education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 34 Number 33 April 7, 2026 ISSN 1068-2341

Should Teacher Observation Systems be Used for Making High-Stakes Decisions?

Michael Strong
Jaehoon Lee
Texas Tech University
John Gargani
Gargani + Co.
Minju Yi
Hyunjin Shim
Texas Tech University
&
Hyunchang (Henry) Moon
Augusta University
United States

Citation: Strong, M., Lee, J., Gargani, J., Yi, M., Shim, H., & Moon, H. (2026). Should teacher observation systems be used for making high-stakes decisions? *Education Policy Analysis Archives, 34*(33). https://doi.org/10.14507/epaa.34.9841

Abstract: This study questions the suitability of teaching observation data for making high-stakes decisions that affect teachers. We define suitability (the extent to which intended purposes are advanced without causing undue harm) and argue that it fundamentally depends on the technical properties of the data produced by an observation system, which, in turn, depend on the attributes

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 6/10/2025
Revisions received: 14/1/2026
Accepted: 5/2/2026

designed into the system. We conducted an experiment to understand better the relationship between the attributes of teaching observation systems and the suitability of their data. We compared three systems with different attributes, including rubrics that impose varying inference loads on raters. Experienced raters were randomly assigned to a system and properly trained. Then, they evaluated the instruction of advanced teacher candidates by viewing videos of their lessons. We considered three criteria when judging the resulting data: the power to predict a teacher's contribution to student learning, the correlation of scores across systems, and rater agreement within systems. We found that a system with a low inference load (along with other attributes) outperformed systems with higher inference loads, but it may still be insufficient for making confident, high-stakes decisions. We maintain that few, if any, widely used observation systems are.
Keywords: teacher observation; high-stakes decisions; inference

¿Deben utilizarse los sistemas de observación docente para la toma de decisiones de alto impacto?

Resumen: Este estudio cuestiona la idoneidad de los datos provenientes de la observación de la enseñanza para la toma de decisiones de alto impacto que afectan al profesorado. Definimos la idoneidad como el grado en que los propósitos previstos se logran sin causar daños indebidos y sostenemos que esta depende fundamentalmente de las propiedades técnicas de los datos producidos por un sistema de observación, las cuales, a su vez, dependen de los atributos diseñados en dicho sistema. Realizamos un experimento para comprender mejor la relación entre los atributos de los sistemas de observación docente y la idoneidad de los datos que generan. Comparamos tres sistemas con atributos diferentes, incluidos instrumentos de rúbricas que imponen distintas cargas inferenciales a los evaluadores. Evaluadores con experiencia fueron asignados aleatoriamente a un sistema y recibieron capacitación adecuada. Posteriormente, evaluaron la enseñanza de candidatos avanzados a docentes mediante la visualización de videos de sus clases. Consideramos tres criterios para valorar los datos resultantes: la capacidad predictiva de la contribución del docente al aprendizaje estudiantil, la correlación de las puntuaciones entre sistemas y el grado de acuerdo entre evaluadores dentro de cada sistema. Encontramos que un sistema con baja carga inferencial (junto con otros atributos) superó a los sistemas con mayores cargas inferenciales; sin embargo, aun así podría resultar insuficiente para tomar decisiones de alto impacto con plena confianza. Sostenemos que pocos, si es que alguno, de los sistemas de observación ampliamente utilizados cumplen con este estándar.
Palabras-clave: observación docente; decisiones de alto impacto; inferencia

Os sistemas de observação docente devem ser utilizados para a tomada de decisões de alto impacto?

Resumo: Este estudo questiona a adequação dos dados de observação da prática docente para a tomada de decisões de alto impacto que afetam professores. Definimos adequação como o grau em que os propósitos pretendidos são alcançados sem causar danos indevidos e argumentamos que ela depende fundamentalmente das propriedades técnicas dos dados produzidos por um sistema de observação, as quais, por sua vez, dependem dos atributos incorporados ao seu desenho. Conduzimos um experimento para compreender melhor a relação entre os atributos dos sistemas de observação docente e a adequação dos dados por eles gerados. Comparamos três sistemas com atributos distintos, incluindo rubricas que impõem diferentes níveis de carga inferencial aos avaliadores. Avaliadores experientes foram

designados aleatoriamente a um dos sistemas e devidamente treinados. Em seguida, avaliaram a prática de candidatos avançados à docência por meio da observação de vídeos de suas aulas. Consideramos três critérios para julgar os dados resultantes: o poder de prever a contribuição do professor para a aprendizagem dos estudantes, a correlação das pontuações entre os sistemas e o grau de concordância entre avaliadores dentro de cada sistema. Constatamos que um sistema com baixa carga inferencial (associada a outros atributos) superou aqueles com cargas inferenciais mais elevadas; ainda assim, pode não ser suficiente para sustentar decisões de alto impacto com segurança. Sustentamos que poucos, se é que algum, dos sistemas de observação amplamente utilizados atendem a esse padrão.
Palavras-chave: observação docente; decisões de alto impacto; inferência

Should Teacher Observation Systems be Used for Making High-Stakes Decisions?

At first glance, the question that motivates our research may not appear worth asking: Are data generated by widely used teaching observation systems suitable for informing high-stakes decisions? After all, school administrators have depended on these data for a variety of high-stakes purposes for a long time (Close et al., 2020). Much effort has been expended by developers and users of teaching observation systems, including the three-year $45-million Measures of Effective Teaching (MET) project (Kane et al., 2013), which argued that observation data can confidently inform high-stakes decisions even when noisy. However, the subsequent $575 million Intensive Partnerships for Effective Teaching (IP) project, designed to implement MET's findings in schools, raises questions about this claim (Baird et al., 2019). After six years, the IP project's final report disappointingly concluded:

> We found that the sites succeeded in implementing measures of effectiveness to evaluate teachers and made use of the measures in a range of human-resource (HR) decisions; overall, however, the initiative did not achieve its goals for student achievement or graduation (p. xxvi)…[and] we found little evidence that the policies designed, in whole or in part, to improve the level of retention of effective teachers had the intended effect. (p. xxvii)

The report suggests several potential factors that may have contributed to the project's failure: incomplete implementation, state-level policy changes, insufficient time for effects to appear, and a flawed theory of action. We zoom in on one that was not considered in the report, and we believe is an entrenched problem threatening many high-stakes decisions: the technical properties of observation data that result from the design of the observation system.

Despite the long history of using observation data for high-stakes decisions, both in the US and internationally (Martinez et al., 2015), what we observe in practice and other researchers suggest (e.g., Madaus & Russell, 2010) is the possibility of a substantial gap between the technical properties that observation data should possess to advance high-stakes purposes and the properties they actually have. Closing the gap is essential to making observation data *suitable* for high-stakes purposes, by which we mean that the data advance intended purposes enough to matter without causing undue harm (Hill & Grossman, 2013). We organize our evaluation of suitability around a property fundamental to high-stakes decisions: the ability to predict teachers' future success.

