education policy analysis archives

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

Arizona State University

Volume 28 Number 128 August 24, 2020 ISSN 1068-2341

Using Multiple Measures of Teaching Quality to Strengthen Teacher Preparation[1]

Jarod Kawasaki
California State University, Dominguez Hills

Karen Hunter Quartz

José Felipe Martínez
University of California, Los Angeles
United States

Citation: Kawasaki, J., Quartz, K. H., & Martínez, J. F. (2020). Using multiple measures of teaching quality to strengthen teacher preparation. *Education Policy Analysis Archives, 28*(128). https://doi.org/10.14507/epaa.28.5001

Abstract: We argue that teacher preparation programs considering approaches to assess teaching quality should choose measures that appropriately represent the complexity of teaching, have formative value in supporting teacher candidates develop as highly qualified teachers and consider the context, mission, and people that the program desires to serve. The authors are part of a research team working with an urban teacher residency program

[1] This research was supported by a grant from the U.S. Department of Education as part of the Teacher Quality Partnership Initiative, Award U405A090159. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Education.

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

Manuscript received: 2/4/2020
Revisions received: 6/26/2020
Accepted: 7/1/2020

housed in a university's teacher education program. The increased focus on clinical experience and mandated accountability that accompany federal grants created a fertile space to experiment with different types of measures and data collection approaches, well beyond what is typical in traditional teacher education programs. In this essay, we discuss the philosophy and considerations that informed the selection of these measures in the program, and the processes that were followed to use this data in ways that consider the complexity of teaching and honor the value of data as a tool for program improvement.
Keywords: program evaluation; multiple measures; teacher learning; urban teacher residency; teacher preparation

Usar múltiples medidas de la calidad de la enseñanza para fortalecer la preparación de los maestros

Resumen: Argumentamos que los programas de preparación docente que consideran enfoques para evaluar la calidad de la enseñanza deben elegir medidas que representen adecuadamente la complejidad de la enseñanza, que tengan un valor formativo para ayudar a los candidatos a docentes a desarrollarse como docentes altamente calificados y considerar el contexto, la misión y las personas que el programa desea. servir. Los autores son parte de un equipo de investigación que trabaja con un programa de residencia de profesores urbanos ubicado en el programa de formación de profesores de una universidad. El mayor enfoque en la experiencia clínica y la responsabilidad obligatoria que acompañan a las subvenciones federales creó un espacio fértil para experimentar con diferentes tipos de medidas y enfoques de recopilación de datos, mucho más allá de lo que es típico en los programas tradicionales de formación docente. Discutimos la filosofía y las consideraciones que informaron la selección de estas medidas en el programa, y los procesos que se siguieron para usar estos datos de manera que consideren la complejidad de la enseñanza y honren el valor de los datos como una herramienta para la mejora del programa.
Palabras-clave: evaluación de programas; múltiples medidas; aprendizaje docente; residencia urbana de maestros; preparación del maestro

Usando múltiplas medidas de qualidade de ensino para fortalecer a preparação de professores

Resumo: Defendemos que os programas de preparação de professores que consideram abordagens para avaliar a qualidade do ensino devem escolher medidas que representem adequadamente a complexidade do ensino, tenham valor formativo no apoio a candidatos a professores a se desenvolverem como professores altamente qualificados e considerem o contexto, a missão e as pessoas que o programa deseja servir. Os autores fazem parte de uma equipe de pesquisa que trabalha com um programa de residência urbana para professores, inserido em um programa de formação de professores de uma universidade. O maior enfoque na experiência clínica e na responsabilidade obrigatória que acompanha os subsídios federais criou um espaço fértil para experimentar diferentes tipos de medidas e abordagens de coleta de dados, muito além do que é típico nos programas tradicionais de formação de professores. Discutimos a filosofia e as considerações que informaram a seleção dessas medidas no programa, e os processos que foram seguidos para usar esses dados de maneiras que considerem a complexidade do ensino e honrem o valor dos dados como uma ferramenta para a melhoria do programa.
Palavras-chave: evaluación de programas; múltiples medidas; aprendizaje docente; residencia urbana de maestros; preparación del maestro

Introduction

In October 2016, new national accreditation standards for teacher preparation were put in motion (U.S. Department of Education, Teacher Preparation Regulations, 2016). These standards require teacher education programs to collect and use P-12 student outcome data along with multiple measures of teaching practice. This mandate raises the bar on what counts as a well-prepared teacher and responds to a chorus of teacher education critics, including leaders of the new federal administration. Immediately after the release of these new standards, a coalition of 35 professional organizations, led by the American Association of Colleges of Teacher Education, took a strong stand[2] against these measures, in particular the student outcome measure. The new standards, they argued, were an unfunded mandate, at odds with the Every Student Succeeds Act, that would impede the diversification of the teaching profession. Vociferous debate continues as these standards make their way into the complex system of higher education, state accreditation offices, and local education agencies.

