AERA Open
October-December 2017, Vol. 3, No. 4, pp. 1–18
DOI: 10.1177/2332858417735526

Examining the Factor Structure Underlying the TAP System for Teacher and Student Advancement

Edward Sloat
Audrey Amrein-Beardsley
Mary Lou Fulton Teachers College
Kent E. Sabo
Clark County School District

In this study, we investigated the factor structure underlying the TAP System for Teacher and Student Advancement using confirmatory and exploratory factor–analytic methods and under conditions of multilevel (nested) data structures and ordinal measurement scales. We found evidence of generally poor fit with the system's posited first-order, three-factor structure with relatively large correlations among measured dimensions. Exploratory analysis suggests one to two interpretable factors, one of which accounts for the majority of explained variance (i.e., a general or common underlying factor). Higher-order modeling confirms the presence of a bifactor structure composed of a single general trait supported by one or two subscales. We use this evidence to question the validity of the inferences drawn from TAP subscale scores. We accordingly discuss implications for low- and high-stakes applications of TAP output, especially when consequential decisions are attached to subscale-level estimates (i.e., teacher compensation based on latent performance as rated through weighted subscales).

Keywords: *accountability, educational reform, evaluation, teacher assessment, performance assessment, factor analysis, multilevel modeling, categorical data estimation*

Over the past four decades, U.S. educational policy makers have enacted multiple legislative reform initiatives in support of student- and teacher-level accountability. This movement began during the minimum competency era in the 1970s (Bracey, 1995), and it picked up pace in 1983 after the release of *A Nation at Risk* (U.S. Department of Education, 1983). Although it appeared to reach its peak after the passage of No Child Left Behind (2002), subsequent federal legislative acts—such as Race to the Top (2011), the No Child Left Behind waivers awarded to states that adopted stronger teacher accountability systems (Duncan, 2009, 2011), and the Teacher Incentive Fund grant competition (U.S. Department of Education, 2012)—have helped to continue the push for stronger accountability in support of educational reform.

During this period, the evolution of policy-based school and teacher accountability reforms has involved two important transitions: first, the transformation of teacher observation systems from a personally reflective mentoring and capacity building activity to that of a metric-driven assessment process; second, the addition of and emphasis on growth in student academic performance (i.e., value-added or growth) as a core measure of instructional quality.

To date, a great deal has been consequently published on the various technical aspects of value-added models, student growth percentiles, and other measures of academic progress (Amrein-Beardsley, 2014; Au, 2010; Betebenner, 2011; Blank, 2010; Chetty, Friedman, & Rockoff, 2014a, 2014b; Hanushek, 2011; Hanushek & Raymond, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Notwithstanding, there remains much controversy over the appropriateness of these test-base metrics as valid representations of teachers' instructional competencies and effects, especially when consequential decisions (e.g., teacher merit pay, tenure, termination) are to be attached to such measures (Baker et al., 2010; Berliner, 2005; Cohen & Goldberger, 2016; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Papay, 2011).

In contrast, technical review of metric-based teacher observation systems has received far less scrutiny. Traditionally, school practitioners have enjoyed widespread use of observation-based evaluation systems to examine teachers' instructional practice (Blank, 2010; Cohen & Goldberger, 2016; T. J. Kane, Kerr, & Pianta, 2014; Steinberg & Garrett, 2016). However, the intent has historically been formative, providing teachers with targeted feedback to improve pedagogical competency (Cohen & Goldberger, 2016; Danielson, 2010, 2011; Steinberg & Garrett, 2016). Not until recently has this focus evolved in response to the same high-stakes policy-based accountability reforms.

Accordingly, these reforms substantively incentivized states to require the use of quantitative metrics based on standardized observational frameworks to evaluate teachers (U.S. Department of Education, 2015). Some of the most widely used frameworks include Charlotte Danielson's framework for teaching (Danielson Group, n.d.), the Classroom Assessment Scoring System (Teachstone, n.d.), Robert Marzano's (n.d.) causal teacher evaluation model, California's Performance Assessment for California Teachers (n.d.), and, of interest in this study, the National Institute for Excellence in Teaching's (NIET's) TAP System for Teacher and Student Advancement (formerly known as the Teacher Advancement Program and hereafter referred to as the TAP System; see NIET, n.d.-a, n.d.-b, n.d.-c, n.d.-d, n.d.-e).

Notably, such measurement systems, especially when they are used for consequential decision-making purposes, require close examination of the psychometric properties that support their inferential warrant (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). While the application of growth and value-added models in evaluation frameworks continue to be rigorously vetted in the published literature (see also AERA, 2015; American Statistical Association, 2014; Baker et al., 2010; Harris & Herrington, 2015), observation-based evaluation systems have received much less attention (see, e.g., Amrein-Beardsley, Holloway-Libell, Montana Cirell, Hays, & Chapman, 2015; Goldring et al., 2015; Lash, Tran & Huang, 2016; Polikoff & Porter, 2014; Weisberg, Sexton, Mulhern, & Keeling, 2009).

Indeed, most published validation studies of teacher evaluation frameworks have focused on criterion approaches contrasting summated evaluator ratings with student achievement (e.g., value added or growth) outcomes (Cohen & Goldhaber, 2016; T. J. Kane et al., 2014; Kimball & Milanowski, 2009; Martínez, Schweig, & Goldschmidt, 2016; Milanowski, 2004, 2011; Milanowski & Kimball, 2005). Yet little attention has been paid to the observational measurement instruments themselves. By default, these types of criterion studies implicitly assume that the validity of the observation systems has been established.

For example, in a chapter titled "How the Framework for Teaching and Tripod 7 Cs Evidence Distinguish Key Components of Effective Teaching," Danielson writes that her system (framework for teaching) is "research-based and [has] been refined over more than a decade based on analysis of prior results and feedback from elementary and secondary practitioners" (Ferguson & Danielson, 2014, p. 99). However, no reference is made to the technical and psychometric characteristics of the system that might serve as evidence of such a claim. Hence, just because it might be "research based" does not mean that its technical and psychometric properties are "research evidenced" or their system uses "research warranted."

Accordingly, we argue that a similar analytic void exists for many, if not most, well-known instructional observation systems currently utilized within policy-prescribed consequential accountability systems (Bill & Melinda Gates Foundation, 2013; T. J. Kane & Staiger, 2012; National Council on Teacher Quality, 2015). We also suggest that use of such systems in high-stakes environments, without supporting scale validation evidence, conflicts with the measurement principles outlined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). This type of research evidence is needed to warrant the use of such systems, both in practice and as the basis of such accountability policies.

We also note that the existing literature base concerning observational rating systems fails to explicitly address substantive methodological and estimation issues characteristic of education-based K–12 observational rating data (Amrein-Beardsley et al., 2015; Cohen & Goldhaber, 2016; Lash et al., 2016). Specifically, in multischool agencies, evaluation data become hierarchically structured, with teachers nested into schools, grades, departments, and so on. When local administrators (i.e., principals and assistant principals) serve as primary evaluators, ratings become interdependent. The nested nature of these data present substantive issues for empirical analysis, affecting results and associated policy inferences (Heck & Thomas, 2015; Luke, 2004; Muthén, 1991, 1994; Raudenbush & Bryk, 2002).

Purpose of the Study

For these reasons, we focus on examining the factor structure posited by one of the most widely used observational evaluation frameworks: NIET's TAP System (see http://www.niet.org; see also, Barnett, Rinthapol, & Hudgens, 2014). Specifically, we investigate whether the TAP System's factor structure is supported by empirical measures of teacher instructional practice, as oft utilized in consequential evaluation settings. Importantly, a search of the literature reveals a dearth of published research regarding this system's measurement attributes. This is worrisome given the extent to which actions are often attached to TAP outcomes, especially subscale-level estimates (i.e., teacher compensation based on performance rated with weighted subscale scores; see NIET, n.d.-d).

However, in our analysis, we do not attempt to reinterpret, redefine, or reformulate the TAP System's structural framework. Rather, we examine the degree to which the current framework is found tenable under an applied empirical context. To our knowledge, this type of foundational analysis has yet to be published internally (e.g., available technical reports) or externally (e.g., peer-reviewed literature). Thus, we believe that this study serves as a critical starting point from which to document some of the technical characteristics of this system and to provide an empirical foundation from

which to further examine, modify, or improve the framework as a consequential teacher assessment tool.

Accordingly, we examine whether the TAP's posited latent factor structure is supported by empirical evidence. We focus on the discriminant validity of the system's subconstructs and the implications that our findings have on the application of the TAP metrics within consequence-based evaluation policies and practices.

TAP System

The Milken Family Foundation developed the TAP System to provide states, districts, and schools "a comprehensive educator effectiveness model" for teacher evaluation purposes (NIET, n.d.-b). The system has two primary purposes: as a summative measure of teacher performance and as a formative tool to help improve teachers' skills through individualized and concentrated professional support (Barnett et al., 2014; Culbertson, 2012; Daly & Kim, 2010). Specifically, the TAP System is intended to provide benchmark and progress measures necessary for teachers to adapt and improve their instructional practice. In addition, yearly improvements in performance are to be aligned with monetary bonuses. Users (e.g., school districts) determine these bonuses based on a weighted combination of teachers' value-added (student growth) measures and classroom observation scores derived from the weighted subscales built into the TAP System (see NIET, n.d.-d).

