Geiger & Amrein-Beardsley, *Cogent Education* (2019), 6: 1602943
**https://doi.org/10.1080/2331186X.2019.1602943**

cogent education

EDUCATION POLICY | RESEARCH ARTICLE

# Student perception surveys for K-12 teacher evaluation in the United States: A survey of surveys

Tray Geiger[1]* and Audrey Amrein-Beardsley[1]

Received: 30 October 2017
Accepted: 29 March 2019
First Published: 3 April 2019

*Corresponding author: Tray Geiger, Arizona State University, USA
E-mail: tjgeiger@asu.edu

Reviewing editor:
Sheng-Ju Chan, National Chung Cheng University, Taiwan, Province of China

Additional information is available at the end of the article

**Abstract:** In this overview piece, we document the six student perception surveys (SPSs) currently available for state, district, and school consumption and use, given SPSs are increasingly becoming one of the more popular "multiple measures" being used to evaluate teacher qualities. We present descriptive information about each of these SPSs, including information related to cost(s), constructs or domains assessed, number of items, response option types, grade level(s) in which the SPS can be administered, etc. Given this information, we also present implications for practice, as well as calls for future research into each SPS individually and writ large, given SPSs' increasing popularity post the passage of the Every Student Succeeds Act (ESSA), and also given what consumers might need to know before SPS purchase or use.

**Subjects: Education; Educational Research; Primary/Elementary Education; Leadership Strategy; School Leaders & Managers; Secondary Education; Education Policy**

**Keywords: student perception surveys; survey methods; teacher effectiveness; teacher evaluation**

## ABOUT THE AUTHORS

Tray Geiger is a Ph.D. Candidate in the Education Policy and Evaluation program in the Mary Lou Fulton Teachers College (MLFTC) at Arizona State University (ASU). His research focuses on teacher evaluation systems, high stakes testing, and educational policy in the United States. Tray and Audrey have collaborated on several peer-reviewed articles and presented at national conferences about teacher evaluation systems and measures, specifically around value-added models (VAMs), high stakes testing, and educational policy.

Audrey Amrein-Beardsley, Ph.D. is a Professor in the MLFTC at ASU. Her research focuses on teacher evaluation systems, high stakes testing, and educational policy in the United States. Audrey and Tray have collaborated on several peer-reviewed articles and presented at national conferences about teacher evaluation systems and measures, specifically around value-added models (VAMs), high stakes testing, and educational policy.

## PUBLIC INTEREST STATEMENT

There has been a notable increase in Student Perception Surveys (SPSs)—surveys given to elementary and secondary school students to evaluate their teachers—since the passage of the Every Student Succeeds Act in 2015. Given their novelty, especially as compared to other teacher evaluation measures (e.g. classroom observation frameworks, student growth measures), little is known about SPSs. This article describes and compares the six known SPSs available for commercial purchase and use to help inform state, district, and school administrators and teachers. How SPSs affect teachers, administrators, schools, districts, and states are also discussed.

## 1. Introduction

Since the 1990s, federal education accountability policies have outlined multiple reforms to improve the state of public education in the U.S. More specifically, a national movement incorporating standards-based reforms and a culture of testing students stemmed from the *A Nation at Risk* report (National Commission on Excellence in Education [NCEE], 1983), after which federal and statewide education policies focused on accountability measures as related to student achievement and teacher quality arose. Such reforms (e.g. No Child Left Behind [NCLB 2001]; Race to the Top [RTTT; U.S. Department of Education, 2009]) most frequently targeted student achievement as measured by standardized tests, and subsequently teacher quality and effectiveness as measured by student achievement, classroom observations of teaching, and most recently, students' perceptions of their teachers via student perception surveys (SPSs). These SPSs are the focus within this piece, in terms of our survey of these surveys.

SPSs are used to obtain students' opinions about different aspects of their teachers' attitudes and pedagogical practices. Ferguson and Danielson (2014) clustered such aspects into the "tripod" of "content, pedagogy, and relationships" with the idea that

> ...in order to deliver instruction effectively, teachers [need] an understanding of the subjects they [a]re teaching (content knowledge), they [need] sufficient skill to help students achieve understanding (pedagogic knowledge and skills), and they [need] to connect with students on a personal level so that students w[ill] be inspired to trust and cooperate (p. 105).

This "tripod," which stemmed from noted psychologist's Erik Erikson's life cycle identity development, was the basis for Ferguson's Tripod SPS (Ferguson & Danielson, 2014), and possibly other SPSs (discussed in more detail, below). Broadly speaking, SPSs are also informed by research on effective parenting, school and classroom environments, a sense of community, and social and emotional support (see Ferguson & Danielson, 2014).

It should be noted that while the SPSs discussed herein indicated in their marketing materials that they were developed per prior research, it was not possible to verify to what extent this was the case due to a lack of technical documents and details (discussed in more depth, below). This shortage of evidence is not necessarily uncommon for teacher evaluation measures, as classroom observation frameworks are also notably lacking in such detail as well. For example, Charlotte Danielson, the developer of one of the most popular classroom observation frameworks, the Danielson Framework for Teaching (FFT), indicates that the FFT is research-based (Ferguson & Danielson, 2014). However, as Author(s) (2017) indicate, there is "no reference...that might serve as evidence of such a claim" (p. 2). While it is highly likely that classroom observation frameworks and SPSs alike indeed developed per established research, it was not possible to provide a robust commentary on this the due to a lack of provided and/or accessible information.

Unlike value-added models (VAMs) or classroom observation measures, SPSs aim to allow students to assess the social, emotional, and instructional qualities teachers bring into the classroom. Teachers showcasing effective observable qualities have been found to positively influence students by transforming students' non-cognitive traits, such as perseverance and self-control. Such non-cognitive traits have been shown to be as critical as academic achievement for future life success (Jackson, 2012).

We conducted a review of the SPSs currently available to K-12 schools, and increasingly being taken up by states and districts as a part of teacher evaluation systems as based on "multiple measures" (i.e. more than VAM and observational indicators) of teacher evaluation (see, for example, Bill & Melinda Gates Foundation, n.d.). We first provide a brief history of SPS use in teacher evaluation systems, acknowledging that SPSs have been in use for decades in higher education (see, for example, Marsh, 1987, 1991, 2007) but not delving into that research for our purposes herein. Second, we review the background research on K-12 SPSs. Third, we review the six

(as of the time of this investigation) SPSs that we identified as being available for any district or state to purchase and use (i.e. vendor-developed SPSs), acknowledging that many states and districts (and perhaps even schools) are developing their own instruments for local purposes (of which we did not explore). The six vendor-developed SPSs include: (1) the iKnowMyClass survey, owned by Corwin, a SAGE company; (2) the K12 Insight Engage survey, owned by K12 Insight; (3) the My Student STeP Survey, owned by My Student Survey; (4) the Panorama survey, owned by Panorama Education; (5) the Tripod survey, owned by Tripod Education Partners; and (6) the YouthTruth student survey, owned by YouthTruth.