This essay details the challenge of measuring teaching quality and describes how one university-based teacher education program is attempting to "reclaim" accountability (Cochran-Smith et al., 2018) by developing a set of locally-sensitive, practical measures and using these measures to inform the learning of teacher candidates and program leaders. Through this case, we examine some of the critical conceptual and methodological issues faced in assessing teacher quality or effectiveness in a teacher education context. We highlight the sources of data and types of indicators to be collected, the extent to which these provide unique or overlapping information, and equally critically, the appropriate ways to use these indicators in combination to assess and improve both teacher preparation and performance.

The Challenge of Measuring Teaching Quality

Teaching and teacher practice are inherently complex, multidimensional constructs. Teaching involves a variety of processes and interactions that take place in the classroom and outside. Some of these processes and interactions are substantive in nature, others related to practical aspects of classroom work (e.g., daily routines, classroom management), yet others pertaining to psychological aspects of teacher-student interactions (e.g., motivation, respect, feedback). Teacher practice more broadly defined further includes a multitude of aspects of the work of a teacher outside the classroom, including among others communication with parents, administrators, and other teachers at the school, school citizenship, and contributions to the broader community. Thus, although the notion of assessing teacher effectiveness has simple intuitive appeal, in practice it involves selecting, defining, collecting information about, and making inferences involving dozens of complex component constructs (Peterson, 1987). Terms like Teaching Quality, Teacher Practice, Teacher Effectiveness, or Teacher Performance are often treated as interchangeable, but they are more appropriately seen as closely related, with important areas of overlap, but also uniqueness. Defining teacher quality or competence depends on the intended uses and context. In many cases, including teacher preparation, the richer the definition of the construct the better. A simple answer to the question of what constructs to evaluate when considering how best to appraise teacher performance could thus be "all of the above" or at least "as many of the above as is practical".

[2] https://secure.aacte.org/apps/rl/res_get.php?fid=3123&ref=rl

A growing number of districts, states, and countries are developing multiple measures evaluation systems to support high-stake inferences and decisions about teachers including hiring and tenure decisions, career advancement, and in some cases compensation (Grissom & Youngs, 2016; Reddy et al., 2015; Steinburg & Donaldson, 2016). Although public debate around these systems has focused on their approach to estimating teacher contributions to student achievement, most systems rely on multiple measures, with the majority of a teacher's rating often resting on indicators other than student achievement. In one example, the Ministry of Education in Chile developed an evaluation system that incorporated teachers' voices about the types of measures that should be included and as a result have positively deemphasized high stakes student test scores as a major determinant for teacher performance (Avalos-Beven, 2018). Yet, this system focused on summative decisions about teacher performance rather than using data to support teacher reflection and development. Multiple measures systems have been found to provide a more complete picture of teacher performance (Goe et al., 2011) and provide information to help teachers adjust and improve instruction and classroom strategies (Duncan, 2011). These and other assumptions have been investigated in the context of student assessment (e.g., Henderson et al., 2003; Schafer, 2003), but the extent to which they collectively hold in practical application for teacher evaluation, or more importantly, teacher learning, is not well understood (Stecher et al., 2018). Ford (2018) examined teachers' use of evaluation data that was specifically collected to give feedback for classroom teaching. Yet, teachers' lack of autonomy to select specific assessments and learning outcomes as well as lack of support on how to use the data prevented them from using the data as initially intended. The tension between summative and formative use of evaluation data is a major issue that underlies many of these evaluation systems.

Part of the complexity lies in understanding how to use and interpret the measures—in particular whether and how to combine them. There are a variety of approaches for combining measures for the purpose of evaluating teachers (e.g., Bell et al., 2018), and which one we choose can be of consequence for the properties of the resulting indicators, and the inferences we ultimately draw about teachers. At least four approaches have been proposed in the literature in psychology and student assessment for combining multiple measures that reflect different attributes of a broader target construct. These include conjunctive and disjunctive evaluation models, a variety of compensatory linear models, and hybrid approaches that combine more than one of these (Henderson et al., 2003).

In conjunctive models, individuals must meet a required standard (i.e., pass) on all individual measures to succeed, whereas the less stringent disjunctive model requires only passing one or more measures. Extending these models to classroom settings, a teacher would need to meet success criteria in all measures (e.g., observations, surveys, value-added models [VAM]) in a conjunctive model of teacher quality, while passing any one of these measures would suffice in a disjunctive model. Compensatory models offer an alternative method that relies on aggregates or linear combinations of measures. These models therefore allow high performance on a measure to compensate for lower performance on another (e.g., a teacher with high observation and survey scores might obtain a successful overall grade despite lower VAM scores; or high observation and VAM scores might compensate for low survey scores). Weighted models can be used to weight indicators according to theoretical importance and reliability (i.e., more reliable measures have greater weight in the composite). Finally, canonical or factor analysis models may be used to examine empirical correlations among indicators and create composites to maximize shared variance, reliability, or stability.

