Research Article

The TAP System for Teacher and Student Advancement: A (Questionable) System of Teacher Accountability and Professional Support

Edward Sloat, EdD
Faculty Associate
Mary Lou Fulton Teachers College
Arizona State University
Tempe, AZ

Audrey Amrein-Beardsley, PhD
Professor
Mary Lou Fulton Teachers College
Arizona State University
Tenpe, AZ

Kent E. Sabo, PhD
Assistant Principal
Clark County School District
Las Vegas, NV

Abstract

In this study, we investigated the factor structure underlying the TAP System for Teacher and Student Advancement used across the nation for increased teacher-level accountability purposes. We found evidence of poor fit based on the factor structure posited and found large correlations among dimensions, suggesting one-to-two factors with one accounting for the majority of explained variance (i.e., a general or common, underlying factor). We use this evidence to question the validity of the inferences drawn from TAP scores, which is of import when users (e.g., principals) use the factors as independent indicators of teacher effectiveness as theorized, and also of concern when users attach consequences (e.g., merit pay) to the indicators as such. This practice is not warranted as evidenced.

Key Words

teacher accountability, teacher evaluation, accountability policy, factor analysis, high-stakes assessment, performance assessment

Background

Over the past decade, federal and state educational policymakers have enacted multiple reform initiatives in support of improving teacher effectiveness, emphasizing teacher-level accountability systems that come along with, typically peripheral and theoretical systems of teacher-level professional support. Federal legislative acts such as Race to the Top (2011) and the No Child Left Behind (NCLB) waivers awarded to states that adopted stronger teacher accountability systems (Duncan, 2011), for example, prioritized accountability mechanisms tied to measurements of teachers' impacts on their students' academic performance over time, with a tangential purpose that these mechanisms also yield objective data that could be used to support teachers' instructional improvements at the same time.

Respectively, these stronger teacher accountability and support mechanisms continue to be highly (and often solely) reliant upon measurements of teachers' value-added and observational dimensions, whereby statisticians calculate the relatively "more objective" value-added measures to assess the "value" a teacher "adds" to (or detracts from) standardized student achievement indicators from the point students enter a teacher's classroom to the point students leave, and whereby practitioners construct the relatively "more subjective" observational system measures to capture latent teacher effects by breaking down teacher effectiveness into a set of tangible and scorable factors (e.g., organization, student engagement, time management). Ideally, these observable factors can also be reduced, quantified, and then used alongside their relatively "more objective" counterparts (i.e., teachers' value-added estimates) for similar teacher accountability and support purposes, although in terms of teacher support observational systems are purposefully designed to provide teachers targeted and timely feedback to help teachers improve their professional practice.

Notwithstanding, and despite the passage of Every Student Succeeds Act (ESSA, 2016) which reinstated state-level control over states' teacher evaluation systems, there remain such "multiple measure" based systems, as well as much controversy over the appropriateness of both measures as valid representations of teachers' effects. This especially of note when consequential decisions (e.g., teacher merit pay, tenure, termination) are to be attached to the output derived via both measures.

Consequently, because not until recently have such observational tools been used within such high-stakes policy environments, have observational systems undergone the research required to support such high-stakes decision-making purposes, or rather warrant the high-stakes decisions to which such observational systems have been increasingly tasked. Put differently, because these systems were not designed for high-stakes accountability but rather informative purposes, whether using observational systems for high-stakes teacher evaluation purposes warrants much more consideration, not to mention research into whether such measurement systems are worthy of their newly elevated tasks.

Teacher Observational Systems

The observational systems now most widely for such increased teacher-level accountability purposes include Charlotte Danielson's Framework for Teaching (Danielson Group, n.d.), the Classroom Assessment Scoring System (CLASS; Teachstone, n.d.), Robert Marzano's Causal Teacher Evaluation Model (Marzano, n.d.), California's Performance Assessment for California Teachers (PACT, n.d.) and, of particular interest in this case, the

National Institute for Excellence in Teaching (NIET) TAP System for Teacher and Student Advancement (hereafter referred to as the TAP; see NIET n.d.a., n.d.b., n.d.c., n.d.d., n.d.e.). These (and really all other) observational systems, if they are to be used for consequential decision-making purposes, require examination of the measurement properties that support their newly charged uses, as again now quite different (i.e., with high-stakes consequences attached) than before (i.e., (in)formal uses meant to support teachers' professional improvements).

In addition, while the application of value-added models in evaluation frameworks continue to be rigorously vetted in the literature, observation-based evaluation systems have received much less empirical attention. Hence, and often by default, many school leaders and practitioners simply assume that just because many of these observational systems have been in use for extended periods of time (i.e., decades), and because they are also habitually advertised as "research-based," this means that they can be used in multiple ways, for multiple purposes, with multiple consequences attached. However, this simply is not true. Just because an observational system might be "tried-and-true" (i.e., used in the past and worked well for formative purposes) and "research-based" (i.e., based on what we know from the research regarding what good teachers should know and be able to do), this does not mean that these observational systems' technical properties are "research-evidenced," or perhaps more importantly "research-warranted" when high-stakes decisions are, quite frankly, at stake.