NIET is the nonprofit organization that oversees and promotes the TAP System, upholding it as a comprehensive model that provides "powerful opportunities for career advancement, professional growth, instructionally focused accountability and competitive compensation for educators" (NIET, n.d.-b). This is facilitated by providing "district and school leadership teams with real-time data to inform decisions," as well as a set of accompanying "best practices" (NIET, n.d.-a). Likewise, the NIET organization states that its initiatives are "impacting over 200,000 educators and 2.5 million students," with "over 90 percent of participating TAP schools [serving] high-need and diverse areas," most notably in Arizona, Arkansas, Indiana, Iowa, Louisiana, Minnesota, Tennessee, and Texas (NIET, n.d.-c). In addition, teacher education programs are increasingly adopting TAP for similar reform purposes (e.g., evaluating student teachers and holding them accountable; see, e.g., Strauss, 2015; Toth, 2015).

However, most published research on the TAP System has focused on whether TAP use increases student achievement (Glazerman & Seifullah, 2012; Mann, Leutscher, & Reardon, 2013; Springer, Ballou, & Peng, 2014), to what extent TAP scores correlate with growth or value-added measures of student achievement (Loeb & Candelaria, 2012; Sartain, Stoelinga, & Brown, 2011), and whether TAP use improves teachers' instructional quality (Armstrong, 2011; Eckert, 2010; Jerald & Van Hook, 2011; Mann et al., 2013). These studies report mixed results regarding TAP's impact on academic outcomes, TAP's low to moderate correlations with growth or value-added measures, and TAP's impact on professional practice within the context of additional professional development and training. Again, these literatures are lacking substantive examinations of the underlying psychometric characteristics of the TAP framework, including verification of the posited latent constructs used to evaluate teacher instructional competency in consequential ways.

Instrument Specifics

In terms of the actual instrument, the TAP rubric is composed of 19 performance indicators situated within three subscales (performance indicators per subscale are noted in parentheses): Instruction (n = 12 indicators), Designing and Planning Instruction (n = 3 indicators), and the Learning Environment (n = 4 indicators). A breakdown of the performance components aligned within each subscale is provided in Table 1.

During the school year, teachers are evaluated by certified evaluators on at least three occasions. Certified evaluators include mentor teachers, master teachers, and school administrators, each of which is local to the teacher's campus. All evaluators are precertified under the TAP protocols based on their rating consistency as compared with national scoring standards. TAP evaluators receive training in application and interpretation of the scoring rubrics with certification based on one's ability to score videotaped anchor lessons "within one point on each indicator and within no more than two points from the national rating on three indicators" (Daly & Kim, 2010, p. 12). TAP certification lasts 1 year, after which an observer must demonstrate interrater consistency per TAP's certification standards to be recertified.

The three classroom observations are independent and occur at different times throughout the school year by different evaluators; hence, multiple rater scores for any single classroom observation are not available (see also McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2015). Following each observation/evaluation, a postconference session is convened between the teacher and the observer to review each teacher's evaluation scores and to identify and discuss instructional strengths and weaknesses. The intent is for teachers to use this information to focus on and improve specific attributes of their professional practice. This process also aligns with the TAP System's intent to provide formative and informative feedback to increase instructional capacity.

Under the TAP System, some observations are unannounced, while others are scheduled to provide teachers with opportunities to demonstrate proficiency on the assessed performance indicators. During an observation session, rating scores are assigned to each of the 19 TAP performance indicators (see Table 1). For each performance

TABLE 1
TAP System Subscales and Components

Instruction (n = 12)	Designing and Planning Instruction (n = 3)	Learning Environment (n = 4)
I1: Standards and Objectives	D1: Instructional Plans	L1: Expectations
I2: Motivating Students	D2: Student Work	L2: Managing Student Behavior
I3: Presenting Instructional Content	D3: Assessment	L3: Environment
I4: Lesson Structure and Pacing		L4: Respectful Culture
I5: Activities and Materials		
I6: Questioning		
I7: Academic Feedback		
I8: Grouping Students		
I9: Teacher Content Knowledge		
I10: Teacher Knowledge of Students		
I11: Thinking		
I12: Problem Solving		

indicator, observers rate teachers on a scale of 1 to 5 (ordinal), with 1 representing unsatisfactory performance, 3 proficiency, and 5 exemplary. In their TAP qualification training, observers are instructed to use ratings of 5 only for teachers who demonstrate "true excellence above and beyond what is expected of a proficient teacher on a certain standard" (Daly & Kim, 2010, p. 11).

At the close of the school year, a teacher's final (i.e., summative) observation score is constructed as a weighted composite of ratings from a mentor teacher, master teacher, and school administrator (see NIET, n.d.-d). Finally, the weighted average observation scores are combined with student academic (i.e., value added or growth) performance measures to derive an overall performance rating for each teacher.

Importantly, the computational methods for aggregating and combining classroom observation scores with student academic growth measures are policy-derived decisions developed by the NIET to compute global performance metrics of instructional quality. Again, these policy-based computations assume that the underlying observational metrics are sound and align with the posited measurement framework.

Methods

In this regard, we focus our study on examining the foundational latent structure of the TAP System's observational instrument. To do so, we utilize a single set of unweighted observational ratings to anchor the analysis to our primary research question: to investigate whether the TAP System's posited factor structure is supported by empirical evidence. Specifically, we focus on examining the factor structure of the second set of unweighted midyear classroom observation scores and its coherence with the stated TAP System framework. The rational for this decision is that assessing structural characteristics based on weighted composite ratings inserts unnecessary policy-imposed noise into the information; that is, this study is not intended as an analysis of the policy environment but, rather, the measurement instrument's psychometric and technical properties.

In addition, the longitudinal progression of the observation schedules suggests that the first formative evaluation scores (early in the school year) may reflect construct-irrelevant variance due to teachers' relative lack of familiarity with the evaluation's component framework and process.[1] This becomes less problematic after the first postconference review, in which all participating teachers have become familiar with the system's process, procedures, and goals. Similarly, the third summative observation ratings (end of the school year) may reflect less variance due to sustained focus on improving component performance from prior evaluation results. Thus, in our opinion, the second observational rating provided the best opportunity to assess the foundational characteristics of the latent factor structure of the posited TAP framework by reducing exposure, training, and policy artifacts.

Study Sample

For this study, we examined teacher observation data collected from a set of 14 school districts in one state. These districts represented a total of 54 schools, including 39 elementary schools (72%), nine middle schools (17%), and six high schools (11%) enrolling a combined 34,055 K–12 students (>3% of the state's total K–12 school enrollment).

TAP observational rating information were available for 1,497 classroom teachers. Almost three quarters (72%) of the teachers were elementary school teachers (n = 1,078), while the remaining 21% were middle school teachers (n = 314) and 7% high school teachers (n = 105). At the time of data collection, 43 schools (80%) had implemented the TAP System for 1 year.

The racial/ethnic makeup of the student population taught by TAP teachers in the sample (approximately) was 18% White, 55% Hispanic, and 27% other, as compared with 42%, 43%, and 15%, respectively, at the state level. We conducted a chi-square test of independence comparing the racial/ethnic compositions of the sample versus the state overall. Results were significant ($p < .05$; $\chi^2 = 8{,}866.433$, $df = 2$, $p < .001$), albeit with a small effects size ($V = .89$; Cohen, 1988). This is likely due to the NIET's focus on serving teachers and students from lower-income communities/schools.

Related, because school participation in the TAP System was voluntary, as based on a majority vote of the school's instructional staff, it should be noted that schools included in this sample likely differ from those not participating in the TAP System. Hence, the sample in this study does not necessarily represent other schools within the district or across the state. This also implies that participants are generally accepting and engaged in the TAP System's evaluation processes and procedures.

Procedures

To investigate whether the TAP System's posited factor structure is supported by empirical evidence, we first applied confirmatory factor analysis (CFA) to evaluate the tenability of the proposition. We followed this with an exploratory factor analysis (EFA) to more explicitly examine attributes of the latent structures inherent in the empirical data. Given the findings, we estimated higher-order CFA models incorporating a general (i.e., common) factor dimension. Each approach provides useful information for understanding the alignment between the hypothesized TAP framework and attributes of the measures from which inferential judgments are derived. We estimated all models using Mplus 7.4 with Multilevel Add-On (released November 2015).

As mentioned, the original data set included evaluation scores for a total of 1,497 teachers distributed among 14 school districts and 54 elementary, middle, and high school campuses. However, the fidelity to the TAP evaluation protocol requires that every teacher be evaluated at least three times per year by a campus-assigned mentor teacher, master teacher, and school administrator. Information for a total of 1,313 teachers satisfied this observational criterion. In addition, estimating two-level factor-analytic models requires a grouping (cluster) variable. Review of the sample data revealed that 232 teacher records contained no school building identification or registered incorrect identification numbers necessary to match against position assignments. This reduced the usable record count to 1,081 teachers (72% of the original data set). Finally, TAP program participation is determined on a site-by-site basis, not at the district level. In each case, the majority of teachers must agree to participate. For this reason, we initially modeled the multilevel structure as teachers nested within schools.

Because the observation rating information nests teachers within schools, we estimated multilevel CFA models to account for the lack of error independence (Bryne, 2012; Heck & Thomas, 2015; Muthén, 1991, 1994; Raudenbush & Bryk, 2002). We recognized that estimating single-level models in the presence of nested data may generate underestimated variances and standard errors and lead to improper inferences based on biased parameter estimates and associated test statistics. To control for this, we used multilevel approaches to partition total variance into components at the individual level (within school) and the group level (between school). Doing so permits modeling group-level latent structures independent of the individual level to obtain unbiased estimates. Accordingly, we used a two-level modeling approach specifying teachers nested within schools. We identified a total of 38 school locations (clusters) with an average cluster size of 28.50 teachers per campus (SD = 7.56; min = 12, max = 41).