It is clear that there are flaws in each of these data combination models discussed above and that no single model will yield the "best" results. We argue that to avoid these potential false positives or negatives, one should consider the specific purposes and uses of the data relative to the priorities, beliefs, and values of the communities in which that data is to serve. For example, a teacher education program might create a composite measure in order to determine which students should graduate, yet to inform practice might require a different approach. Mehrens (1989) suggests that before committing to a model for combining data, one must first ask whether the data should be combined at all. It may be better to consider how each of the measures provides some specific insight into the rich and complex picture of classroom teaching and can be used to inform efforts to improve teacher performance (Schmidt & Kaplan, 1971). We turn now to our local effort to develop and use multiple measures in combination to improve and document the impact of the program.

Defining Teaching Quality for a Local Context

Urban teacher residency programs have emerged as a promising hybrid of university-based and alternative preparation programs, with the potential to transform teacher preparation in viable, transformative ways to promote teaching quality, student learning and educational equity (Berry et al., 2008; Guha et al., 2016; Klein et al., 2013). Inspiring Minds through a Professional Alliance of Community Teachers (IMPACT) was created in 2009 in partnership with the UCLA Teacher Educatin Program with the goal of preparing highly qualified community teachers and urban school teacher-leaders for high-need subject areas of elementary and secondary math, science, and early childhood education. Student teachers—referred to as residents—work in cohort teams, engaging in a variety of courses including methods, learning theory, language acquisition, and others. In addition, residents concurrently participate in a yearlong residency with a mentor teacher in one of 32 IMPACT residency schools and early childhood centers—sites chosen based on their commitment to collaboration, teacher learning, and personalized education. The creation of the urban teacher residency program provided an opportunity to reimagine our approach to measuring teaching quality. Our research team of teacher educators, researchers, and evaluators, wanted to complement summative evaluations of teachers required by the state (and federal funding sources) with more formative approaches that used multiple sources of information about teaching for learning and development, both by individual teachers and the program as a whole.

Denzin's (1978) foundational work on triangulation guided our effort to define, and subsequently approach the measurement of the variety of complex processes and interactions comprised in a robust conceptualization of teaching quality. Denzin proposes four types of triangulation—data, investigator, theory, and methodological—to help researchers capture complex phenomena. We used methodological triangulation to choose methods and measures that had different strengths and weaknesses, thereby increasing the credibility of our findings. In choosing this approach for assessing teaching practice or effectiveness, we sought to balance the strengths and limitations of each type of instrument or measure, to ensure that they collectively and appropriately represented the key aspects of teaching of interest. Although complementarity of information was a key factor in selecting data collection tools, it was also important to ensure that the measures individually conformed to minimal accepted standards of measurement quality. Table 1 presents five standards and accompanying guiding questions that were proposed by Goe et al. (2008) and adopted for developing instruments to measure and evaluate teaching in IMPACT.

Table 1

Measurement standards and guiding questions

Measurement Standard	Guiding Questions
Reliability	How accurately do scores reflect individual standing with respect to the qualities being measured?
Validity	Is there empirical evidence that the measures relate in expected ways among themselves and with important external variables, such as student achievement? To what extent can we generalize from a specific measure to inferences about a teacher's overall practice or effectiveness?
Credibility	Will stakeholders accept the measures as a reflection of dimensions of teaching they find meaningful and critical to the profession? Do the measures have face value?
Coherence	Do the measures work together to provide mutually reinforcing information? Are measures of teaching quality integrated into a system of feedback, professional development and classroom practice?
Consequences	What evidence do we have to support using these measures in the intended context? What are the intended and some potential unintended consequences of using them in this way?

Answering these measurement questions requires careful planning, and thoughtful, cooperative, and challenging work by researchers, teacher educators, and other stakeholders. We began this process in our local context with a discussion about whether to use an existing framework to define teaching quality (e.g., Danielson, 2013; La Paro et al., 2004) or develop a more contextually-sensitive definition. In the end, we decided to privilege the value of common, local understandings about equitable teaching and humanizing pedagogy (e.g., Bartolome, 1994; Freire, 2000) as well as research-based knowledge about science and mathematics instruction. We aimed for a rich definition of teaching quality that aligned with the values and principles of the social justice-oriented teacher residency program and a commitment to capture as many of the relevant teaching constructs as was practical. Our definition focused on four dimensions: 1) teaching with academic rigor, 2) promoting content discourse, 3) ensuring equitable access to content, and 4) creating a safe and positive classroom ecology. We invested significant effort in refining an observation rubric developed from these four dimensions and conducted a series of generalizability studies to establish its reliability (Nava et al., 2019). It was vitally important to identify and articulate these dimensions in a way that could be tracked and assessed over time in order to support the early career learning of new teachers as well as help program leaders understand and be accountable for the quality of teaching practice their graduates take into the field. The observation rubric and its definition of good teaching grounded our selection and development of six additional measures.