As these and additional options become available for teacher evaluation, it is important to track everything possible about these sets of measures to hopefully make information about these instruments, their options, their strengths/weaknesses, their levels of reliability and validity (e.g. validation studies, if applicable), and the like more accessible to the educated consumer, and especially the consumer who might not know what the right questions are to ask when contemplating any such vendor or instrument. Related, because there has not yet been any contributions to the field that adequately describe and compare the SPSs currently on the market, this piece is critical and also timely in that such a resource should be made available to help others make more informed decisions should they be looking to purchase a proprietary SPS. Thus, our survey of these surveys provides such a review and synthesis of the six SPSs listed prior.

## 2. Background

### *2.1. History of SPSs in teacher evaluation systems*

Surveys to evaluate teachers for formative purposes have been used for decades, primarily in higher education, albeit only sporadically for the evaluation of K-12 teachers, and not typically for high-stakes decisions (Peterson, Wahlquist, & Bone, 2000). It has only been of late that SPSs have been used in more formative or summative ways as integral parts of states' and districts' formal teacher evaluation systems as based upon "multiple measures" (Bill & Melinda Gates Foundation, n.d.; Chester, 2003; Darling-Hammond, 2012; The New Teacher Project, 2011; U.S. Department of Education, 2009). Such systems as based upon "multiple measures" are to permit more facets of teaching to be evaluated as part of the whole (i.e. overall teacher quality construct), all the while helping to, at least theoretically, alleviate some of the concerns associated with using each individual measure in isolation (e.g. observer subjectivities, inter-rater reliability, and comprehensiveness issues with classroom observations; reliability, validity, and fairness issues with VAMs; observer subjectivities, internal reliability, and response rate issues with SPSs).

In the early 1990s, it was estimated that SPSs were being used in under five percent of K-12 school districts nationwide (Educational Research Service, 1988; Loup, Garland, Ellett, & Rugutt, 1996). It was just after the turn of the century when SPSs began to make their entrance into K-12 education when Harvard economist and adjunct lecturer in public policy, Ronald Ferguson, constructed what would eventually become the leading SPS in the country (LaFee, 2014)—the Tripod Survey. At the time, Ferguson was helping a small school district in Ohio with an unexplained issue of "uneven student achievement." He ended up anonymously surveying the districts' students about their experiences with their teachers after none of the current methods of teacher assessment led to any conclusions about the "uneven" achievement data. As previously mentioned, Ferguson developed this initial survey based on Erikson's cycle of identity development (Ferguson & Danielson, 2014). Ferguson found the students' survey answers to be consistent and accurate in that students were able to recognize effective and ineffective teaching (LaFee, 2014). From there, Ferguson (2012) worked with the Ohio district for several years to further refine his instrument. By 2012, nearly one million students had completed it.

The growth of SPSs accelerated yet again after the infamous *Widget Effect* report was released in 2009, in which Weisberg, Sexton, Mulhearn, and Keeling (2009) highlighted the measurement troubles with America's contemporary teacher evaluation systems. Of explicit concern was that

upwards of 95% of teachers in Weisberg et al.'s sample were rated as "effective" or "highly effective." Weisberg et al., accordingly, argued that so many teachers receiving such high evaluations was illogical given student achievement across the U.S. was average, at best, in comparison to other industrialized nations. Hence, this too increased calls for teacher evaluation reform. The U.S. was purportedly not performing well, and the country's education system as reliant upon an antiquated teacher evaluation system, allegedly continued to permit mediocre teachers to hide behind the curtains of obsolete teacher evaluation systems, so the logic went.

The combination of Weisberg et al.'s (2009) findings, the aforementioned $4.35 billion in RTTT funds incentivizing states to adopt reformed teacher evaluation systems, and also the congressional authorization of the NCLB waivers excusing states from not meeting NCLB's prior 100% student proficiency by 2014 goals if states reformed their teacher evaluation systems, led states to reform their teacher evaluation systems. Adding to these policies was the rhetoric surrounding teacher accountability that led states to explore, adopt, and implement more teacher evaluation indicators (i.e. "multiple measures") to contribute to the new and improved teacher evaluation systems that states were being incentivized to develop.

Around the same time, the Bill & Melinda Gates Foundation funded the Measures of Effective Teaching (MET) study (i.e. 2009 to 2011). SPSs were an integrated part of the "multiple measures" used within and across MET studies, which also increased the interest in and use of SPSs across the nation. The goal of the MET study was to identify measures of effective teaching that could inform teachers, schools, and districts about teachers' strengths and weaknesses (Bill & Melinda Gates Foundation, n.d.); hence, SPSs also fit this goal. Results indicated that the Tripod—the SPS used as a part of the MET studies—was a valid and reliable way to measure teacher effectiveness, and potentially more so than classroom observation measures (Bill & Melinda Gates Foundation, 2012). This finding also contributed to the general rise of and interest in SPSs. Consequently, one year after the MET study findings were published, SPSs were increasingly being publicized as another viable "multiple measure" option (e.g. Butrymowicz, 2012; Heitin, 2012; O'Donnell, 2014). It was then that SPSs truly began to be investigated, adopted, and incorporated into states' reformed teacher evaluation systems.

As of 2015, seven states[1] required SPSs to be used as part of the state's/districts' reformed teacher evaluation systems, and an additional 25 states[2] and the District of Columbia allowed for SPS use within their formal teacher evaluation systems (Doherty & Jacobs, 2015). The relative weight of SPSs data as compared to the weights attributed to the other now common measures of teacher effectiveness (e.g. teachers' value-added data, teachers' observational ratings) continue to vary per state, and it remains unknown how all of this might progress, or regress, as per the Every Student Succeeds Act (ESSA, 2015). Given that the use of teacher-level value-added data seems to be diminishing or even disappearing from some states' teacher evaluation systems (Will, 2016; see also state-level policies in Connecticut, Nevada, and Texas), it is possible that additional states will begin to utilize SPSs more; although, this would also likely be true even if states were not to abandon their value-added initiatives. In other words, it presently seems that SPSs will continue to be integrated into states'/districts' teacher evaluation systems, regardless of the other indicators already in use. If SPSs are not currently in use, they may be soon adopted, given SPSs are apparently being viewed as the trending measures that allow previously unevaluated aspects of teaching to be assessed.

