Examining the Validity of a State Policy-Directed Framework for Evaluating Teacher Instructional Quality: Informing Policy, Impacting Practice

By: Edward F Sloat, Ed.D (esloat@asu.edu)

Introduction

Public education agencies throughout the country have utilized varied approaches for evaluating teachers' instructional practice (McClellan, 2012; Learning Sciences Marzano Center, 2012; Marzano, 2011; Strong, 2012; MET Project, 2010). Until recently, however, these evaluation approaches were not directly integrated into state or national accountability systems. In 2009, the role of teacher evaluation in high stakes accountability environments substantively changed under the federal Race to the Top (RTTT) program (USDOE, 2009). To qualify for funding under this program, states were required to implement systemic teacher evaluation systems based, in part, on quantitative measures of student learning. Arguably, these legislative initiatives fundamentally altered the context by which local education agencies (LEAs) conceptualize, implement, and utilize teacher evaluation systems. That is, the structural components for evaluating classroom teachers became codified into law, breaking a long tradition of local oversite and community decision-making (Marzano, Frontier, & Livingston, 2011; McNergney, 2002; Weisberg, Sexton, Mulhern, & Keeling, 2009; Wise, 1984).

In spring 2011, the Arizona state legislature passed Senate Bill 1040 (Arizona Revised Statutes §15-203 (A) (38)) requiring all public school districts to evaluate teacher instructional quality (TIQ) using (at minimum) quantitative measures of academic progress and measures of teachers' professional practice. Additionally, evaluation outcomes became part of each teacher's permanent record, subject to review by future employers. Results were required to be integrated into local pay for performance compensation plans (House Bill (HB) 2823, 2012). In this regard,

the policy action transformed teacher evaluation into an inherently high stakes consequential activity with the potential of impacting tenure, professional stature, and personal well-being.

Importantly, Senate Bill 1040 did not mandate a specific methodology or set of measures for evaluating teacher effectiveness. Rather, it directed the State Board of Education (SBE) to develop a general framework under which districts would design and implement their own evaluation systems and related metrics within loosely defined data quality guidelines (Arizona Department of Education (ADE), 2012). As a result, LEAs throughout Arizona were simultaneously placed in the position of legislative compliance, while at the same time assuming responsibility for developing, implementing, and justifying their measurement systems.

Problem Statement

This context presents considerable ethical, organizational, and technical problems of practice where change initiatives and accountability requirements originate externally, while implementation details are locally carried out through internal design and decision-making. To the extent that policy-directed legislative frameworks fail to align with local values, important issues of practice may occur. For this reason, examining the warrants ascribed to policy-imposed evaluation systems becomes an ethical necessity demanding methodological rigor, reflection, and a concern for individual well-being. Arguably, as consequential stakes increase, the need to critically evaluate the inferential authority of such systems also increases (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014; Linn, 2008).

To this end, it seems reasonable to suggest that the examination of validity evidences is foundational for informing localized decision-making and ensuring efficacy of such systems (Danielson, 2010).[1] Unfortunately, little evidentiary criteria regarding acceptable levels of

reliability and validity currently exist in the literature at either the local or state policy levels (ADE, 2012; Amrein-Beardsley, 2014; Herlihy, Karger & Pollard, 2014).

Study Focus

To this end, the researcher examined multi-faceted validity evidences associated with implementing Arizona's policy-directed teacher evaluation framework.[2] The data originated from within a large suburban school district in Phoenix, Arizona.[3] The primary purpose of this examination was to (1) assess the inferential warrant for identifying and distinguishing good/effective teachers implemented under the state-imposed policy framework, (2) inform local decision-makers in the design, implementation, and efficacy of their system, (3) affect changes/improvements in the local evaluation process, and (4) contribute to the larger discourse regarding this particular state's education accountability policy, as well as state policy in general.

Research Questions

Four primary and twenty-two supporting research questions were advanced within this study. The four primary questions were: (1) *To what degree do validity evidences of the LEA's teacher evaluation system support inferences on Teacher Instructional Quality (TIQ)?* (2) *How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?* (3) *To what degree have the perspectives and behaviors of stakeholders been incorporated into, and influenced, the system's implementation?* and (4) *How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader?*

Theoretical Frameworks

The researcher framed this study within multiple theoretical and conceptual frameworks including epistemology/world views, the social context of education in society, contemporary perspectives of validity theory, and theories of organizational change and sustainable innovation.