In addition to the nested nature of the data set, we recognized that the observation ratings recorded under the TAP System were based on a five-option (ordinal) Likert-type scale. This lack of a continuous measurement scale poses substantive estimation issues in latent variable modeling (Brown, 2015; Byrne, 2012; Heck & Thomas, 2015; Muthén & Muthén, 2008–2012), including attenuated indicator correlations, possible emergence of "pseudofactors," and biased standard errors (Brown, 2015). Thus, ignoring the existence of noncontinuous measures may lead to incorrect inferences based on model output. To mediate the impact of ordinal measures, we invoked the Mplus WLSMV estimator (weighted least squares with mean- and variance-adjusted chi-square test) for all factor-analytic models (Muthén & Muthén, 2008-2012). Here, Brown (2015) noted that "WLSMV procedures produce accurate test statistics, parameter estimates, and standard errors of CFA models under a variety of conditions," including conditions of "nonnormality and model complexity" (p. 355).

Table 2 provides summary distribution statistics for the TAP evaluation components. The information is organized by the three TAP behavioral domains and their measured components: Designing and Planning Instruction (D1–D3), Learning Environment (L1–L4), and Instruction (I1–I12).

As shown, the data contain no missing values, and all variables represent the full range of possible values (1 to 5). Skew and kurtosis statistics suggest generally well-behaved "close-normal" variability. The median for all measured variables is 3, while the means range from 2.86 (I11–Thinking and I12–Problem Solving) to 3.67 (L3–Environment and L4–Respectful Culture). While the five-item Likert-type scale displays relatively normal distributions, we decided to conservatively treat the measured data as categorical for model estimation, especially since Mplus provides a robust estimator (WLSMV) sensitive to distributional assumptions. The TAP-component polychoric correlation matrix generated

TABLE 2
TAP System Evaluation Components Descriptive Statistics

	M	*Mdn*	*SD*	Variance	Skew	*SE*	Kurtosis	*SE*
Design and Planning Instruction								
D1: Instructional Plans	3.24	3.00	.804	.647	.039	.068	.264	.135
D2: Student Work	3.06	3.00	.749	.562	.112	.068	.561	.135
D3: Assessment	2.91	3.00	.794	.630	.158	.068	.507	.135
Learning Environment								
L1: Expectations	3.32	3.00	.832	.692	.048	.068	−.172	.135
L2: Managing Student Behavior	3.53	4.00	.899	.808	−.179	.068	−.320	.135
L3: Environment	3.67	4.00	.823	.677	−.026	.068	−.527	.135
L4: Respectful Culture	3.67	4.00	.823	.677	−.122	.068	−.331	.135
Instruction								
I1: Standards and Objectives	3.18	3.00	.844	.712	.006	.068	−.158	.135
I2: Motivating Students	3.28	3.00	.821	.673	.117	.068	−.160	.135
I3: Presenting Instruct Content	3.24	3.00	.882	.779	.003	.068	−.303	.135
I4: Lesson Structure and Pacing	3.11	3.00	.868	.753	.100	.068	−.196	.135
I5: Activities and Materials	3.19	3.00	.814	.662	−.064	.068	.115	.135
I6: Questioning	3.00	3.00	.821	.675	.102	.068	.018	.135
I7: Academic Feedback	3.04	3.00	.824	.680	.099	.068	−.009	.135
I8: Grouping Students	3.03	3.00	.779	.607	.086	.068	.296	.135
I9: Teacher Content Knowledge	3.41	3.00	.811	.658	.031	.068	.002	.135
I10: Teacher Knowledge of Students	3.19	3.00	.787	.619	.049	.068	.051	.135
I11: Thinking	2.86	3.00	.829	.686	.390	.068	.074	.135
I12: Problem Solving	2.86	3.00	.918	.842	.573	.068	.016	.135

Note. For all components, valid $N = 1{,}311$; missing values = 0, minimum = 1, maximum = 5.

by Mplus is provided in Table 3. By default, Mplus computes polychoric correlations when ordered categorical data and WLSMV estimators are declared (Muthén & Muthén, 2008–2012).

The mean intercomponent correlation of the sample data is .612, ranging between a minimum of .476 to a maximum of .861. In addition, scale (Cronbach alpha) reliability indices report $\alpha = .960$ for all 19 TAP items, $\alpha = .848$ for Designing and Planning Instruction (D1–D3), $\alpha = .896$ for Learning Environment (L1–L4), and $\alpha = .938$ for Instruction (I1–I12). No items were flagged for removal based on potential negative impacts to scale reliabilities.

Acknowledging the nested nature of the data set, we computed intraclass correlation coefficients (ICCs) using Mplus 7.4 to assess the degree of variability in the measures attributable to the clustering of teachers within schools (i.e., variability in the teacher ratings explained by school assignment). In general, ICC values in excess of .05 (or 5%) warrant application of multilevel modeling methods (Brown, 2015; Byrne, 2012). For the sample data, we found the mean ICC to be above this threshold ($M = .122$; min = .066, max = .273). Indeed, 12 of the 19 indicators (63%) report ICC values >.10, suggesting that multilevel modeling approaches be utilized to obtain unbiased parameter estimates and model fit statistics.

For each of the CFA models, we relied on the following model fit statistics to guide our analysis: chi-square (χ^2), comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR; Brown, 2015; Bryne, 2012, Heck & Thomas, 2015; McCaffery et al., 2015).[2] We adopted general fit criterion thresholds as follows: chi-square, $p < .05$; CFI/TLI ≥ .95; SRMR ≤ .08; RMSEA ≤ .06 (Hu & Bentler, 1999). We directed Mplus to report modification indices (MIs) >3.84 to assist in identifying areas of problematic fit (Brown, 2015).

When generating EFA models, we again recognized the categorical nature of the measured variables (allowing Mplus to invoke the WLSMV estimator) and the nested structure of the data set. For the latter attribute, we estimated two-level EFA models specifying ordered extraction of one to four latent factors at the within-school level (individual) while leaving the between-school level (group) unrestricted. For all EFA rotations, we utilized the Oblimin (oblique) procedure. We based our warranted factor extractions for the EFA models on review of scree plots, Kaiser criterion (eigenvalues >1.00), size of rotated factor loadings, and factor interpretability. As an extraction procedure, parallel analysis was not available in Mplus for categorically measured variables with the WLSMV estimator. In addition, this option is

TABLE 3
TAP System Sample Data Component Polychoric Correlation Matrix (Mplus)

	D1	D2	D3	L1	L2	L3	L4	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11
D2	.693																	
D3	.707	.753																
L1	.657	.693	.620															
L2	.528	.555	.498	.760														
L3	.583	.593	.520	.698	.744													
L4	.611	.604	.530	.738	.813	.846												
I1	.710	.651	.608	.674	.521	.573	.586											
I2	.606	.691	.556	.708	.638	.683	.670	.620										
I3	.673	.658	.624	.705	.568	.593	.611	.709	.668									
I4	.627	.622	.572	.721	.665	.592	.610	.644	.648	.715								
I5	.651	.713	.642	.705	.547	.608	.579	.646	.721	.694	.641							
I6	.596	.593	.568	.592	.520	.574	.515	.569	.625	.633	.563	.634						
I7	.557	.612	.551	.619	.527	.570	.564	.562	.633	.575	.546	.579	.636					
I8	.560	.584	.558	.653	.613	.585	.611	.593	.629	.575	.657	.603	.569	.558				
I9	.654	.681	.623	.678	.553	.547	.599	.669	.647	.740	.626	.67	.592	.58	.579			
I10	.623	.655	.571	.705	.589	.624	.655	.619	.691	.607	.599	.62	.546	.583	.661	.634		
I11	.591	.627	.546	.583	.518	.607	.559	.554	.614	.581	.558	.612	.609	.552	.534	.559	.546	
I12	.558	.580	.521	.525	.476	.569	.527	.536	.548	.561	.502	.555	.595	.531	.502	.512	.477	.861

Note. See Table 1 for indicator codes.

not available as an extraction method for two-level EFA models. For these reasons, we conducted a parallel analysis based on a single-level EFA model applying an maximum likelihood (ML) estimator.

Based on results obtained from the EFA analysis, inclusion and examination of a primary common factor seemed warranted. In this regard, we reformulated four additional CFA models to evaluate the appropriateness of second-order and bifactor solutions, including a single common-factor model.

Results

Confirmatory Factor Analysis

As discussed, the posited factor structure of the TAP observational framework identifies three subscales (see Table 1). It is assumed that these three domains covary to some degree. However, it is also presumed that each construct independently measures unique attributes of instructional practice sufficient to permit inferential analysis of the subscales and the aligned indicators. That is, the intent of the evaluation framework is to compare component and subscale scores for the purpose of identifying areas of instructional strength/weakness, to direct targeted interventions, to promote improvement in professional practice, and to assign weights for merit pay purposes. To do so requires that correlations among latent factors remain relatively low and display substantive and statistically significant component loadings that are interpretable.

To evaluate the tenability of the posited TAP factor structure, we performed three initial CFA analyses. Each specified a correlated three-factor model with no cross loadings between measured variables. We estimated CFA 1 as a standard single-level model (nonnested). We estimated CFA 2 using the Mplus TYPE = COMPLEX procedure, which adjusts model fit statistics and parameter estimates for error dependencies due to the nested structure of the data (Brown, 2015). Here, CLUSTER = SCHOOL identified the grouping level. We estimated CFA 3 by explicitly modeling identical within-school (individual) and between-school (group) three-factor frameworks. We used the Mplus TYPE = TWOLEVEL procedure for this model. For all models, we used a WLSMV estimator. To establish the latent measurement scales for CFA 1 and 2, each factor's first measured indicator served as a marker variable—the default setting in Mplus. For CFA 3, we established scaling by fixing the variances of each latent factor to 1 while leaving all variable loadings unrestricted.[3] Table 4 reports the fit statistics for each estimated CFA model.

