disabilities. If students' disabilities are an important influence on student achievement gains that are beyond the teacher's control, then excluding this from the available predictors penalizes programs that serve students with disabilities.

Another challenge in evaluating programs is that, in many cases, the number of students with disabilities who take the general assessment and can be linked back to a specific preparation program may not be large enough to permit a meaningful disaggregated analysis of special education preparation. Students who take the alternate assessment will not contribute to VAM scores because the technical properties of the assessment do not support use of VAM. It is worth observing that many general educators will have one or more students with disabilities contributing to their results.

Co-teaching by both general and special educators in a single subject for students with disabilities is particularly common. In such cases, it is important to include controls for peer and school context effects; but it is not clear how to attribute student scores in co-teaching situations (Hock & Isenberg, 2012). If both teachers are graduates of the same program, this is a nonissue. When the teachers are from different programs, there is no way to separate out their effects. For individual teacher evaluation, Hock and Isenberg support the full roster method in which a student's scores are fully counted in estimating VAM scores for every teacher who taught the student. This approach could also be applied in program evaluation.

Student learning targets

Although VAMs are the best available approach to estimate the link between teaching and learning, VAM scores are only available for a small percentage of teachers: typically those whose students are in Grades 4-8 who have taken the annual state assessment in mathematics or English language arts or— in some states—science, social studies, and high school end-of-course examinations. To evaluate the effect of teachers on students in untested grades and subjects, states are increasingly using student learning targets[2] (SLTs): goals created by teachers, schools, or districts for individual students or entire classes based on locally or externally created assessments or other classroom-based measures (Buckley & Marion, 2011). Among the benefits, SLTs can be available for essentially all teachers; and teachers may more readily understand these targets and assessments. There is also an inherent appeal in using SLTs for teachers working with students with disabilities because the measures can better represent the learning experiences of their students. However, SLTs require substantial professional work to develop and are not currently standardized. It is unclear whether research involving SLTs will be able to distinguish effective instruction from ineffective instruction. Also no research has evaluated the validity or reliability of SLTs. With these limitations and without controls for students' prior knowledge and student or school characteristics, SLT studies are at risk of inaccurately estimating teachers' contributions to student learning. Consequently, because the current value of SLTs in evaluating teacher preparation programs is limited, we do not include SLTs in the following sections.

2 Student learning objectives (SLOs) is another common term that can be used interchangeably with SLTs. We use SLTs here to avoid confusion with *student learning outcomes*, a term that refers to measures of knowledge, skills, and noncognitive measures acquired by students in institutions of higher education.

Using student learning outcomes to assess initial preparation for teaching students with disabilities.

There are threats to the validity of inferences drawn about teachers from student performances on standardized tests, but VAM scores appear to hold the most promise relative to other measures. Threats to validity and additional challenges carry over into teacher preparation program evaluation. In the face of such challenges, it is prudent for decision makers to proceed cautiously and evaluate the intended and unintended consequences that arise from using any particular approach in an evaluation system. Policy makers will have to weigh the types of decisions they seek to make against the types of evidence available to develop sound public policy that will benefit K-12 students. See Braun (2013) for a thoughtful discussion of how stakeholders should proceed cautiously and iteratively, incorporating audits and feedback given the challenges of VAM for teacher (and program) evaluation.

Student data can be used in program evaluation only when teachers have left the training program and have begun to practice their profession. One important question for researchers and policy makers is the best time to collect the student data—one year out? three years out? ten years out? New opportunities for debate and research will also arise with the administration of new assessments aligned with the Common Core State Standards. This will provide the potential for linking student outcomes to teacher preparation programs for teachers who teach in different states because the student assessments will be the same. This can be beneficial in particular for evaluating special education preparation. Larger student samples will be available as more students with disabilities take equivalent assessments. Researchers will also be able to track teachers who teach in a different state from where they received their pre-service training.

Classroom Practice Assessments

Teacher education programs would benefit greatly from assessments of classroom instruction that provide both formative and summative performance data. These assessments must reflect key dimensions of effective teaching for students with disabilities. Special education has a long history of research that has generated a wealth of information about effective teaching practices.

Improving teachers' use of these practices, i.e., explicit strategy instruction, during content instruction is likely to raise the achievement of students with disabilities, particularly in the areas of reading, writing, and mathematics. If assessments of practice are used to evaluate and improve teacher preparation programs, then faculty in those institutions will need access to assessments of classroom instruction that are valid for that purpose. Additionally, program evaluation data regarding practice will be most useful if it examines the range of critical responsibilities that special educators have, such as collaborating with general education teachers, parents, and other service providers; managing Individual Education Programs; and providing instruction.

We give an overview of two classroom observation systems that are commercially available and may be considered for use in evaluating teacher education programs: FFT—Charlotte Danielson's Framework for Teaching (Danielson, 2007) and CLASS—Classroom Assessment Scoring System (Pianta, La Paro, & Hamre, 2008). Both observation systems are supported by research on their reliability and validity. Applicable across content areas, these offer advantages over subject-specific protocols. Examples are MQI—the Mathematical Quality of Instruction (Hill, Ball, Goffney, & Rowan, 2008) protocol— and PLATO—the Protocol for Language Arts Teaching Observations (Grossman et al., 2010). Specifically, these observation systems can be used to assess all teachers and use their information to gauge how teacher preparation programs writ large are performing.

To assess whether CLASS and FFT would be appropriate for measuring effective teaching for students with disabilities in teacher preparation programs, we organize this section in the following way. First, we introduce each observational system and summarize existing research on reliability and validity. Second, we consider how appropriate the CLASS and FFT observation systems are for evaluating the extent to which teacher preparation programs equip teachers with the skills needed to work with students with disabilities.

In addition to CLASS and FFT, two emerging tools for assessing classroom practice that are currently being validated may prove useful for a better understanding of how well programs are preparing their candidates. The edTPA tool was developed by researchers at Stanford University in collaboration with national professional organizations (Stanford Center for Assessment, Learning, and Equity [SCALE], n.d.). MyiLogs was designed to assess how well general and special education teachers' instruction was aligned with the standards for teaching mathematics and reading and how well that instruction incorporated the use of evidence-based practices (Kurz & Elliott, 2012). We have selected edTPA and MyiLogs because these tools have been developed with instructional

practice for students with disabilities in mind and should provide information to teacher educators who are aligned with what we know about effective teaching for students with disabilities. Unlike the CLASS and FFT, edTPA and MyiLogs can: (a) consider the subject matter being taught, (b) be more useful in providing information at a finer level of detail, and (c) be used to revise the course work and field experiences of teacher preparation programs focused on students with disabilities.