Epistemology/World Views: The researcher embraced a pragmatic world-view, addressing postpositivist, constructivists, and advocacy/participatory perspectives in the research design (Creswell, 2009; Gergen, 2009; Stringer, 2007). The approach assembled evidences across multiple methodological traditions in order to rigorously examine the many-faceted dimensions of construct validity (AERA et al., 2014; Kane, 2001; Messick, 1989). For the questions posed herein, this necessitated examination of evidences grounded in measurement theory (postpositivist), multi-level stakeholder perspectives (constructivists), and exposure of teacher 'voices' as recipients of the evaluation policy, activity, and consequence (advocacy).

Social Context of Education in Society: The act of teacher evaluation is embedded within a larger context in which teachers, educators, and policy makers internalize the purpose, practice, and outcomes associated with public schooling (Dewey, 1900; Gergen, 2009; Good, 1999; McGee-Banks & Banks, 1997). Here, the researcher hypothesized that different stakeholders held different perspectives and these perspectives affect the nature and structure of the evaluation process. As such, these dimensions were explicitly embedded in the research context.

Validity Theory: The researcher adopted a contemporary view of validity (AERA et al., 2014; Cronbach & Meehl, 1995; Kane, 2001; Messick, 1989; Shepard, 1993, 1997). Here, validity is seen as a single unifying construct under which many forms of evidences collectively contribute to an understanding of the evaluation context. It is this assembly of evidences that permit assessment of the inferential warrant for making claims of instructional quality.

Organizational Change & Innovation: The implementation of new, high stakes, evaluation systems also concerns aspects of sustainable innovation and organizational change. In this regard, the researcher also used these theories to examine stakeholder perspectives on evaluation purposes, processes, efficacy, and the integration of teacher 'voice' (i.e. influence) in the design process (Fullan, 2009; Hall, Loucks, Rutherford, & Newlove, 1975; Hall & Hord, 2011, Kotter, 1996; 2010; Rogers, 2003).

Measurement Frameworks

Consistent with state policy requirements, the LEA's teacher evaluation system adopted the Danielson Framework for Teaching (FFT) as the basis for assessing the professional practice (PP) of teachers (Danielson, 2010).[4] Under this framework, all site principals were certified as trained evaluators using the Teachscape Focus on-line training and calibration system (Teachscape, 2016).[5] In addition, the researcher constructed hierarchical value-added residual gain models (VAMs) to statistically estimate student academic growth in reading, mathematics, and science using data from state-administered standardized tests. These two metrics (PP & VAM) were then combined to form a single weighted composite Teacher Instructional Quality (TIQ) score.[6] Based on a teacher's placement along the TIQ composite scale, one of four interpretive ratings were assigned: Ineffective, Developing, Effective, or Highly Effective.[7]

Methods

The researcher applied a single-phase, concurrent, mixed-methods design triangulating multiple sources of qualitative and quantitative evidence onto a single validation construct: Teacher Instructional Quality (TIQ; Gelo, Braakmann, & Benetka, 2008; Creswell, 2009; Plano-Clark & Creswell, 2010; Greene, 2007). Facets of criterion, reliability, content, consequential, and construct articulation evidence of validity were explicitly examined. Collectively, these data

were intended to inform on the inferential qualities of measures used to assess TIQ and suitability of the system to attain its stated policy goals (i.e. improving instructional competence and increasing student learning).

More specifically, quantitative approaches included correlational, scale reliability, ANOVA, and exploratory/confirmatory factor analyses, as well as survey-based statistical methods. Qualitative approaches included coding and construct development techniques adopted primarily from grounded theory (Creswell, 2009; Stake, 2010; Saldana, 2009). Qualitative data were derived via personal interviews (n = 22) across four stakeholder groups: Teachers as Recipients (n = 7), Principals as Implementers (n = 8), District Policy Makers as Decision Makers (n = 4), and State Policy Makers as Reformers (n = 3). The researcher selected interview participants using random and purposeful sampling methods depending on group affiliation, while using semi-structured interview protocols for each of the four stakeholder groups. Each individual interview lasted approximately 30-45 minutes (≅ 15 hours of audio).