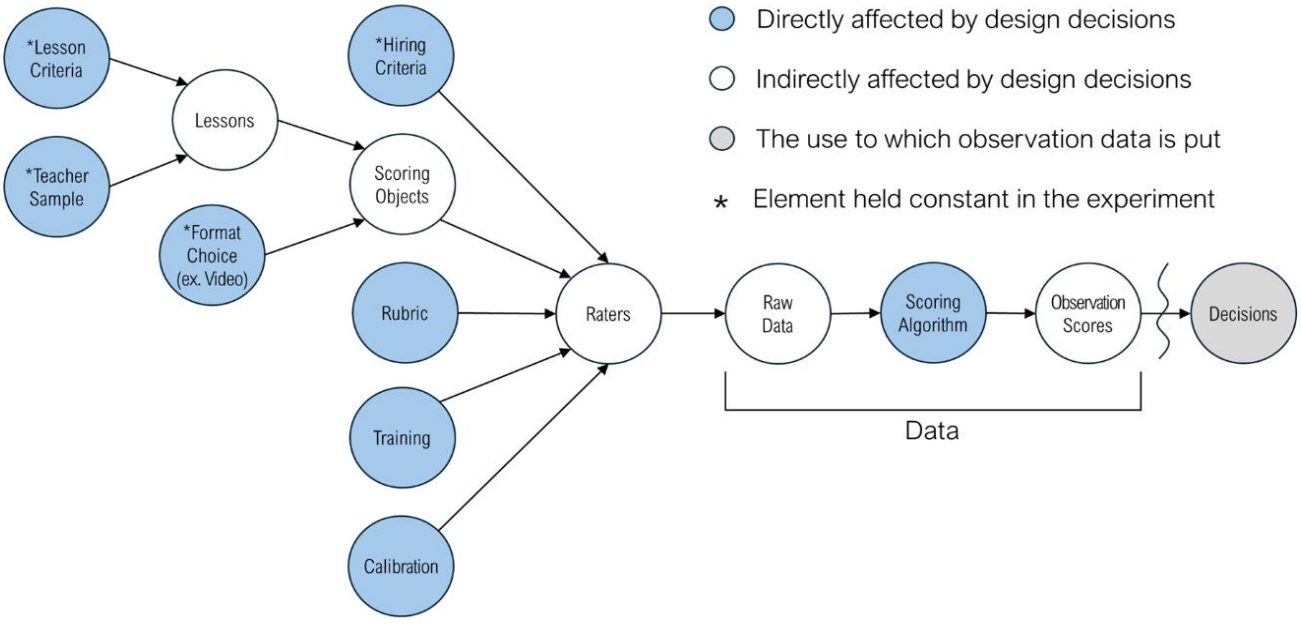
Teachers are not hired for what they have accomplished, but rather for what they are expected to accomplish; their instruction is observed and evaluated in order to encourage improvement in the future, not merely document what was done; and they are paid bonuses not only

to reward past success, but also to encourage them to remain in the profession and repeat their accomplishments. Prediction in this context is sometimes based on the assumption that past performance is indicative of future performance, because effectiveness is a relatively stable teacher attribute. Alternatively, it may be supported by empirical evidence that predictions about future performance hold to a meaningful degree (e.g., Blazar, 2015). In either case, identifying past effectiveness is key, and this is what teacher observation systems report as scores. The predictive power of observation scores depends on other properties of the data, in particular score reliability, which in turn are supported or compromised by the design of the teaching observation system.

We conceptualize a teaching observation system as having a number of attributes that are determined by design decisions, some directly, others indirectly (Figure 1). Attributes directly affected by design decisions include lesson criteria, teacher sample, review format (e.g., video versus in-person), hiring criteria for evaluators, observation rubric, evaluator training, evaluator calibration procedures, and the scoring algorithm. Those indirectly affected by design decisions include the lessons to be evaluated, the scoring objects, the raters, the raw data, and the observation scores. The line of reasoning we investigate here is that a system designed to have certain attributes related to its rubric (lower instead of higher inference load), calibration procedures (continuous instead of yearly), and scoring algorithm (summing normalized counts instead of summing ordinal categories) will tend to produce more reliable data that in turn amplify predictive power, making them more suitable for high-stakes decisions. We argue that developers and users of teaching observation systems should establish that data possess these properties (among others) to a sufficient degree before trusting them to guide high-stakes decisions. If data fall short of being suitable (as we and others suspect may be common), developers and users should look to the design of the teaching observation system to help explain and reduce the gap.

Figure 1

The Attributes of a Teaching Observation System



To establish that a gap between the desired and actual properties of the data may exist for the reasons outlined above, we conducted a randomized experiment comparing data produced by three teaching observation systems. The data describe the quality of a lesson taught by advanced teacher candidates who also administered tests to their students before and after the lesson. We use the pre-post difference in test scores as a direct measure of instructional effectiveness for the lesson against which observation scores are compared. Two of the observation systems we evaluate, T-TESS (Texas Education Agency, 2016) and TAP (NIET, 2012), are used throughout the state of Texas with teacher candidates to determine graduation and certification, and with practicing teachers at the classroom, school, district, and state level to inform decisions regarding employment, retention, salary, and bonuses. The third system, RATE (Gargani & Strong, 2014), was explicitly designed to produce data that are predictive of the future performance of teachers, in part because it imposes a lower inferential load on evaluators, employs continuous calibration, and uses a scoring algorithm that sums normalized counts of teaching behaviors rather than tallying ordinal categories that characterize teaching behaviors. There is published evidence that RATE can produce data that are at least as predictive of future effectiveness as widely used instruments, and often more so, with inter-rater reliabilities that typically exceed those reported by the developers and users of other systems (Gargani & Strong, 2014). Notably, RATE was not developed for high-stakes purposes but rather to support teachers early in the school year if the evidence it generates suggests they may struggle to be effective. T-TESS and TAP, which are used for high-stakes purposes, have no published data regarding their ability to predict teacher effectiveness or their inter-rater reliability. This raises concerns because educators cannot determine whether basing high-stakes decisions on their data will lead to poor outcomes with negative consequences for teachers and students. With two points of comparison—the pre- and post-tests and the data produced by RATE—we set out to determine whether there is cause for concern about the use of observation data for high-stakes purposes in teacher preparation programs that employ similar observation systems.

Literature Review

High Stakes in the Real World

Teachers face high-stakes decisions from the moment they begin preparing for grades, graduation, and certification, and throughout their teaching careers, regarding employment, salary, dismissal, promotion, and bonuses. Educational systems cannot operate unless high-stakes decisions like these are routinely made, and it would be better to base them on evidence rather than intuition (Braun, 2005; Hunter, 2024). Often, the evidence comes in part or in whole from teaching observation systems. A recent survey found that at least 61% of practicing teachers across 36 U.S. states are evaluated using formal observation systems (Close et al., 2020). Regarding states that did not report using teacher-level observations in the survey, Strong and colleagues note that a further 12% may use them but do not specify this. An additional 10% allows for local control, which may also include the use of teacher observations. That raises the likely possible use of teacher observations to well over 80% of the states. Grissom and Bartanen (2021) suggest that formal observations are "the largest component of evaluation ratings" (p. 131), Garrett and Steinberg (2015) describe observation as "a critical component of teacher ratings" (p. 225), and Steinberg and Donaldson (2016) find that "classroom observation remains the dominant method of evaluation in U.S. schools." (p. 344). Internationally, teaching observation systems are used widely, and a study by Martinez, Taut, and Schaaf (2015) reviewed observation systems in Singapore, Japan, Chile, Australia, Germany, and the US. Furthermore, nearly 100% of teacher candidates in the US receive high-stakes

evaluations of this kind to graduate and earn a credential (CAEP, 2020; NCTQ, 2018; Putman et al., 2024)[1].

