Including Students With Disabilities and English Learners in Measures of Educator Effectiveness

Nathan D. Jones[1], Heather M. Buzick[1], and Sultan Turkan[1]

The purpose of this essay is to provide an overview of the challenges of accounting for students with disabilities (SWDs) and English learners (ELs) in the evaluation of mainstream teachers. We focus on the two prominent indicators of teaching quality—classroom observations and value-added scores. We begin by describing each indicator and outlining the specific challenges related to the inclusion of SWDs and ELs in mainstream teacher evaluation. We then suggest recommendations for states and districts to ensure that teacher evaluation systems adequately and fairly account for these students. Finally, we provide researchers with a set of recommendations for improving the evidence base surrounding the validity of teacher evaluation measures with regard to SWDs and ELs.

Keywords: assessment; measurements; school/teacher effectiveness; special education

Within the U.S. K–12 education system, there is consensus among policy makers, administrators, and educators that students' opportunities for learning depend on the quality of teaching they receive in schools. Existing educator evaluation systems have been criticized for a lack of fidelity and rigor (e.g., Weisberg, Sexton, Mulhern, & Keeling, 2009). Consequently, the federal Race to the Top program has emphasized the need for multiple measures in educator evaluation systems, with a specific emphasis on student growth (U.S. Department of Education, 2010). Emerging teacher evaluation efforts have focused mainly on two indicators of teaching quality—observations and student test scores. States and districts are including local indicators as well to create weighted evaluation systems that avoid overreliance on any one measure, particularly student achievement. Examples of weighted indicators in state/district evaluation systems are displayed in Table 1.

Despite advances in research on teacher evaluation (for summaries, see Harris, 2011; Bell et al., 2012), there has been virtually no attention given to whether teachers are effectively educating exceptional populations—namely students with disabilities (SWDs) and English learners (ELs). For all the criticism of No Child Left Behind, one of its achievements was requiring that districts attend to the achievement of subgroups of students, including SWDs and ELs. In contrast, although Race to the Top was designed to ensure that students have highly effective teachers, there is no explicit mention of teachers' effectiveness at differentiating their instruction. In this essay, we argue that if teacher evaluation systems fail to acknowledge the presence of SWDs and ELs in teachers' classrooms, it is problematic in terms of validity (as teachers' evaluations may represent an incomplete and potentially inaccurate picture of teachers' instruction) and equity (by providing disincentives for attending to these students' needs).

We discuss these two student populations together for several reasons. First, both represent critical subgroups in U.S. schools. Among K–12 students, approximately 12% receive special education services and 11% are ELs (U.S. Department of Education, 2007, 2008). Most SWDs and ELs educated in mainstream classrooms—the majority of SWDs spend 80% or more of their time in regular classrooms (Table 2 shows the distribution by disability subtype), and, although there are no national statistics on the time ELs spend in mainstream classrooms, trends suggest that more ELs are being included in regular classroom instruction for longer periods of time (e.g., Zehler et al., 2003). Second, SWDs and ELs in teachers' classrooms can contribute meaningfully to teachers' practices because they often require teachers to modify or supplement their instruction. Third, many have questioned whether the performance of SWDs and ELs on standardized assessments supports valid inferences about their learning and academic growth. At the same time, we acknowledge that there is wide variation within and between these two populations in terms of classroom context and accessibility needs and that a small percentage of students both have disabilities and are ELs. We therefore outline recommendations for how evaluation

[1]Educational Testing Service, Princeton, NJ

Educational Researcher, Vol. 42 No. 4, pp. 234–241
DOI: 10.3102/0013189X12468211

Table 1
Examples of State/District Evaluation Systems (as of August 2012)