### *2.2. Research on K-12 SPSs*

#### *2.2.1. Benefits when using SPSs*

Although relatively new to the formal teacher evaluation scene, proponents of SPSs see many benefits to their use. From a logistical standpoint, as also related to the broader picture of teacher evaluation, one of the biggest advantages of utilizing SPSs is that such surveys target the populations 1) with whom teachers interact the most as 2) the intended consumers of teachers'

instruction (Bill & Melinda Gates Foundation, 2012; Ferguson, 2012; Follman, 1992; Goe, Bell, & Little, 2008; LaFee, 2014).

Kane and Staiger (2012), using data from the MET study, also purported that SPS data are more reliable and stable than classroom observational data. This may be because students spend multiple hours with their teachers, compared to trained classroom observers who might spend only a few hours with each teacher every year. In addition, student evaluations of teachers are averaged across many students, where observational data is typically only gathered by one or two people, which helps to alleviate issues with inter-rater reliability (i.e. issues that plague teacher observational systems and scores).

Likewise, evidence suggests that SPS data may also be more reliable and stable than teachers' value-added data, also keeping in mind how relatively unstable value-added indicators continue to be (Ballou & Springer, 2015; Schochet & Chiang, 2013; Yeh, 2013), with researchers demonstrating the overall stability (i.e. reliability) of SPS data to be acceptable, or more specifically moderate[3] in the case of the Tripod data included in the aforementioned MET studies. Polikoff (2015), however, found that the overall stability of this SPS measure was low to moderate,[4] and therefore more or less on par with the reliabilities observed across teachers' value-added output (see also Bill & Melinda Gates Foundation, 2012; Fauth, Decristan, Rieser, Klieme, & Buttner, 2014; Ferguson, 2008; Follman, 1995; Peterson et al., 2000; Wagner, Gollner, Helmke, Trautwein, & Ludtke, 2013; Wallace, Kelcey, & Ruzek, 2016).

There are also purported benefits to SPSs from a measurement perspective, especially in relation to other measures. Findings taken from the MET studies evidenced that the Tripod—both overall and each individual construct—had high predictive validity with student achievement gains, which implied that SPSs might also be predictive of outcomes in subjects and grades that are untested (Bill & Melinda Gates Foundation, 2012; Raudenbush & Jean, 2014).

Notwithstanding, other pragmatic benefits to SPSs are that they are relatively cost-effective (Bill & Melinda Gates Foundation, 2012; Schulz, Sud, & Crowe, 2014), especially as compared to classroom observations and the test- and data-based infrastructures and systems required for the calculations and analyses of teachers' value-added (e.g. via VAMs). Surveys are also quick and easy to administer to students, and allow schools (and districts/states) to gather relatively quick impressions of students' opinions about their teachers' qualities (Schulz et al., 2014). This is to also potentially allow for teachers to more promptly adjust their pedagogical practices and behaviors in the classroom in real time (e.g. given the time delays thwarting the use of similar feedback taken from the aforementioned VAM-based estimates; Stevens, Harris, Liu, & Aguirre-Munoz, 2013; Whitehurst, Chingos, & Lindquist, 2014).

Also, given the multiple constructs—the relevant pedagogical concepts or characteristics that SPSs are designed to measure, as defined within the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014)—of which SPSs are comprised, teachers are also potentially able to receive specific feedback that focuses on a variety of facets of teaching, also facilitating the real time improvement of teacher practice (Bill & Melinda Gates Foundation, 2012). This also stands in stark contrast to teachers' VAM data, where teachers might be rated as "ineffective" but have no idea what part of their instruction needs improvement (Whitehurst et al., 2014).

### *2.2.2. Concerns with using SPSs*

As is typical in all areas of research, including teacher evaluation inquiry, with benefits also come some noteworthy concerns. One of these is that all students are not able to evaluate all facets of what teaching actually entails, and all instruments are incapable of effectively capturing these facets; hence, here, too, is an issue with representativeness. For example, while students may be able to assess whether a teacher is engaging (however "engaging" may be defined), or presents new material in interesting ways (however "new" and "interesting" may be defined), students may

not be able to determine how well a teacher knows a certain set of content standards or performance objectives (Goe et al., 2008; Peterson, Stevens, & Ponzio, 1997).

Additionally, it can be difficult to get buy-in from teachers for the use of SPSs, especially for summative purposes, as many teachers fear that students' evaluations could be biased based on how their students personally feel about them, versus their qualities as teachers (Kauchak, Peterson, & Driscoll, 1985; Schulz et al., 2014). Teachers also fear that their SPS-based output may be biased given the possibility that students may not take the surveys very seriously with or without fully understanding the potential implications survey output might have on their teachers' professional (and personal) lives (e.g. informing consequential personnel decisions regarding tenure, merit pay, termination; Nott, 2014; Schulz et al., 2014). From the higher education literature, while not knowing to what extent this research might generalize into the K-12 sector, we know that there are many biasing factors at play that are significantly correlated with SPS output. These potentially biasing factors include but are not limited to how students' grades, perceptions of course rigor, class sizes, genders (and whether an individual student's gender matches or varies from an individual teacher's gender), whether the course is a required core course or an elective, and the like (see Marsh, 2007, for a list of potentially biasing factors that are also theoretically in play). Hence, teachers likely have empirically grounded causes for pause.

Related, measurement error is also an empirical reality for SPSs, with concerns centering around reliability, validity, bias, fairness, and more. Regarding reliability, while inter-rater reliability might be of relatively less concern (e.g. in comparison to teachers' observational estimates), internal reliability is certainly of concern in that it is imperative that any SPS that is used is demonstrated to be internally reliable or, rather, yield internally consistent responses across student respondents. Regarding validity, validity of inference remains important, especially given the potential threats to validity also prevalent with surveys of any kind. Selection bias, for example, is something with which all SPS users must contend; that is, if "enough" students do not respond, there are few if any valid inferences that can be drawn on behalf of the whole. Halo-rating errors are also common, as characterized by students forming a general impression of their teachers or a teacher's instructional effectiveness, and simply marking a teacher high on all marks without discriminating among items. Despite the best efforts of instrument developers to construct SPSs with items reflecting important dimensions of teaching, the general impression each student brings to such evaluations too often drives the rating process, often yielding very high internal consistency reliability estimates ($\alpha \geq 0.90$). The inverse is also true with severity errors, with general negative impressions skewing teachers' SPS output downwards, while also thwarting inferential validity (see Hobson & Talbot, 2001; Pike, 1999; Schmelkin, Spencer, & Gellman, 1997).