The justification for the widespread use of observation data is straightforward—either alone or in combination with other data (in particular, value-added measures)—observation data are believed to improve high-stakes decisions and the actions they precipitate (e.g., Kane et al., 2012). In other words, the data are purportedly suitable. However, it is surprising how little evidence developers make available regarding the suitability of observation data for high-stakes purposes (Bell et al., 2019; Dobbelaer, 2019; Luoto, 2023), and there is research to suggest we should be skeptical of their suitability in real-world contexts. Baker et al. (2013) have also warned about the legal consequences of using low-quality information to make high-stakes decisions. There are validity issues when the observer has prior knowledge of the teacher (Whitehurst et al., 2014), when administrator observers are tempted to game the system (Geiger & Amrein-Beardsley, 2017), and the related well-known "widget effect," where most teachers are evaluated as satisfactory or better (Weisberg et al., 2009) or, at least, poorly differentiated from one another (Kraft & Gilmour, 2017). More recently, race and gender bias has been observed by some researchers (Grissom & Bartanen, 2021), a topic that has received less attention than it deserves.

Even under favorable conditions, observation data may not fare well. Disappointing correlations with student learning metrics have led scholars and practitioners to question the validity of widely used observation systems, such as the Danielson Framework for Teaching (FFT; Danielson, 2007). In their review of the literature on this topic, Basileo and Toth (2019) found, from the 20 studies in their sample, that most observation rubrics yielded only "small to moderate [mostly positive] correlations" (p. 5), suggesting they are of limited utility even though 64% were statistically significant in math, and 43% in ELA. Marzano and colleagues conducted their own study of the Marzano Teacher Evaluation Model (MTEM; Marzano et al., 2013), widely used in Florida but previously under-researched. In this large-sample study, they found that MTEM produced what they characterize as only small, positive correlations with value-added measures, again suggesting that they are of limited practical utility, despite being statistically significant.

From Berliner's (2014, 2018) perspective, there are problems with evaluating teachers using standardized achievement test data due to validity concerns and with classroom observational systems due to reliability concerns. When both are used together, he notes (as do Basileo and Toth, 2019), the correlation between the two is "quite low," suggesting that "either one or both methods are failing to assess the intended construct" (Berliner, 2018, p. 1). In his 2018 paper, Berliner describes the two methods as the Scylla and Charybdis of teacher evaluation because neither represents a safe course of action. Therefore, he advocates an alternative means of appraising teachers, who he claims are usually evaluated to identify and eliminate "bad" performers. He does not, however, mention the increasing number of pay-for-performance programs that seek to identify high-quality teachers for merit pay. Like targeting underperforming teachers, the stakes are still high, but a salary increase is on the line rather than a job.

Berliner (2018) asserts that the reliability issues he associates with observation systems can be overcome by adequately training observers. However, there is little substantiating evidence for this claim (except perhaps under certain research conditions), for most users of observation systems rarely publish their reliability statistics (Patrick et al., 2020). When they do, there exists no consensus on what level of agreement needs to be attained (Wilhelm et al., 2018), whether it persists after training (see Myford & Wolfe, 2009, on rater drift), or precisely how modifying training that is

[1] Not all states require performance assessment (e.g., edTPA or CalTPA), but NCTQ and CAEP strongly recommend TPPs to provide opportunities for teacher candidates to apply knowledge and skills they learned in their program to K-12 classroom and be evaluated by teacher educators and mentor teachers, etc.

currently judged to be adequate would improve any of this. Even if observers are well-trained, there is reason to believe that their ratings remain vulnerable to biases (e.g., Liu et al., 2019) that may emerge across the various contexts in which they work. If these biases, which we discuss later, reduce the validity or reliability of observation data, then the high-stakes decisions that rely on them may not be improved and may be undermined.

RATE and the Philosophy of Subtraction

In developing RATE, Strong and colleagues adopted a framework derived from the philosophy of subtraction, which emphasizes removing extraneous elements to reveal the essence of what is being evaluated. Most often encountered in the fields of architecture and design, the philosophy of subtraction is well summed up in the following aphorism from Antoine de Saint-Exupery (1939):

> In anything at all, perfection is finally attained not when there is no longer anything to add, but when there is no longer anything to take away… (Chapter 3, translation by L. Galantière, p. 42)

The process for developing RATE involved several iterations. Strong and colleagues began with 12 items of teaching behavior frequently observed in other systems, which they subsequently reduced to six by excluding those that contributed little or nothing to predicting student learning gains. Their goal was to enhance consistency and fairness by providing a more accurate reflection of a teacher's effectiveness. Ultimately, they contend that this philosophy ensures that the evaluation process is both equitable and focused on meaningful indicators of teaching quality. They do not maintain that these indicators are the only important teaching behaviors, merely that they are predictive of effective teaching.

RATE and Cognitive Psychology

The second theoretical orientation driving this work stems from studies in cognitive psychology, information processing, and perception that describe what happens when people make judgments about human behavior, particularly regarding the effects of heuristics, biases, and inference. As discussed in Gargani & Strong (2014), cognitive operations such as confirmation bias (Wason, 1960), motivated reasoning (Kunda, 1990), and inattentional blindness (Simons & Chabris, 1999) contribute to biases in our observations of human behavior. These and other phenomena have been conceptualized within a framework of cognitive operations that distinguish judgments made quickly and with little conscious deliberation from those that are slower and more reflective (labeled System 1 and System 2 processes by Stanovich & West, 2000). Kahnemann (2002) noted that much of the unreliability in human judgment stems from humans' inability or disinclination to use System 2 to monitor and correct System 1 judgments. Experts are no exception. Their judgments are often produced using System 1 processes that rely on shortcuts and heuristics; yet, because they are experts, one may falsely infer that their ad hoc intuitions result from systematic System 2 processes. We suggest that evaluators using teaching observation systems will likewise make error-prone intuitive judgments when the observation systems at their disposal impose a sufficiently high inference load. If such errors are not mitigated by thorough training, calibration, rubric structure, and other attributes of the teaching observation system, then the reliability and validity of the observation system may be compromised.

Inference refers to the process of drawing conclusions or making assumptions based on incomplete or indirect evidence. In the context of observing teacher behavior, inference plays a role in how evaluators interpret and make sense of what they observe. We hypothesize that the more assumptions or judgments an observer must make, the higher the inference load and the more likely

raters are to rely on heuristics, express biases, and produce error-prone judgments. We characterize rubrics as providing systematic guidance to raters, and inference is the mechanism by which they fill gaps in that guidance. In other words, evaluative judgments about teaching are a function of both inference and systematic guidance.

We distinguish five types of judgment that require inference in any observation rubric (not necessarily made sequentially). One, the observer must determine that the observed behavior is an example of the teaching dimension to be scored. Two, the observer is called upon to judge which sub-behaviors identified in the rubric are present. Three, the observer must contend with a class of judgments related to any descriptors such as adjectives, adverbs, and other qualifiers (e.g., "appropriate", "most", "effectively") included in the rubric's language. Four, the observer must infer unseen behaviors or motives present in the rubric (such as "teacher anticipates possible misunderstandings"). Five, the observer must determine where the behavior falls on the scale described in the rubric (often a four- or five-level scale reflecting increasing levels of proficiency).