Seven Measures Measuring Teaching Quality

To help teacher educators understand and assess the teaching quality of IMPACT residents, the research group decided to collect information from seven different sources:1) observation rubrics, 2) teaching artifacts, 3) instructional logs, 4) VAM, 5) pedagogical content knowledge, 6)

surveys of teachers and mentors, and 7) teacher portfolios (see also Table 2). These seven measures were designed to capture different types of information about teaching practice and quality and were aligned with the four dimensions of the IMPACT framework for teaching and learning.

Table 2

Date collection timeline for multiple measures of teaching quality

Residency Year (pre-service)			Year 1 Teaching (in-service)		
Fall	Winter	Spring	Fall	Winter	Spring
Observation	Observation	Observation	Observation	Observation	Observation
PCK*		Mentor Evaluation	Instructional Quality Assessment	PCK*	
	Logs	PACT/ edTPA			Logs
Resident Survey 1		Resident Survey 2		Resident Survey 3	Value-added scores

*Only two rounds of pre- and post- Pedagogical Content Knowledge (PCK) Assessments was administered to cohort 1 & 2 math and science residents: Mathematics Knowledge for Teaching (MKT) and Assessing Teacher Learning About Science Teaching (ATLAST).

We describe each of the measures and briefly discuss how they advanced program improvement in a residency context.

Classroom Observations

The observation framework was developed to operationalize the four dimensions of teaching quality in terms of eleven aspects of teacher classroom practice (Nava et al., 2019). For example, one of the content discourse sub-dimensions focuses on teachers' facilitation of participation structures, based on research that getting students to talk about mathematics or science takes careful orchestration of tasks, norms, and fluent facilitation from teachers (Franke et al., 2007).

Resident Survey and Mentor Evaluation

Residents completed an initial survey, an end of the residency year survey, and then an end of the program survey. Each of them consisted of items that asked about residents' beliefs about teaching and experiences in the program. The mentor evaluation survey was administered at the end of the residency year and asked the mentor to evaluate their resident on a series of items aligned with the four dimensions of teaching quality. Mentors were also asked about their experiences in the program.

Instructional Logs

Logs consisted of a two-week series of daily short surveys. In these surveys, residents self-reported their use of formative assessment strategies emphasized in their university methods course. All courses in IMPACT were designed or refined based on the four dimensions of teaching quality and thus, the formative assessment strategies in the logs reflected these dimensions as well. The logs were administered once during their resident year and again during their first full year of teaching in order to see if there was any change in strategies teachers used.

Instructional Quality Assessment (IQA)

The IQA (Matsamura et al., 2006) was adapted for use at the end of the residency program, when residents were in their first full-year of teaching. The IQA is intended to promote integration of theory and practice in learning "rigorous content and pedagogy" (Crosson et al., 2006, p. 1). Residents identified an assignment they gave to students, completed a questionnaire detailing the teaching context for this assignment, and attached six associated samples of student work. This evidence is scored by trained raters using a rubric adapted for the residency program's definition of teaching quality and comprising four dimensions: (1) Rigor-Potential of the task, (2) Expectations-Clarity, (3) Expectations-Communication, (4) Equitable Teaching-Relevance.

Pedagogical Content Knowledge (PCK) Assessments

We adopted two measures, one for math and one for science to assess residents' PCK, 1) the Mathematical Knowledge for Teaching (MKT) developed by the University of Michigan and the Assessing Teacher Learning About Science Teaching (ATLAST) developed by Horizon Research. It was expected that residents might show growth in their math or science PCK over the 18 months in the program. The pretest was administered at the beginning of the program and the posttest was administered three months into teachers' first full year of teaching.

Performance Assessment for Credentialed Teachers (PACT)

The PACT (now called the edTPA) is a teacher performance assessment that pre-service teachers must pass in order to earn their teaching credential. Pre-service teachers design a series of lessons and select specific moments to video record. An external assessor watches the videos, with writings from the pre-service teachers' lessons and classroom artifacts (e.g., student work, handouts, powerpoint slides) to assess their skills in planning, instruction, assessment, academic language and reflection.

California Standards Test Scores

Test scores were collected from the residents employed by our local district partner. The scores were collected from the residents' classes during their first full year of teaching. A value-added model called "academic growth over time" was used to examine the individual progress for each student from the standardized test from the previous year. The model also considers contextual factors that might influence test scores. All of the scores were given to us by our local district partner.

Each measure tells us something about how a resident is performing in one or more of the four dimensions of teaching quality. Collecting data through direct observation in the classroom can generally yield rich evidence of instruction, and in principle has a high face value for assessing teaching practice (e.g., teacher educators documenting the frequency and quality of residents' questioning strategies or the extent to which residents use questions that promote student discourse). In practice, however, the value of this approach for specific programmatic uses is directly mediated by factors such as the knowledge, background, and training of the observers, the number of observations, and the specific lesson and times chosen for the observation. If an observation takes place on an atypical day or is recorded by a novice observer, we may not get an accurate representation of a teacher's practice.