Likewise, SPS results can be negatively impacted due to a variety of other factors, such as the nonrandom sorting of students into classrooms (which is a preeminent cause for concern, also with student value-added data; Rothstein, 2009, 2010) that yields classroom compositions, relatively and sometimes purposefully, homogenized by student-level factors, such as race/ethnicity, achievement level, or socioeconomic status (Desimone, Smith, & Frisvold, 2010; Driscoll, Peterson, Crow, & Larson, 1985). These confounding factors might also cause error, or "noise," also provided whatever content the teacher may teach.

Regardless, and as noted by Goe et al. (2008), most important is that that users of SPSs heed caution as SPSs should only be used in summative teacher evaluations if and when they have been validated for such intended purposes (see also AERA et al., 2014). This also carries serious policy implications not to be overlooked at any level.

### *2.2.3. Research on the six vendor-developed SPSs*

As SPSs have only recently become more popularized, however, the majority of empirical studies to date use survey data from the MET study (i.e. Tripod survey data). The other five vendor-developed surveys we review in this article have not yet been analyzed, either alone or in conjunction with

other measures of teaching effectiveness, with the same frequency of the Tripod, if at all. Hence, we present next what we did in this survey of these surveys to add to this area of research.

## 3. Data collection

To conduct our survey of the six SPSs mentioned prior, we reviewed all survey companies' websites to determine what material they made publicly accessible, if any. More specifically, we reviewed each website to determine: 1) whether the full survey instrument was provided to review; 2) whether the survey's constructs (e.g. the big latent concepts to be assessed) were accessible; 3) whether the survey's items per construct were accessible; 4) whether survey-specific details, such as the number of items and item format (e.g. Likert scale) were detailed; 5) what, if any, developmental, technical, or other related reports (e.g. validation studies) were available; 6) whether the cost associated with purchasing, implementing, and/or analyzing survey results were made explicit; and 7) what, if any other information pertinent to the survey's background, development, use or use requirements, implementation or administration directions, analytical notes or recommendations, and the like were made known.

Thereafter, we contacted each company that did not publically provide this information to request copies of their survey instruments, as well as any other relevant material that we might have been missing after our initial search for the abovementioned information. We did not request any actual survey data (i.e. students' responses) from any of the companies.

Of the six SPSs, the Panorama survey was the only one that was freely accessible online; however, we were also able to obtain the iKnowMyClass Survey and the YouthTruth Survey from the Corwin and YouthTruth companies, respectively. Tripod Education Partners required Institutional Review Board (IRB) approval to access their instrument, but even with IRB approval, they denied granting us access because this study was not a part of a student's dissertation. MyStudentSurvey failed to respond to several emails, and K12 Insight immediately refused to participate, citing the proprietary nature of its survey as the reason why.

Regardless, and while our attempts to access these instruments and information about these instruments might serve as study findings in and of themselves, for those who did not respond to either our initial emails or the opportunity to review the article draft, we still did our best to analyze all that we could without these three (50%) of the actual survey instruments in hand. Likewise, we provided a final draft of this article to all six SPS companies to allow them to review our findings, as well as to provide each with an opportunity to confirm or refute any of the information we asserted or advanced. Three of the six companies (K12 Insight, Tripod, YouthTruth) responded with feedback, generally speaking to information about technical development and validation, or brief details, revisions regarding wording, or other minor details.

## 4. Survey review

In this section we provide a general overview of the six surveys as a group. We also discuss several commonalities across all surveys, as well as some points of difference. However, it is also worth noting at the onset of this discussion that no information was publically available online regarding the K12 Insight Engage Survey regarding its development, content, or related details, and, as noted above, no additional information was provided to us when we contacted the K12 Insight company. Hence, what we provide next is everything we accessed about five of the six (83%) of the actual SPSs available on the market. Each survey's specific features and components can be found in Table 1 (i.e. parent company, developmental history, highest degree of lead developer, cost(s) associated with use, constructs or domains, number of items, response option types, grade level(s) in which it can be administered and used, whether online or paper administration is available, the primary website, and each company's response to our request for the survey and other related information).

**Table 1. Characteristics of Currently Available Vendor-Developed SPSs**

| | iKnowMyClass Survey | K12 Insight Engage Survey | My Student Survey STeP Survey | Panorama Survey | Tripod Survey | YouthTruth Student Survey |
|---|---|---|---|---|---|---|
| Parent Company | Corwin (SAGE) | K12 Insight | My Student Survey | Panorama Education | Tripod Education Partners | YouthTruth |
| Development | Dr. Russell Quaglia (Quaglia Institute for Student Aspirations) | *Unknown* | Dr. Ryan Balch (Vanderbilt University for doctoral dissertation) | Dr. Hunter Gehlbach and researcher team (Harvard Graduate School of Education) | Dr. Ronald Ferguson (Harvard University Adjunct Lecturer in Public Policy) | Phil Buchanan, Ellie Buteau, Valerie Threlfall (Center for Effective Philanthropy) |
| Highest Degree of Lead Developer | Dr. Russell Quaglia: Ed. D.,Educational Administration, Columbia University | *Unknown* | Dr. Ryan Balch: Ph.D., Leadership and Policy Studies, Vanderbilt University | Dr. Hunter Gehlbach: Ph. D., Educational Psychology, Stanford University | Dr. Ronald Ferguson: Ph.D, Economics, Massachusetts Institute of Technology (MIT) | Phil Buchanan: MBA, Harvard University; Ellie Buteau: Ph.D., Social-Personality Psychology; Valerie Threlfall: MBA, MPP, Northwestern University, Harvard University |
| Cost | $250/school plus $2.50/ student | *Unknown* | *Unknown* | $2.50/student*** | $3/classroom plus $50/ school ($2,000 minimum)**** | *Unknown* |
| Constructs/Domains | Class Efficacy, Cooperative Learning, Critical Thinking, Discipline Problems, Engagement, Positive Pedagogy, Relationships, Relevance | *Unknown* | Coach, Content Expert, Counselor, Manager, Motivator, Presenter | Classroom Climate, Classroom Engagement Classroom Learning Strategies, Classroom Mindset [Dispositional, Behavioral], Pedagogical Effectiveness, Classroom Rigorous Expectations, Classroom Belonging, Classroom Teacher-Student Relationships, Valuing of the Subject | Captivate, Care, Challenge, Clarify, Classroom Management, Confer, Consolidate | Academic Rigor and Expectations, Classroom Culture, Instructional Methods, Personal Relationships, Relevance, Student Engagement |
| Grade Level(s) | Grades 3–5; Grades 6–12 (short and long versions) | *Unknown* | Grades 6–12* | Grades 3–5; Grades 6–12 | Grades K-2; Grades 3–5; Grades 6–12 | Grades 3–5; Grades 6–12 |