We use a simple formula to produce what we call Inference Load (IL). We count the number of inferential judgments in the rubric and divide by the number of observational indicators (sometimes referred to as dimensions) as follows:

$$\text{IL}=\frac{N_{\text{inferences}}}{N_{\text{indicators}}} \quad (1)$$

Here, all judgments are weighted equally. Observation rubrics are commonly described as being high- or low-inference. This usually means that observers are either required to make judgments (high-inference) or count behaviors (low-inference). Most teacher observation systems in use today are high-inference. However, the distinction is not binary since the total number of inferences required of an observer may vary across rubrics according to the IL formula, thus obscuring differences among similarly classified high- or low-inference rubrics. We contend that it is more helpful to think of observation rubrics as a continuum from highest to lowest inference, and that IL should be considered before designing the rubric, rather than categorizing it as high- or low-inference post hoc. Absent such forethought, training may help observers align their judgments when the IL is high, but most training manuals place little emphasis on this aspect of preparation. Moreover, developers and users of observation systems deem their current training adequate, and it is unclear how they could modify it to mitigate the adverse effects of inferences. More frequent or continuous calibration is another possible solution, but it is arduous when IL is high.

The Logic of How Data Affect Decisions

Hypothesis

We hypothesize that a teacher observation system designed to have specific attributes, including lower IL, will be more suitable for high-stakes decisions than a system without these attributes. Our logic is as follows: (a) observation scores have less random error when the raw data contain less error and scoring algorithms do not add error; (b) raters produce raw data with less random error when rubrics, training, and calibration support agreement; (c) data with less random error have greater predictive power, all else equal; and (d) more predictive data are more appropriate for high-stakes decisions. Underlying the logic is the assumption that high-stakes decisions are made by ranking teachers by their effectiveness; thus, the correlation between observation scores and student learning gains matters more than placing teachers in fixed categories of absolute ability (aims that need not be mutually exclusive).

Research Questions

To test our hypothesis that teaching observation systems designed to have certain attributes will produce data more suitable for high-stakes decisions, we posed three research questions about observation systems with different bundles of attributes.

RQ1: How well do T-TESS, TAP, and RATE predict a teacher's contribution to student learning?

RQ2: How similarly do T-TESS, TAP, and RATE rank teachers?

RQ3: How well do raters using the same observation system agree with each other?

If the answers to these questions suggest that RATE performs better than T-TESS and TAP, they provide evidence that RATE's bundle of attributes produces data that better inform high-stakes decisions.

Methods[2]

Context

The study was conducted in a teacher education program in Texas, which employed over 376,000 teachers in the 2021-22 school year, of whom 43,000 were newly prepared, according to the Texas Education Agency (TEA). Texas has introduced pay-for-performance programs over the years that reward teachers based on their classroom performance and evidence of student learning. One, the Texas Incentive Allotment, is active at the time of writing.

Conditions: Teaching Observation Systems

We compare data from three observation systems: T-TESS, TAP, and RATE. The first two are widely used in Texas but lack published supporting research. RATE, on the other hand, has considerable associated research (Gargani & Strong, 2014) and to date has been used almost exclusively in research settings. Figure 1 illustrates the generic structure of a teaching observation system, which informs system developers' design decisions. The eight blue attributes depend directly on design decisions, and those marked with an asterisk were held constant in our experiment. The constant attributes are:

- *Lesson Criteria.* Lessons took the form of direct instruction, lasting at least 20 minutes, ensuring that all targeted behaviors in every observation system were observable.
- *Teacher Sample.* We selected 30 advanced teacher candidates conducting 4th-grade math lessons.
- *Format Choice.* Evaluators viewed 20 minutes of video of direct instruction for each lesson, so their judgments were based on the same amount of information, and they could review teaching behaviors if necessary.
- *Hiring Criteria.* We hired teacher educators as raters who had at least five years of experience as classroom teachers, at least three years of experience as full-time teacher educators, and at least one year of experience evaluating teachers with one or more observation systems.

The remaining four attributes varied as a package across T-TESS, TAP, and RATE, and are summarized in Table 1.

[2] IRB approval was obtained from the first author's university's research office

Table 1

Four Attributes that Varied across the Three Observation Systems

		Teaching Observation System		
Attribute		**T-TESS**	**TAP**	**RATE**
Rubric	*N* Items	8	12	6
	IL Score	3.6	3.1	1.0
Training	Format	Online	Online	In Person
	Length	Self-Paced	Self-Paced	One Day
Calibration	Process	Certification	Certification	Comparison
	Frequency	Yearly	Yearly	Continuous
Scoring Algorithm	Item Scale/Measure	5-Point Scale	5-Point Scale	Count of Behaviors
	Lesson Score	Average Item Score for a Lesson	Average Item Score for a Lesson	Sum of Normalized Counts

Rubric. For both T-TESS and TAP, we included only rubric items that describe observable classroom behaviors, consistent with how the instruments are used in a statewide pay-for-performance program in Texas. All the items on RATE describe behaviors that are observable in the classroom. See Appendix A for a complete list of the items we used in each rubric. The T-TESS rubric has eight observable behaviors we label as items, five in the instructional domain and three in the learning environment domain. The TAP rubric has 12 items that measure constructs similar to those in the T-TESS rubric. There are eight items in the instructional domain and four in the learning environment domain. RATE has six items. The inference load of the rubrics, which we calculated using Equation 1, varied by rubric as 3.6 for T-TESS, 3.1 for TAP, and 1.0 for RATE.

Training. Raters using T-TESS and TAP completed a self-paced, online training at an earlier date. Those randomly assigned to the RATE system received a one-day in-person training session before the experiment began.

Calibration. The three systems incorporate calibration activities in their training. T-TESS and TAP suggest that users recalibrate annually. To ensure that evaluators using T-TESS and TAP were fully prepared, they completed an online recalibration during the two days prior to participating in the research. Evaluators using RATE also calibrated immediately before the experiment started, in their case, during their in-person, one-day training. Additionally, the RATE system incorporates continuous recalibration into its scoring process. Every lesson is scored *independently* by two raters simultaneously. After completing their independent scores, they compared them and made adjustments as needed, producing revised scores. In this study, two pairs working independently of each other scored each lesson. It should be noted that only independent scores are used to measure reliability, and revised scores are used to measure predictive power.

Scoring Algorithm. For T-TESS and TAP, raters assign a score to each rubric item on a 5-point scale (i.e., a whole number between 1 and 5, inclusive). These item scores are the "raw data" in Figure 1. Lesson scores are calculated by averaging the item scores associated with a lesson and a rater. This resulted in four lesson scores from T-TESS and four from TAP. The overall score is

calculated by averaging the four lesson scores. For this study, "overall scores" are the "observation scores" shown in Figure 1 (in practice, observation scores may combine scores from multiple raters observing multiple lessons per teacher). In contrast, RATE requires raters to count the number of instances of behaviors described by each item; thus, the raw data are counts. Later, the count assigned by a rater to a given item for a lesson is adjusted by subtracting the minimum count for that item across all lessons and then dividing by the range, defined as the difference between the maximum and minimum counts for that item across all lessons. This normalizes the item scores (ensures that every normalized item score falls between the inclusive range of 0 to 1). Normalizing in this manner makes each item equally important (it aligns counts with different base rates) and maximally variable (it helps squeeze as much information as possible from each item). A lesson score is calculated as the sum of all the normalized item scores for that lesson produced by a rater. Two lesson scores were produced for each lesson, and the overall score for the lesson is the sum of the two lesson scores.