To account for the potential error associated with observers and times common to classroom observations, we included measures that offered a different balance of strengths and limitations. Specifically, residents completed daily logs that kept track of formative assessment practices over a complete two-week instructional unit. The instructional logs provide insight into the

corpus of discourse strategies that a resident might use across an instructional unit. This approach faces its own particular limitations and concerns related to the depth of evidence obtained from survey measures and the veracity of self-reports more generally. Yet on the other hand, it allowed the program to monitor for all residents a specific set of instruction practices of interest every day over a substantively meaningful period of time. This volume and granularity of evidence is not as viable in practice with classroom observations.

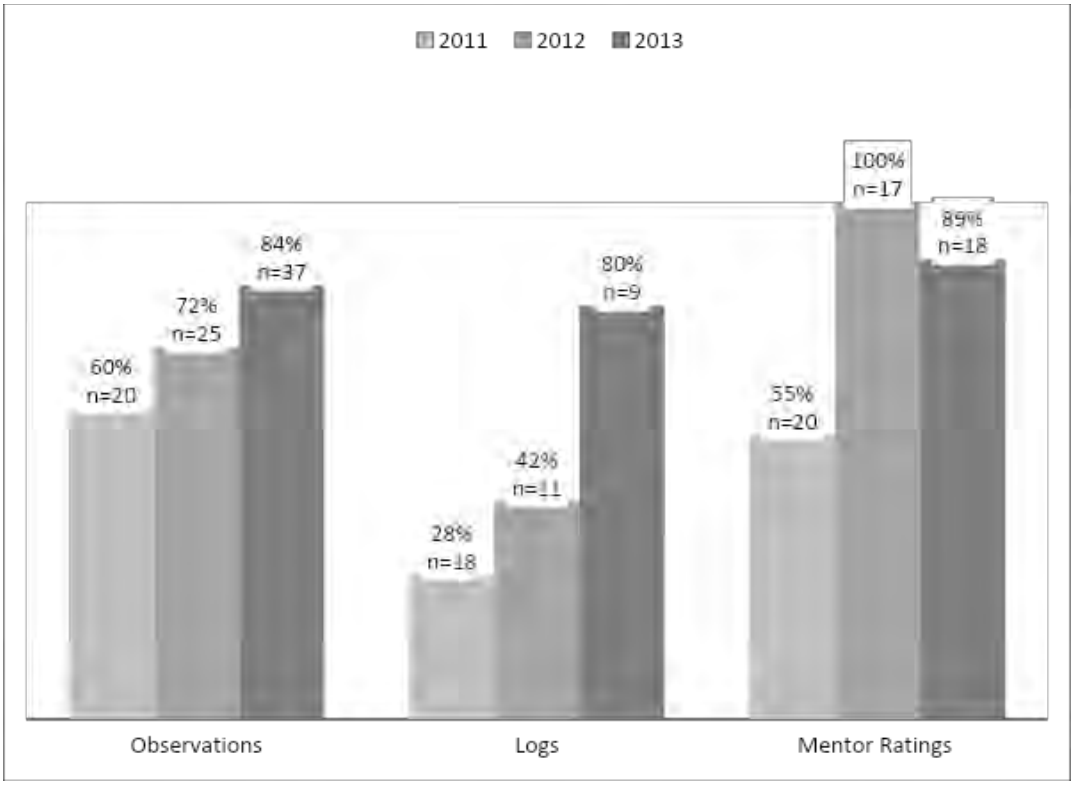
Furthermore, we collected artifacts of residents' teaching (i.e., IQA). By collecting a classroom lesson from a resident along with samples of the student work generated from the lesson, a researcher can evaluate how well the resident is promoting content discourse evidenced in the details of a lesson plan and accompanying student work samples. We also administered surveys for mentors to evaluate their residents and for residents to self-assess on the different dimensions of teaching quality. The surveys provide holistic summative judgements about teaching quality that many of the other measures do not provide.

Collecting and Using Data for Teacher Development

To illustrate how IMPACT used local measures for formative use within the program, Figure 1 shows how we analyzed the observation ratings, instructional log data, and mentor evaluation in combination to draw inferences on how residents promoted student talk about math and science in classrooms (see Quartz et al., 2017 for an in-depth discussion of this example).

Figure 1

Measures of student teachers' performance on promoting content discourse, % proficient across 3 cohorts (2011-13)



After collecting data on the seven multiple measures over time, we learned about the unique challenges associated with each measure. For example, we struggled with missing data on the value-added measure because several residents taught untested subjects or had insufficient test score data. We also struggled with consistently collecting the instructional log data. Despite daily reminders, resident response rates were often low due, in part, to the demands of teaching. Faculty and residents shared that it was difficult to sustain completing the logs for two full weeks. In addition, residents indicated that the daily log had too many questions that deterred them from completing it after the first few days. In response, we shortened the data collection period from two weeks to one week and we narrowed down the log items to simply include a list of the core assessment practices that were emphasized in their methods class. Our response rates improved dramatically allowing us to include the log data in our evaluation of the program and discuss these data in residents' methods class to inform their learning about classroom assessment practices.