(*Continued*)



cogent education

**Table 1. (*Continued*)**

| | iKnowMyClass Survey | K12 Insight Engage Survey | My Student Survey STeP Survey | Panorama Survey | Tripod Survey | YouthTruth Student Survey |
|---|---|---|---|---|---|---|
| Number of Items | Grades 3–5 survey: 27; Grades 6–12 survey: 20 (short), 50 (long) | *Unknown* | Unknown** | Grades 3–5 survey: 43; Grades 6–12 survey: 50 | Grades K-2: 36; Grades 3–5: 36; Grades 6–12: 36 | Grades 3–5: 25; Grades 6–12: 30 |
| Response Option(s) | Five-point Likert scale | *Unknown* | Five-point Likert scale | Five-point Likert scale; Scale adjusted for content of each item | Five-point Likert scale | Grades 3–5: Three-point Likert scale; Grades 6–12: Five-point Likert scale |
| Online Administration | Yes | *Unknown* | Yes | Yes | Yes | Yes |
| Paper Administration | No | *Unknown* | Yes | Yes | Yes | Per Request |
| Main website | iKnowMyClass Survey | K12 Insight Engage Survey | My Student Survey STeP Survey | Panorama Survey | Tripod Survey | YouthTruth Student Survey |
| Response When Contacted | Provided survey items and additional documents | Refused to provide survey or additional information; cited its proprietary nature | No response | N/A; Survey items and additional documents publicly shared online | Required approved IRB proposal and use of data for student's dissertation to share survey items | Provided survey items and additional documents |

*Note*: Any cell containing *"Unknown"* signifies that such information was either not available, or not evident on public-facing websites, publicly posted documents, or documents shared directly with me.
*It is not clear if a separate version exists for elementary grades, as the company website does not specify grade levels; however, the Bill & Melinda Gates Foundation (2012) and English, Burniske, Meibaum, and Lachlan-Hache (2015) indicated there is a separate version for Grades 4–5, yet other documents do not support this (Voight & Hanson, 2012). Further, English et al. (2015) also reference a separate survey for Grades 3–5.
**It is not clear how many items the survey contains due to conflicting documents. The Bill & Melinda Gates Foundation (2012) indicates there are 55 items, while Voight and Hanson (2012) indicate there are 63.
***This is the cost for survey administration and data analysis; the surveys (i.e. individual items) are freely available online at https://www.panoramaed.com/panorama-student-survey
****This is the estimated cost for the Ohio Department of Education (ODOE), per a publically available service summary. For more information, see (Ohio Department of Education, 2017, p. 32).



cogent education

Although most surveys are similar to others, overall, each is also unique in terms of its constructs, items, and other relevant characteristics (e.g. cost(s), implementation procedures, question format). In terms of their similarities, across all survey instruments, all but the Panorama survey items were proprietary in nature in that copies of the instruments were not publicly accessible (i.e. the average user would not be able to access five of six (83%) of these instruments without engaging with the company, or rather the sales or customer service department of the company, first). However, most companies provided general survey information online (either directly on their websites or in publically available developmental or informational documents), such as constructs measured, grade levels, number of items, response options, and the like. Although, it should also be noted that not all publically available information was easy to find, as we found some documents and reports via search engines, such as Google Scholar.

All surveys except one, the K12 Insight Engage survey, were developed by leaders or teams of people who had doctoral degrees. Only two developers, from My Student Survey and Panorama Education, had degrees related to education (Leadership and Policy Studies and Educational Psychology, respectively). Interestingly, while one of the three YouthTruth developers had a Ph.D. (in social psychology), the others had master's degrees in business or public policy rather than advanced degrees in education, measurement, or another related field, as some might have assumed or expected.

Cost information for survey purchase, implementation, and/or analysis was also not often readily available, and when it was listed, it was difficult to find. Rather, it seemed that companies wanted interested parties (e.g. potential customers) to contact the company for this information. Regardless, the costs associated with these SPSs included not just access to a company's survey instrument, but also detailed data analyses of the responses (e.g. descriptive/summary information; response breakdowns for each teacher by student, classroom, grade, etc.; breakdown of student responses by student demographics by race/ethnicity) and, depending on the company, use of an online portal to view both their raw and analyzed data. Some companies offer extra services for additional costs, such as in-person training or professional development to better understand survey results, or the option to disseminate surveys via paper versus online.

Arguably, the most important feature of each survey instrument are the constructs that the instrument developer purports the instrument measures. Related, all survey instrument developers assert that their instruments measure constructs that are related to the socio-emotional aspects of teaching which appears to be a significant selling point of SPSs. Compared to other measures of evaluating teaching, such as VAMs or classroom observations, SPSs offer the best option to provide feedback about the more affective aspects of teaching, as also noted prior (see also Jackson, 2012). Accordingly, while each SPS does include constructs that focus more distinctly on teachers' pedagogical practices and skills, such as classroom efficacy or management (iKnowMyClass, MyStudentSurvey, Tripod), classroom or academic expectations (Panorama, YouthTruth), and instructional methods (YouthTruth), all SPSs include constructs that focus on the socio-emotional aspects of teaching (see Table 2).

While all surveys included constructs focusing on both pedagogical skills and affective aspects to teaching (see Table 2), it was difficult to determine exactly to what extent commonalities existed in these constructs or domains across surveys. Other than the Panorama survey, full items were not provided or accessible for the other SPSs. Thus, there was no way to do a more in-depth analysis. Further, the way each construct was named across surveys differed greatly. For example, the iKnowMyClass survey contains a construct titled "Positive Pedagogy," the Panorama survey contains a construct titled "Pedagogical Effectiveness," and the YouthTruth Student survey contains a construct titled "Instructional Methods." While these construct sound similar, and while one could posit that they measure the same latent construct, without access to the full spectrum of items per construct per survey, and ideally knowing how and why each of these constructs were created, it was not possible to accurately assess to what extent these constructs measured the same things.