State/district	Evaluation system	Source
Louisiana	Professional practice, including observation score (50%); value-added (50%).	http://www.louisianaschools.net/lde/uploads/20118.pdf
Ohio	Teacher performance on standards, including observations (50%); student growth measures, including value-added (50%)	http://education.ohio.gov/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=128955
Tennessee	Observation score (50%), student outcomes (50%—value-added, 35%; student achievement, 15%)	http://team-tn.org/assets/educator-resources/Calculating_the_Effectiveness_Rating.pdf
Washington, D.C., Impact Model	Value-added (35%), observations (40%), student achievement (15%), school commitment (10%)	http://www.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/IMPACT-2012-Grp1-Aug13-web.pdf

Note. These examples focus on components of the evaluation system for teachers in tested grades and subjects. Each evaluation system has other components for evaluating educators who do not have sufficient data from students to compute value-added scores.

Table 2
Percentage of Time Spent by Students With Disabilities in the General Education Classroom

	Less than 40%	40%–79%	More than 79%
Speech or language impairments	5	6	86
Developmental delay	16	21	62
Visual impairments	12	14	62
Specific learning disabilities	9	28	61
Other health impairments	11	25	60
Hearing impairments	16	17	53
Orthopedic impairments	25	17	51
Traumatic brain injury	23	23	45
Emotional disturbance	23	19	39
Autism	36	18	36
Deaf-blindness	29	17	30
Intellectual disability	49	27	16
Multiple disabilities	46	17	13
All students with disabilities	15	22	58

Note. Percentages include students served under IDEA who attend regular schools. Students in homebound/hospital placement, a correctional facility, or a separate public facility, or those who attend private school but receive publically funded special education are not included in the table.
Source: U.S. Department of Education, National Center for Education Statistics (2011). *Digest of Education Statistics, 2010* (NCES 2011-015), Chapter 2.

systems can better attend to the specific disability categories of teachers' students, as well as different levels of students' English proficiency, and, where appropriate, we differentiate challenges and recommendations that may be unique to either subgroup.

The remainder of this essay is organized around two common indicators of teaching quality—student achievement, namely value-added scores, and classroom observations.[1] For each indicator, we first briefly describe the measure and outline the challenges in accounting for SWDs and ELs in mainstream teacher evaluation. We then suggest recommendations for states and districts to ensure that teacher evaluation systems adequately and fairly account for the inclusion of SWDs and ELs in mainstream classrooms. Finally, we provide researchers with a set of recommendations for improving the evidence base surrounding the validity of teacher evaluation measures with regard to SWDs and ELs. These suggestions are also summarized in Table 3.

The relative importance of the challenges we describe below depends largely on local context and may also depend on the number of teachers with SWDs and/or ELs in a state/district or, for individual teachers, the proportion of students from special populations in a given classroom. Our goal in highlighting these challenges is to encourage states/districts/researchers to explore these challenges and consider short- and long-term solutions for those that are found to be the most threatening to the credibility of their evaluation systems. We advocate for a system that holds teachers accountable for educating all students and contributes to professional development while minimizing unintended consequences, including, for example, not differentiating instruction and discouraging teachers from entering the profession.

Value-Added Scores

Challenges of Accounting for SWDs and ELs in Teacher Value-Added Models

Value-added scores are derived from statistical models that attempt to explain the contribution of individual teachers to

Table 3
Summary of Suggestions for Researchers and Practitioners in Evaluating Teachers Who Educate Students From Special Populations

Student Academic Progress	
Researchers	**Practitioners**
• Include students with disabilities (SWDs) and English learners (ELs) in all research on model building and validation • Consider the heterogeneity within each of the subgroups of SWDs and ELs • Conduct sensitivity studies with SWD and EL specific variables (i.e., accommodation use, entry/exit from special education, EL classification, English proficiency scores) • Test assumptions regarding the presence of SWDs and ELs in general educators' value-added scores	• Support accessible assessments that offer precise measurement along the entire score scale (e.g., adaptive assessment, universal design) • Create a standardized system to accurately assign, monitor, and record the use of testing accommodations • Work with teachers to understand the quality of their individual value-added score given their particular classroom context • Adopt a roster validation system to monitor instruction for students are shared across mainstream and special education or English as a second language (ESL) teachers.
Classroom Observations	
Researchers	**Practitioners**
• Provide guidance on how to modify rubrics to include items or response categories specific to both SWDs and ELs • Conduct research on validity and reliability for modified rubrics or SWD- and EL-specific observation protocols • Evaluate observer performance specific to SWDs and ELs	• Consider adopting an observation protocol designed specifically for use with both SWDs and ELs • Consider supplementing an existing observation protocol with a subset of items or response categories specific to teaching both SWDs and ELs • Adopt a scoring support system to assist raters in using existing protocols to assess teachers' instruction to SWDs and ELs • Ensure observers are knowledgeable about and/or trained in the instructional needs of both SWDs and ELs