The researcher coded transcriptions of all qualitative information, (≅ 83,100 words). The researcher also used a variety of structured reliability and validation techniques including member checking, randomized data/multiple-coder triangulation, and stakeholder review, reflection, and discussion. An adaptation of the Lawshe Content Validity Ratio (CVR) questionnaire was administered to twenty-one *Subject Matter Experts* to inform the content adequacy of the Danielson framework (Lawshe, 1975; Wilson, Pan, & Schumsky, 2012). Finally, the researcher maintained a personal journal as a primary source of data collection (Corbin & Strauss, 2008), to record impressions of informal conversations, meetings, training sessions, policy discussions, and related decision pathways. These data informed aspects of all analyses.

Findings

RQ 1: To what degree do validity evidences of the LEA's teacher evaluation system support inferences on Teacher Instructional Quality (TIQ)?

Criterion Evidences (RQ1a): Table 1 summarizes the criterion associations observed between primary and secondary quantitative components used in constructing TIQ outcomes.

Table 1

Criterion Associations Between Primary and Secondary Evaluation Components

Analysis Category	Method	Finding	Significance
VAM x PP Score	Correlation	r = .25	(p < .05)
VAM x Subject (Same Yr.)	Correlation	r = .26 to .46	(p < .05)
VAM x Subject (Cross Yrs.)	Correlation	r = Not Sig. to .16	(p < .05)
VAM x Year	Correlation	r = Not Sig.	(p > .05)
VAM x PP Growth Groups	ANOVA/Tukey HSD	Low (10^{th} %'ile):	Sig. to Mid/High (p < .05)
		Mid (45th-55th %'ile):	Sig. to Low; Not Sig. to High (p< .05)
		High (90^{th} %'ile):	Sig. to Low; Not Sig. to Mid (p < .05)

(VAM) Value-Added Scores; (PP) Professional Practice Ratings;

As illustrated, correlations between evaluation components (PP versus VAM) were shown to be generally weak (most ≤ .30) or statistically insignificant. These weak component associations raise concerns of construct coherence: whether each component (VAM and PP) similarly inform and/or contribute to the posited latent TIQ construct. In addition, VAM measures were unstable over time (i.e., VAM x Subject (Cross Yrs.); r ≤ .16)). Finally, test of mean differences in PP across high, middle, and low VAM group differentiated teachers only in the bottom 10^{th} percentile.

Reliability Evidences (RQ1b): Table 2 summarizes the reliability information associated with the PP and VAM model scales.

Table 2

Reliability Indices for VAM Models and PP Scales

Analysis Category	Method	Finding	Significance
PP Composite	α, Ord. α, Ord. θ	α = .95 to .98	
PP Sub-Domain (4)	α, Ord. α, Ord. θ	α = .81 to .94	

VAM (Level-1)	Psuedo-r^2	r^2 = .71 to .77 ($p < .05$)
VAM (Level-2)	ICC	ICC = .05
True Score Item-Scale Range	SEM	22% to 37%*
VAM/PP Reliability/Bias	Qualitative: Stakeholder Interviews (n = 22)	Teacher, District, State (n = 15): *Concerns*: Rater Consistency, Bias, Construct-Irrelevant Variance (External Influences); *Causal factors*: insufficient observation time, attribute omissions, test scores as inadequate] Principals (n = 8): Limited substantive concerns

(*) SEM expressed as proportion of raw score scale; (α) Cronbach Alpha; (Ord. α) Ordinal Alpha; (Ord. θ) Ordinal Theta; (ICC) Interclass Correlation Coefficient; (SEM) Standard Error of Measure

As illustrated, reliabilities associated with the PP total score scales were strong, equaling or exceeding α = .95. Subscale reliabilities were lower, but above α = .80. In contrast, VAM model (level 1) pseudo-r^2 measures ranged between .71 and .77, and yielded relatively large standard errors when re-expressed as proportions of total score (item) scale ranges. Additionally, stakeholder narratives revealed substantive concerns regarding the potential for rater inconsistency, bias, and impact of construct-irrelevant factors on both achievement measures and evaluator ratings. Inversely, principals expressed more trust in score reliability than did teachers, district, and state representatives.