Models CFA 1–3. The standard CFI/TLI model fit statistics for the three initial models (CFA 1–3) marginally fell within acceptable range (>.95). In contrast, the RMSEA index fell outside acceptable levels for CFA 1 (single-level model) and CFA 2 (nested model with unrestricted between-school-level structure) at .109 and .067, respectively. This improved for CFA 3 (the fully specified multilevel model), with the RMSEA dropping to .059 but nevertheless remaining on the

TABLE 4
CFA/EFA Model Fit Statistics

Model	Description	χ^2 (*df*)[a]	RMSEA			CFI	TLI
			Value	90% CI	CFit		
CFA 1	Three-factor, SL, CAT, WLSMV	2,465.60 (149)	.109	[.105, .113]	0.000	.962	.957
CFA 2	Three-factor, ML, CAT, COMPLEX, WLSMV	867.62 (149)	.067	[.063, .071]	0.000	.968	.964
CFA 3	Three-factor, ML, CAT, TWOLEVEL, WLSMV	1,437.58 (298)	.059	n/a	n/a	.959	.953
			SRMR: WI (.044), BT (.087)				
EFA 1	One-factor, ML, URB, CAT, TWOLEVEL, WLSMV	1,622.37 (152)	.095	[.090, .099]	0.000	.947	.881
EFA 2	Two-factor, ML, URB, CAT, TWOLEVEL, WLSMV	654.91 (134)	.060	[.055, .065]	0.000	.981	.952
EFA 3	Three-factor, ML, URB, CAT, TWOLEVEL, WLSMV	405.83 (117)	.048	[.043, .053]	0.756	.990	.970
EFA 1	Four-factor, ML, URB, CAT, TWOLEVEL, WLSMV	306.42 (101)	.043	[.038, .049]	0.975	.993	.975
CFA 4	Bifactor (3), ML, CAT, COMPLEX, WLSMV	327.20 (133)	.037	[.032, .042]	1.000	.991	.989
CFA 5	Second order, LM, CAT, COMPLEX, WLSMV	894.85 (150)	.068	[.065, .072]	0.000	.967	.963
CFA 6	Bifactor (2), ML, CAT COMPLX, WLSMV	752.99 (145)	.062	[.058, .067]	0.000	.973	.969
CFA 7	One-factor, SL, ML, CAT, COMPLEX, WLSMV	1,173.60 (152)	.079	[.075, .083]	0.000	.955	.950

Note. Criteria: chi-square, reject $p < .05$; RMSEA, reject >.60, 90% CI does not include .60, and CFit significance ($p < .05$) → $P(H_o$ RMSEA < .05) significant at $p < .05$ level; SRMR, reject >.08; CFI/TLI, reject <.95. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; 90% CI = 90% confidence interval; CFit = significance (*p* value) testing that RMSEA <.05; SL = single-level, nonnested model; CAT = declared categorical indicators; WLSMV = Mplus estimator—weighted least squares with mean and variance-adjusted χ^2 statistic; ML = multilevel; COMPLEX and TWOLEVEL = Mplus multilevel modeling procedures; n/a, not applicable; SRMR = standardized root mean square residual; WI = within; BT = between; EFA = exploratory factor analysis; CFA = confirmatory factor analysis; URB = unrestricted between-level structure.
[a]For each chi-square value, $p < .001$.

edge of the threshold criterion. The SRMR index for the full multilevel framework (CFA 3) indicated good fit at the individual level (within school) but poor fit at the group level (between school). This suggests that the factor structure at the cluster level may be different from that of the individual. Overall, this mix of information did not present a consistent interpretation of acceptable model fit.

Review of the individual factor loadings for CFA 1–3 revealed substantive and statistically significant ($p < .05$) values. Table 5 reports the standardized loadings for the three models. The factor loadings for CFA 1–3 (within school) are similar. In addition, the CFA 3 between-school-level loadings are generally larger than those found at the within-school level. It is not unusual to find different magnitudes within- and between-school-level structures are compared in that the parameter estimates are derived independently.

Across Models 1–3, the latent factor correlations were substantively large ($M = .867$, range = .774–.940), suggesting poor discriminant inference across the subscale scores; that is, scores in one domain are good predictors of scores in the other domains, making it unclear what specific trait is being captured by the data. Low discrimination might suggest the presence of a general common factor that could be modeled with alternative factor frameworks, such as second-order or bifactor models.

MIs also provide useful information for evaluating model fit. The size of an MI reflects the approximate amount that the model's chi-square value would decrease if the particular parameter was freely estimated. Generally, well-fitting models produce MIs that are small. MIs >3.84 are considered substantive (Brown, 2015), but our focus is on values that suggest a sizable effect. For poor or marginally fitting models, review of large MIs may suggest a need for structural adjustments. In addition, presence of numerous large MIs may indict the fidelity of the theoretical framework being tested. Importantly, adjusting model specifications should be

TABLE 5
CFA Model Factor-Indicator Standardized Loadings

			CFA 3		CFA 4: Bifactor 3[a]		CFA 5:	CFA 6: Bifactor 2[b]		CFA 7
Factor: Indicator	CFA 1	CFA 2	Within	Between	Specific	General	Second order	Specific	General	General
Design by										
D1	0.871*	0.857*	0.832*	0.980*	0.193*	0.794*	0.877*	0.205*	0.790*	0.797*
D2	0.895*	0.885*	0.869*	1.006*	0.240*	0.815*	0.874*	0.253*	0.812*	0.825*
D3	0.833*	0.809*	0.810*	0.789*	0.644*	0.734*	0.801*	0.631*	0.731*	0.765*
Learning by										
L1	0.930*	0.929*	0.929*	0.924*	0.213*	0.848*	0.929*	0.217*	0.848*	0.869*
L2	0.840*	0.830*	0.838*	0.900*	0.485*	0.720*	0.831*	0.492*	0.716*	0.780*
L3	0.880*	0.868*	0.857*	0.863*	0.456*	0.751*	0.869*	0.460*	0.749*	0.820*
L4	0.901*	0.881*	0.887*	0.968*	0.578*	0.752*	0.880*	0.584*	0.748*	0.833*
Instruct by										
I1	0.821*	0.796*	0.775*	0.905*	−0.053	0.802*	0.765*		0.797*	0.786*
I2	0.828*	0.835*	0.816*	0.911*	−0.005	0.839*	0.838*		0.837*	0.823*
I3	0.840*	0.830*	0.816*	0.898*	−0.059*	0.839*	0.834*		0.831*	0.819*
I4	0.810*	0.796*	0.782*	0.965*	−0.082*	0.805*	0.799*		0.796*	0.787*
I5	0.827*	0.814*	0.816*	0.907*	0.012	0.817*	0.817*		0.814*	0.799*
I6	0.782*	0.743*	0.732*	0.843*	0.145*	0.742*	0.746*		0.744*	0.736*
I7	0.766*	0.730*	0.717*	0.970*	0.062*	0.732*	0.733*		0.731*	0.723*
I8	0.780*	0.760*	0.753*	0.825*	−0.049	0.764*	0.762*		0.761*	0.751*
I9	0.812*	0.806*	0.812*	0.697*	−0.082*	0.812*	0.808*		0.807*	0.798*
I10	0.773*	0.790*	0.774*	0.924*	−0.092*	0.798*	0.793*		0.792*	0.780*
I11	0.857*	0.811*	0.807*	0.780*	0.545*	0.738*	0.813*		0.812*	0.806*
I12	0.824*	0.768*	0.770*	0.683*	0.640*	0.692*	0.770*		0.771*	0.762*
Common by										
Design							0.943*			
Leader							0.891*			
Instruct							0.974*			

Note. See Table 1 for indicator codes. CFA = confirmatory factor analysis.
[a]Bifactor 3 = general plus three subfactors
[b]Bifactor 2 = general plus two subfactors.
*Coefficients significant at the $p < .05$ level.

based on theoretical grounds and not simply driven by sample-based information.

Hence, we requested MIs for CFA 1 (single level) and 2 (multilevel, COMPLEX).[4] However, MIs are not available from Mplus when TYPE = TWOLEVEL procedures are utilized (CFA 3). For CFA 1 and 2, the magnitude of the L1 indicator distinguished itself (L1 is the first measured variable in the Learning Environment domain, concerning a teacher's ability to establish learning expectations). That is, the mean MI in the CFA 1 and 2 models excluding L1 was approximately 33 and 11, respectively. In contrast, the MIs for Design by L1 were 184 in CFA 1 and 88 in CFA 2, while those for Instruct by L1 were 213 (CFA 1) and 105 (CFA 2).

Given that the chi-square values for CFA 1 and CFA 2 were 2,466 and 868, respectively, permitting L1 to freely load onto these domains would improve the model's chi-square between 8% and 12%. However, making these adjustments is not supported by theory, since the L1 component was established to measure aspects of the Learning Environment and not attributes of Designing and Planning Instruction or Instruction. Finally, review of the remaining MIs (26 in CFA 1 and 21 in CFA 2) revealed that each potential adjustment would necessitate aligning measured components to latent dimensions that they were not intended to measure.