Purpose

Subsequently, we argue that a research void exists surrounding most (if not all) of the well-known observational systems currently being used across most (if not all) teacher-level accountability and support systems. We also suggest that use of such systems in high-stakes consequential environments, without supporting research evidence warranting high-stakes use, counts as educational malpractice, and more specifically conflicts with the measurement principles outlined in the *Standards for Educational and Psychological Testing* developed by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME; see AERA, APA, & NCME, 2014). Should research evidence not warrant a high-stakes use, in other words, a state or district may be liable for misuse. See, for example, *Education Week* (2015) for the approximately 15 lawsuits surrounding the alleged misapplications of teachers' high-stakes teacher evaluation data (i.e., teachers' value-added and observational data) for high-stakes decision-making purposes.

Hence, to set forth one example of what might *not* be warranted when using such observational systems, as per our research on one of the aforementioned and most widely used systems marketed and used for high-stakes decision-making purposes, we studied whether the aforementioned TAP should be used for high-stakes purposes including the distribution of teacher merit pay. More specifically, we investigated whether the factors (i.e., the overall concepts, competencies, and characteristics meant to capture teacher effectiveness) and items (i.e., the individual items meant to be observed in order to capture the overall factors) included within the TAP observational rubric function as intended. We also investigated whether the factors advanced by TAP should be, therefore, weighted and used to allocate consequences, including the monetary incentives advanced (see, for example, Jerald & Van Hook, 2011; NIET n.d.d.). We also did this because to our knowledge this type of investigation does not yet exist, although it is necessary, again, to warrant any such evaluative judgments or decisions.

The TAP System

The TAP is advertised and promoted 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.), that is in use and “impacting over 200,000 educators and 2.5 million students,” with “[o]ver 90 percent of participating TAP schools [serving] high-need and diverse areas” (NIET, n.d.c.). TAP is built upon three-factors and 19 items: *Instruction* (n=12 items), *Designing and Planning Instruction* (n=3 items), and the *Learning Environment* (n=4 items), all of which are used to evaluate teacher instructional competency, especially in consequential ways (see also Table 1). These factors and items are also, at least in theory, to help support teachers’ professional development.

Table 1

TAP Factors and Subscales (Items Per Subscale Not Included)

TAP Subscales and Components

Classroom 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		

During the school year, teachers are evaluated by certified evaluators on at least three different occasions. Some observations are unannounced while others are scheduled, with certified evaluators including mentor/master teachers and school principals, each of which are to be local to each evaluated teacher’s campus. All evaluators are certified under TAP protocols, and during observational sessions rating scores are assigned to each of the 19 TAP performance items (see Table 1) using a 1 to 5 scale with a rating of 1 representing unsatisfactory performance, 3 representing proficiency, and 5 representing exemplary performance, after which items are collapsed and then weighted in order to make overall summative decisions about the evaluated teachers and their measured effects.

Following each observation, a post-conference session is also convened between the teacher and observer to review each teacher's evaluation scores and identify/discuss instructional strengths and weaknesses. The intent here (i.e., the formative function) is for teachers to use this information to focus on and improve their professional practice. At the close of each school year, however, a teacher's final (i.e., summative) observational score is also constructed as a weighted composite for the year. It is this composite score with which we were explicitly concerned.

While this weighted measure is also combined with each teacher's academic (i.e., value-added) indicator or estimate, the overall computational measure assumes that the underlying observational metric and its weighted subcomponents are also sound and empirically defensible. While we are certainly also concerned about the soundness and defensibility of the value-added component, as are many other scholars in this area of research, of priority here was whether the intended and marketed uses of TAP's observational system, as "research-based," were also "research-warranted," or rather sound, defensible, and also valid.

Methods

Hence, we assessed the foundational characteristics of the TAP observational system's factor structure within using confirmatory factor analysis (CFA) and exploratory factor analysis (EFA) approaches. More specifically, we utilized 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 was supported by empirical evidence.

Sample

We examined teacher observation data for 1,081 teachers collected from a set of 14 school districts in one state. These districts represented a total of 54 schools including 39 elementary (72%), nine middle (17%), and six high schools (11%) enrolling a combined 34,055 K-12 students (just over 3% of the state's total K-12 school enrollment). The race/ethnicity of the student population taught by TAP teachers in the sample included students representing higher proportions, that were statistically significant as compared to state averages, of students who were from racial minority and poor backgrounds. This is likely due to NIET's focus on serving teachers and students from lower income communities/schools.