Of these four attributes, we do not consider training as an explanatory factor. This may introduce uncertainty, but we believe it amounts to only a minor threat to validity because (1) all training was conducted to a standard of quality that system developers judged to be adequate, (2) we have no reason to doubt the effectiveness of the intended training, and (3) no superior training exists. We acknowledge that training modality (online vs. in-person) could indirectly affect rater behavior through differences in engagement, opportunities for clarification, or social calibration dynamics. However, isolating such effects would require a separate experimental design. Thus, we focus on how variations in the remaining three attributes—rubric, calibration, and scoring algorithm—affect data.

Selection and Random Assignment

We made a request through Texas Tech's Teacher Education Department for videos of 30 elementary-level mathematics lessons taught by advanced teacher candidates in the department's program. We stipulated that a lesson-level pre- and post-student assessment accompany each lesson. We then hired 18 experienced raters and randomly assigned six to each teacher observation system (i.e., condition). We randomly assigned videos from the pool of 30 to raters as follows. For T-TESS and TAP, judges were assigned 20 videos each. As a group, they produced four independent scores per video. For RATE, judges were assigned 10 videos each. As a group, they produced two independent scores and two revised scores for each video. We count this as two scores because we use the two independent scores to estimate reliability and the two revised scores to estimate predictive power.

The discrepancy in the number of randomly assigned videos ensured that all training and rating was accomplished within 16 hours. All evaluators were previously trained to use T-TESS and TAP, so evaluators randomly assigned to these observation systems required only an hour to complete recalibration training. Evaluators randomly assigned to RATE, however, required a full day for training. A flowchart illustrating the randomized design is provided in Appendix B.

Statistical Analysis

Score Calculation

As described in the above section *Scoring Algorithm*, raters randomly assigned to T-TESS and TAP assigned a score to each item on a 5-point scale. These item scores were averaged within lesson and rater to produce a lesson score—four lesson scores per lesson for T-TESS and four for TAP. Lesson scores were averaged to produce a single overall score for each lesson. Raters randomly assigned to RATE provide counts of behaviors described by each item. Item counts were "normalized" by subtracting, for each item, the minimum count observed across all lessons from the

count produced by a rater, and then dividing the result by the difference between the maximum and minimum counts for the item across all lessons. The resulting values, ranging from 0 to 1, were summed within each lesson to produce a lesson score—two lesson scores per lesson. Finally, lesson scores were summed to yield a single overall score for each lesson.

Predicting a Teacher's Contribution to Student Learning (RQ1). The overall lesson score for each observation system was used to predict the teacher's contribution to learning. We estimated the teacher's contribution to student learning on the lesson as a pre-post difference in scores on assessments administered immediately before and after the lesson. Specifically, we calculated this as a *weighted change in proportions*—the pre-post change in proportion of students in the class who scored at or above the "proficient" level, weighted by the inverse of the proportion of students who scored less than proficient at the pre-test. This method assigned greater weight to lessons in which a larger share of students were already proficient before the lesson, ensuring comparability across lessons. The change in proportion could not be determined for eight lessons because data were missing for the entire class. To address this, we employed multivariate imputation by chained equations (MICE; van Buuren & Groothuis-Oudshoorn, 2011). Then, we calculated Pearson correlations between the weighted change in proportions and the T-TESS, TAP, and RATE scores, using both the original censored dataset (N = 22 lessons) and the full imputed dataset (N = 30 lessons in each of the five imputed datasets). We conducted a two-sided hypothesis test to assess whether the correlations based on the raw and imputed data differed from 0. In addition, we performed Ly et al.'s (2016) Bayes factor test for correlations to measure the strength of evidence supporting the alternative hypothesis ($0 < \varrho < 1$) against the null ($-1 < \varrho < 0$). Then, we applied the qualitative categories for evidence strength described in Held and Ott (2018).

Because other authors have characterized correlations and reliabilities for teaching observation systems as small, medium, or large using various standards, we provide similar interpretations. As we describe in our *Discussion*, this is a problematic practice. We use Cohen's (1988) guidelines for correlations of small (0-.29), moderate (.30-.49), and strong ($\geq$ .50), which characterize how discernible they are (how large the true correlation is in relation to the random error of its measurement), not how suitable a correlation may be for a given purpose.

Calculating the Similarity of Ranked Scores Across Observation Systems (RQ2). To gauge the similarity of ranked scores, we estimated two types of correlation for all pairs of systems (T-TESS with TAP, T-TESS with RATE, and TAP with RATE). We used Pearson correlations to gauge the similarity of the T-TESS, TAP, and RATE lesson scores upon which the lesson ranks were based, and Spearman's rho correlations to gauge the similarity of the ranked lesson scores. In ranking the scores, lessons with higher scores were assigned lower ranks, and any lessons with tied scores were assigned the average of the ranks they would have occupied. For example, lessons with identical scores at the 3rd and 4th positions would both be ranked 3.5th. We conducted two-sided hypothesis tests to assess whether the correlations differed from 0 and again use Cohen's (1988) classification of small, medium, and large to characterize the size of both types of correlations (acknowledging they were intended for Pearson correlations).

Calculating the Agreement of Scores Within an Observation System (RQ3). We explored multiple indices of rater agreement. We ranked the T-TESS, TAP, and RATE lesson scores produced by each rater, using the full imputed dataset[i]. Any lessons with tied scores were assigned the average rank, as described in the previous section for RQ2. Using the ranked lesson scores, we computed Kendall's *W,* which ranges from 0 to 1, where 0 = perfect disagreement and 1 = perfect agreement. We then computed the average of Spearman's rho correlations across all pairs of raters.

Correlations were Fisher z-standardized, averaged, and then unstandardized to express the result on the original scale.

We also estimated the agreement on the lesson or item sum scores upon which the rankings were based. We measured agreement as a type of intraclass correlation (ICC) based on a "two-way random effects model of absolute agreement between multiple raters" (McGraw & Wong, 1996, p. 32). The ICC was calculated as

$$\mathrm{ICC}(2,k) = \frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R} + \frac{\mathrm{MS_C} - \mathrm{MS_E}}{n}}, \tag{2}$$

where $\mathrm{MS_R}$ = mean square for rows, $\mathrm{MS_E}$ = mean square for error, $\mathrm{MS_C}$ = mean square for columns, n = number of subjects (lessons), and k = number of raters. The model, denoted ICC(2, k), assumes that participants were randomly selected from a larger population of raters with similar characteristics. Consequently, our findings can be generalized to any raters with the same characteristics as those chosen for the current study.