Exploratory Factor Analysis

We used EFA to further explore dimensions of the sample data in the context of the posited TAP structure. As with CFA, we conducted all EFA procedures using Mplus 7.4 with Multilevel Add-On. For the main EFA analysis, we invoked the WLSMV estimator, categorical measured variables, and a

two-level nested structure, and we directed Mplus to sequentially estimate one- to four-factor extractions at the within-school level (individual) while leaving the structure at the between-school level (group) unrestricted. Because the latent factors are permitted to correlate, we utilized Oblimin (oblique) rotation. Model fit reflections are based on chi-square, scree plot, Kaiser criterion (eigenvalues >1.00), factor loadings, and factor interpretability. As mentioned, Mplus does not provide parallel analysis for TYPE = TWOLEVEL or use of WLSMV estimators. For this reason, we estimated a single-level EFA model employing ML estimators to generate parallel extraction information.

Model fit statistics. Model fit statistics for a one- through four-factor extraction procedure are presented in Table 4 (EFA 1–4). For all models, chi-square indices suggest poor ability of the estimated factor loadings to reproduce the sample component correlations. However, the indices drop substantively when moving from a one- to two-factor extraction solution and then decline less when moving to three- and four-factor extraction models. In addition, RMSEA indices suggest that three- and four-factor extraction models are adequate. CFI/TLI figures support two-, three-, or four-factor solutions, and the SRMR within-school indices support all four factor options.

Factor identification. Based on Kaiser criterion (eigenvalues >1.00), review of eigenvalues for the within-school level of the EFA model reveal two factors with values >1.00 (F1 = 11.804, F2 = 1.112). Similarly, eigenvalues for the between-school level also suggest the presence of two factors (F1 = 14.511, F2 = 1.645). Scree plots for the within- and between-school level EFA information are presented in Figures 1 and 2, respectively. Each suggests the presence of one or two factors.

We also examined the Oblimin (oblique) rotated factor loadings (i.e., pattern matrix, within-school level) to provide insight on the dimensions present in the data. These loadings are displayed in Table 6. The data suggest that a two-factor solution may be most tenable in that the three- and four-factor extraction models do not provide sufficient interpretable loading patterns beyond two dimensions. We conducted the same analysis on the factor structure matrixes and observed similar loading patterns. Finally, the consistently high loadings reported for the one-factor model also suggest the presence of a common overarching dimension. The pattern of loadings suggests that one dimension may be defined by the Learning domain (L2–L4) and a second by a combination of all three Design components (D1–D3) and a mix of the Instruction indicators.

EFA factor correlations. As with the CFA analysis, an important tenet of the TAP System is that the three posited dimensions are each identifiable and sufficiently independent to afford meaningful inference on the subscale scores. Examining the EFA factor correlations helps examine this presumption. Importantly, only F1 and F2 revealed interpretable dimensions (albeit, the factor interpretations may differ among models). Review of the information indicated that the correlation between F1 and F2 is .67 under the two-factor model and .72 under the three- and four-factor EFA models. In addition, F3 reports similarly sized association with F1 (r = .73 for three-factor model, r = .77 for four-factor model) but substantively lower correlations with F2 and F4. F4 also reports generally lower correlations with the other factors (r = .42–.54).

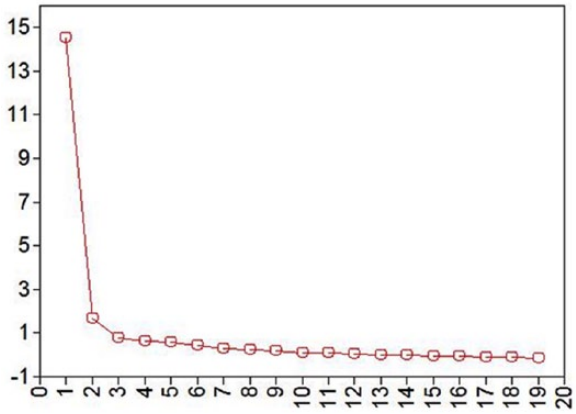

FIGURE 1. *Exploratory factor analysis: two-level scree plot—between school.*

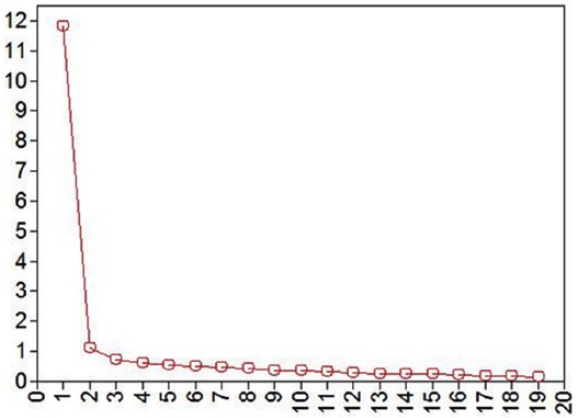

FIGURE 2. *Exploratory factor analysis: two-level scree plot—within school.*

Parallel analysis. Mplus does not provide parallel analysis for TYPE = TWOLEVEL EFA models or models declaring categorical data employing WLSMV estimators. However, the TAP rating information did not display substantive signs of nonnormality and were based on a five-item ordinal scale. Because of this, we felt it appropriate to estimate a standard

TABLE 6
EFA Rotated (Oblimin Pattern Matrix) Loadings for Factor Extraction Models (Within-School Level)

	One-factor model	Two-factor model		Three-factor model			Four-factor model			
Indicator	F1	F1	F2	F1	F2	F3	F1	F2	F3	F4
D1	0.777*	0.214*	0.625*	0.867*	−0.063	−0.024	0.851*	0.055	−0.054	−0.033
D2	0.807*	0.154	0.711*	0.799*	−0.060	0.101*	0.725*	0.038	0.123*	−0.011
D3	0.751*	0.092	0.709*	0.791*	−0.122*	0.103*	0.884*	−0.048	0.063	−0.128
L1	0.858*	0.566*	0.386*	0.561*	0.409*	−0.037	0.252	0.498	0.042	0.232
L2	0.784*	0.806*	0.068	0.134*	0.761*	0.006	−0.119	0.845	0.039	0.156
L3	0.804*	0.745*	0.153	0.071*	0.730*	0.155*	0.075	0.779*	0.089	−0.059
L4	0.835*	0.929*	−0.006	−0.020	0.889*	0.114*	0.051	1.011*	−0.015	−0.178
I1	0.767*	0.227*	0.603*	0.846*	−0.038	−0.037	0.637*	0.060	0.023	0.151
I2	0.805*	0.388*	0.500*	0.581*	0.239*	0.061	0.240	0.304	0.175*	0.267
I3	0.805*	0.246*	0.628*	0.857*	−0.006	−0.029	0.530*	0.076	0.082	0.272*
I4	0.774*	0.425*	0.431*	0.672*	0.226*	−0.087*	0.289	0.297	0.035	0.329*
I5	0.810*	0.197*	0.674*	0.755*	0.006	0.083	0.467*	0.062	0.196	0.229*
I6	0.723*	0.104	0.666*	0.571*	−0.005	0.220*	0.213	0.006	0.386*	0.304*
I7	0.707*	0.223*	0.546*	0.511*	0.112*	0.158*	0.192	0.154	0.281*	0.248*
I8	0.741*	0.403*	0.416*	0.485*	0.283*	0.048	0.184	0.339	0.150*	0.227
I9	0.804*	0.241*	0.630*	0.774*	0.038	0.019	0.480*	0.113	0.122	0.233*
I10	0.764*	0.420*	0.426*	0.591*	0.250*	−0.019	0.379	0.335	0.029	0.147
I11	0.799*	−0.215*	1.003*	0.119*	0.066*	0.773*	0.012	0.026	0.878*	−0.010
I12	0.763*	−0.270*	1.016*	0.015	0.054	0.874*	−0.023	0.000	0.958*	−0.080

Note. See Table 1 for indicator codes. Oblique (Oblimin) pattern matrix loadings represent unbounded regression coefficients. EFA = exploratory factor analysis.
*$p < .05$.

single-level EFA model based on maximum likelihood methods and generate a parallel analysis to further identify substantive latent factors in the data (Ledesma & Valero-Mora, 2007).

Figure 3 provides the scree plot of the model with the eigenvalues for the randomized (n = 1,000) parallel analysis overlaid. The information suggests that the TAP information contains a single latent factor. While a second factor might be suggested based on Kaiser criterion (eigenvalues >1.00) and other criteria, the parallel analysis suggests that this value is no larger than would be expected by random chance.

EFA calibration samples. To explore the stability of a possible two-factor solution, we randomly divided the original data (n = 1081) set into two subsamples (n_1 = 543 and n_2 = 538; Izquierdo, Olea, & Abad, 2014) and estimated a new set of EFA models on each subsample and compared the various model output with the full-data EFAs to assess parameter stability and extraction decisions (i.e., each specified in Mplus as TYPE = TWOLEVEL with categorical measures employing WLSMV estimators specifying sequential extractions from one to four at the within-school level with unrestricted covariances at the between-school level). For both subsamples, Kaiser criterion (eigenvalues >1.00) at the within-school level suggested presence of one primary dimension with a less impactful secondary factor. Chi-square model fit statistics remained significant but declined in magnitude similar to the full data set. The remaining model fit indices matched the full data set closely. Review of rotated factor loadings yielded similar perspectives and component clustering as found in the full model, suggesting that a two-factor solution was tenable.

The biggest difference occurred at the between-school level for Subsample 1. Here the two-factor extraction model failed to provide an interpretable second factor, unlike that found in Subsample 2 and the full data set. However, the three- and four-factor extraction frameworks for Subsample 1 retained the component clustering/ordering, supporting two interpretable dimensions. In addition, eigenvalues for the between-school-level Subsample 1 identified four possible dimensions instead of the two.