Procedures

We first applied CFA approaches to evaluate whether the TAP System's posited factor structure was supported by empirical evidence. 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 followed this with EFA approaches to more explicitly examine attributes of the latent structures inherent in the empirical data. When generating EFA models, we again recognized both the categorical nature of the measured variables 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 (individual) level while leaving the between-school (group) level unrestricted. For all EFA rotations we utilized the Oblimin (oblique) procedure and based our warranted factor extractions on review of scree plots, Kaiser criterion (eigenvalues greater than 1.00), size of rotated factor loadings, and factor interpretability.

Based on results obtained from the EFA analysis, inclusion/examination of a primary common factor seemed warranted. In this regard, we reformulated four additional CFA models to evaluate the appropriateness of both second order and bi-factor solutions including a single common factor model. All other sampling, procedural, and other methodological details of our study can be found in Sloat, Amrein-Beardsley, and Sabo (2017).

Findings

As noted, our findings suggest that the posited three-factor TAP observational framework (see Table 1) yields a poor-to-marginal fit (i.e., the factor and items do not function or "hold together" per factor as posited). Rather, a dominant first- or sole factor dimension was present suggesting that the TAP observational rubric is measuring one versus three dominant factors as marketed and claimed. That is, an overall "teacher effectiveness" factor was observed, as measured by the 19-items when combined or collapsed together, that should *not* be separated or much less weighted by factor. Put differently, using the TAP to yield a common (i.e., general) sense of whether a teacher is effective or not might very well be a defensible use of the TAP (and perhaps other) observational system(s), but the factors or subcomponents postulated to more distinctively capture what it means to be an effective teacher as per the TAP (and perhaps other) observational system(s), do not hold, empirically speaking. From an application point of view, this also means that taking consequential actions (e.g., making merit-based decisions) based on the factor scores as conceived is not warranted as per the evidence.

Moreover, one should not simply assume that without empirical evidence factor-level scores are uniquely measuring factor-level teacher effectiveness behaviors, when instead they might be contributing to a larger, more general, definition of what it means to be an effective teacher, or what it means to *not* be an effective teacher, neither of which can be justifiably apportioned as desired in at least this case (e.g., in terms of weights and monies or other consequences attached to inappropriately weighted measures). Herein exist concerns in both policy and practice, for this observational system and perhaps others.

Conclusions

As noted, classroom observations serve as critical components of many federal and state educational reform initiatives because they appear to provide summative as well as actionable formative information to practitioners. On the latter point, it seems reasonable to expect that teachers use evaluation information in a formative manner to improve targeted areas of professional practice. On the former point, it stands to reason that the use of summative measures within pay-for-performance and other high-stakes decision-based systems may provide incentives (and disincentives) that may motivate teachers to improve specific competencies and increase their overall performance, not to mention student performance, over time. Indeed, TAP developers presume this type of causal pathway whereby such summative and formative evaluation measures should lead to improved instructional competence, and increased student academic performance over time, again as incentivized (Jerald & Van Hook, 2011; NIET, n.d.d.).

However, results from this study suggest that reliance on different factor-level scores to identify targeted practices, initiate interventions, and consequentially infer on attributes of teachers' professional effectiveness may be suspect, in this and perhaps other cases.

Due to TAP's widespread use this is certainly important to note, however, also given the potential pragmatic implications (e.g., teachers who might contest not receiving a merit pay sum given an unjustifiably weighted score), policy implications (e.g., school leaders who might via local policy require the attachment of high-stakes consequences to one or more factors over other(s)), and potential legal ramifications (e.g., teachers who might be terminated, at least in part, due to performing poorly on one or more factors over other(s)).

At the same time, while the three-factor structure of the TAP may not be empirically supported, this does not mean that the summative scale constructed from the individual indicators (i.e., representing the general or common factor) does not capture essential elements of quality instructional practices.

Indeed, and accordingly, school leaders, policymakers, and the like might be wiser (and safer) to simply attach high-stakes decisions (and low-stakes decisions for that matter) to the overall scores derived via this, and perhaps other observational systems, until the empirical evidence supports such partitioning practices otherwise.

Author Biographies

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. Email: esloat@asu.edu

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. Email: audrey.beardsley@asu.edu

Kent 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. Email: saboke@nv.ccsd.net

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.

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

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

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

Duncan, A. (2011). *Winning the future with education: Responsibility, reform and results. Testimony given to the U.S. Congress.* Washington, DC: Retrieved from http://www.ed.gov/news/speeches/winning-future-education-responsibility-reform-and-results

Education Week. (2015). *Teacher evaluation heads to the courts.* Retrieved from http://www.edweek.org/ew/section/multimedia/teacher-evaluation-heads-to-the-courts.html

Every Student Succeeds Act (ESSA) of 2015, Pub. L. No. 114-95, § 129 Stat. 1802. (2016). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

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

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 (NIET). Retrieved from http://files.eric.ed.gov/fulltext/ED533382.pdf

Marzano. (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/

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.

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

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

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

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

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

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 Publishing, Inc.

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. Retrieved from http://journals.sagepub.com/doi/full/10.1177/2332858417735526

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