There is no universally accepted rule regarding what constitutes a "good enough" agreement, but there are some guidelines to follow. Landis and Koch (1977) offered benchmarks for interpreting Kendall's *W*: slight (< .20), fair (.20-.39), moderate (.40-.59), substantial (.60-.79), and almost perfect (≥ .80). For Speaman's rho, we adopt Cohen's (1988) guidelines (acknowledging they were intended for Pearson correlations). For the ICC values, we follow Cicchetti's (1994) classifications of poor (< .40), fair (.40-.59), good (.60-.74), and excellent (.75-1.00).

Results

RQ1. How well do T-TESS, TAP, and RATE predict a teacher's contribution to student learning?

Table 2 presents Pearson correlations for our measure of student learning attributable to instruction and the overall scores from T-TESS, TAP, and RATE. The correlations were calculated using both the original dataset with missing observations and the MICE method that imputed missing observations. In both cases, the correlations for T-TESS and TAP were effectively 0, ranging from –.05 to –.04, while the correlations for RATE were substantially higher, .29 using either the original dataset or imputed datasets. The correlation for RATE would be characterized as at the upper boundary of small, according to Cohen (1988), and none of the correlations was statistically significant. However, the Bayes factor test indicates that RATE provides moderate evidence (Bayes factor = 6.93) in favor of the alternative hypotheses (a correlation between 0 and 1). In contrast, T-TESS and TAP provide weak evidence (Bayes factors of 0.71 and 0.78, respectively) in favor of the null hypothesis (a correlation of -1 and 0). These findings suggest that, compared to T-TESS and TAP, RATE is more predictive of a teacher's contribution to student learning gains and better suited for high-stakes decisions, albeit not well-suited given its small correlation. Tables 2, 3, and 4 present results based on the full sample of 30 lessons; however, eight lessons had missing student test data, yielding N = 22 for correlations with student learning gains in the original (non-imputed) analysis. The MICE imputation procedure restored the sample to N = 30 for all imputed analyses.

Table 2

Correlations of Lesson Scores Produced by T-TESS, TAP, and RATE with Student Learning Gains (RQ1)

System	Original data			Imputed data				Bayes factor		
	r	*T*	*p*	*r*	*SE*	*t*	*p*	Strength of evidence		Hypothesis Supported
T-TESS	–.05	0.24	.811	–.05	0.19	0.28	.778	0.71	(weak)	Null
TAP	–.04	0.18	.860	–.04	0.19	0.21	.836	0.78	(weak)	Null
RATE	.29	1.34	.195	.29	0.18	1.53	.125	6.93	(moderate)	Alternative

Note: Null hypothesis: $-1 < \varrho < 0$; alternative hypothesis: $0 < \varrho < 1$.

RQ2. How similarly do T-TESS, TAP, and RATE rank teachers?

Table 3 presents two types of correlation for pairs of teaching observation systems—Pearson correlations for overall scores and Spearman's rho correlations for ranked scores. For both, the correlations between T-TESS and TAP were characterized as large by Cohen (1988) ($r = .79$, $\varrho = .85$) and statistically significant ($p < .001$). In contrast, both types of correlation between RATE and these observation systems ranged from small ($\varrho = .15$) to moderate ($r = .37$)—only the Pearson correlation with TAP was statistically significant ($r = .37$, $p = .046$). This suggests that RATE may be measuring a different construct or dimension of instruction, therefore ranking lessons differently than T-TESS and TAP. Because RATE also correlates more strongly with learning gains attributable to instruction, it suggests that the construct or dimension it measures may be better suited for high-stakes purposes (albeit still not highly suited given the small correlation).

Table 3

Correlations in Scores Across Teaching Observation System (RQ2)

Scores from			Pearson Correlation			Spearman's Rho Correlation		
			r	*t*	*p*	ϱ	*S*	*p*
T-TESS	&	TAP	.79	6.81	< .001	.85	656.45	< .001
T-TESS	&	RATE	.30	1.66	.109	.15	3801.90	.416
TAP	&	RATE	.37	2.09	.046	.33	3028.50	.078

Note: Null hypothesis: $-1 < \varrho < 0$; alternative hypothesis: $0 < \varrho < 1$.

RQ3. How closely do raters using the same observation system agree with each other?

Table 4 presents measures of rater agreement. Applying the standard suggested by Landis and Koch (1977), Kendall's *W* indicates there is fair agreement in ranked lesson scores for T-TESS (.36), TAP (.38), and RATE (.27). All values were statistically significant at the .001 level. Following Cohen's (1988) standard, the average of Spearman's rho correlations across all pairs of raters was small for T-TESS (.25), TAP (.26), and RATE (.14), and all were statistically significant ($p < .001$). Using Cicchetti's (1994) standard, the ICC estimates indicate fair agreement in (not ranked) lesson scores among raters using T-TESS (.54), TAP (.43), and RATE (.47). The 95% confidence intervals show that only the ICC for T-TESS was statistically significant at the .05 alpha level.

Table 4

Agreement of Scores and Ranked Scores Among Raters by Teaching Observation System (RQ3)

	Kendall's *W*			Spearman's Rho			Intraclass Correlation		
								95% CI	
System	*W*	χ^2	*p*	ϱ	χ^2	*p*	ICC	Lower	Upper
T-TESS	.36	62.61	< .001	.25	420.93	< .001	.54	.18	.78
TAP	.38	65.64	< .001	.26	456.71	< .001	.43	.00	.73
RATE	.27	47.66	< .001	.14	345.38	< .001	.47	-.89	.99

Note: Null hypothesis: $-1 < \varrho < 0$; alternative hypothesis: $0 < \varrho < 1$.

These results suggest that (1) estimates of agreement vary by the measure used and (2) across all measures, the agreement of lesson scores produced by RATE is on par or lower than that of the other observation systems. However, measures of agreement can be sensitive to the number of ratings. Perhaps the RATE scores would be made more accurate by simply adding more ratings from more raters.

Discussion

In this study, we examined how bundles of attributes present in three teacher observation systems—T-TESS, TAP, and RATE—may affect the suitability of their data for high-stakes decisions. Data are suitable when they advance intended purposes enough to matter without causing undue harm. Our analysis suggests that systems with attributes similar to those of RATE produce data that are more frequently suitable than systems with attributes similar to those of T-TESS and TAP. Our evidence is that in a head-to-head experiment, RATE better predicts a teacher's contribution to learning (the results for RQ1), may measure a more relevant construct or dimension of instruction (the results for RQ2), and promotes nearly equal agreement among the raters using the system (the results for RQ3). In the end, however, we question whether any of the systems we studied are suitable. Given their similarity to others in widespread use, none may be.

We cannot untangle the bundle of attributes that varied across systems and partial out the marginal effect of each. They come as a package with each system, thus defying random assignment. This is why we introduced the framework represented in Figure 1. It operationalizes our conception of a teaching observation system as a collection of attributes that interact and ultimately inform the observation scores that guide decisions. Through this lens, we can determine which attributes of a teaching observation system are within the designer's control and may make the data more suitable. These are presented in Table 1 and the surrounding text, and may be summarized as (1) rubrics with the fewest possible items and the lowest possible inference load per item, (2) calibration that is continuous and uses a process of comparison that holds raters accountable for their reasoning for assigning scores, and (3) a scoring algorithm that starts with data that are more granular than ordinal levels and normalizes item scores to make them equally important and maximally variable. We cannot say with certainty that these are the attributes that matter most, but they provide clues that may help those who design observation systems for high-stakes purposes, as well as those who must choose among them to guide decisions with real-world consequences.