student achievement. To estimate a teacher-specific effect on achievement, value-added models take into account student prior achievement on standardized tests and may also control for other student and school characteristics.[2] There are potential benefits to using value-added scores, particularly relative to other indicators based on student outcomes. Value-added scores (a) can provide a standardized, common metric for estimating teacher effects, (b) are intended to support causal inferences about a teacher's impact on student growth, (c) are based on large-scale, standardized assessments that may have more desirable psychometric properties than other student assessments, and (d) are able to be evaluated for validity. Nonetheless, their appropriateness for making high-stakes decisions about teachers has been controversial because of a number of strong limitations. One concern is that the year-to-year correlation of value-added scores appears to be "small" to "moderate," though there is some evidence that some component of performance persists within teachers over time (Goldhaber & Hansen, 2010; McCaffrey, Sass, Lockwood, & Mihaly, 2009). Concerns have been raised that value-added scores may lead to the misclassification of teachers, or that they do not adequately account for the selection of teachers and students into schools. However, evidence varies regarding the scope of this bias in value-added scores (e.g., Briggs & Dominique, 2011; Chetty, Friedman, & Rockoff, 2011; Rothstein, 2010). There are also a number of logistic challenges that must be addressed, including how to develop a fair evaluation system when only a small percentage of teachers can have individual value-added scores estimated.[3] We refer the reader to Baker et al. (2010) and Glazerman et al. (2010) for more details and additional technical concerns related to value-added models in general.

We suggest that in addition to general concerns, SWDs and ELs present unique challenges that can affect the quality of value-added scores. Broadly, these challenges include (a) factors that threaten the validity of inferences about academic achievement over time for SWDs and ELs and (b) the complex instructional contexts in which SWDs and ELs are taught. In any given classroom, as the number of SWDs or ELs increases, these challenges will have greater impact on value-added scores. Although value-added models can statistically control for time-invariant student characteristics, factors outside of the control of the teacher that change over time can present challenges for estimating teacher effects.

One threat to the validity of inferences about academic progress for both SWDs and ELs is inconsistent use of testing accommodations. There is evidence that some accommodations used by SWDs and ELs are associated with changes in performance, though the direction of this impact varies (for ELs: Pennock-Roman & Rivera, 2011; for SWDs: Sireci, Scarpeti, & Li, 2005). Although accommodations are intended to remove barriers to students' ability to access test items, they can be inappropriately assigned or ineffective (e.g., ELs: Abedi, Hofstetter, & Lord, 2004; SWDs: Ketterlin-Geller, Alonzo, Braun-Monegan, & Tindal, 2007), which can increase measurement error and misrepresent students' true academic growth. Despite the increased emphasis on including ELs and SWDs in large-scale assessments and providing them with testing accommodations, states have not systematically completed the work of assigning appropriate

accommodations based on ELs' history of formal schooling in the United States and in their home country, as well as their literacy skills in English and in their native language (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007), nor have they resolved debates about controversial accommodations for SWDs that may interact with the measured construct (e.g., read aloud). The uncertainty regarding the allocation of appropriate accommodations, limited resources, or students' changing needs can result in accommodations being delivered inconsistently across years, which can inflate or deflate measures of student growth, potentially affecting value-added scores.