CLASS—Classroom Assessment Scoring System

CLASS is designed to measure classroom quality, concentrating on the interactions between teachers and students in classrooms. Although initially developed for use in pre-K through third-grade classrooms (Pianta, La Paro, et al., 2008), CLASS has since been adapted and validated for use in upper elementary grades and a version of the instrument (CLASS-S) has been developed for secondary classrooms (Pianta, Hamre, Hayes, Mintz, & La Paro, 2008). CLASS is premised on the idea that "the structure and nature of teacher-child interactions likely . . . contribute positively to students' development as a consequence of experience in the classroom" (Pianta & Hamre, 2009, p. 112). Thus, the assessment system quantifies teacher-student interactions in three domains believed to influence students' academic and social outcomes:

- Emotional Support focuses on the degree to which a teacher is able to establish a positive climate, is responsive and sensitive to student needs, and shows regard for students' perspectives.
- Classroom Organization focuses on how well teachers manage behavior in their classrooms, have clear expectations, organize their instruction for learning, and make use of instructional time.
- Instructional Support focuses on how teachers help students develop knowledge of concepts, the quality of feedback teachers provide to students, and how teachers provide support for developing more complex language through their discussions with students.

Researchers who developed CLASS drew on findings from studies funded through the National Institute of Child Health and Human Development (NICHD) that were designed to understand predictors of health, behavior, language, and academic outcomes for preschool through elementary children. NICHD's Early Child Care Research Network (2002, 2005) developed the Classroom Observation System (COS) to examine teacher-student interactions in a series of studies conducted with students and their teachers in prekindergarten, kindergarten, first-grade, third-grade, and fifth-grade classrooms. The COS was comprised of two separate instruments. One was a time-sampling instrument that captured setting and activities (e.g., a teacher-managed activity vs. a child-managed activity, literacy activity); teacher behaviors (e.g., reads aloud, interacts with whole class, interacts with small group); and child engagement. The second was a rating system in which teachers were rated on global behaviors that captured certain child-teacher interactions, such as positive emotional support, classroom management, literacy instruction, and evaluative feedback.

Through a series of large-scale studies, scores on the COS were shown to predict positive student outcomes in language development, reading, mathematics, social competence, and behavior, demonstrating that certain teacher-student interactions were indicative of teaching quality (Hamre & Pianta, 2005; Pianta, La Paro, Payne, Cox, & Bradley, 2002).

Research on the COS then became the basis for developing CLASS, which has also been evaluated in studies involving literacy and mathematics instruction in prekindergarten through high school grades. We organize findings from this research around several key questions important to establishing the instrument's validity.

Does existing research provide evidence of validity and reliability for using CLASS?

The CLASS and its underlying constructs have been studied in multiple rigorous studies. Additionally, researchers have examined the relationship between CLASS and other important outcomes and whether or not training on the behaviors represented in the CLASS results in improved teaching. Moreover, researchers have established the degree to which raters can be trained to rate the CLASS reliably and if performance on the CLASS is a stable indicator of teaching quality within lessons, across lessons, and for different groups of students taught.

Validity. The three CLASS domains (Emotional Support, Classroom Organization, and Instructional Support) were established in NICHD studies (Hamre & Pianta, 2005; Pianta et al., 2002) and then again in studies of CLASS (Bell et al., 2012; Clifford et al., 2005; Downer et al., 2012; Hamre & Pianta, 2005; Howes et al., 2008; La Paro, Pianta, & Stuhlman, 2004; Hamre, Pianta, Mashburn, & Downer, 2007). Researchers used factor analytic techniques to determine whether or not individual items on CLASS represented the constructs of interest (i.e., the three main types of teacher-student interactions). Factor analytic techniques enable researchers to determine if performance items, such as positive climate, correlate with the construct of interest (e.g., Emotional Support). Across the NICHD studies and CLASS studies, researchers found that items on the CLASS correlated with the three constructs of interest. For example, in the Bell et al. study (2012) the correlations ranged from .42 to .92, and from .52 to .95 in the Hamre et al. study (2007). Thus, Hamre et al. argued that these three domains represent critical teacher-student interactions.

Performance on the three CLASS domains has also been linked to desirable student outcomes; and when teachers are supported to change their interactions with students in these three domains, students' academic, social, and behavioral outcomes improve. Numerous studies have demonstrated relationships between performance on the CLASS and student outcomes in prekindergarten, elementary grades, and secondary grades. However, the relationships are complex; and findings are not always consistent across subjects and grade levels (Curby, Rimm-Kaufman, & Ponitz, 2009; Howes et al., 2008; Mashburn et al., 2008; Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008; Ponitz, Rimm-Kaufman, Brock, & Nathanson, 2009). For example, Hamre and Pianta (2005) found that first-grade students who were identified as at functional risk

(i.e., poor performance on academic and social indicators at 54 months) or at demographic risk (i.e., born to mothers who had less than a college education) made achievement gains similar to their peers who were not at risk when they had teachers who scored higher on the CLASS domains of Instructional and Emotional Support. Additionally, these researchers found that students at functional risk experienced fewer teacher-student conflicts when placed with teachers who scored high on Emotional Support. Ponitz et al. (2009) noted more inconsistent findings between CLASS domains and students' reading and mathematics performances. Teachers' scores on the Classroom Organization dimension of the CLASS along with teachers' perceptions of classroom chaos predicted reading achievement gains in first-grade classrooms but not gains in mathematics achievement. Similarly, other researchers have established that performance on CLASS domains, particularly Emotional Support, predicted achievement for students in third, fifth, and secondary grades. However, once again, findings were not consistent across subjects (Allen et al., 2013; Bill & Melinda Gates Foundation, 2012; Pianta, Belsky, et al., 2008). Only one study has failed to establish relationships between CLASS domains and student achievement. Bell et al. (2012) did not find relationships between scores on CLASS domains and scores on students' end-of-year algebra course exams.

Meta-analyses of intervention study findings have shown that explicit strategy instruction best predicts the magnitude of treatment outcomes for students with disabilities; see Swanson (2001) for a review. This instruction involves: (a) orienting students to a task using advance organizers, (b) modeling new skills and strategies for students, (c) practicing skills and strategies over time and in explicit ways, (d) sequencing the difficulty of tasks, (e) probing student responses systematically, (f) prompting and cuing strategy use, (g) elaborating on student responses and materials used, and (h) providing small interactive group instruction. See Swanson (2001) for a summary of findings. These practices should be highlighted in any assessment of instruction for students with disabilities.

Further validation for the CLASS comes from research on My Teaching Partner, a web-mediated professional development program focused on improving teacher-student interactions identified in the CLASS. Pianta, Mashburn, Downer, Hamre, and Justice (2008) showed that teacher-student interactions, as assessed by the CLASS, could be improved when secondary teachers participated in a professional development effort using My Teaching Partner. Additionally, the benefits of participation were greater in classrooms with more students at academic risk. Allen, Pianta, Gregory, Mikami, and Lun (2011) showed that participation in My Teaching Partner not only improved secondary teacher-student interactions on the CLASS but also improved student achievement across subjects. Quality of teacher-student interactions mediated the impact of the intervention on student achievement. Specifically, teachers who exhibited stronger teacher-student interactions had more positive student achievement gains as a result of their participation in My

Teaching Partner. Mikami, Gregory, Allen, Pianta, and Lun (2011) compared teachers randomly assigned to My Teaching Partner and a control group on observations using the CLASS and a self-report of classroom peer interactions. Students of teachers in the My Teaching Partner group demonstrated more positive peer interactions in observations (but not on self-report measures) and moderated the impact of highly disruptive behavior on classroom peer interactions. Thus, developing stronger teacher-student interactions provided a protective effect for students at risk for emotional and behavioral disorders.