Content Evidences (RQ1c): Table 3 summarizes content suitability evidences obtained from a variety of methods and stakeholders.

Table 3

Summary of Content Evidences

Analysis Category	Method	Finding
FFT Representation (Danielson Framework for Teaching)	Lawshe CVI/R (SME: n = 33)	23% (5 out of 22) FFT components not substantive indicators of TIQ; 0% FFT components are inappropriate indicators *Criteria for Significance*: (n = 21, $p < .05$, one tail) = .359
	Confirmatory Factor Analysis	*Uncorrelated Factor Model*: Poor Fit Indices; High Factor Score Correlations (i.e. Low Discriminant Inference): r = .70 to .86, n = 238, $p < .05$

	Correlated Factor Model: Mixed/Improve Fit Indices; High Factor Score Correlations (i.e. Low Discriminant Inference): r = .82 to .92
Exploratory Factor Analysis (PAF, Oblique Rotation)	Substantive extracted factors = 2; Number of factors inconsistent with theoretical FFT framework; Within factor loadings inconsistent with theoretical framework with exception of FFT Domain 4 (Professional Responsibilities); Suggestion of a single dominant latent factor; Factor correlations: r = .80 (P < .001); Factor loading structures differ between experienced & less experienced teachers
Qualitative: Stakeholder Interviews (n = 23)	Teacher, District, State (n = 15): Concerns: Omitted/underrepresented attributes of good/effective teaching Principals (n = 8): Limited substantive concerns

(SME) Subject Matter Experts; (CVI/R) Lawshe Content Validity Index/Ratio Questionnaire; (FFT) Danielson Framework for Teaching; (PAF) Principle Axis Factoring

As illustrated, exploratory and confirmatory factor analyses did not support the posited four-domain (factor) structure attributed to the Danielson framework. In addition, statistical evidence revealed strong covariance (r > .70) between the four Danielson domains suggesting a lack of discriminant inference. In addition, internal factor structures differed between more/less experienced teachers. Stakeholder concerns also suggested that important attributes of good/effective teaching were omitted/underrepresented in the evaluation framework, and narratives reaffirmed test scores as providing an inadequate representation of TIQ. Finally, subject matter experts identified 23 percent of the Danielson components as not being critical elements for identifying/distinguishing good/effect teaching.

Consequential Evidences (RQ1d): Table 4 summarizes findings related to consequential evidences obtained from stakeholder perspectives.

Table 4

Summary of Consequential Evidences

Analysis Category	Method
Intended-Unintended Impacts	Qualitative: Stakeholder Interviews (n = 23)

Findings: *Concerns (Negative):*

- The evaluation system is viewed as imposed, external, and policy driven (Teachers, District)
- Test scores viewed as inadequate/incomplete primary measure of instructional quality (Teachers, District, State)
- Teachers express compliance/conformity to, rather than acceptance of, measured attributes
- Teachers, District, State participants express negative attitudes regarding overall impact on improving instruction or increasing student learning
- Omission/reductionism raises stress/fear, lowers trust, harms professional identity (Teachers, District, State)
- *Clarity/Focus/Structure*:
    - Leads to reductionism, narrowing of curriculum & instruction, reduces instructional creativity (Teachers, District)
    - Important learnings are crowded out: non-tested content, affective (Teachers, District, State)
- System may lead to increased difficulty in attracting/retaining teachers, harm school climate, lost focus on other important administrative/instructional leadership duties (District, State)

Findings: *Positive Affirmations*:

- Principals express substantively positive views of impact on both instruction and learning; all impacts are viewed as positive (even stress/fear)
- *Clarity/Focus/Structure*:
- Leads to increased opportunity for communication, dialog, reflection on measured attributes (Teachers, Principals, State)

As illustrated, Clarity/Focus/Structure emerged as a dominant theme, incorporating both positive and negative perspectives of system consequence. However, within these facets, negative sentiments dominated the narratives. Positive affirmations focused on the clarity provided by the system in terms of what and how teachers are evaluated. Here, *clarity* supported focused communication, dialog, and understanding of required competencies and outcomes. In contrast, participants also viewed *clarity* as reductionist leading to attribute omission and underrepresentation. Stakeholder narratives suggest this contributed to mistrust, higher stress, and potential harm to professional climate. Throughout, an increased emphasis on test scores (compared to prior evaluation systems) was viewed as narrowing instruction and restricting learning. Overall, participants expressed skepticism that the collective system would promote instructional improvement and attain higher student outcomes.