Inference load is a rubric attribute we believe has been overlooked by designers, offering them an opportunity to improve the suitability of the data. Rubrics sit at the heart of teaching observation systems, and they should be designed for the people who use them. They, like all of us, are more effective when the cognitive demands of their work are moderated. The measure of

inference load we introduce offers a simple way for designers to assess the cognitive burden their systems place on raters and to determine how to reduce it.

Our study has limitations that fall into two categories. *Design limitations* include a small sample of teacher candidates from a single teacher preparation program, a narrow focus on 4th-grade mathematics lessons, and observation systems with different training modalities. *Analytic limitations* include student test data with missing scores (addressed through multiple imputation) and fewer lessons scored by observers assigned to the RATE system. These limitations should not be dismissed, but neither should they overshadow the practical consequences our results have for designers and users of teaching observation systems.

Our findings have implications for both teacher preparation programs and state-level evaluation policy. Teacher preparation programs should scrutinize the observation systems they use for certification decisions, prioritizing instruments with demonstrated predictive validity over those adopted by convention. State policymakers designing or mandating evaluation frameworks should require evidence that observation data are suitable for their intended high-stakes purposes before approving systems for widespread use.

The designers of teaching observation systems have a responsibility to make validity evidence widely available, and users have a responsibility to verify suitability in their specific contexts. The cost and effort required to gather and disseminate such evidence is high, but so is the cost of making high-stakes decisions poorly—costs borne not by designers and decision-makers but by teachers, students, and parents. This misalignment of incentives helps explain why evidence of suitability remains scarce.

Gathering and disseminating evidence of suitability imposes technical demands that many developers and school administrators may find challenging to meet. Thus, many rely on rules of thumb about reliability or predictive strength to demonstrate that teaching observation data are suitable. This is a dangerous path. Consider our results for RQ2. Small changes in estimates of rater agreement resulted in significant changes in how the estimates were characterized. For example, the score agreement for T-TESS, TAP, and RATE is characterized as fair using Kendall's *W*. To make matters worse, different terms sometimes mean the same thing across standards (e.g., small and weak, large and exceptional), and sometimes the meaning of the same term varies across standards (e.g., small may mean not useful, difficult to discern given the noise, or smaller than typically reported). Rules of thumb must also be interpreted with the understanding that different measures of the same construct, like agreement, may yield quite different estimates. In this study, calculating agreement estimates from ranked scores and average scores may help account for some of the variation we observed because ranking "throws out" some potential information.

The responsible course of action is to demonstrate the suitability of data in each context. This requires administrators, teachers, parents, and others to agree on intended purposes, what it means to advance them to a meaningful degree, what potential harm may be done, and what it means for harm to be "undue." This will result in operationalizations of suitability that vary from context to context, as well as different evidence requirements and standards against which the evidence will be compared. Rules of thumb will not suffice, even if widely reported.

We contend that our original question was indeed worth asking and worth repeating whenever educators seek guidance for high-stakes decisions from teaching observation data. By definition, high-stakes decisions have a substantial effect on the lives of teachers and those planning to become teachers. A credential may be earned or forfeited, practicing teachers may be hired, promoted, or fired, and merit bonuses may be won or lost. In all these cases, data from teaching observation systems will likely be relied upon, in whole or in part, to make the decision. It is our duty to ensure that the systems that generate these data satisfy the highest standards possible. We are

attempting to point a way toward such a standard while acknowledging that we still have a long way to go.

References

Baker, B. D., Oluwole, J., & Green, P. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*, *21*(5), 1-71. https://doi.org/10.14507/epaa.v21n5.2013

Baird, M. D., Steiner, E. D., Robyn, A., Gutierrez, I. A., Brodziak de los Reyes, I., Stecher, B. M., ... & Fronberg, K. (2019). *Intensive partnerships for effective teaching enhanced how teachers are evaluated but had little effect on student outcomes.* Rand Corporation and American Institutes for Research. https://doi.org/10.7249/RB10009-1

Basileo, L. D., & Toth, M. (2019). A state level analysis of the Marzano teacher evaluation model: Predicting teacher value-added measures with observation scores. *Practical Assessment, Research, and Evaluation*, *24*(1), 6.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3-29. https://doi.org/10.1080/09243453.2018.1539014

Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record*, *116*(1), 1-31. https://doi.org/10.1177/016146811411600102

Berliner, D. C. (2018). Between Scylla and Charybdis: Reflections on and problems associated with the evaluation of teachers in an era of metrification. *Education Policy Analysis Archives*, *26*, 54. https://doi.org/10.14507/epaa.26.3820

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review, 48*, 16-29. https://doi.org/10.1016/j.econedurev.2015.05.005

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Educational Testing Service. https://www.ets.org/Media/Research/pdf/PICVAM.pdf

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284. https://doi.org/10.1037/1040-3590.6.4.284

Close, K., Amrein-Beardsley, A., & Collins, C. (2020). Putting teacher evaluation systems on the map: An overview of state's teacher evaluation systems post-every student succeeds act. *Education Policy Analysis Archives*, *28*(58). https://doi.org/10.14507/epaa.28.5252

Cohen, J. (1988). *Statistical power for the behavioural sciences* (2nd ed.). Lawrence Erlbaum.

Council for the Accreditation of Educator Preparation (CAEP). (2020). *Standards for accreditation of educator preparation*. CAEP. https://caepnet.org/standards/2022-itp/introduction

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching.* AsCD.

Dobbelaer, M. J. (2019). *The quality and qualities of classroom observation systems.* [PhD Thesis - Research UT, graduation UT, University of Twente]. Ipskamp Printing. https://doi.org/10.3990/1.9789036547161

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Gargani, J., & Strong, M. (2014). Can we identify a successful teacher better, faster, and cheaper? Evidence for innovating teacher observation systems. Journal of Teacher Education, 65(5), 389-401. https://doi.org/10.1177/00224871145425

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224-242. https://doi.org/10.3102/01623737145375

Geiger, T. J., & Amrein-Beardsley, A. (2017). Administrators gaming test-and observation-based teacher evaluation methods: To conform to or confront the system. *AASA Journal of Scholarship & Practice*, *14*(3), 45-53. https://link.gale.com/apps/doc/A511788140/AONE?u=anon~df5af0d1&sid=googleScholar&xid=75d4aaa8

Grissom, J. A., & Bartanen, B. (2022). Potential race and gender biases in high-stakes teacher observations. *Journal of Policy Analysis and Management*, *41*(1), 131-161. https://doi.org/10.1002/pam.22352

Held, L., & Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application*, *5*(1), 393-419. https://doi.org/10.1146/annurev-statistics-031017-100307

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83(2)*, 371-384. https://doi.org/10.17763/haer.83.2.d11511403715u376

Hunter, S. B. (2024). High-leverage teacher evaluation practices for instructional improvement. *Educational Management Administration & Leadership*, *52*(4), 991-1013. https://doi.org/10.1177/17411432221112995

Kahneman, D. (2002). *Maps of bounded rationality: A perspective on intuitive judgment and choice*. http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains*. MET Project, Bill & Melinda Gates Foundation. Retrieved from ERIC: http://eric.ed.gov/?id=ED540960