A second measurement challenge is that a large proportion of SWDs and ELs exhibit low performance on state assessments (ELs: Perie, Grigg, & Donahue, 2005; SWDs: Thurlow, Bremer, & Albus, 2011). This puts into question the quality of value-added scores for teachers with large numbers of SWDs or ELs because extreme scores have higher measurement error and are less reliable, properties that are compounded when using multiple test scores. In addition, student learning gains may not be realized because of floor effects. Although these issues are not confined to SWDs and ELs, the systemic and predictable nature of these low-scoring subgroups may engender the perception of unfairness.

In addition to the aforementioned measurement challenges, additional complications for modeling value-added scores include heterogeneity within the subgroups of ELs and SWDs, as well as the context in which SWDs and ELs are taught. Students in both subgroups vary in terms of student characteristics, opportunity to learn, accessibility, and special services. For example, for some students with disabilities, such as those with learning disabilities or emotional and behavioral disorders, movement into or out of special education can change the services students receive and the amount of time spent in the regular classroom, making it more difficult to isolate teacher effects. For ELs, late-arrival students with low English proficiency pose different challenges from those ELs who are designated to be Fluent English Proficient but are underachieving on state content assessments. For the latter group, the proportion of time spent learning content and the English language changes as students progress; this may lower students' test performance temporarily, independent of a teacher's efforts. There may also be peer effects associated with nondisabled, English-proficient students who share the classroom with SWDs and/or ELs that may differ depending on disability subtype or level of English proficiency. That is, the performance of all students in a classroom may be affected—positively or negatively—by the presence of a co-teacher, extra funding support for special services, peer behaviors, or other factors not directly related to an individual teacher.[4] Simply including a discrete variable for SWD or EL status may not account for such sources of heterogeneity in a value-added model.

A final challenge for incorporating SWDs and ELs into value-added models is attributing students' growth to individual teachers. For mainstream teachers, it is not uncommon that they share responsibility for instruction with special education teachers, particularly for students with high incidence disabilities. Similarly, for ELs, there is often shared responsibility between mainstream teachers and English as second language (ESL) teachers. For SWDs and ELs, academic growth in a given year is likely dependent on the quality (and content) of instruction received in both settings, as well as the degree to which this instruction aligns, factors that cannot be directly estimated by value-added models.

Value-Added Scores: Suggestions for Practitioners and Researchers

There are various strategies for practitioners to attribute SWD and EL performance to individual teachers when responsibility for individual students is shared across mainstream and special education or ESL teachers. We propose that districts adopt a roster validation system, which can increase both the face validity of value-added scores as well as the accuracy of estimates (Hock & Isenberg, 2011). The Houston Independent School District, for example, uses a system where teachers can regularly log in and verify the accuracy of their rosters. As part of the roster system, Hock and Isenberg (2011) recommend that both the general education and special education teachers receive 100% responsibility of their shared students. Although roster validation will not address the question of whether general education teachers have adequately differentiated their instruction for SWDs or ELs, it will decrease the likelihood that these students are viewed as the sole responsibility of special education or EL teachers.

To improve the quality of value-added scores relative to SWDs and ELs, we suggest that state personnel support accessible assessments that offer precise measurement along the score scale, such as a multistage adaptive assessment,[5] and develop systems to accurately assign, record, and monitor the use of testing accommodations.[6] Also, we recommend that administrators work with teachers to understand the validity of their value-added score given their particular classroom context, in order to minimize unintended consequences and perceived unfairness. As we suggest below, more research is necessary to better understand how decisions about including or excluding SWDs and ELs affect the validity of teachers' value-added scores. In the short term, we recommend that states and districts balance the unintended consequences of not including test scores from SWDs and ELs (e.g., teachers focusing their reading instruction only on students who will be included in their value-added estimates) against the various threats to validity when scores are included.