We continued this practice of asking residents for feedback on the measures and our data collection practices. For example, we interviewed a few residents about the IQA and found that overall, they felt that the IQA ratings were fair and accurate. Yet, they discussed the overwhelming burden with preparing the portfolio and felt the ratings were not that useful because the ratings and feedback were received over a month after the lesson was taught. This led us to make the IQA reports more detailed and incorporate support from teacher educators to debrief the reports.

The most consistently collected measure was the observation rubric because it was part of the daily work of the teacher educators. Residents also used the observation rubric during methods class to rate video recordings of their own and their peers teaching. Using the rubric in class was a way for teacher educators to further support residents' understanding of the dimensions and their associated instructional strategies. As we have argued, these measures capture different parts of teaching quality at different depths and granularity. Thus, using these measures *in combination* get at the rich complexity of teaching quality in ways that are theoretically and empirically justified and can be used for program improvement.

Validity and Multiple Measures

IMPACT made the decision to look at data collected from these seven measures with an eye towards their formative use in assessing teacher quality for program improvement. We argue that in this context, the data are messy and there a number of constraints that prevent us from using traditional psychometrics where validity is commonly seen as unitary and purpose dependent, and *validation* entails formulating an *interpretive argument* for the intended inferences derived from measures, and providing sufficient evidence to support this argument (Kane, 2006). This evidence includes both theoretical and conceptual justification for the constructs involved, and empirical evidence of the properties of the indicators (i.e., reliability and accuracy, expected patterns of inter-correlation, predictive power over criterion measures). As with individual measures, assessing validity in a multiple measure context requires assumptions about and careful operationalization of the theoretical construct being measured (i.e., teacher quality). Different uses require different validity arguments, and evidence--uses that carry serious consequences require the greatest extent of theoretical and empirical support.

Importantly, this traditional approach to validation is notoriously hard to implement in practice when the measures are locally developed and administered, and used and refined continuously for formative purposes in dynamic contexts. Thus, in developing a system of multiple measures for local use in IMPACT, we considered how to retain the core logic of the validity argument, but broaden our conceptions of reliability, rigor, evidence, and triangulation from both a quantitative and qualitative perspective, with a focus on sustained, systematic formative uses.

Our validity argument is a reinterpretation of Kane's notion where *validation* entails the collection of quantitative and qualitative evidence tied to particular uses and a specific context (i.e., program improvement). IMPACT measures were grounded in a conceptual framework about high quality math and science teaching, and further conceptualized through the lens of equitable teaching and humanizing pedagogy. Although we were able to conduct a pilot generalizability study with the observation framework (Nava et al., 2019), this process took a substantial amount of time and resources to complete. The IQA, PCK assessments, and PACT/edTPA were established measures and have in principle documented their own validation warrants (e.g., Pecheone & Chung, 2006). Importantly, however, standard measurement practice establishes that these warrants do not carry over to new and different uses and contexts. Moreover, with local measures, data is often unavailable to assess patterns of intercorrelation or predictive power among measures due to limited sample sizes, missing data, inconsistent granularity and units of measurement, and adaptations to the measures themselves. Because of this complexity, we focused on evidence that the measures were sensitive to change and behaved in ways that were consistent with expert local knowledge and perceptions on the ground.

Implications

Our case study depicts one teacher education program's effort to navigate the tension between collecting and using data for compliance versus learning purposes. We have described this program's attempt to design an assessment system that meets state and national standards while also supporting professional learning for teacher candidates. Large scale evaluation systems (e.g., Measures of Effective Teaching; Kane et al., 2013) are well funded and designed to meet the highest standards for measurement quality. Yet, these large-scale systems are not designed to provide information to support teacher learning. With the politically heightened challenge for teacher preparation programs to be held accountable for student outcomes, many programs devote their resources to collecting data for accountability purposes, but lack the capacity to use this data for program improvement and teacher learning (Tatto et al., 2016). Designing an assessment system that informs pre-service teacher learning requires careful attention to the types of data that will facilitate thoughtful reflection and the processes that will help pre-service teachers engage with that data. This may include attention to how teachers can have input and agency in deciding what and how to measure their own learning (Lavigne & Good, 2020). Our explicit aim was to design an assessment system that considered the standards for measurement quality, yet prioritized measures and data that informed program, teacher educator, and resident learning.

Our case study highlights two key considerations for programs considering a redesign of their assessment system to support teacher learning. First, it is important to consider the assumptions and consequences (intended and unintended) of the various approaches for combining measures. There is no *best, fully scientific* and objective way to weight or otherwise combine multiple measures to evaluate teachers. A certain degree of arbitrariness is involved in any of the frameworks discussed; the question is not whether subjective, non-scientific considerations are involved, but where, how, and to what extent. Making explicit the assumptions and judgments that informed the design of a teacher evaluation system, its' goals, components, and procedures will enable us to better monitor the operation of the system, make necessary adjustments and improvements, and ultimately offer evidence supporting the validity of the inferences about teacher effectiveness, and the usefulness of the system for improving teacher practice.