**Table 2. Categorization of Constructs per SPS**

| SPS | Pedagogy/Skills-Focused Constructs | Affective Constructs |
|---|---|---|
| iKnowMyClass Survey | Class Efficacy, Cooperative Learning, Critical Thinking, Discipline Problems, Positive Pedagogy, Relevance | Engagement, Relationships |
| K12 Insight Engage Survey | *Unknown* | *Unknown* |
| My Student Survey STeP Survey | Content Expert, Manager, Presenter | Coach, Counselor, Motivator |
| Panorama Survey | Classroom Climate, Classroom Learning Strategies, Classroom Mindset [Behavioral, Dispositional], Pedagogical Effectiveness, Valuing of the Subject | Classroom Engagement, Classroom Belonging, Classroom Teacher-Student Relationships, Classroom Rigorous Expectations |
| Tripod Survey | Captivate, Clarify, Classroom Management, Consolidate | Care, Challenge, Confer |
| YouthTruth Student Survey[1] | Academic Rigor and Expectations, Instructional Methods, Relevance | Classroom Culture, Personal Relationships, Student Engagement |

*Note*: Any cell containing "*Unknown*" signifies that such information was either not available, or not evident on public-facing websites, publicly posted documents, or documents shared directly with me.

[1]Per YouthTruth representatives

Most survey companies provide surveys for students in grades 3–12, though the Tripod survey is also available for students in Kindergarten through grade 2. Likewise, survey instruments appear to come in different versions for students in different grades, which also allows survey developers to adjust the length of the survey and language and word choice to be age appropriate. Related, the surveys range between 20 and 50 Likert-scale items, with the number of items varying per survey company and per grade ranges of the students. All companies provide online administration of the survey, while several offer hard copy administration upon request.

Lastly, one surprising finding, and possibly the most important regarding potential implications for current and potential SPS users, was that the survey instruments, overall, were greatly lacking in vetting and review by external academic/research communities. While technical documents, user guides, white papers, and some externally conducted analyses were available for and about several of the survey instruments, along with numerous magazine and newspaper articles also used for marketing purposes, there was an overwhelming absence of external academic/research (e.g. peer-reviewed) articles surrounding all of these SPSs.

Hence, not only might this serve as a call for more research to be done in all areas of SPSs in K-12 education, this might also serve as a call out to potential buyers and users to not simply trust that which is advertised or delivered to potential consumers without asking for research evidence (even if just internal at this point) to support the claims being made by the companies offering (and selling) these SPSs. For example, if an SPS is advertised for its capacities to "improve teachers' teaching," to help teachers "make stronger relationships," or to "improve your school," potential SPS consumers and users might ask SPS companies for the evidence backing such claims, ideally including some of the psychometric evidence noted prior.

## 5. Discussion

As mentioned, the main draw of SPSs is that they are able to allow important dimensions of teaching and teachers' purported or perceived impacts on students that are often overlooked by other measures of teaching effectiveness to be evaluated. Such affective or socio-emotional dimensions of teaching should be captured and included in teacher evaluation systems, whether

for summative purposes and perhaps more importantly for formative purposes, including teacher reflection and professional and instructional development. However, while the benefits of SPSs are many, there are several important points of note that must be discussed before we (or others) might advocate for the widespread use of SPSs in teacher evaluation systems.

One main concern is the lack of peer-reviewed research, as noted above. This dearth of independently or blindly verified research raises serious concerns about the quality of these specific surveys, as well as their overall levels of validity and reliability. Even in cases where statistical relationships have been demonstrated between SPS data and other measures of teaching quality, it remains unknown as to what actually causes those relationships (Wallace et al., 2016). As SPSs are required or allowed to be a part of 65% of states' (i.e. 33 out of 50 states plus DC) formal teacher evaluation systems (Doherty & Jacobs, 2015), SPS use could accordingly be problematic in that the instruments have not been properly vetted or deemed as appropriately valid and reliable, perhaps internally but more importantly as per the general academic community at large.[5] Thus, it is possible that schools, districts, and states are evaluating teachers and, as the case may be, making consequential personnel decisions based on unreliable and/or invalid SPS-based information.

Related to the lack of externally vetted research are many unknowns surrounding the technical development and validation of the SPSs, as well. For example, and as briefly mentioned prior, nowhere did the majority of survey developers (i.e. all but YouthTruth) note in publicly accessible or provided documents that when they adjusted their surveys for multiple users and purposes (e.g. different grade levels), that they also empirically made sure the adjusted instruments still functioned as marketed and theorized (e.g. conducting analyses of internal reliability and factor structures, as revised and adjusted to fit consumers' needs or demands). Related, while some of the SPSs were developed by people with educational and/or methodological training, it is unclear to what extent the SPS developers and related staff have the adequate and/or necessary psychometric knowledge and skills to conduct such complex analyses, as needed to support the variety of uses and users of each survey, accordingly.

Another related topic worthy of discussion is whether children, especially young children, are able to provide valid and/or reliable feedback (for example, see Scott, 1997; Vaillancourt, 1973). Most survey companies provide their SPSs for students as young as grade 3, while the Tripod survey is also available for students as young as Kindergarten through grade 2. While some surveys companies, such as YouthTruth, have developed their surveys using best practices for surveying young children (for example, see de Leeuw, 2011), it is still possible that younger students are not able to provide as valid or as reliable responses as survey developers or school administrators might like, which is also a potential cause for concern, especially if SPS results are to inform consequential decisions for teachers in these grade levels as well.

Lastly, all surveys with the exception of the Panorama survey, are proprietary in nature. Further, all companies developing and overseeing each survey are for-profit entities, save for YouthTruth, which is a non-profit company. The proprietary and for-profit nature of the majority of SPS companies is of special concern as these surveys are being marketed to states and school districts that are funded by limited public taxpayer funds. Given the importance of holistically assessing teachers' effectiveness or quality, including such latent constructs that, to date, only SPSs can capture, it seems only ethical and fair that SPS companies are more open to making their survey instruments and related developmental and technical reports and other documents more accessible, given this is also a public venture. Further, while some of the survey costs might appear reasonable on the surface, for schools in some of the country's larger districts or states, the use (i.e. purchase) of an SPS can easily run hundreds of thousands of dollars. School administrators have previously indicated that their budgets simply do not have the funds to pay for such evaluation instruments and accompanying data analyses, on top of the already established classroom observation protocols and VAMs, without additional

fiscal support (e.g. grants) (Bailis, Homana, & Melchior, 2010). Hence, this too is a noteworthy consideration when contemplating the costs of evaluating teacher effectiveness using "multiple measures" such as these, as also in line with current policy and pragmatic trends.