Studies of the CLASS have provided evidence that the instrument measures teacher-student interactions underlying effective teaching and that teachers can learn to engage in more effective interactions if provided appropriate learning opportunities. Additionally, the CLASS has appeared to be a tool that can be used across grade levels and different content areas to identify effective teacher-student classroom interactions that are likely to lead to stronger achievement gains.

Reliability. For the CLASS to support valid conclusions about teaching quality, the instrument and accompanying training must be able to produce consistent scores under similar conditions. Researchers must determine if different observers can rate the same teacher similarly and how many lessons and lesson segments they must rate in order to achieve a stable indicator of teacher quality. Further, if the CLASS is to be used across grade levels and content areas to rate teaching quality, it must be able to produce similar results. For example, teachers providing mathematics instruction must be able to obtain scores that are similar to those teaching English, and teachers providing instruction in fourth grade must be able to obtain scores that are similar to those teaching high school; otherwise, scores on the CLASS may result in biased judgments about certain groups of teachers.

Most studies examining the CLASS initially focused primarily on whether or not researchers could be trained to rate teachers on video. Observers were trained to use the CLASS before being released to rate teachers independently (Curby et al., 2009; Downer et al., 2012; Howes et al., 2008; La Paro et al., 2004; Mashburn et al., 2008; Pianta et al., 2005). In these studies, the observers' ratings were compared to those of a master rater. Mean weighted kappas used to calculate the level of rater agreement ranged from .65 to .73, which is considered an acceptable level of agreement (Landis & Koch, 1977). Further, 83% to 93% of raters' individual ratings were exactly the same or within one point of the expert's responses, which is considered excellent for an instrument that requires a fair degree of judgment to score.

More rigorous studies of reliability, however, have yielded somewhat inconsistent findings. Six studies examined the degree to which observers could rate the CLASS similarly over time. Researchers calculated intraclass correlations to determine if there were differences between raters after accounting for how raters varied in scoring individual lessons. In four of the studies, intraclass correlations for each of the three domains and dimensions within those domains were good to excellent, ranging from .60 to .93, with the majority of correlations between .70 and .80 (Allen et al.,

2013; Dominquez, Vitiello, Fuccillo, Greenfield, & Bulotsky-Shearer, 2011; Reyes, Brackett, Rivers, White, & Salovey, 2012). Two other studies found major variations in raters. Bell and colleagues (2012) in their study of algebra instruction found that intraclass correlations were low to moderate, ranging from .24 to .58 for the three domains and dimensions of instruction within those domains. Further, research conducted under the auspices of the MET project (e.g., Bill & Melinda Gates Foundation, 2013) also established considerable scoring variation among raters, but intraclass correlations were not reported. Instead, MET researchers described the amount of variation in teacher performance that was due to differences in raters, differences in how individual raters scored different lessons taught by the same teacher, and differences due to unexplained sources of variation.

To make valid inferences about teaching quality from the CLASS, researchers need to determine how much of the variation in CLASS scores is due to real differences in teaching quality versus variations in raters, different segments of a lesson, and different lessons over time. Bell et al. (2012) found that the amount of variation due to real differences between teachers on the CLASS ranged from 13% for Instructional Support to 35% for Classroom Organization. The remaining variation was attributed to the rater, the lesson segment, the lesson, the time of year the lesson was conducted, and measurement error. In the MET Project, researchers found similar variations in scores due to real differences between teachers' instruction. Further, they established the need to rate four lessons using different observers to achieve a more stable indicator of teaching quality. Under these conditions, they demonstrated that 65% of the variation among CLASS scores was due to persistent differences between teachers. Thus, the MET researchers concluded that in order to use the CLASS reliably, at least four observations of each teacher using different raters must be conducted over the course of the year.

FFT—The Framework for Teaching

The Framework for Teaching Observation Survey (Danielson, 1996, 2007), or FFT, was developed out of PRAXIS III: Classroom Performance Assessments, created by Educational Testing Service to assess teaching skills and classroom performance (Dwyer, 1994). The Framework for Teaching Evaluation Instrument, developed in 2011 and updated in 2013, has been enhanced to support administrators in the evaluation of classroom teachers; it also has been updated to examine the Common Core standards.

The FFT consists of four domains:

- Domain 1: Planning and Preparation;
- Domain 2: The Classroom Environment;
- Domain 3: Instruction; and

- Domain 4: Professional Responsibilities.

These four domains are comprised of 22 components made up of 76 elements. FFT has been adopted in the evaluation systems of some large school districts, such as the Los Angeles Unified School District, and some states, such as Illinois, Rhode Island, and Delaware. Some states like Rhode Island are using modified versions of FFT that only include Domains 2 and 3, the only domains actually observed in practice.

FFT, which was based on an extensive empirical and theoretical literature, is intended to reflect the complexity of teaching. Although CLASS focuses specifically on teachers' interactions with their students, FFT is designed to be comprehensive in nature, including not only teachers' instructional activities but also their other responsibilities. In addition, FFT also aligns with the Interstate New Teachers Assessment and Support Consortium (INTASC) standards, (Council of Chief State School Officers, 2011), the set of competencies the organization recommended for new teachers.

The FFT has theoretical underpinnings in the constructivist approach to learning. From this perspective, individuals develop an understanding of concepts by doing the intellectual work themselves. Individuals interpret new experiences through their existing cognitive structures, so the work of teachers is not simply to provide explicit instruction but is instead to design activities that engage students in constructing their own knowledge. As Danielson (2007) describes:

> It is the premise of the Framework for Teaching that it is important for students—all students—to acquire deep and flexible understanding of complex content, to be able to formulate and test hypotheses, to analyze information, and to be able to relate one part of their learning to another (p. 15).

In short, from the constructivist perspective, teaching is complex and instruction requires a clear instructional purpose. Thus, the complexity of teaching is reflected in the FFT's four domains.

Does existing research support the validity and reliability of FFT?

The research base underlying FFT is not nearly as extensive as the research supporting the validity and reliability of CLASS. However, researchers have examined the degree to which FFT predicts student achievement, and there is also emerging research on whether raters can score FFT reliably and how stable scores are within individuals (relative to other sources of variance). Despite the limited research base supporting FFT, the acceptance of the instrument is indicated by the large number of districts and states that have chosen to adopt FFT over other measures in their teacher evaluation systems.