Construct Articulation (RQ1e): Table 5 summarizes findings related to stakeholder perspectives regarding construct articulation.

Table 5

Construct Articulation

Analysis Category	Method
Attributes of Good Effective Teachers	Qualitative: Stakeholder Interviews (n = 23)

Findings: *Defining Good/Effective Teacher*:

- Content delivery not seen as dominate attribute
- Test score as inadequate representation of quality
- Affective attributes most important (i.e. affective impacts on students and affective dimensions of professional practice)
- *Teacher Practice*: passionate, attitude, motivation, commitment, emotionally invested, facilitate, build/nurture relationships, architects of personalized learning environment; role models (microcosm of society, develop citizens); concern with individual, personal, emotional well-being
- *Students Affective Impacts*: inspire, motivate, transform, instill desire, trust, sense of self-worth, confidence, future, hope,
- *Teacher as Architect*: purposefully engineer the environment, conditions (affective), and context of learning; actively transform/enhance student connection to learning experience
- *Indirect Curriculum*: non-content centered, affective goals and objectives; causal link between short-term affective and long-term personal well-being
- Content learning longer-term is an artifact of good/effective teaching;

As illustrated, stakeholder narratives reflected the sentiment that standards-based pedagogical competencies and standardized test scores fail to fully capture important defining attributes that distinguish good/effective instruction. Indeed, stakeholders conceptualized instructional quality in terms of affective dimensions, including non-academic impacts teachers have on students as well as a teacher's personal/emotional commitment to their professional practice. Additional dimensions not fully represented in the evaluation context included concepts of teachers as architects of the learning, transformational influence (on students), and attention to indirect (non-academic) curriculums.

RQ 2: How does providing an evolving dialog on validity evidence influence organizational policy decisions regarding implementation of the LEA's teacher evaluation system?

Study findings suggest that providing LEA decision makers' access to high quality, relevant information was a critical component of the design process. The narratives and

evidences examined indicate the following: (a) Decision maker access to empirical information acted to challenge a priori perspectives, facilitate critical reflection, and promote dialog; (b) The provision of accurate, complete, unbiased information was a critical condition for empowering confidence and authority in policy makers; (c) Information brokers served an important role in the decision-making process, bringing specialized skills, knowledge, and expertise to the policy environment; and (d), open communication, dialog, and critical reflection were valued activities within the decision-making processes.

The flow of information into the evaluation system's development process substantively impacted implementation pathways by facilitating changes in professional development strategies, enhanced stakeholder communication, recognition of measure limitations (mitigating high stakes consequences), and promoting positive cultural attitudes toward the evaluation activity. Early sharing of empirical data lead to the implementation of a formal organizational "program evaluation" activity, increased emphasis on (rubric) scoring training, implementation of additional interrater reliability studies, the need for additional evaluator training in the areas of teacher mentoring and support, increased emphasis on information sharing/transparency, and an adjustment of criteria for identifying *Ineffective* teachers.

RQ 3: To what degree have the perspectives and behaviors of stakeholders been incorporated into, and influenced, the system's implementation?

Organizational leaders made substantive attempts to provide opportunities for teacher input, feedback, and communication. However, active involvement by teachers was generally limited with participant narratives reflecting a general sense of "non-involvement." Here *lack of voice* was coincident with *lack of influence and empowerment*. Teacher narratives suggest these perceptions served, in part, as catalysts for mistrust and attitudes of compliance rather than acceptance. As a result, delivery of information and opportunity for engagement were revealed

as necessary, but insufficient requirements to attain active inclusion of teacher voice in the decision-making processes.

A similar lack of inclusion and empowerment was reflected within the narratives of educators throughout the successive levels of policy leadership and decision-makers (i.e., principal, district, and state participants). This suggests the evaluation framework may be disconnected from the perspectives of those it was designed to assess. In addition, lack of inclusion suggests stakeholder acceptance of the initiative might continue to be tenuous, threatening the long term sustainability of the system outside of influence exerted by external power centers (i.e., state legislative authority, government mandates and guidelines).