## 6. Conclusion

As noted in this survey of surveys, SPSs are another useful tool for states, districts, and schools to adopt and implement, as SPSs in conjunction with other "multiple measures" such as classroom observations and VAMs allow for more holistic evaluations of teachers. As such, it is the affective domains that SPSs purport to capture that should be included alongside the primary two measures —classroom observations and VAMs—being used across teacher evaluation systems, as these socio-emotional qualities of teaching likely affect and transform students' academic lives and psycho/social-cognitive selves as much, if not more than, teachers' pedagogical and instructional skills and capacities.

Related, as per ESSA (2015), states must include four school-level indicators in their newly proposed accountability systems, though only three are required to be academic in nature (e.g. state test proficiency, graduation rates). The fourth indicator is allowed to be a non-academic factor, which might include aspects akin to some of the socio-emotional constructs that make up the SPSs discussed in this article. With states' freedom in determining this fourth indicator, it is possible that more states will turn to assessment and accountability measures unrelated to teacher quality (at the aggregate school, district, or state level) as measured in the traditional sense (i.e. via student value-added data and classroom observations), given the numerous and well-documented logistical, pragmatic, and statistical issues with such measures. That then leaves the door open for SPSs (among other surveys, such as peer or parent surveys) to be incorporated into state-level, teacher evaluation and accountability policies and plans in a more widespread manner.

However, before consumers might simply adopt these surveys and blindly trust their empirical value, it is important that said consumers recognize and fully understand the following before SPS purchase and, especially, high-stakes use: (1) SPSs lacks peer-reviewed research, so they might not be ready for prime time, especially if high-stakes consequences are to be attached to SPS output; (2) related, there are many unknowns about the technical development and also empirical validation of SPSs, overall and by sub construct or domain (e.g. Author(s), 2017); (3) whether young children are able to provide valid and reliable feedback is a serious concern, as it has been for decades (e.g. Popham, 2011), even though some still encourage the use of SPSs in early elementary grade levels (e.g. Ferguson, 2008); and (4) all SPSs except one are proprietary in nature, which should make users more cautious when considering a given SPS, given these surveys are typically not open for external vetting and empirical review. At minimum, potential users might ask all SPS companies, regardless of the surveys' proprietary natures, for the technical and empirical evidence in general and as per the four points above prior to purchase and use.

Likewise, it is even more imperative at this point that much more research is conducted and made accessible to the consumer, as well as to the public, about these SPSs individually and as a whole, given their documented strengths and barriers, empirically and in practice. Without this critical information, users might continue to use these instruments, and more importantly SPS output, for consequential decision-making purposes, somewhat blindly. Likewise, consumers might continue to make claims as based on SPS output that are not necessarily warranted by the evidence. Consequently, it is incumbent for us as a research community to contribute analyses related to SPSs that are impactful and meaningful to and within educational policy and practice in this regard.

**Funding**
The authors received no direct funding for this research.

**Author details**
Tray Geiger[1]
E-mail: tjgeiger@asu.edu
ORCID ID: http://orcid.org/0000-0002-0544-8109
Audrey Amrein-Beardsley[1]
E-mail: audrey.beardsley@asu.edu
[1] Educational Policy and Evaluation, Mary Lou Fulton Teachers College, Arizona State University, PO Box 871811, Tempe, AZ 85287-1811, USA.

**Citation information**
Cite this article as: Student perception surveys for K-12 teacher evaluation in the United States: A survey of surveys, Tray Geiger & Audrey Amrein-Beardsley, *Cogent Education* (2019), 6: 1602943.

**Notes**

1. CT, GA, HI, IA, KY, MA, UT.
2. AK, AZ, AR, CO, DC, FL, ID, KS, MN, MS, MO, NV, NM, NC, ND, OH, OK, OR, PA, SC, TN, VT, VA, WA, WI, WY.
3. Stability coefficients for Tripod data were .41 and .41 for mathematics and English/language arts teachers, respectively, and both coefficients were significant at the $p < .001$ level (Polikoff, 2015). We interpret these coefficients as per Merrigan and Huston (2008) definition: $r$: $.8 \leq r \leq 1.0$ = a very strong correlation; $.6 \leq r \leq .8$ = a strong correlation; $.4 \leq r \leq .6$ = a moderate correlation; $.2 \leq r \leq .4$ = a weak correlation; and $0 \leq r \leq .2$ = a very weak correlation, if any at all.
4. Sub-scale stability coefficients for Tripod data ranged from .32-.41 for mathematics teachers and from .32-.45 for English/language arts teachers, depending on the scale, and all coefficients were significant at the $p < .001$ level (Polikoff, 2015). We interpret these coefficients as per Merrigan and Huston (2008) definition: $r$: $.8 \leq r \leq 1.0$ = a very strong correlation; $.6 \leq r \leq .8$ = a strong correlation; $.4 \leq r \leq .6$ = a moderate correlation; $.2 \leq r \leq .4$ = a weak correlation; and $0 \leq r \leq .2$ = a very weak correlation, if any at all.
5. While several peer-reviewed articles do exist (Polikoff, 2015; Polikoff & Porter, 2014), they are based on the Measures of Effective Teaching (MET) study (which used Tripod survey data); the MET data has yet to be psychometrically validated in a peer-reviewed study, though Wallace et al. (2016) did test the factor structure of the Tripod instrument.

**References**

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bailis, L., Homana, G., & Melchior, A. (2010). *Formative evaluation of youthTruth - Final report*. Waltham, MA: Center for Youth and Communities, Heller School for Social Policy and Management, Brandeis University.

Ballou, D., & Springer, M. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher*, 44(2), 77–86. doi:10.3102/0013189X14474904

Bill & Melinda Gates Foundation. (2012). *Asking students about teaching: Student perception surveys and their implementation*. Seattle, WA: Author.

Bill & Melinda Gates Foundation. (n.d.). Frequently asked questions. Retrieved from http://k12education.gatesfoundation.org/teacher-supports/teacher-development/measuring-effective-teaching/why-met-additional-resources/frequently-asked-questions/

Butrymowicz, S. (2012, May 3). Student surveys may help rate teachers. *The Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/student-surveys-may-help-rate-teachers/2012/05/11/gIQAN78uMU_story.html?utm_term=.05de45df7cf8.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice*, *22*(2), 32–41. doi:10.1111/j.1745-3992.2003.tb00126.x

Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education. Retrieved from http://www.smmcta.com/uploads/9/9/4/2/9942134/evaluation_research_stanford_2012.pdf

de Leeuw, E. (2011, May). *Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting*. Naantali, Finland: Paper presentation at the annual meeting of the Academy of Finland.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy*, *24*(2), 267–329. doi:10.1177/0895904808330173

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015. Evaluating teaching, leading and learning*. Washington, DC: National Council on Teacher Quality.