Validity. The predictive validity of FFT has been examined in several studies. Many have shown that FFT scores correlate with student achievement gains to a modest degree, e.g., Gallagher (2004); Holtzapple (2003); Kimball, White, Milanowski, and Borman (2004); Milanowski (2004). However,

the strength of these relationships has varied across grades and subject areas (Gallagher, 2004; Milanowski, 2004). For example, Holtzapple (2003) found that the correlations between composite FFT scores (summed across the four domains) and student gains on state assessments ranged from .28 to .37, depending on the subject. Other studies have also found that the association between FFT and student achievement gains varies by grade and subject area. This variation in findings may be in part due to the differences in the ways research studies have implemented FFT in practice. For example, studies have varied in their approaches to training raters, the number of times teachers were observed, their adherence to the proposed use of the instrument, and the student tests used. In more recent work, researchers in the MET Study (Bill & Melinda Gates Foundation, 2012) found that the correlation between FFT scores and teachers' underlying value-added scores[3] was .18 in math and .11 in English language arts, in comparison to .25 and .12 respectively for CLASS. Generally, the two measures appear to correlate similarly with teachers' value-added scores, although the slightly lower correlations for FFT and VAM scores might be attributable to the restricted range for FFT scores (FFT uses four scoring categories on its rubrics, and CLASS uses seven).

Reliability. Implementing FFT in practice raises many of the same concerns that were raised in reference to CLASS. To establish rater reliability, FFT requires substantial training on the part of raters: raters need to be initially certified against master raters' scores and then regularly recalibrated. The rationale for the training is that if FFT is to be an appropriate tool for evaluating teacher effectiveness, ratings of teachers must reflect true differences in teacher effectiveness rather than differences attributable to raters, lessons, teaching assignments, or other factors.

In comparison to CLASS, which has an extensive research base supporting its reliability, fewer studies have been conducted on the reliability of FFT. In fact, the only large-scale assessment of FFT's reliability that we identified was the MET Study. Analyzing lessons in which more than one rater scored the same lesson, researchers examined the degree to which variation in scores was attributable to differences among teachers and how much was attributable to other factors, such as raters, lessons, sections, and times of year. Overall, approximately 37% of the variation in FFT scores was attributable to teacher differences; and the individual components (e.g., questioning, managing student behavior) ranged from 15% to 33%. There was less lesson-to-lesson variation in FFT than in CLASS (10% vs. 27%); yet there was greater unexplained variation in FFT scores than in CLASS scores (43% vs. 34%). As with CLASS, the overall reliability of FFT scores increased as the number of lessons observed increased; and the authors suggested that districts score at least four lessons for a given teacher.

Using CLASS and FFT to assess initial preparation for teaching students with disabilities

Although researchers have examined the validity and reliability of FFT and CLASS, no existing studies have focused specifically on teachers' effectiveness in educating students with disabilities

3 The MET Study defined "underlying value-added scores" as the persistent differences in measured student achievement gains.

(Jones et al., 2013); and only two studies have employed FFT in evaluating teacher preparation routes (Nourgaret et al., 2005; Sindelar et al., 2004). Thus, we lack guidance on how FFT and CLASS can be used to support the preparation of special educators or general educators serving students with disabilities. One primary challenge in using these observation systems in the context of teacher preparation is that neither CLASS nor FFT assesses essential dimensions of instruction that are necessary to meet the specific and heterogeneous needs of students with disabilities. These dimensions are likely to receive considerable emphasis in preparation programs. For an overview of such practices, see the recent practice guides in reading and math published by the What Works Clearinghouse (Gersten et al., 2008, 2009).

For example, explicitness has been a defining feature of many studies of effective teaching for students with learning disabilities (Brownell et al., 2009; Gersten, Baker, Haager, & Graves, 2005; Vaughn, Gersten, & Chard, 2000; Vaughn et al., 2009; Wanzek, Vaughn, Roberts, & Fletcher, 2011). Explicit instruction involves building a rationale for learning a concept, strategy, or skill; modeling how to use the strategy or skill or showing examples; giving clear explanations of concepts and connections between concepts; and practicing with students until they understand a concept and how to apply it or use a strategy or skill with novel tasks. Despite the clear benefits of explicit instruction for students with disabilities, neither FFT nor CLASS assesses these practices. The instructional support domain of the CLASS does not contain criteria for rating direct, explicit, systematic instruction; and FFT is based on a constructivist view of instruction that emphasizes student-centered teaching in lieu of direct, explicit, systematic instruction. Some students with disabilities require considerable explicit teacher support to engage in cognitively complex tasks. Further, these students may need considerable repetition on key basic skills to develop the fluency they need to successfully comprehend and analyze texts and solve mathematical problems. Special education teachers could be disadvantaged if raters do not understand the issues some of these students have interacting with peers and their teachers.

Although scholars have raised substantive concerns about the capacity of these two instruments to evaluate teachers working with students with disabilities, some research has suggested that within the population of special educators, teachers who have gone through formal teacher preparation programs score higher on FFT than those with minimal preparation. For instance, Nourgaret et al. (2005) used a modified version of the FFT to compare special education teachers who completed formal preparation programs with teachers who had less than six hours of preparation. These teachers provided instruction to high school students with high-incidence disabilities. One researcher who had considerable experience in special education conducted all observations. Mean differences between the two groups of special education teachers were significant and large with prepared graduates outperforming those with minimal preparation.

Sindelar et al. (2004) also used the PRAXIS III on which the FFT was based to examine differences between special education teachers prepared through traditional campus-based routes, alternative routes offered by school districts, and routes that involved close collaboration between a school

district and college of education. Analysis of teacher means showed some differences between the groups on certain dimensions of the four domains, but not on others. Effect sizes demonstrating the magnitude of differences were not reported. The collaborative preparation group outperformed the district alternative and campus-based teacher education group on one dimension of Domain 4. The collaborative preparation group and campus-based route outperformed the district alternative on one dimension of Domain 1 and one dimension of Domain 2. The campus-based group outperformed the other two groups on Domain 3. These findings and those from the Nourgaret et al. study (2005) suggested that performance on the various domains measured by the FFT can be influenced by teacher preparation in special education, but exactly how special education preparation affects these various dimensions within domains is unclear. It is also important to note that researchers in both studies were trained in and had experience teaching special education. It remains to be seen whether or not raters with general education experience would be able to rate special education teachers similarly and how variability in special education teachers' performance on the FFT compares to variability in general education teachers' performance. Similarities in score distributions on the FFT are essential to ensuring that it is free of scoring bias.

The CLASS may also hold promise as a formative tool in teacher preparation for students with disabilities, as research shows that general education teachers' performance on the CLASS can be changed as a result of professional development (Allen et al., 2011; Pianta, Mashburn, et al., 2008). Findings from these professional development studies have suggested that the CLASS can be used to assess pre-service teachers' development of more effective teacher-student interactions in course work and field experiences. What we do not know at this point is how applicable the CLASS is for assessing effective teacher-student interactions for teachers instructing students with disabilities and capturing change in those interactions that can be related to preparation. Although the Classroom Organization and Emotional Support domains seem appropriate for general and special education teachers serving students with disabilities, as noted earlier, important aspects of effective teaching practice for students with disabilities are missing from the Instructional Support domain.