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. MET Project, Bill & Melinda Gates Foundation. Retrieved from ERIC: https://files.eric.ed.gov/fulltext/ED540959.pdf

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249. https://doi.org/10.3102/0013189X17718

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498. https://doi.org/10.1037/0033-2909.108.3.480

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*, 61-95. https://doi.org/10.1007/s11092-018-09291-3

Luoto, J. M. (2023). Comparative education and comparative classroom observation systems. *Comparative Education*, *59*(4), 564-583. https://doi.org/10.1080/03050068.2023.2173917

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19–32. https://doi.org/10.1016/j.jmp.2015.06.004

Madaus, G., & Russell, M. (2010). Paradoxes of high-stakes testing. *Journal of Education*, *190*(1-2), 21-30. https://doi.org/10.1177/0022057410190001-205

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation*, *49*, 15-29. https://doi.org/10.1016/j.stueduc.2016.03.002

Marzano, R. J., Carbaugh, B., Rutherford, A., & Toth, M. D. (2013). Marzano center teacher observation protocol for the 2014 Marzano teacher evaluation model. *Palm Beach Gardens, Florida: Marzano Center/Learning Sciences International.*

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30-46. https://doi.org/10.1037/1082-989X.1.1.30

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

National Council on Teacher Quality (NCTQ) (2018). *2018 Teacher Prep Review.* Author.

National Institute for Excellence in Teaching (2012). *The effectiveness of TAP: Research summary 2012*. https://www.niet.org/research-and-policy/show/research/the-effectiveness-of-tap-research-summary-2012

National Institute for Excellence in Teaching (2017). *TAP evaluation and compensation (TEC) guide*. https://www.niet.org/assets/Resources/b3e88c1d57/niet-educator-effectiveness-services-resources-catalog.pdf

Patrick, H., French, B. F., & Mantzicopoulos, P. (2020). The reliability of Framework for Teaching scores in kindergarten. *Journal of Psychoeducational Assessment*, *38*(7), 831-845. https://doi.org/10.1177/073428292091

Putman, H., Ellis, C., Noble, R., & Peske, H. (2024). *Clinical Practice Framework.* National Council on Teacher Quality. https://files.eric.ed.gov/fulltext/ED646730.pdf

Saint-Exupéry, A. D., Galantière, L., & Cosgrave, J. O. H. (1939). *Wind, sand and stars.* Reynal & Hitchcock.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059-1074. https://doi.org/10.1068/p2810

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, *23*(5), 645-665. https://doi.org/10.1017/S0140525X00623439

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., Robyn, A., Baird, M. D., Gutierrez, I. A., Peet, E. D., Brodziak de los Reyes, I., Fronberg, K., Weinberger, G., Hunter, G. P., & Chambers, J. (2018). *Improving teaching effectiveness. Final report: The intensive partnerships for effective teaching through 2015–2016*. RAND. https://doi.org/10.7249/RR2242

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*(3), 340-359. https://doi.org/10.1162/EDFP_a_00186

Texas Education Agency (2016). *Appraiser handbook: T-TESS Texas teacher evaluation and support system*. https://teachfortexas.org/Resource_Files/Guides/T-TESS_Appraiser_Handbook.pdf

Texas Education Agency (n.d.). *Teacher guidelines: T-TESS*. Teach for Texas. https://teachfortexas.org/Teachers/TeacherGuidelines

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1-67. https://doi.org/10.18637/jss.v045.i03

Warrens, M. J. (2017). Transforming intraclass correlations with the Spearman-Brown formula. *Journal of Clinical Epidemiology*, *85*, 14-16. https://doi.org/10.1016/j.jclinepi.2017.03.005

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129-140. https://doi.org/10.1080/17470216008416717

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our national failure to acknowledge and act of differences in teacher effectiveness* (2nd ed.). The New Teacher Project (TNTP). http://tntp.org/ideas-and-innovations/view/the-widget-effect

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Brown Center on Education Policy, Brookings Institute.

Wilhelm, A. G., Rouse, A. G., & Jones, F. (2018). Exploring differences in measurement and reporting of classroom observation inter-rater reliability. *Practical Assessment, Research, and Evaluation*, *23*(1), 4.

About the Authors

Michael Strong
Texas Tech University
Michael.Strong@ttu.edu
https://orcid.org/0000-0003-2661-4775
Michael Strong is a research scientist in the College of Education at Texas Tech University. He is the former Director of Research at the New Teacher Center at the University of California, Santa Cruz, and at the Center on Deafness at the University of California, San Francisco. His current interest centers on the observation of teaching behavior and the performance of teacher pay-for-performance programs.

Jaehoon Lee
Texas Tech University
Jaehoon.Lee@ttu.edu
https://orcid.org/0000-0002-5040-843X
Jaehoon Lee, Ph.D., is an associate professor of educational psychology at Texas Tech University. His scholarly expertise spans advanced research methodologies, statistical modeling techniques, and measurement instruments applied across education, psychology, public health, and related disciplines. His recent work focuses on the evaluation and application of mixed-effects models, mixture models, Bayesian approaches, and propensity score methods for analyzing complex data sets.

John Gargani
Gargani + Co.
john@gcoinc.com
https://orcid.org/0009-0007-1409-6276
John Gargani is president of Gargani + Co., a company that helps organizations around the world achieve their social and environmental missions by measuring their impact, designing new programs, and improving performance. He is an evaluator, professor, speaker, and writer with nearly 30 years of experience directing evaluations of every type.

Minju Yi
Texas Tech University
Minju.Yi@ttu.edu
https://orcid.org/0000-0003-0491-331X
Dr. Minju Yi is an assistant professor of practice in the Department of Teacher Education at Texas Tech University. Her research is informed by classroom teaching, professional practice, and policy work in mathematics teacher preparation and development. Her scholarship focuses on: (1) developing and evaluating interventions to strengthen preservice teachers' mathematical content

knowledge and pedagogical skills; (2) investigating how teachers' mathematical understanding translates into instructional practice; (3) designing curriculum and instructional modules that integrate STEM disciplines to enhance both teacher development and student learning; and (4) examining the impact of teacher evaluation systems on teaching quality to inform equitable, evidence-based policy decisions.

Hyunjin Shim
Texas Tech University
yeshj80@gmail.com
Hyunjin Shim, Ph.D., is an associate researcher specializing in educational research on the development of knowledge and skills among preservice mathematics teachers. Her scholarly interests include curriculum and instruction, teacher preparation, and evidence-based approaches to mathematics education. She brings decades of experience as an elementary school teacher, which informs and strengthens her research.

Hyunchang (Henry) Moon
Augusta University
hymoon@augusta.edu
https://orcid.org/0000-0003-0973-2593
Dr. Hyunchang (Henry) Moon is an assistant professor of pediatrics at the Medical College of Georgia, Augusta University, and a senior editor for the *Canadian Medical Education Journal*. His scholarly work centers on learning design, educational research, and the integration of AI and emerging technologies, with a strong emphasis on evidence-based and learner-centered approaches. He contributes to the design and continuous improvement of medical curricula, supports the effective and ethical use of educational technologies, and collaborates with educators and researchers to advance teaching quality and learning outcomes.

education policy analysis archives

Volume 34 Number 33 April 7, 2026 ISSN 1068-2341