Finally, CFA models based on the calibration samples were not estimated, because the purpose of our analysis was not to respecify the TAP framework nor fully reinterpret factor meaning. Rather, the CFA and EFA analyses were intended to examine suitability of the posited factor structure

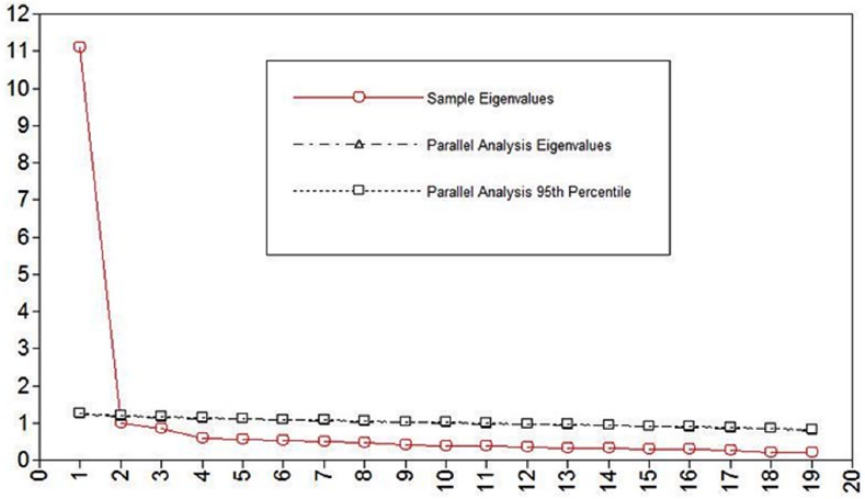

FIGURE 3. *Exploratory factor analysis plot of eigenvalues with overlay of parallel-analysis values. Parallel-analysis eigenvalues generated from 1,000 randomized samples.*

advanced by the TAP System as used in this study setting. That said, there is some indication from the EFA analysis that two dimensions represent a plausible within-school-level structure for the sampled data set.

Higher-order CFA. Review of the initial CFA (1–3) and EFA (1–4) models suggests the presence of a general (common) dimension influencing variation in the measured variables. This is apparent from the magnitude of the CFA factor correlations, the aggregated impact of cross-factor influences (as evidenced by MIs), the factor extraction analysis, and the interpretation of factor loadings. In addition, application of the TAP framework is intended to assess overall instructional competence, which then serves as the basis for establishing merit-based rewards (NIET, n.d.-d).

To this end, we specified three additional CFA modeling frameworks (CFA 4–6) that attempt to evaluate the contribution of a general factor. The first (CFA 4) is a bifactor model, followed by a second-order model (CFA 5). Each incorporates a general factor to account for variation in the measured variables (Chen, West, & Sousa, 2006). Finally, we specified a single general-factor model (CFA 6) to evaluate and compare model fit and factor loading information.

Bifactor models presume the existence of a general (common) factor that accounts for a substantive proportion of variance in all observed measures. However, bifactor specifications also assume that multiple domain-specific factors account for additional unique variance beyond that of the general factor. Here, bifactor models specify direct effects of the general factor on each measured variable independent of the variance accounted for by the domain-specific factors. Brown (2015) notes that bifactor models are most appropriate where a substantive unidimensional factor is posited with numerous but less substantive subdomains.

Second-order models also assume the presence of a general factor but specify its effect in a different manner. Here, the general factor is accounting for the covariance between the related latent factors and not variation in the observed variables. Second-order models presume that each latent factor is influenced by a higher dimension.

We estimated the general factor CFA models as multilevel frameworks using the TYPE = COMPLEX procedure in Mplus with declared categorical measures employing the WLSMV estimator. We left latent factor scaling at the default setting in Mplus except for the first measured variable (I1) in the Instruction domain under the bifactor framework. Here, we unrestricted the indictor and set the factor variance to 1.0 to establish the measurement scale. This adjustment enabled the model to estimate properly with no warnings or errors. For the second-order model, we fixed the Instruction loading parameter on the Common factor (Common by Instruct) at .90 to resolve estimation issues.[5] Table 4 reports the fit statistics for each of the higher-order factor (CFA 4 and CFA 5) and single general-factor (CFA 6) models.

Models CFA 4 and 5. For the bifactor (CFA 4) and second-order (CFA 5) models, the chi-square values remain significant, consistent with the previously estimated CFA models. Here, the chi-square for the second-ordered model (CFA 5; $\chi^2 = 895$, $df = 150$, $p < .001$) was slightly higher than the first-order model (CFA 4; $\chi^2 = 868$, $df = 149$, $p < .001$) previously discussed. The RMSEA suggests good fit for the bifactor specification (CFA 4; RMSEA = .037) and poor fit for the second-order framework (CFA 5; RMSEA = .068). CLI/TLI indices suggest good model fit for both specifications; however, the values for the second-order model are below those of the bifactor estimates.

Bifactor models permit a common (general) factor to account for the direct effects of all measured variables. In addition, subdomains are specified to account for direct effects on a subset of indicators. In this configuration, it is possible that one or more subfactors may be rendered irrelevant (Brown, 2015; Chen et al., 2006). That is, the subdomain factor loadings may become low or insignificant because most of the measured variance is accounted for by the general factor. Indeed, review of the factor loadings for the estimated bifactor model (CFA 4) suggests that this is occurring for the Instruction subdomain. Table 5 reports the standardized factor loadings for the bifactor (CFA 4) and second-order (CFA 5) models.

For the bifactor model (CFA 4), loadings for seven of 12 Instruction indicators reported negative magnitudes. In addition, the magnitude on 10 indicators was relatively low, and four were not significant. In contrast, all the remaining subdomain loadings (Design, Leader, and Common) were substantive and significant. This suggests that the components of the Instruction subdomain were better accounted for by the general (i.e., common) factor. Six components reported substantive cross loadings between the specific and general factor: one in Design (D3: Assessment), three in Learning Environment (L2: Managing Student Behavior, L3: Environment, L4: Respectful Culture), and two in Instruction (I11: Thinking, I12: Problem Solving). Finally, all parameter estimates reported for the second-order model were well behaved. That is, all the standardized domain loadings were substantive in magnitude and significant ($p < .05$).

Because the bifactor (CFA 4) model identified the Instruction subdomain as being captured by the general factor, we estimated an adjusted bifactor model (CFA 6) eliminating the indicators for this dimension. The model fit statistics (Table 4) provided mixed results: $\chi^2 = 752.99$, $df = 145$, $p < .001$; RMSEA = .062, 90% CI [.058, .067], CFit = 0.00 (for definition, see Table 4); CFI = .973; TLI = .969. The chi-square value more than doubled that reported by the unadjusted bifactor (CFA 4) framework; the RSMEA moved from being acceptable to poor; and the CFI and TLI indices both declined. However, all the standardized factor loadings were substantive and significant (see Table 5 to review the factor loadings for the adjusted bifactor model, CFA 6). As with CFA 4, lower loading values on the Designing and Planning Instruction and Learning Environment domains are noted, as are similar cross loadings when compared with the general factor. This suggests, again, that the common factor is accounting for most of the variation in the measured variables.

Finally, we estimated a one-factor CFA model (CFA 7) based on the premise that a single dimension might be explaining the variation in the empirical data. The model fit statistics (Table 5, CFA 7) for this simplified framework reported generally poor results: $\chi^2 = 1{,}173.60$, $df = 152$, $p < .001$; RMSEA = .079, 90% CI [.075, .083], CFit = 0.00; CLI = .955; TLI = .950. The standardized factor loadings (Table 5) were large (.723–.869) and significant ($p < .05$). No MIs >3.84 were reported.

Summary of Findings

The principle research question that we investigated in this study concerned the tenability of the TAP System's posited factor structure. To assess this question, we conducted multiple confirmatory and exploratory factor analyses to examine whether the proposed three-factor structure was supported by empirical data. Our additional interest focused on the discriminant validity afforded by the three TAP subconstructs (Designing and Planning Instruction, Instruction, and Learning Environment) and their ability to uniquely inform on targeted areas of professional practice. As our analysis evolved, we explored additional factor structures examining the presence of a single general dimension and its association with secondary factors. The data set originated from a large sample of elementary schools engaging in the TAP System.

Our findings suggest that the posited three-factor framework provides a poor to marginal fit with the empirical data. Exploratory examination of factor structures suggests that a two-factor solution may be more tenable. Subsequent bifactor and second-order CFA models seem to provide a better conceptualization of the TAP structure where a common (i.e., general) factor accounts for most of the variance in the measured variables. A bifactor structure where the general factor substantively aligns to all measured indicators performs best from a fit perspective, as compared with a second-order framework. Here, three of the four subcomponents of the Learning Environment domain seem to retain their posited meaning as a interpretable secondary (supporting) dimension. In addition, large factor correlations revealed throughout the modeling activities suggest that the discriminant validity among the posited latent constructs is low.

Put differently, results from bifactor specifications suggest the presence of a general latent construct supported by a subset of components from the Learning Environment domain. This is consistent with the exploratory factor information that does not distinguish Instruction as a fully independent dimension. The reason may be that the TAP framework itself is capturing predominantly instructional competency behaviors as the primary behavioral trait with components related to developing the Learning Environment in secondary (perhaps supporting) roles. Fittingly, this has implications for low- and high-stakes applications of TAP output.