Second, it is imperative to consider how the system, measures, and data answer important questions about various points along the trajectory of teacher development. Multiple measures

systems have the potential to support a culture of reflection, improvement, and accountability among teachers, teacher educators, and the many other educators seeking to deepen student learning. These measures and data need to address teacher learning at various time points across the academic year. Then, the data need to be collected, organized, and visualized in ways that speak directly to the questions around teacher learning that coursework and fieldwork are aiming to support. One such example comes from Yeager & colleagues (2013) who argue for measurement for improvement—specifically practical measures of everyday processes that can evaluate whether a change led to an improvement.

The possibility of replacing compliance-focused evaluation systems with more meaningful efforts to assess and improve teacher practice and performance is a welcome development in education policy (Richmond et al., 2019). Yet, good measures take time to develop, solid systems based on these measures take longer to test and implement, and the consequences of specific uses of these systems are largely unknown and will take longer to assess. As we have argued in this essay, measuring teaching quality in ways that inform and improve practice involves making careful theoretical considerations, and methodological decisions. As Lewin aptly stated, "There's nothing so practical as a good theory."

Acknowledgements

We would like to acknowledge all of the IMPACT residents and mentors for their commitment, capacity, and resilience to enacting social justice theories and pedagogies within their classrooms and schools. We thank them for allowing us to learn about formative feedback with them. We also want to thank the IMPACT research team, Kathryn Anderson-Levitt, Mollie Appelgate, Helen Davis, Danny Dockterman, Noelle Griffin, Sunanda Kushon, Jaime Park, Imelda Nava, Nichole Rivera, Jon Schweig, and Jia Wang, for the contribution and feedback on this work.

References

Avalos-Bevan, B. (2018). Teacher evaluation in Chile: Highlights and complexities in 13 years of experience. *Teachers and Teaching, 24*(3), 297-311. https://doi.org/10.1080/13540602.2017.1388228

Bartolomé, L. (1994). Beyond the methods fetish: Toward a humanizing pedagogy. *Harvard Educational Review, 64*(2), 173-195. https://doi.org/10.17763/haer.64.2.58q5m5744t325730

Bell, C., Santibañez, L., & Taylor, E. S. (2018). *Improving teacher practice. Getting down to facts II.* Stanford University.

Berry, B., Montgomery, D., & Snyder, J. (2008). *Urban teacher residency models and institutes of higher education: Implications for teacher preparation.* Center for Teaching Quality.

Cochran-Smith, M., Carney, M. C., Keefe, E. S., Burton, S., Chang, W. C., Fernández, M. B., Miller, A. F., Sanchez, J. G., & Baker, M. (2018). *Reclaiming accountability in teacher education*. Teachers College Press. https://doi.org/10.1007/978-981-13-2026-2

Crosson, A. C., Boston, M., Levison, A., Matsumura, L. C., Matsumura, L. C., Resnick, L. B., Kim, M., & Junker, B. W. (2006). *Beyond summative evaluation: The Instructional Quality Assessment as a professional development tool*. CSE Technical Report 691. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Danielson, C. (2013). *A collection of performance tasks & rubrics: Primary mathematics*. Routledge. https://doi.org/10.4324/9781315853154

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. McGraw-Hill.

Duncan, A. (2011) *Duncan Tells Teachers: Change is Hard*. Homeroom, Retrieved December, 09, 2012 from http://www.ed.gov/blog/2012/08/duncan-tells-teachers-change-is-hard/

Ford, T. G. (2018). Pointing teachers in the wrong direction: Understanding Louisiana elementary teachers' use of *Compass* high stakes teacher evaluation data. *Educational Assessment, Evaluation, and Accountability, 30*(3), 251-283. https://doi.org/10.1007/s11092-018-9280-x

Franke, M. L., Kazemi, E., & Battey, D. (2007). Mathematics teaching and classroom practice. *Second Handbook of Research on Mathematics Teaching and Learning*, *1*, 225-256.

Freire, P. (2000). *Pedagogy of the oppressed*. Bloomsbury Publishing.

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

Goe.L, Holheide, L., & Miller, T (2011) *A practical guide to designing comprehensive teacher evaluation systems*. National Comprehensive Center for Teacher Quality.

Grissom, J. A., & Youngs, P. (2016). *Improving teacher evaluation systems: Making the most of multiple measures*. Teachers College Press.

Guha, R., Hyler, M.E., & Darling-Hammond, L. (2016*). The teacher residency: An innovative model for preparing teachers*. Learning Policy Institute.

Henderson-Montero, D., Julian, M, & Yen, W (2003) Multiple perspectives on multiple measures. *Educational Measurement: Issues and Practice*, *22*(2), 6-6. https://doi.org/10.1111/j.1745-3992.2003.tb00121.x

Kane, M. T. (2006). Validation. *Educational Measurement*, *4*, 17-64.