Although CLASS and FFT may be appropriate for evaluating teacher preparation for working with students with disabilities, the feasibility of their implementation is a different challenge. For one, teacher educators and school district personnel will need to be trained extensively on either instrument, particularly if high-stakes decisions such as credentialing a teacher or program approval are to be based on FFT and CLASS scores. Common certification procedures for both instruments include participation in intensive training to achieve reliability with a master scorer as well as ongoing rater training for recalibration purposes. To complicate matters, the training process may need to be more extensive for personnel who are not special educators, as they may not understand as well what should define effective instruction for students with disabilities. An additional complication presented in the context of teacher preparation is the case where faculty are asked to rate their student teachers for the purposes of evaluating their own programs. Observation data generated by faculty could be useful for candidate or formative program evaluation; however,

independent evaluations of at least a subsample of program teachers will be necessary for program approval and accreditation purposes.

Emerging practices in classroom assessment for teachers working with students with disabilities

Two classroom practice instruments designed specifically to describe the instructional practice of general and special education teachers working with students with disabilities may hold some potential for evaluating teacher education. The Education Teacher Performance Assessment (edTPA) developed by researchers at the SCALE (n.d.) was designed to assess the instructional practice of teacher education graduates. The intent of the edTPA is to develop a more rigorous initial licensure for teacher education graduates in general and special education to certify their competence for providing all students with rigorous content instruction. The edTPA was developed with the Common Core Standards for Student Learning (National Govern ors Association Center for Best Practices and Council for Chief State School Officers, 2010) in mind. Like other performance assessments, edTPA was designed to provide a broad view of classroom instruction that evaluates teachers' ability to plan, provide instruction, assess student learning, analyze their instruction, and support students' acquisition of academic language. Candidates submit videotapes of three- to five-lesson segments from an instructional unit that show evidence of their teaching ability for one particular group of students. They also submit artifacts from a clinical experience that includes assessments of student learning and commentaries about their planning, instruction, and evaluation of student learning. These sources of information about teaching are scored using 15 analytic rubrics. It is worth noting that general and special education teachers must describe how they addressed the needs of students with disabilities according to goals on each Individualized Education Plan and how they incorporated evidence-based strategies for these students into their instruction.

The edTPA requires teachers to provide evidence of how they are addressing the needs of students with disabilities. Yet, neither has its capacity for discriminating between more effective and less effective teaching for students with disabilities been determined nor has performance on the edTPA been correlated with achievement gains for students with and without disabilities. Further, it is not clear how different visions of effective teaching in general and special education will be reconciled. Major questions must be addressed before the edTPA can be used to evaluate the outcomes of teacher education for students with disabilities. For instance, how will more social constructivist approaches to teaching in general education be reconciled with what are considered more behavioral and information-processing approaches to teaching in special education? Additionally, how will raters be trained to distinguish between effective and ineffective instructional practices for students with disabilities and those without? Answers to these questions and others must be answered through research. Although the edTPA has been field-tested at universities in 21 states, psychometric data on the instruments' reliability and validity has not been published. Further, it is

unclear what resources will be required to implement the edTPA reliably across teacher education programs. Thus, we are uncertain about whether or not the edTPA will be a viable option for evaluating teacher preparation quality for teachers who will provide instruction to students with disabilities.

MyiLOGS is a tool developed to assess language arts and mathematics instruction for students with disabilities (Kurz & Elliott, 2012). The system is designed to assess the opportunities that students with disabilities have to learn content standards; thus, it has some potential as a self-report mechanism for assessing the focus of teachers' instructional practice. MyiLOGS measures students' opportunities to learn by asking teachers to self-report on four aspects of their instruction using an electronic recording system:

- the content they are teaching that is aligned with content standards
- the cognitive level at which they are teaching that content
- the evidence-based practices they are using to convey the content
- the time they spend on specific content

Only one published study has examined the use of MyiLOGs in general and special education. Kurz, Elliott, Wehby, and Smithson (2010) found that general and special education teachers who reported providing more coverage of mathematics content than represented in the state standards and providing instruction at higher levels of cognitive demand were more likely to have students with disabilities who scored higher on tests of mathematics achievement. Additionally, independent observers were able to establish that teachers could accurately report their own instruction—initial evidence that MyiLOGS can be used reliably. Whether or not MyiLOGS holds potential as an evaluation tool for teacher education, however, remains to be seen. The developers of MyiLOGS are only now in the earliest stages of validating its use as an evaluation tool in Arizona, Pennsylvania, and New Jersey.

Information from validation studies of edTPA and MyiLOGS will be helpful in determining if these assessments can be used to evaluate the effectiveness of teacher education programs and individual teacher candidates. It is encouraging to see assessments that consider effective instructional practice for students with disabilities emerge. Such assessment tools may support the types of practices and content that should be taught to teacher candidates preparing to work with students with disabilities. All teachers are responsible for teaching students with disabilities in schools. The edTPA and MyiLOGS—if demonstrated to be valid and reliable evaluations tools—may be more productive than instruments such as the CLASS and FFT and provide information that can add to information collected with more generic observation systems for improving teacher preparation programs.

Perception Surveys

A survey of cooperating teachers, school principals, and teacher education graduates is a commonly used method for evaluating teacher education graduates and programs. In most of these surveys, researchers ask cooperating teachers and school principals to rate beginning teachers on a set of instructional and professional behaviors, or graduates are asked to rate the quality of certain aspects of their preparation. Such surveys, if valid, are attractive for two important reasons:

- Surveys are relatively easy for teacher preparation programs to administer and analyze.
- If well designed, surveys can help teacher educators better identify areas of classroom practice where their candidates excel and areas that could use strengthening.

To date, only five research studies have examined the use of perception surveys as an evaluation tool for general education teachers, and none have looked specifically at teachers' practices for educating students with disabilities. Three studies examined whether or not principals can discriminate between more effective and less effective teachers. Jacob and Lefgren (2008) as well as Harris and Sass (2010) asked principals to complete surveys where they rated teachers on broad dimensions, e.g., overall effectiveness, classroom management, raising student achievement in mathematics. In both studies, researchers established that principals' ratings of teachers were correlated with the teachers' value-added scores in reading and mathematics; and these correlations were stronger than correlations between experience and degrees earned and teachers' value-added scores. Harris and Sass also found that correlations were stronger in the elementary grades and when principals had more experience working with the teacher. Further, principals could effectively predict the future value-added scores of teachers new to their school. Rockoff et al. (2010) also found that principals' ratings of teachers' success in teaching reading and mathematics predicted teachers' value-added scores. These researchers concluded that principals' ratings could be useful in identifying effective teachers, including those who had just begun their teaching careers. These researchers also acknowledged, however, that principals' ratings might not perform similarly (i.e., have the same capability to identify effective teachers) if the ratings were part of a teacher's evaluation.