Policy and Measurement Considerations

The TAP framework posits that instructional quality may be assessed across three distinct behavioral domains: Instruction, Learning Environment, and Designing and

Planning Instruction. However, questionable model fit suggests that this proposition may not be tenable. Large latent factor correlations further indict the suitability of interpreting instructional performance at the subscale level due to poor discriminant validity. That is, unique interpretation of specific subscales is brought into question when indicators from other dimensions are highly correlated, which seems to be the case here.

The combination of poor model fit and substantive latent factor correlations subsequently suggests that consequential actions (e.g., interventions and/or professional status decisions) based on subscale scores may not be warranted (or, at best, should be done with caution), since explicit identification of the measured behavioral constructs remains unclear. Put differently, if the validity of the inferences to be drawn from TAP subscale scores does not hold as posited, especially when consequential decisions are attached to subscale-level estimates (i.e., teacher compensation based on latent performance as rated with weighted subscales), this may be problematic in policy and practice.

Correspondingly, when the independence (discriminant inference) across subscales is not empirically established, it may be more prudent to utilize the unweighted summated score to distribute individuals along the primary trait for which the instrument was originally designed. Indeed, the empirical implication may be to devalue consequential decisions based on subscale scores and rely more heavily on full-scale summated measures to identify and distinguish relative distributions of instructional competency.

Indeed, while the TAP instruments are posited to reflect the most important tasks, skills, knowledge, and abilities relevant for witnessing high-quality instructional practices, Haladyna (2013) noted that "the most important feature of a test is the validity of its test score interpretation and use" (p. 4; see also M. T. Kane, 2013; Messick, 1989, 1998). This implies that claims of validity require a range of supporting evidences, most certainly including the empirical assessment of the instrument's measurement characteristics. As evidenced herein, again, some foundational examinations suggest that the measured constructs posited for the TAP instrument may not neatly align with the proposed constructs of interest. This raises questions concerning score interpretation and the proper use of scores for specific policy and pragmatic purposes.

Within a larger context, states and districts throughout the United States have increasingly adopted policy-driven educational reform initiatives based on student- and teacher-level accountability measures (e.g., standardized observational inventories such as the TAP System and growth- or value-added models). We believe that these policy initiatives have encouraged what might be considered to be hasty implementation of evaluation tools and instruments at the expense of affording sufficient attention to the technical aspects of the measures utilized. Again, while more attention has been paid to the technical attributes of using growth and/or value-added modeling to assess teachers' impact on student learning, little empirical attention has been paid to the observational systems meant to be used alongside and theoretically complement these measures (see, e.g., Bill & Melinda Gates Foundation, 2013; Chester, 2003; Chin & Goldhaber, 2015; Martínez et al., 2016; National Council on Teacher Quality, 2015; U.S. Department of Education, 2015).

We assert that when the foundational psychometric properties and validity evidences of any instrument remain unexamined, it is difficult for policy leaders to warrant their use within multiple policy settings (AERA et al., 2014; Haladyna & Rodriguez, 2013; Linn, 1993; Messick, 1989, 1998). This is especially true when interpretation of evaluation measures potentially affect the personal and professional identities of those being evaluated (M. T. Kane, 2006, 2013; Messick, 1998; Sheppard, 1993, 1997; Slomp, Corrigan, & Sugimoto, 2014).

In this regard, the ubiquitous application of standardized observational frameworks used to evaluate teachers elevates the need for empirical evidence supporting their technical design and application (Amrein-Beardsley et al., 2015; Goldring et al., 2015; Polikoff & Porter, 2014; Weisburg et al., 2009; Lash et al., 2016). Without such evidence, the validity threats surrounding misapplication of these measures may be both profuse and profound. Our contention, therefore, is that the greater the consequence, the greater the need for extensive and exhaustive validity evidence (AERA et al., 2014; Lane 2014; see also M. T. Kane, 2013; Messick, 1989, 1998).

The findings presented herein underscore the need to rigorously examine the psychometric characteristics of all observational teacher evaluation frameworks inclusive of how such systems are currently being implemented and utilized. In turn, these evidences should be used to shape the extent to which consequential educational evaluation and accountability policies might also be implemented.

Limitations and Opportunities for Further Research

In this study, we did not attempt to address all of the varied types of validity evidences possible for evaluating the psychometric attributes of the TAP System's evaluation instrument. Rather, we purposefully focused on an initial examination of the coherence between the framework's posited three-factor structure and empirical measures. In addition, most of our analytic attention was directed at within-school scale characteristics, with little effort made toward examining structures at the group level (between school). However, we believe that is an important extension of our analyses.

In addition, most of the sample participants reported evaluation scores for three independent time points (start of year,

midyear, and end of year). As discussed previously, we chose to focus on midyear evaluation ratings as a starting point for examining attributes of the TAP instrument. Replicating the analysis for these three periods and conducting temporal invariance testing would also represent an important extension to this study. In addition, within each of the three periods, teachers were evaluated by any one of three types of certified raters: mentor teachers, master teachers, and school administrators. Additional invariance studies across rater type and time should be conducted. Finally, parametric invariance test results might be examined between levels and across teacher characteristics (e.g., years of experience, grade level, instructed subject area), as this would likely be of added empirical value.

In this study, we also to utilize the full set of sample data to evaluate the alignment between the hypothesized factor structure and the empirical data. Here, we estimated numerous CFA models, followed by a variety of EFA procedures. An alternative way to utilize the sample data would have been to initially split (via stratified random sampling by school) the participants into two groups and conduct a variety of comparative analysis and invariance testing across subsamples.

Related, we initially focused on evaluating two-level nested models that grouped teachers within schools. This was initially decided per the volunteer basis by which schools choose to participate, as noted prior. Hence, extending the modeling to three levels (teachers within schools within districts) would add another perspective to the analysis and permit the analysis of comparative factor specifications at the school and district levels. In addition, because the sample schools represented a more diverse racial/ethnic student population than that found at the state level, this compromises more generalized inferences beyond the localized context, suggesting that replication with expanded samples is warranted.

While not strictly an analysis of the evaluation instrument itself, the TAP System imposes a number of policy restrictions regarding how evaluation scores are derived, aggregated, and combined to arrive at overall teacher performance. This includes weighted ratings based on rater type, methods for combining the three time-period scores, and procedures for deriving overall performance classifications based on integrating observational scores with value-added measures. Each of these reflects important policy dimensions inherent in the TAP evaluation system that should be examined within the context of program purpose and consequential outcomes.

Conclusions

Classroom observations serve as critical components of many federal and state educational reform initiatives because they appear to provide actionable formative and summative information to practitioners and policy makers. That is, it seems reasonable to expect that teachers use evaluation information in a formative manner to improve targeted areas of professional practice (Goldring et al., 2015) and that school and district leadership use evaluation information in a summative manner for policy-directive purposes (e.g., pay-for-performance incentives, retention/promotion, hiring, or other high-stakes policy decisions). Indeed, TAP System developers presume this type of causal pathway whereby formative and summative evaluation measures should lead to improved instructional competence, boost retention of highly effective teachers in high-needs schools, and ultimately incentivize and increase student academic performance over time (Jerald & Van Hook, 2011; see also NIET, n.d.-d).

However, fidelity to these types of outcomes requires pedagogically specific feedback aligned to component scores that uniquely assess discernable attributes of teachers' instructional practice. Importantly, results from this study suggest that reliance on subdimension scores to identify targeted practices, initiate interventions, and consequentially infer attributes of instructional competency may be suspect. At the same time, while the three-factor structure of the current TAP System framework may not be conclusively supported, this does not mean that summative scales constructed from the individual indicators (i.e., representing the general or common factor) does not capture essential elements of quality instructional practices. However, warrant for this claim requires evidence not currently available in the technical literature. We believe that this needs to be part of the evolving evidence that may help warrant the use of this and similar observational systems for low- and high-stakes uses and decision-making purposes.

Acknowledgments

This research was made possible by a U.S. Department of Education Teacher Incentive Fund grant (S385A100163).

Notes

1. In this study, approximately 80% of participating campuses were new to the TAP System. Thus, the first observation event (i.e., near the start of the school year) represented a teacher's first exposure to the evaluation.

2. In Mplus, SRMR is reported for TYPE = TWOLEVEL only when WLSMV estimators are used with declared categorical indicators.

3. We encountered estimation errors when we left the scaling markers at the default setting. Our resetting of the markers to the latent factors permitted full model estimation with no warnings or errors.

4. Complete tabulations of MIs for each CFA model are not provided due to space considerations.

5. Default scaling settings caused estimated factor correlations >1.0.

ORCID iD

A. Amrein-Beardsley https://orcid.org/0000-0001-6924-3025

References

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, *44*, 448–452. doi:10.3102/0013189X15618385

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

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. New York, NY: Routledge.

Amrein-Beardsley, A., Holloway-Libell, J., Montana Cirell, A., Hays, A., & Chapman, K. (2015). "Rational" observational systems of educational accountability and reform. *Practical Assessment, Research and Evaluation*, *20*(17). Retrieved from http://pareonline.net/getvn.asp?v=20&n=17

Armstrong, A. (2011). Rising above the floodwaters: TAP helps Louisiana school rebuild professional learning program. *Journal of Staff Development*, *32*(6), 46–51.

Au, W. (2010). Neither fair nor accurate: Research based reasons why high-stakes tests should not be used to evaluate teachers. *Rethinking Schools*, *25*(2), 34–38.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute. Retrieved from http://www.epi.org/publication/bp278/

Barnett, J. H., Rinthapol, N., & Hudgens, T. (2014). *TAP research summary: Examining the evidence and impact of TAP. The System for Teacher and Student Advancement*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED556331.pdf

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, *56*(3), 205–213.