For researchers, there is a need to conduct more nuanced validity investigations for value-added modeling, testing various assumptions regarding the presence of scores from SWDs and ELs in mainstream teachers' value-added scores and exploring variables that account for heterogeneity within these subgroups. Rather than including a single variable for percentage of SWDs and/or ELs in a class, we encourage research on value-added models with respect to inconsistent accommodation use across years, disability subtype, and changes in English proficiency or classifications for ELs (e.g., limited English proficiency, reclassified English proficient). Examples of analyses include computing the correlation of teacher rankings based on several value-added models and estimating whether these finer-grained categories change the inferences made about teachers.

We believe that if research finds that individual teachers' value-added scores are robust to these theoretically important

variables, SWDs and ELs should be included in the models in practice, which would promote the perception that educators in mainstream classrooms are being held accountable for the quality of instruction for SWDs and ELs. Alternatively, if the proposed research finds that additional information on SWDs and ELs improves the validity of value-added scores, we encourage states/district to use variables specific to special populations in their value-added models, with attention given to the interpretation and practical consequences of their use.

Observations of Classroom Instruction

Challenges of Accounting for SWDs and ELs in Observations of Teachers' Instruction

Although there has been considerable research and commentary on the quality of value-added models, there has been comparatively little scrutiny given to observation protocols in the context of teacher evaluation, despite the fact that most new models of teacher evaluation require their use. There is also little research on whether observation systems adequately measure teachers' effectiveness with regard to SWDs and ELs. In the following, we describe approaches for observing classroom instruction, discuss considerations related to SWDs and ELs, and provide recommendations for practitioners and for future research.

Teacher observation systems are commonly based on a set of theoretical dimensions that are intended to define critical aspects of teaching (e.g., classroom climate, quality of feedback to students); all operate with a working definition of what constitutes "good teaching." Observation protocols typically fall into one of two categories—either they are intended to be used across all settings (e.g., Framework for Teaching [FFT]; Danielson, 2007) or they are developed for use in specific subject areas (e.g., the Mathematical Quality of Instruction assessment; Hill et al., 2008).

With states and districts feeling pressure to simultaneously develop and implement multiple teacher evaluation measures, most districts and states have decided to limit administrative burden by adopting a single observation protocol, with the use of FFT being the most common.[7] This approach eliminates concerns about differences in validity and reliability across various subject-specific and student population–specific protocols; however, it presents challenges for holding teachers accountable for the performance of SWDs and ELs in their classroom. That is, for observation protocols to support valid inferences about teacher effectiveness, they must adequately address whether teachers are using practices that are identified as effective for SWDs and ELs. The observation protocols being considered for teacher evaluation (e.g., FFT) typically do not outline expectations for the instruction provided to SWDs and ELs; rather, the interactions between teachers and their students are described in general terms. Further, as we outline below, there is evidence that the instructional needs of SWDs and ELs vary from their peers in important ways. If instructional practices deemed effective for SWDs or ELs are not represented in the observation systems used by states/districts, it may provide disincentives for teachers to adopt such practices in their teaching.

For SWDs, much of the literature on effective instructional practice has been developed around the inclusion of students with high-incidence disabilities (such as learning disabilities) in general education settings, particularly in the delivery of reading instruction. For example, Vaughn and Linan-Thompson (2003) suggest that the most promising instructional approaches for students with learning disabilities are those that are "characterized as being well specified, explicit, carefully designed, and closely related to the area of instructional need (e.g., reading, spelling, math)" (p. 142). In addition to an emphasis on instruction that is direct and explicit, Brownell and colleagues have emphasized the need for general educators to demonstrate knowledge related to basic reading skills, peer learning, and self-management techniques (Brownell et al., 2009).

Likewise, there are specific instructional practices that benefit ELs—and which may not be accounted for in general observation protocols. This might be due to the multitude of approaches to defining effective instructional practice for ELs. Some argue that ELs benefit from the basic instructional practices that have been proven to be effective for native English speakers but they need modifications to this instruction (Gersten & Baker, 2000). In line with this thinking, ELs often benefit from a sustained instructional emphasis on vocabulary development, including, for example, the multiple meanings of words in English (e.g., August, Carlo, Dressler, & Snow, 2005). It has also been argued that effective teaching of ELs goes beyond vocabulary instruction (Schleppegrell, 2004).