Rockoff and Speroni (2010) examined whether or not mentor teachers could discriminate between more and less effective beginning teachers based on their evaluations of those teachers on a detailed set of teaching standards. Ratings across these standards were averaged to produce one score that was used in the analysis. Both experienced and less experienced mentors were able to predict teacher effectiveness after teachers had completed their first year. It was clear from the analysis that mentor teachers did not rate the teachers against consistent standards—some raters were harsher and others were more lenient. Thus, the researchers examined the degree to which variability within raters and across raters predicted student achievement and found that both were predictive. These findings suggest that mentor teachers may be able to provide useful information about the quality of a teacher education program's graduates even after accounting for differences in how mentors

rate teachers.

Using perception surveys to evaluate preparation programs for teachers working with students with disabilities

Findings from the small number of studies described in this report suggest that principals' and mentors' ratings of beginning teachers may be useful for determining if certain programs produce more effective teachers than others. However, more research substantiating these findings is necessary if these surveys are to be a viable tool for evaluating teacher education programs and graduates. Additionally, research examining use of these tools in context—where the results have consequences for teachers and programs—is needed. Researchers must identify those dimensions of classroom practice that are most predictive of student achievement gains and other important outcomes. We also need studies that determine if surveys of classroom practice that provide more detailed information about the dimensions of classroom practice, such as those used in the Rockoff and Speroni (2010) study, are more predictive of student outcomes than surveys that provide more global ratings (e.g., rating teachers on overall classroom management skill or instructional skill). Clearly, surveys that provide more information about teachers' practices would likely be more useful to teacher educators who are considering those aspects of their programs that need strengthening.

Most importantly, researchers need to determine whether or not principals and mentor teachers are capable of discriminating between effective and ineffective teachers working with students with disabilities. Qualitative interviews of principals in New Mexico suggested that many principals do not have the skill to rate special education teachers or to determine if general education teachers were providing students with disabilities with effective instruction (Fix, Steinbrecher, Mahal, & Serna, 2013). Mentor teachers, particularly those trained in special education, may be more capable than principals of identifying effective teaching practices for students with disabilities. Thus, researchers must establish the types of educational backgrounds principals and mentors must have in order to accurately evaluate general and special education teachers working with students with disabilities. The positive findings from the studies reported in this paper suggest that this is an avenue of research worth exploring.

Summary and Recommendations

Intervening to strengthen teacher preparation for all teachers, including teachers of students with disabilities, is a compelling approach to improving students' educational outcomes for a variety of reasons: including the following:

- Teachers have been found to be the most important in-school factor related to student achievement (Aaronson et al., 2003; Rivkin et al., 1998).
- Changes to preparation will affect a large number of teachers immediately and an increasing percentage of all teachers over time (Feistritzer, 2011).
- Improving teacher preparation is a proactive solution that will benefit teachers in training before they are teachers of record.
- Despite the potential importance of teacher preparation as a point of intervention in education, current decision making by policy makers and teacher educators is based on study results lacking evidence that the data are predictive of educational outcomes for students with or without disabilities.

A fundamental challenge confronting teacher educators and policy makers is: What is the best available evidence that can be used to make decisions now? This synthesis illustrates that the best available measures—validated observational measures (CLASS and FFT) and VAM scores—can provide useful information for program evaluation. The CLASS and FFT each assesses multiple dimensions of teaching, can be scored reliably, and has been predictive of positive outcomes for students. The strength of correlations between CLASS and FFT and VAM scores was similarly low (r ranging from .11 to .25 across content areas) in the only study examining both measures (Bill & Melinda Gates Foundation, 2012). The MET study also found that both FFT and CLASS require observations of at least four lessons by different raters to obtain reliable results.

The use of VAM scores to evaluate teacher preparation has intuitive appeal because it links teaching and learning. Many students with disabilities participate in standardized state assessments, which means that some special education teachers and many general education teachers will have effectiveness indicators that reflect the learning outcomes of many students receiving special education services. Although there is some inconsistency in findings, there are studies demonstrating that VAM scores can provide a standardized, objective measure of student learning that enables comparisons across preparation programs (Gansle et al., 2012; Goldhaber & Liddle, 2012; Mihaly et al., 2012). VAM for program evaluation mitigates some of the challenges that emerge in assessing individual teachers by using aggregates of teachers across schools, districts, and years; see Baker et al. (2010) and McCaffrey et al. (2004). However, VAM raises new concerns, such as the impact of program selectivity.

Although currently available measures of classroom practice and student learning outcomes can

substantially improve the assessment of teacher preparation, some limitations require appropriate cautions in their use. These limitations relate both to the needs of special educators and children with disabilities as well as the informational needs of policy makers concerned with making decisions for all students. VAM is based on standardized tests that will not assess some important educational outcomes (Baker et al., 2010). This shortcoming is most pronounced for students with disabilities for whom the most critical academic goals may be outside the range assessed by state tests and for whom nonacademic outcomes may be critical. An additional challenge for the use of VAM to assess teacher preparation for students with disabilities is the large number of these students who do not participate in standardized testing. This limits the number of students included in teachers' VAM scores and, for many teachers, makes calculating VAM scores impossible. Even though many students with high-incidence disabilities participate in standardized testing, estimating the impact of their teachers on student learning is more challenging due to the inconsistent use of test accommodations and the decreased precision of extreme scores (Jones et al., 2013). Focus on program, rather than educator evaluation, provides the advantage that the results for students with disabilities across multiple teachers can be aggregated.

The observational assessments of classroom practice also present substantive limitations in addressing the instructional needs of students with disabilities. The most obvious limitation is the absence of studies examining the validity of using CLASS and FFT with teachers of students with disabilities. This is a particularly serious concern for students with significant disabilities whose learning contexts diverge substantially from the types of classes in which the extant research has been conducted. Additionally, the absence of direct assessment of direct, explicit instruction in both CLASS and FFT raises content validity concerns regarding the use of these instruments to assess special education teachers or general education teachers working with students with disabilities (Brownell et al., in press; Vaughn et al., 2009; Wanzek et al., 2011). Practice measures, as potential proxies for predictors of student outcomes, have an additional limitation: these measures do not directly assess positive outcomes for students.

In addition to the challenges specific to the needs of students with disabilities and their teachers, both sets of measures present specific challenges for policy makers. The most critical challenge in deploying observational measures to assess teacher preparation may be capacity and cost. As described above, there are consequential human capital costs to training and implementing observational systems like CLASS and FFT. To be used systematically to evaluate teacher preparation, these measures would need to be available in all schools in a state or at least a strong representative sample. This is a consequential decision for a state to undertake at political, policy, and fiscal levels. Although implementation of CLASS or FFT is more practical for implementation within a preparation program, consequential start-up and training costs remain.