RQ 4: How did the process of engaging in this action-research study impact the investigator as a scholarly researcher and organizational leader?

The researcher examined the challenges and barriers encountered during the research process and how these experiences impacted the researcher's leadership and scholarly abilities. Substantive challenges involving data integrity, stakeholder communication, researcher positionality, and the general complexity of the study itself conspired to make the activity both challenging and interesting. This context necessitated on-going reflection, adjustment in thinking, multiple rounds of analysis and writing, and patience. In addition, personal relationships evolved along with the research activity, and trust-building became a central focus for engaging in the evaluation development process.

Simultaneously navigating through these factors tested the researcher's capacity to design and conduct large-scale mixed-method research projects, forced reflection on the dual role that applied researchers occupy in organizations, and permitted the exercise and enhancement of personal leadership skills. As a result of the dissertation activity, the researcher believes he is better equipped to serve in the role of scientific policy advisor and organizational leader.

Concluding Reflections

The purpose of this research study was to evaluate the warrant for making high stakes, consequential, decisions under a state-policy-imposed framework for evaluating teacher instructional quality. Because the policy framework potentially impacts the professional, personal, and economic well-being of the individuals being evaluated, rigorous validation of the system in context was warranted.

In this regard, the empirical evidence compiled herein suggests that the current measurement framework lacks sufficient warrant for making high stakes consequential decisions of teacher instructional competence based on the following: (a) The measured components used in the evaluation framework displayed poor internal coherence. Here, weak component associations indict the proposition that each measured component independently informs, and contributes to the latent TIQ construct; (b) VAM scale reliabilities seem insufficient to warrant their inclusion in highly consequential decision-making processes. Stakeholders also questioned aspects of rater consistency, bias, and impacts of non-instructional factors (i.e. construct-irrelevant variance); (c) Professional practice ratings data did not support the four-domain (factor) structure posited by the Danielson framework. In addition, strong factor covariances indict this measure's suitability for discriminating explicit areas of pedagogical strength/deficiency; (d) The evaluation framework suffers from omission/underrepresentation of attributes important for distinguishing instructional quality. These factors include affective dimensions of both instructional impact and professional practice; (e) Evidence suggests presence of numerous unintended, negative consequences resulting from the evaluation activity that impede progress toward stated and well-intended policy goals; (f) Stakeholders held generally negative sentiments regarding the potential for the evaluation system to substantively

improve instructional practice and student learning (a primary goal of the state-imposed policy framework). Here, trust in the system is negatively impacted by a perceived lack of educator voice (influence) with regard to policy development and system implementation; and (g) The state policy-imposed teacher evaluation framework lacks a clearly articulated representation of the latent TIQ construct that it purports to assess.

Significance and Limitations

In this study the researcher applied a comprehensive approach for evaluating construct validity within a high stakes consequential context. Based on findings, the researcher proposed nine recommendations for improving the LEA's current evaluation process, offered six generalized policy recommendations, and advanced eight areas of additional research. In addition, numerous limitations were explicitly examined, each also potentially impairing the reliability and generalizability of study findings.[8] These included but were not limited to (a) the influence of researcher positionality, (b) limited stakeholder subgroup membership, (c) restricted teacher representations, (d) use of unadjusted multi-level observational (Danielson) rating data, (e) single district representation, and (f) lack of empirical interrater reliability information.

In closing, the research discussed herein documented substantive issues concerning LEA implementation of Arizona's state-policy framework (which is akin to many other state-policy frameworks). At the same time, however, it facilitated substantive positive changes regarding system design, implementation, and application within this local setting. Accordingly, research results have been presented at numerous state and national research conferences and local public policy meetings.[9] In this way, it is hoped that the findings have made (and will continue to make) a contribution to the larger policy discourse on teacher evaluation and school accountability.