Driscoll, A., Peterson, K. D., Crow, N., & Larson, B. (1985). Student reports for primary teacher evaluation. *Educational Research Quarterly*, *9*(3), 43–50.

Educational Research Service. (1988). *Teacher evaluation: Practices and procedures*. Arlington, VA: Author.

English, D., Burniske, J., Meibaum, D., & Lachlan-Hache, L. (2015). *Uncommon measures: Student surveys and their use in measuring teaching effectiveness*. Washington, DC: American Institutes for Research.

Every Student Succeeds Act of 2015. Pub. L. No. 114-95, Stat. 1177. (2015).

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Buttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–29. doi:10.1016/j.learninstruc.2013.07.001

Ferguson, R. F. (2008). *The Tripod project framework*. Cambridge, MA: Harvard University.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28. doi:10.1177/003172171209400306

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 98–143). San Francisco, CA: Jossey-Bass.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal*, *75*(3), 168–178.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal*, *25*(1), 57–78.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Heitin, L. (2012, July 11). Next up in teacher evaluations: Student surveys [Web log post]. *Education Week*. Retrieved from http://blogs.edweek.org/teachers/teaching_now/2012/07/next_up_in_teacher_evaluations_student_surveys.html

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching*, *49*(1), 26–31. doi:10.1080/87567550109595842

Jackson, C. K. (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (Working Paper No. 18624). Cambridge, MA: National Bureau of Economic Research. doi:10.1094/PDIS-11-11-0999-PDN

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.

Kauchak, D., Peterson, K. D., & Driscoll, A. (1985). An interview study of teachers' attitudes toward teacher evaluation practices. *Journal of Research and Development in Education*, *19*(1), 32–37.

LaFee, S. (2014). Students evaluating teachers. *School Administrator*, *3*(71), 17–25.

Loup, K., Garland, J., Ellett, C., & Rugutt, J. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education*, *10*(3), 203–226. doi:10.1007/bb00124986

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253–388. doi:10.1016/0883-0355(87)90001-2

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*(2), 285–296. doi:10.1037//0022-0663.83.2.285

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). New York: Springer.

Merrigan, G., & Huston, C. L. (2008). *Communication research methods*. New York, NY: Oxford University Press.

National Commission on Excellence in Education. 1983. *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office. Retrieved from http://www2.ed.gov/pubs/NatAtRisk/index.html

The New Teacher Project. (2011). *Rating a teacher observation tool: Five ways to ensure classroom observations are focused and rigorous*. Brooklyn, NY: Author.

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

Nott, R. (2014, May 8). Perception surveys give students say on teachers. *The New Mexican*. Retrieved from http://www.santafenewmexican.com/news/education/perception-surveys-give-students-say-on-teachers/article_6922d76b-9665-539f-859b-ba7b0a668db5.html

O'Donnell, P. (2014, May 11). Ohio students soon could be grading their own teachers. *The Plain Dealer*. Retrieved from http://www.cleveland.com/metro/index.ssf/2014/05/ohio_students_could_soon_be_gr.html

Ohio Department of Education. (2017). Student survey instrument for teacher evaluation service provider publicly-available service summary. Retrieved from https://education.ohio.gov/getattachment/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System/Alternative-Components/Tripod-Student-Survey-Form-C-2017-2018.pdf.aspx

Peterson, K. D., Stevens, D., & Ponzio, R. (1997). Variable data sources in teacher evaluation. *Journal of Research and Development in Education*, *31*(3), 123–132.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Study surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 135–153. doi:10.1023/A:1008102519702

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education*, *40*(1), 61–86. doi:10.1023/A:1018774311468

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183–212. doi:10.1086/679390

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teacher quality. *Educational Evaluation and Policy Analysis*, *36*(4), 399–416. doi:10.3102/0162373714531851

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th ed. ed.). Boston, MA: Pearson.

Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 170–202). San Francisco, CA: Jossey Bass.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571. doi:10.1162/edfp.2009.4.4.537

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214. doi:10.1162/qjec.2010.125.1.175

Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, *38*(5), 575–592. doi:10.1023/A:1024996413417

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142–171. doi:10.3102/1076998611432174

Schulz, J., Sud, G., & Crowe, B. (2014). *Lessons from the field: The role of student surveys in teacher evaluation and development*. Sudbury, MA: Bellweather Education Partners.

Scott, J. (1997). Children as respondents: Methods for improving data quality. In L. E. Lyberg, P. P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 331–350). Hoboken, NJ: John Wiley & Sons.

Sloat, E., Amrein-Beardsley, A., & Sabo, K. E. (2017). Examining the factor structure underlying the TAP System for Teacher and Student Advancement. *AERA Open, 3*(4), 1–18. doi:10.1177/2332858417735526.

Stevens, T., Harris, G., Liu, X., & Aguirre-Munoz, Z. (2013). Students' ratings of teacher practices. *International Journal of Mathematical Education in Science and Technology*, *44*(7), 984–995. doi:10.1080/0020739X.2013.823250

U.S. Department of Education. 2009. *Race to the Top Program executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Vaillancourt, P. M. (1973). Stability of children's survey responses. *The Public Opinion Quarterly*, *37*(3), 373–387. doi:10.1086/268099

Voight, A., & Hanson, T. (2012). *Summary of existing school climate instruments for middle school.* San Francisco, CA: REL West at WestEd.

Wagner, W., Gollner, R., Helmke, A., Trautwein, U., & Ludtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*(x), 1–11. doi:10.1016/j.learninstruc.2013.03.003

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, *53*(6), 1834–1868. doi:10.3102/0002831216671864

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

Whitehurst, G. J. R., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts.* Washington, DC: Brown Center on Education Policy at Brookings.

Will, M. (2016, December 30). Assessing quality of teaching staff still complex despite ESSA's leeway. *Education Week*, *36*(16), 31–32.

Yeh, S. S. (2013). A re-analysis of the effects of teacher replacement using value-added modeling. *Teachers College Record*, *115*(12).



***Cogent Education*** **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**