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. In *Research Paper. MET Project*. Bill & Melinda Gates Foundation.

Klein, E. J., Taylor, M., Onore, C., Strom, K., & Abrams, L. (2013). Finding a third space in teacher education: Creating an urban teacher residency. *Teaching Education, 24*(1), 27-57. https://doi.org/10.1080/10476210.2012.711305

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, *104*(5), 409-426. https://doi.org/10.1086/499760

Lavigne, A. L., & Good, T. L. (2020). Addressing teacher evaluation appropriately: A research brief for policymakers. *APA Division 15 Policy Brief Series, 1*(2), 1-7.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment*. CSE Technical Report 681. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). http://cresst.org/publications/cresst-publication-3052/

Mehrens, W. (1989) Combining evaluation data from multiple sources. In Millman and L. Darling-Hammond (Eds) *The new handbook of teacher evaluation: Assessment of elementary and secondary school teachers* (pp. 322-336). Sage. https://doi.org/10.4135/9781412986250.n19

Nava, I., Park, J., Dockterman, D., Kawasaki, J., Schweig, J., Quartz, K.H., & Martinez, J.F. (2019). Measuring teaching quality of secondary mathematics and science residents: A classroom observation framework and pilot generalizability study. *Journal of Teacher Education, 70*(2), 139-154. https://doi.org/10.1177/0022487118755699

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education*, *57*(1), 22-36. https://doi.org/10.1177/0022487105284045

Peterson, K. (1987) Teacher evaluation with multiple and variable lines of evidence. *American Education Research Journal*, *24*(2). 311-317. https://doi.org/10.3102/00028312024002311

Quartz, K.H., Martínez, J.F., & Kawasaki, J. (2017). *Using multiple measures of teaching quality to improve the preparation of urban teachers.* Sage Research Methods Case. https://doi.org/10.4135/9781473980143

Reddy, L. A., Kettler, R. J., & Kurz, A. (2015). School-wide educator evaluation for improving school capacity and student achievement in high-poverty schools: Year 1 of the school system improvement project. *Journal of Educational and Psychological Consultation*, *25*(2-3), 90-108. https://doi.org/10.1080/10474412.2014.929961

Richmond, G., Salazar, M. D. C., & Jones, N. (2019). Assessment and the future of teacher education. *Journal of Teacher Education, 70*(2), 86-89. https://doi.org/10.1177/0022487118824331

Schafer, W. D. (2003), A state perspective on multiple measures in school accountability. *Educational Measurement: Issues and Practice*, *22*, 27–31. https://doi.org/10.1111/j.1745-3992.2003.tb00125.x

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24,* 419-434. https://doi.org/10.1111/j.1744-6570.1971.tb00365.x

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

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., Abby, R., Baird, M.D., Gutierrez, I. A., Peet, E. D., Brodziak de los Reyes, I. Fronberg, K., Weinberger, G., & Hunter, G. P. (2018). *Intensive partnerships for effective teaching enhanced how teachers are evaluated but had little effect on student outcomes.* [RB-10009-1-BMGF, 2019]. RAND Corporation, https://doi.org/10.7249/RB10009

Tatto, M. T., Savage, C., Liao, W., Marshall, S. L., Goldblatt, P., & Contreras, L. M. (2016). The emergence of high-stakes accountability policies in teacher preparation: An examination of the US Department of Education's proposed regulations. *Education Policy Analysis Archives*, *24*(21). https://doi.org/10.14507/epaa.24.2322

Yeager, D., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement.* Carnegie Foundation for the Advancement of Teaching.

About the Authors

Jarod Kawasaki
California State University, Dominguez Hills
jakawasaki@csudh.edu
http://orcid.org/0000-0001-9076-4108
Jarod Kawasaki is an assistant professor of teacher education in the College of Education at California State University, Dominguez Hills. He is also the assessment coordinator responsible for managing the assessment and evaluation work for the five divisions within the college. His research interests are teacher learning, social justice in science education, program evaluation, and evaluation data use.

Karen Hunter Quartz
University of California, Los Angeles
quartz@ucla.edu
https://orcid.org/0000-0003-2605-4042
Karen Hunter Quartz is Director of the UCLA Center for Community Schooling, Research Director of the UCLA Community Schools, and Adjunct Professor in the UCLA Graduate School of Education and Information Studies. Her research interests include community schooling, teacher quality, retention, and career development, and urban school reform.

José Felipe Martínez
University of California, Los Angeles
jfmtz@ucla.edu
José Felipe Martínez is an associate professor of research methodology at the Graduate School of Education and Information Studies, University of California, Los Angeles. His research focuses on measurement issues in program, school, and teacher evaluation, and developing and using indicators of instruction and classroom climate.

education policy analysis archives

Volume 28 Number 128 August 24, 2020 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.