References

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

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

Arizona Department of Education (ADE). (2012). *Arizona framework for evaluating educator effectiveness fact sheet*. Phoenix, Arizona: Arizona Department of Education. Retrieved from http://www.azed.gov/.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research, 3rd edition*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage Publications.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. doi:10.1037/h0040957

Danielson, C. (2010). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Danielson, C. (2011). *The framework for teaching evaluation instrument: 2011 edition*. Princeton, NJ: The Danielson Group. Retrieved from http://www.danielsongroup.org/framework/

Darling-Hammond, L. (1997). School reform at the crossroads: confronting the central issues of teaching. *Educational Policy*, 11(2), 151-156.

Dewey, J. (1900). *The school and society*. Chicago, IL: The University of Chicago Press. Retrieved from http://books.google.com/books .

Fast, E. F., & Hebbler, S. (2004). *A framework for examining validity in state accountability systems.* Washington, D.C.: Council of Chief State Schools Officers. Retrieved from http://www.ccsso.org/Documents/2004/Framework_For_Examining_Validity_2004.pdf.

Fullan, M. (2009). Have theory, will travel: A theory of action from system change. In A. Hargreaves & M. Fullan (Eds.), *Change Wars*. Bloomington, IN: Solution Tree.

Gelo, O., Braakmann, D., & Benetka, G. (2008). Quantitative and qualitative research: Beyond the debate. *Integrative Psychological and Behavioral Science*, *42*(3), 266-290. doi:10.1007/s12124-008-9078-3

Gergen, K. J. (2009). *An invitation to social construction* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Gibson, S, Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, *76*(4), 669-682. doi:10.1037/0022-0663.76.4.569

Good, T. L. (1999). The purpose of schooling in America. *The Elementary School Journal*, *99*(5), 383-389.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Wiley and Sons.

Hall, G., & Hord, S. (2011). *Implementing change: Patterns, principles, and potholes*. New York, NY: Ablongman Pearson Education.

Hall, G. E., Loucks, S. F., Rutherford, W. L., & Newlove, B. W. (1975). Levels of use of the innovation: A framework for analyzing innovation adoption. *Journal of Teacher Education, 26*(1), 52-56. doi:10.1177/002248717502600114.

House Bill (HB) 2823 (2012). *Schools; teachers; principals; evaluation system*. Fiftieth Legislator, Second Regular Session, Phoenix, AZ. Retrieved from http://www.azed.gov/.

Herlihy, C., Karger, E., & Pollard, C. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1).

Kane, M. T. (1992). The assessment of professional competence. *Evaluation and Health Professions, 15*(2), 163-182. doi:10.1177/016327879201500203

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2010). Validity and fairness. *Language Testing, 27*(2), 177-182. doi:10.1177/0265532209349467

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34-70. doi:10.1177/0013161x08327549

Kotter, J. (1996). *Leading change*. Boston, MA: Harvard Business School Press.

Kotter, J. (2010). Before you can get buy-in, people need to feel the problem. *Harvard Business Review.* Retrieved from https://hbr.org/2011/02/before-you-can-get-buy-in-peop.html.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575. doi:10.1111/j.1744-6570.1975.tb01393.x

Learning Sciences Marzano Center. (2012). *The Marzano causal teacher evaluation model*. White Paper Prepared for the Oklahoma State Department of Education. Retrieved from http://www.marzanoevaluation.com/files/Oklahoma-Marzano-Teacher-Evaluation-White-Paper.pdf

Linn, R. (2008). *Validation of uses and interpretations of state assessments*. Washington, D.C.: Council of Chief State Schools Officers.

Loucks, S. F, & Hall, G. E. (1979). *Implementing innovations in schools: A concerns-based approach.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria, VA: Association for Supervision & Curriculum Development.

Marzano, R. (2011). *Research base and validation studies on the Marzano evaluation model*. Retrieved from http://www.marzanoevaluation.com/files/Research_Base_and_Validation_Studies_Marzano_Evaluation_Model.pdf

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

McClellan (2012). Teacher evaluator training & certification: Lessons learned from the Measures of Effective Teaching project. Retrieved from http://www.teachscape.com/resources/teacher-effectiveness-research/2012/02/teacher-evaluator-training-and-certification.html

McGee-Banks, C. A., & Banks, J. A. (1997). Reforming schools in a democratic pluralistic society. *Educational Policy, 11*(2), 183-193. doi:10.1177/0895904897011002004