At a policy level, critical challenges for using VAM are the realities that it will not include the majority of teachers and it is not currently available in many states. The extent to which it is possible to obtain sufficient data for programs to examine program effects for either special educators or

students with disabilities is an open question. Finally, although value-added measures may serve as a useful global outcome indicator, VAM will not provide the sort of fine-grained detail that teacher educators would need to revise and improve programs. VAM may signal the need to revise, but it will not signal how the program needs to be revised.

From a policy and programmatic perspective, the critical decision-making challenge will inevitably involve how to obtain the most meaningful outcome data and how to use those data in an appropriately cautious manner. The most obvious caution is to limit the use of measures to decisions for which the measure is relevant and available. To state the obvious, value-added measures cannot contribute to decisions about program accreditation or design for a special education preparation program for students with severe disabilities. Similarly, it is unclear at present how valid FFT or CLASS can be for assessing a special education program. Even in domains where relevant practice or student learning data are available, appropriate checks and balances in the decision-making process are needed to ensure that evidence is given due weight and that reasonable standards are set for the repeatability of results before consequential decisions are made.

A critical challenge going forward will be to develop a data infrastructure that can best support strengthening teacher preparation for all students. It is apparent that the most valuable advances in this domain are likely to occur at the state level because no other entity has the appropriate jurisdictional reach. Coordinated statewide action will be necessary to obtain consistent, comparable, and meaningful data across schools, districts, and preparation programs. In states that choose to work to improve teacher preparation, it appears that the two greatest needs are measures of student growth and of educator practice. In the student growth domain, measures of student progress including as many students as practical and systematically including students with diverse disabilities by design are needed. Required are (a) measures of practice (i.e., observations or supervisor ratings) that meet the rigorous technical standards raised in this paper and (b) measures which are assured to capture critical practices that enable teachers to meet the needs of diverse students. These challenges are certainly daunting, both technically and at a policy level; however, the quality of the decisions that leaders make about educator preparation will be bounded in a very real way by the quality of the information they have to inform those decisions. If we do not improve the information systems guiding decision making, progress on outcomes may be more a matter of chance than of skill or will.

References

Aaronson, D., Barrow, L., & Sander, W. (2003). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*, 95-136.

Allen, J. P., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the Classroom Assessment Scoring System-Secondary. *School Psychology Review, 42*, 76-98.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*, 1034-1037. doi: 10.1126/science.1207998

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn,Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn,Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper No. 278). . Retrieved from The Economic Policy Institute website: http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17,* 62-87.

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains*. Seattle, WA: Author.

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Seattle, WA: Author.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wycoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31,* 416-440.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added* (Research Report No. 11-18). Retrieved from Educational Testing Service website: www.ets.org/

research/pic

Braun, H. (2013). *Magical thinking and the use of value-added models for educator accountability.* Manuscript in preparation.

Brownell, M. T., Bishop, A. G., Gersten, R., Klingner, J. K., Penfield, R., Dimino, J., . . . Sindelar, P. T. (2009). The role of domain expertise in beginning special education teacher quality. *Exceptional Children, 75,* 391-421.

Brownell, M. T., Steinbrecher, T., Kimerling, J., Park, Y., Bae, J., & Benedict, A. (in press). Dimension of teacher quality in general and special education. In P. T. Sindelar, E. D. McCray, M. T. Brownell, & B. Lignugaris-Kraft (Eds.), *Handbook of research on special education teacher preparation*. New York, NY: Routledge.

Buckley, K., & Marion, S. (2011). *A survey of approaches used to evaluate educators in non-tested grades and subjects*. Retrieved from National Center for the Improvement of Educational Assessment website: http://www.relcentral.org/research_alliances/a-survey-of-approaches-used-to-evaluate-educators-in-non-tested-grades-and-subjects/

Buzick, H. M., & Jones, N. D. (2013). *Using test scores from students with disabilities in teacher effectiveness indicators*. Manuscript submitted for publication.

Buzick, H., & Laitusis, C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 9*, 537-544.

Buzick, H. M., & Stone, E. A. (2013*). A meta-analysis of research on the read aloud accommodation for K-12 students with disabilities.* Manuscript submitted for publication.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood.* (Working Paper No. 17699). Washington, DC: National Bureau of Economic Research.

Clifford, R., Barbarin, O., Chang, F., Early, D., Bryant, D., Howes, C., . . . Pianta, R. (2005). What is pre-kindergarten? Characteristics of public pre-kindergarten programs. *Applied Developmental Science*, *9*, 126-143. doi:10.1207/s1532480xads0903_1

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement*? Retrieved from National Center for Analysis of Longitudinal Data in Education Research website: http://www.caldercenter.org/PDF/ 1001058_Teacher_Credentials.pdf

Council of Chief State School Officers. (2011). *Interstate Teacher Assessment and Support*

Consortium (InTASC) model core teaching standards: A resource for state dialogue. Washington, DC: Author. Retrieved from http://www.ccsso.org/documents/ 2011/intasc_model_core_teaching_standards_2011.pdf

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher-child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology, 101*, 912-925.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Dominguez, X., Vitiello, V., Fuccillo, J., Greenfield, D., & Bulotsky-Shearer, R. (2011). The role of context in preschool learning: A multilevel examination of the contribution of context-specific problem behaviors and classroom process quality to low-income children's approaches to learning. *Journal of School Psychology, 49*, 175-195. doi: 10.1016/j.jsp.2010.11.002

Downer, J. T., López, M. L., Grimm, K. J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher–child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly*, *27*, 21-32.

Dwyer, C. A. (1994). Criteria for performance-based teacher assessments: Validity, standards, and issues. *Journal of Personnel Evaluation in Education*, *8*, 135-150.

Feistritzer, C. E. (2011). *Profile of teachers in the U. S. 2011.* Washington, DC: National Center for Education Information.

Feng, L., & Sass, T. R. (2010). *What makes special-education teachers special? Teacher training and achievement of students with disabilities* (CALDER Working Paper No. 49). Washington, DC: The Urban Institute.

Fix, R., Steinbrecher, T., Mahal, S., & Serna, L. (2013). *Administrator knowledge: How it relates to hiring and retaining special education teachers.* Paper presented at the meeting of the Council for Exceptional Children, San Antonio, TX.

Floden, R. E. (2012). Teacher value added as a measure of program quality: Interpret with caution. *Journal of Teacher Education, 63*, 356-360.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*,

79, 79-107. doi:10.1207/s15327930pje7904_5

Gansle, K. A., Noell, G. H., & Burns, J. M. (2012). Do student achievement outcomes differ across teacher preparation programs? An analysis of teacher education in Louisiana. *Journal of Teacher Education*, *63*, 304-317.