In addition to providing instruction focused on vocabulary and reading skills, successful teachers of EL students are attentive to students' home language and culture in their instruction (e.g., Paneque & Barbetta, 2006). Others also argued that effective EL teaching requires specific knowledge, in that teachers of ELs should be able to engage ELs in using the academic language of a particular discipline (Brisk & Zisselsberger, 2011). It is likely that successful teachers of ELs are able to integrate both language and content objectives, as is outlined in the sheltered academic instructional practices proposed by Echevarria, Vogt, and Short (2008).[8] In summary, if the goal of observation systems is to assess the quality of instruction provided to all students, then a major challenge moving forward will be how to account for the kinds of instructional practices described above that appear uniquely beneficial for SWDs and ELs.

One final issue relates to the reliability of observers' scores. If observation protocols are to support valid interpretations about teachers' instructional quality for ELs and SWDs, researchers must attend to whether observers themselves can reliably differentiate between teachers who do and do not make use of effective instructional practices for these populations. A handful of recent large-scale research studies suggest it can be challenging to teach observers to score in reliable ways (Bell et al., 2012). Researchers have identified multiple sources of variation in observation scores for teachers, such as differences across observers or across lessons or class periods (Bill & Melinda Gates Foundation, 2012; Hill, Charalambous, & Kraft, 2012). We argue that in addition to these sources of error, an observer's training and background specific to ELs or SWDs may lead to additional variation, and there is evidence that principals and other school administrators often lack the expertise necessary to evaluate teachers' instruction of SWDs (Blanton et al., 2006); the same appears true for ELs.

Observations: Suggestions for Practitioners and Researchers

Observation systems should ideally reflect teachers' responsiveness to their students' needs, as effective teachers would make different instructional decisions (e.g., grouping practices, emphasis on beginning reading skills) depending on the characteristics of the SWDs and ELs in their classroom. One option for districts would be to adopt observation protocols designed specifically for use in classrooms with SWDs and ELs; the literature provides several examples.[9] However, given the enormous implementation challenges that come with introducing even a single protocol (e.g., training and certifying school and district personnel to reliably score the protocol, calibrating raters' scoring over time, scheduling observations around administrators' many time commitments, combining observation scores with other sources of evaluation data), it is unlikely—at least in the short term—that districts and states would be in a position to adopt multiple observation systems. A second option, which we believe is more likely to be adopted by districts/states, is to supplement an existing observation protocol (e.g., FFT) with a subset of items specific to teaching SWDs and ELs.[10] Alternatively, existing response categories on observation protocols could be adapted to more appropriately reflect teachers' interactions with SWDs and ELs.

One viable short-term solution would be to develop "scoring support documents" to assist observers in the scoring process, with an emphasis on the kinds of evidence-based practices that have proven to be effective for each of these populations. In the absence of such direction, it will be unlikely that observers will examine whether teachers attend to the needs of SWDs and ELs. If districts are to adopt such modified rubrics, they would benefit from guidance from researchers regarding how to help develop such assessments and whether these revisions have implications for the validity and reliability of the instruments. Thus far, however, researchers have not addressed these issues.

To improve observer familiarity with instruction for SWDs and/or ELs, districts could ensure that observers have some training or background specific to special student populations. It may be difficult for them to attend to whether teachers are providing appropriate instruction for all of their students. There is also a need for research that more closely examines observer performance specific to SWDs and ELs. For example, future research could investigate whether observer experience with SWDs and ELs improves their ability to assess teachers' practice reliably, whether there are strategies that districts can adopt to train observers to attend to teachers' effectiveness in educating SWDs and ELs, and whether modified versions of existing protocols improve observers' reliability in scoring teachers' instruction for SWDs and ELs.