McNergney, R. F., Imig, S. R., & Pearlman, M. A. (2002). Teacher evaluation. In J. W. Guthrie (Ed.), *Encyclopedia of Education* (2nd ed., Vol. 7, pp. 2453-2461). New York, NY: Macmillan

Mehrens, W. A. (1977). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18. doi:10.1111/j.1745-3992.1997.tb00588.x

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027. doi:10.1037/0003-066x.35.11.1012

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

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

Measures of Effective Teaching (MET) Project. (2010). *A composite measure of teacher effectiveness*. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Value-Add_100710.pdf

Milanowski, T. (2011, April 10). *Validity research on teacher evaluation systems based on the framework for teaching*. Paper presented at the 2011 annual meeting of the American Education Research Association, New Orleans, Louisiana. Retrieved from http://www.aera.net/repository

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

Plano Clark, V. L., & Creswell, J. W. (2010). *Understanding research: A consumer's guide*. New York, NY: Pearson Education, Inc.

Popham, J. W. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13. doi:10.1111/j.1745-3992.1997.tb00586.x

Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press.

Saldana, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage Publications.

Shepard, L. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450. doi:10.2307/1167347

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-24. doi:10.1111/j.1745-3992.1997.tb00585.x

Stake, R. E. (2010). *Qualitative research: Studying how things work*. New York, NY: Guilford Press.

Stringer, E. T. (2007). *Action Research (3rd ed.)*. Thousand Oaks, CA: Sage Publications.

Stronge, J. H. (2012). *Stronge evaluation system report.* Williamsburg, Virginia: Stronge & Associates Educational Consultants, LLC. Retrieved from http://www.strongeandassociates.com/files/Stronge%20Evaluation%20System%20Report.pdf

Teachscape Focus. (2016). Teachscape Focus. Retrieved from https://www.teachscape.com/products/focus#overview

U.S. Department of Education. (2009). *Race to the top program: Executive summary*. Washington, D. C. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

Weisberg, D., Sexton S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *The Education Digest, 75*(2), 31-35.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210. doi:10.1177/0748175612440286

Wise, A. E. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: Rand Corporation. Retrieved from http://www.rand.org/pubs/reports/R3139.html.

Endnotes

[1] Danielson (2010) argues that, "… credibility in an evaluation system is essential. A principal or a superintendent must be able to say to the school board and the public, "… everyone who teaches here is good and here's how I know" (p. 36). Fast and Hebbler (2004) argue that in any type of accountability system "… validation evidence is necessary to support the accountability claims made about individuals and agencies and the accompanying imposition of stakes." (p. 2).

[2] During the course of the study, the researcher was employed as the LEA's Director of Research and Accountability, charged with developing methodological approaches for measuring teacher instructional effectiveness consistent local values and legislative policy. Positionality issues were examined as part of the research activity and discussed under study limitations and reflections.

[3] Enrollment ≅ 25,500, 24 elementary and high school campuses, and approximately 1,400 certified classroom teachers. School/classroom representations (for teacher quantitative and qualitative measures) were purposefully restricted to a population of 238 self-contained, general education teachers in grades 3 to 6 reporting a full complement of instructionally aligned assessment and professional practice measures. This helped reduce non-standard instructional contexts and settings lacking direct measures of instructed content.

[4] The Danielson framework contains of four domains: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities. For each teacher, PP ratings are assigned to a total of twenty-two pedagogical components.

[5] The Teachscape Focus on-line training and calibration system is directed aligned to the components and rubrics used by the Danielson Framework for Teaching.

[6] Academic achievement measures constitute 33 percent of the composite TIQ score while observations of professional practice represent the remaining 67 percent.

[7] Application of these descriptive labels was mandated by state legislation, stipulating that the descriptive classifications become part of the teacher's permanent evaluation record and available to prospective future employers. Annually, LEAs are required to submit aggregated reports to the Arizona Department of Education concerning the distribution of teacher evaluation ratings.

[8] The full dissertation including discussion of study limitations may be downloaded from the following URL: https://drive.google.com/file/d/0B7yiR233u-FwcFNKdmNTazgwUlE/view?usp=sharing.

[9] This research has also been presented to the Arizona Task Force on Teacher Evaluation (August, 2015), and at the Arizona School Boards Association (December 2015).