Gersten, R., Baker, S., Haager, D., & Graves, A. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners. *Remedial and Special Education, 26*, 197-206.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools* (NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/ practiceguides/

Gersten, R., Compton, D., Connor, C., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. (2008). *Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades: A practice guide.* (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U. S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Goldhaber, D., & Liddle, S. (2012). *The gateway to the profession: Assessing teacher preparation programs based on student achievement* (Working Paper 65). Retrieved from National Center for Analysis of Longitudinal Data in Education Research website: http://www.caldercenter.org/upload/Goldhaber-et-al.pdf

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (Working Paper No. 45). Retrieved from National Center for the Analysis of Longitudinal Data in Education Research, Urban Institute website: http://www.caldercenter.org/upload/CALDERWorkPaper_45.pdf

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, *76*, 949-967. doi:10.1111/j.1467-8624.2005.00889.x

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U. S. early childhood and elementary classrooms*. Retrieved from Foundation for Child Development website: http://fcd-us.org/resources/building-science-classrooms-application-class-framework-over-

4000-us-early-childhood-and-e

Harris, D. N., & Sass, T. R. (2010). *What makes for a good teacher and who can tell?* (CALDER Working Paper No. 30). Retrieved from The Urban Institute website: http://www.urban.org/url.cfm?id=1001431&RSSFeed=UI_Education.xml

Hill, H. C., Ball, D. L., Goffney, I. M., & Rowan, B. (2008). Validating the ecological assumption: The relationship of measure scores to classroom teaching and student learning. *Measurement: Interdisciplinary Research and Perspectives, 5*(2-3), 107-117.

Hock, H., & Isenberg, E. (2012). *Methods for accounting for co-teaching in value-added models.* Washington, DC: (Report No. 7482). Mathematica Policy Research.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*, 207-219.

Howes, C., Burchinal, M., Pianta, R., Brant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*, 27-50. doi:10.1016/j.ecresq.2007.05.002

Jacob, B. A., & Lefgren, L. (2008). Can principles identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*, 101-136. doi:10.1086/522974

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

Kimball, S., White, B., Milanowski, A., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*, 54-78. doi:10.1207/s15327930pje7904_4

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012). *Teacher preparation programs and teacher quality: Are there real differences across programs?* (Working Paper No. 1204). Department of Economics, University of Missouri.

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *The Journal of Special Education*, *44*, 131-145.

Kurz, A., & Elliott, S. N. (2012). My instructional Learning Opportunity Guidance System (MyiLOGS) Version 2.0. Tempe, AZ: Arizona State University.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*,159–174.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. The Elementary School Journal, 104, 409-426. doi:10.1086/499760

Mashburn, A. J., Pianta, R., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. Child Development, 79, 732-749.

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, 67-101.

Mihaly, K., McCaffrey, D., Sass, T., & Lockwood, J. R. (2012). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. (Research paper series 12-12). Andrew Young School of Policy Studies.

Mikami, A., Gregory, A., Allen, J., Pianta, R., & Lun, J. (2011). Effects of a teacher professional development intervention on peer relationships in secondary classrooms. School Psychology Review, 40, 367-385.

Milanowski, A. (2004). Relationships among dimension scores of standards-based teacher evaluation systems, and the stability of evaluation score-student achievement relationships over time. Madison, WI: Wisconsin Center for Education Research, Consortium for Policy Research in Education, University of Wisconsin-Madison.

National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network (2002). The relation of global first-grade environment to structural classroom features and teacher and student behaviors. The Elementary School Journal, 102, 367-387.

National Institute of Child Health and Human Development Early Child Care Research Network. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. The Elementary School Journal, 105, 305-323. doi:10.1086/428746

National Governors Association Center for Best Practices and Council for Chief State School Officers. (2010). …Need more information…

Noell, G. H., Porter, B. A., Patt, R. M., & Dahir, A. (2008). Value added assessment of teacher preparation in Louisiana: 2004-2007. Louisiana Board of Regents. Retrieved from http://www.regents.state.la.us/Academic/TE/Value%20Added.htm

Nourgaret, A., Scruggs, T., & Mastropieri, M. (2005). Does teacher education produce better special education teachers? Exceptional Children, 71, 217-229.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. American Educational Research Journal, 45, 364-397. doi:10.3102/000288,31207308230

Pianta, R. C., & Hamre, B. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. Educational Researcher, 3, 109-119.

Pianta, R. C., Hamre, B., Hayes, N., Mintz, S., & La Paro, K. M. (2008). Classroom Assessment Scoring System - Secondary (CLASS-S). University of Virginia.

Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? Applied Developmental Science, 9, 144-159. doi:10.1207/s1532480xads0903_2

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom Assessment Scoring System: Manual K-3. Baltimore, MD: Paul H. Brookes.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten environment to teacher, family, and school characteristics and child outcomes. The Elementary School Journal, 102, 225-238. doi:10.1086/499701

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. Early Childhood Research Quarterly, 23, 431-451. doi; 10.1016/j.ecresq.2008.02.001

Plecki, M. L., Elfers, A. M., & Nakamura, Y. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. Journal of Teacher Education, 63, 318-334.

Ponitz, C. C., Rimm-Kaufman, S. E., Brock, L., & Nathanson, L. (2009). Early adjustment, gender differences, and classroom organizational climate in first grade. The Elementary School Journal, 110, 142-162.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional

climate, student engagement, and academic achievement. Journal of Educational Psychology, 104, 700-712. doi:10.1037/a0027268

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (1998). Teachers, schools, and academic achievement. Econometrica, 73, 417-458.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. Labour Economics, 18, 687-696. doi:10.1016/j.labeco.2011.02.004

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2010). Information and employee evaluation: Evidence from a randomized intervention in public schools (Working Paper No. 16240). Cambridge, MA: National Bureau of Economic Research.

Sindelar, P., Daunic, A., & Rennells, M. S. (2004). Comparisons of traditionally and alternatively trained teachers. Exceptional Children, 12, 209-223.

Stanford Center for Assessment, Learning, and Equity [SCALE] (n.d.). The Education Teacher Performance Assessment (edTPA). Retrieved at https://scale.stanford.edu/teaching/edtpa

Swanson, H. L. (2001). Searching for the best model for instructing students with learning disabilities. Focus on Exceptional Children, 34(2), 1-16.

Vaughn, S., Gersten, R., & Chard, D. (2000). The underlying message in LD intervention research: Findings from research syntheses. Exceptional Children, 67, 99-114.

Vaughn, S., Wanzek, J., Murray, C. S., Scammacca, N., Linan-Thompson, S., & Woodruff, A. L. (2009). Response to early reading intervention: Examining higher and lower responders. Exceptional Children, 75, 165-183.

Wanzek, J., Vaughn, S., Roberts, G., & Fletcher, J. M. (2011). Efficacy of a reading intervention for middle school students with disabilities. Exceptional Children, 79, 73-87.

Wineberg, T. W. (2006). Enacting an ethic of pedagogical vocation: Pursuing moral formation in responding to the call of sacrifice, membership, craft, memory, & imagination (Unpublished doctoral dissertation). Simon Fraser University, Burnaby, Canada.