Betebenner, D. W. (2011, April). *Student growth percentiles*. National Council on Measurement in Education Training Session presented at the Annual Conference of the American Educational Research Association, New Orleans, LA.

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable Measures of Effective Teaching: Culminating findings from the MET project's three-year study*. Retrieved from http://www.gatesfoundation.org/press-releases/Pages/MET-Announcment.aspx

Blank, R. F. (2010). *State growth models for school accountability: Progress on development and reporting measures of student growth*. Washington, DC: Council of Chief State School Officers.

Bracey, G. W. (1995). *Final exam: A study of the perpetual scrutiny of American education. Historical perspectives on assessment, standards, outcomes, and criticism of US public schools*. Bloomington, IN: TECHNOS Press.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Bryne, B. M. (2012). *Structural equation modeling with Mplus*. New York, NY: Routledge.

California's Performance Assessment for California Teachers. (n.d.). Retrieved from http://www.pacttpa.org/_main/hub.php?pageName=Home

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189–225. doi:10.1207/s15327906mbr4102_5

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. doi:10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679. doi:10.1257/aer.104.9.2633

Chin, M., & Goldhaber, D. (2015). Exploring explanations for the "weak" relationship between value added and observation-based measures of teacher performance. Cambridge, MA: Center for Education Policy Research. Retrieved from http://cepr.harvard.edu/files/cepr/files/sree2015_simulation_working_paper.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., & Goldberger, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387. doi:10.3102/0013189X16659442.

Culbertson, J. (2012). Putting the value in teacher evaluation. *Phi Delta Kappan*, *94*(3), 14–18. doi:10.1177/003172171209400304

Daly, G., & Kim, L. (2010). *A teacher evaluation system that works*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED533380.pdf

Danielson, C. (2010). Evaluations that help teachers learn. *Educational Leadership*, *68*(4), 35–39.

Danielson, C. (2011). *The framework for teaching evaluation instr*ument. Princeton, NJ: Danielson Group.

Danielson Group. (n.d.). *The framework*. Princeton, NJ: Danielson Group. Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15.

Duncan, A. (2009). *The race to the top begins: Remarks by Secretary Arne Duncan*. Washington, DC: U.S. Department of Education. Retrieved from http://www.ed.gov/news/speeches/2009/07/07242009.html

Duncan, A. (2011). *Winning the future with education: Responsibility, reform and results. Testimony given to the U.S.*

Congress. Retrieved from http://www.ed.gov/news/speeches/winning-future-education-responsibility-reform-and-results

Eckert, J. (2010). *Performance-based compensation: Design and implementation at six Teacher Incentive Fund sites*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.niet.org/assets/Publications/performance-based-compensation-tif.pdf?processed=1

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.

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years*. Washington, DC: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/our-publications-and-findings/publications/an-evaluation-of-the-chicago-teacher-advancement-program-chicago-tap-after-four-years

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value-added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96–104. doi:10.3102/0013189X15575031

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*, New York, NY: Routledge.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*, 466–479. doi:10.3386/w16606

Hanushek, E. A., & Raymond, M. (2005). Does school accountability lead to improved school performance? *Journal of Policy Analysis and Management*, *24*(2), 297–329. doi:10.3386/w10591

Harris, D. N., & Herrington, C. D. (2015). Editors' introduction: The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher*, *44*(2), 71–76. doi:10.3102/0013189X15576142

Heck, R. H., & Thomas, S. L. (2015). *A introduction to multilevel modeling techniques: MLM and SEM Approaches Using Mplus*. New York, NY: Routledge.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Izquierdo, I., Olea, J., & Abad, F. A. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema*, *26*(3), 395–400. doi:10.7334/psicothema2013.349

Jerald, C. D., & Van Hook, K. (2011). *More than measurement: The TAP System's lessons learned for designing better teacher evaluation systems*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://files.eric.ed.gov/fulltext/ED533382.pdf

Kane, M. T. (2006). Validation. In Robert L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–66). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. San Francisco, CA: Jossey-Bass.

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. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*(1), 34–70. doi:10.1177/0013161X08327549

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127–135. doi:10.7334/psicothema2013.258

Lash, A., Tran, L., & Huang, M. (2016). *Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system* (Report No. REL 2016-135). Washington DC: Regional Education Laboratory West.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, *12*(2), 1–11. Retrieved from http://pareonline.net/pdf/v12n2.pdf

Linn, R. L. (Ed.). (1993). *Educational measurement* (3rd ed.). Phoenix, AZ: Oryx Press.

Loeb, S., & Candelaria, C. A. (2012). *How stable are value-added estimates across years, subjects, and student groups?* Retrieved from the Carnegie Knowledge Network website: http://www.carnegieknowledgenetwork.org/briefs/value-added/value-added-stability/

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Mann, D., Leutscher, T., & Reardon, R. M. (2013). *Findings from a two-year examination of teacher engagement in TAP schools across Louisiana*. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved from http://www.niet.org/assets/PDFs/interactive-louisiana-student-achievement.pdf

Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance. *Educational Evaluation and Policy Analysis*, *38*(4), 738–756. doi:10.3102/0162373716666166

Marzano, R. (n.d.). *Dr. Robert Marzano's causal teacher evaluation model*. Blairsville, PA: Learning Sciences International. Retrieved from http://www.iobservation.com/Marzano-Suite/dr.-robert-marzanos-causal-teacher-evaluation-model/

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models of teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurements: Issues and Practice*, *34*(2), 34–46. doi:10.1111/emip.12061

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Phoenix, AZ: American Council on Education and Oryx Press.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(*1–3*), 35–44.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and students achievement evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33–35.

Milanowski, A. (2011, April). *Validity research on teacher evaluation systems based on the framework for teaching*. Paper presented at the Annual Meeting of the American Education Research Association, New Orleans, LA. Retrieved from http://files.eric.ed.gov/fulltext/ED520519.pdf

Milanowski, A., & Kimball, S. M. (2005, April). *The relationship between teacher experience and student achievement: A synthesis of three years of data*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*(2), 376–398.

Muthén, B. O., & Muthén, L. K. (2008–2012). *Mplus: Statistical analysis with latent variables, users guide*. Los Angeles, CA: Muthén & Muthén.

National Council on Teacher Quality. (2015). *State of the states 2015: Evaluating teaching, leading and learning*. Washington, DC. Retrieved from http://www.nctq.org/dmsView/StateofStates2015

National Institute for Excellence in Teaching. (n.d.-a). *Educator effectiveness*. Retrieved from http://www.niet.org/what-we-do/educator-effectiveness-tools/

National Institute for Excellence in Teaching. (n.d.-b). *Elements of success*. Retrieved from http://www.niet.org/tap-system/elements-of-success/

National Institute for Excellence in Teaching. (n.d.-c). *NIET impact overview*. Retrieved from http://www.niet.org/our-impact/niet-impact-overview/

National Institute for Excellence in Teaching. (n.d.-d). *TAP evaluation and compensation guide*. Retrieved from https://www.gpisd.org/cms/lib01/TX01001872/Centricity/Domain/6651/TEC handbook.pdf

National Institute for Excellence in Teaching. (n.d.-e). *TAP System CORE training*. Retrieved from http://www.tapsystemtraining.org/

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 20 U.S.C. § 6319 (2002).

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163–193. doi:10.3102/0002831210362589

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

Race to the Top Act of 2011, S. 844, 112th Congress (2011). Retrieved from http://www.govtrack.us/congress/bills/112/s844

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

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago*. Chicago, IL: Consortium on Chicago School Research. Retrieved from https://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19, pp. 405–450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24

Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating. *Research in the Teaching of English*, *48*(3), 276–302.

Springer, M. G., Ballou, D., & Peng, A. (2014). Estimated effect of the Teacher Advancement Program on student test score gains. *Education Finance and Policy*, *9*(2), 193–230. doi:10.1162/EDFP_a_00129

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317. doi:10.3102/0162373715616249

Strauss, V. (2015). Gates Foundation puts millions of dollars into new education focus: Teacher preparation. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2015/11/23/gates-foundation-put-millions-of-dollars-into-new-education-focus-teacher-preparation/

Teachstone. (n.d.). *Classroom Assessment Scoring System (CLASS)*. Charlottesville, VA: Teachstone Training. Retrieved from http://www.teachstone.org/about-the-class/

Toth, H. (2015). *College of education wins $7 million grant for teacher prep reform*. Lubbock, TX: Texas Tech University. Retrieved from http://today.ttu.edu/posts/2015/11/college-of-education-wins–gates-foundation-grant-for-teacher-prep-reform

U.S. Department of Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: US Government Printing Office.

U.S. Department of Education. (2012). *U.S. Department of Education boosts district-led efforts to recognize and reward great teachers and principals through the 2012 Teacher Incentive Fund*. Retrieved from http://www.ed.gov/news/press-releases/us-department-education-boosts-district-led-efforts-recognize-and-reward-great-teachers-and-principals-through-2012-teacher-incentive-fund

U.S. Department of Education. (2015). *Race to the Top fund: Awards*. Retrieved from http://www2.ed.gov/programs/racetothetop/awards.html

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

Authors

EDWARD SLOAT is a faculty associate at Arizona State University. His research focuses on value-added modeling, education accountability and evaluation systems, measurement and validity theory, assessment design, and applied statistical methods.

AUDREY AMREIN-BEARDSLEY is a professor at Arizona State University. Her research focuses on educational policy, educational measurement, quantitative research methods, and high-stakes tests and value-added methodologies and systems.

KENT E. SABO is an assistant principal in Las Vegas. His research focuses on educational research and measurement, the effectiveness and efficiency of teaching and learning interventions, and intelligent and adaptive learning systems.