Conclusion

The challenges related to developing useful measures of teaching effectiveness have received significant attention from researchers and practitioners. In this essay, we have presented challenges and potential solutions for using indicators of teacher effectiveness from two subgroups of students that have received little attention in teacher evaluation. As states and districts develop and modify teacher evaluation systems, we encourage them to attend to the unique challenges associated with including SWDs and ELs in each of the indicators. Not accounting for these challenges would undermine the validity of inferences about teachers' effectiveness, particularly in cases of teachers who have a large proportion of either SWDs or ELs in their classroom, and it would be counterproductive to the goal of providing a high-quality education to all students.

NOTES

[1]The challenges we outline are also applicable to other indicators. Student learning objectives (SLOs) represent one increasingly common strategy for measuring student learning when teachers work in grades or subject areas for which test scores are not available. States/districts are also exploring the use of student surveys or instructional artifacts to evaluate teachers. Although there are incentives to roll out these measures quickly, we urge districts and states to consider the challenges related to SWDs and ELs.

[2]For a summary of models, see, e.g., McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004).

[3]Value-added scores are typically estimated using data from students in Grades 4 through 8 who have taken the state general assessment in reading and mathematics (and possibly other subjects including science and social studies). Value-added scores can be calculated for approximately 30% of teachers, as reported on state websites.

[4]See, e.g., Chin, Daysal, and Imberman (2011) for ELs, and Hanushek, Kain, and Rivkin (2002) for SWDs.

[5]See, e.g., Russell and Kavanaugh (2011).

[6]See Thompson, Morse, Sharpe, and Hall (2005) for practical guidelines.

[7]FFT has been adopted in the evaluation systems of large school districts such as the Los Angeles Unified School District and states such as Illinois, Delaware, and Rhode Island.

[8]These practices are reflected in Echevarria, Vogt, and Short's (2008) Sheltered Instruction Observation Protocol (SIOP). Although the SIOP has empirical reliability, it is criticized for making strong assumptions about content teachers' training in language teaching and learning.

[9]For SWDs, one notable example is the observation protocol developed by Brownell et al. (2009), which was designed for use in special education reading instruction. For ELs, see the example of Echevarria et al.'s (2008) SIOP protocol.

[10]The state of Idaho, for example, is exploring the use of a modified version of FFT that would be more sensitive to the instructional needs of SWDs.

REFERENCES

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*, 1–28.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research and Practice*, *20*, 50–57.

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

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. Retrieved from http://metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Blanton, L. P., Sindelar, P. T., & Correa, V. I. (2006). Models and measures of beginning teacher quality. *Journal of Special Education*, *40*, 115–127.

Briggs, D., & Dominique, B. (2011). *Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times*. Boulder, CO: National Education Policy Center.

Brisk, M., & Zisselsberger, M. (2011). "We've let them in on the secret": Using SFL theory to improve the teaching of writing to bilingual learners. In T. Lucas (Ed.), *Teacher preparation for linguistically diverse classrooms: A resource for teacher educators* (pp. 111–126). New York, NY: Routledge.

Brownell, M., Bishop, A., Gersten, R., Klingner, J., Penfield, R., Dimino, J., Haager, D., . . .Sindelar, P. T. (2009). The role of domain expertise in beginning special education teacher quality. *Exceptional Children*, *75*, 391–411.

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

Chin, A., Daysal, N. M., & Imberman, S. (2011, October). *Impact of bilingual education programs on limited English proficient students and their peers: regression discontinuity evidence from Texas*. Working paper. Retrieved from http://www.uh.edu/~achin/research/bilingual_ed_texas_oct2011.pdf

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

Echevarria, J., Vogt, M., & Short, D. J. (2008). *Making content comprehensible for English-language learners: The SIOP model*. Boston, MA: Allyn & Bacon.

Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English-language learners. *Exceptional Children*, *66*, 454–470.

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

Goldhaber, D., & Hansen, M. (2010). Is it just a bad class? Assessing the stability of measured teacher performance. *CEDR* Working Paper 2010-3. University of Washington, Seattle, WA.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Inferring program effects for specialized populations: Does special education raise achievement for students with disabilities? *Review of Economics and Statistics*, *84*, 584–599.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard University Press.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56–64.

Hock, H., & Isenberg, E. (2011). *Methods for accounting for co-teaching in value-added models*. Princeton, NJ: Mathematica Policy Research. Retrieved from http://www.aefpweb.org/sites/default/files/webform/Hock-Isenberg%20Co-Teaching%20in%20VAMs.pdf

Ketterlin-Geller, L. R., Alonzo, J., Braun-Monegan, J., & Tindal, G. (2007). Recommendations for accommodations: Implications of (in) consistency. *Remedial and Special Education*, *28*, 194–206.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice*, *26*(3), 11–20.

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.

McCaffrey, D. F., Sass, T., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*, 572–606.

Paneque, O., & Barbetta, P. (2006). A study of teacher efficacy of special education teachers of English language learners with disabilities. *Bilingual Research Journal*, *30*, 171–193.

Pennock-Roman, M., and Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, *30*(3), 10–28.

Perie, M., Grigg, M., & Donahue, P. (2005). The nation's report card: Reading 2005 (NCES 2006-451). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*, 175–214.

Russell, M., & Kavanaugh, M. (2012). *Assessing students in the margin: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Mahwah, NJ: Lawrence Erlbaum.

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

Thompson, S. J., Morse, A. B., Sharpe, M., & Hall, S. (2005). Accommodations manual: How to select, administer, and evaluate use of accommodations for instruction and assessments of students with disabilities. *The Council of Chief State School Officers*. Retrieved from http://www.ccsso.org/Documents/2005/Accommodations_Manual_How_2005.pdf

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

U.S. Department of Education. (2007, April 9). Final Rule 34 CFR Parts 200 and 300: Title I—Improving the academic achievement of the disadvantaged; Individuals With Disabilities Education Act (IDEA). *Federal Register*, *72*, 17748–17781.

U.S. Department of Education, National Center for Education Statistics, Schools and Staffing Survey (SASS). (2008). "Public School, BIE School, and Private School Data Files," 2007–08. Retrieved from http://nces.ed.gov/pubs2009/2009321/tables/sass0708_2009321_s12n_02.asp

U.S. Department of Education. (2010). Race to the top. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Vaughn, S., & Linan-Thompson, S. (2003). What is special about special education for students with learning disabilities? *Journal of Special Education*, *37*, 140–147.

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

Zehler, A. M., Fleischman, H. L., Hopstock, P. J., Stephenson, T. G., Pendzick, M. L., & Sapru, S. (2003). Policy report: Summary of findings related to LEP and SpEd-LEP students. Retrieved from www.ncela.gwu.edu/files/rcd/BE021195/policy_report.pdf

AUTHORS

NATHAN D. JONES is an assistant professor in the Boston University School of Education, Two Silber Way, Boston, MA 02215; *ndjones@bu.edu.* His research focuses on teacher quality and teacher development, and his most recent research has examined the measurement of special educators' practice and their knowledge for teaching.

HEATHER M. BUZICK is a research scientist at Educational Testing Service–Center for Foundational and Validity Research, 660 Rosedale Road, MS 10-R, Princeton, NJ 08541; *HBuzick@ets.org*. Her research involves validity and fairness issues related to test takers from special populations.

SULTAN TURKAN is an associate research scientist at Educational Testing Service–Center for Foundational and Validity Research, 660 Rosedale Road, MS 10-R, Princeton, NJ 08541; *STurkan@ets.org*. Her research focuses on understanding, measuring, and improving teacher knowledge for teaching content to English learners.

Manuscript received June 22, 2012
Revision received September 18, 2012
Accepted October 5, 2012