education policy analysis archives

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

Arizona State University

Volume 31 Number 117 October 24, 2023 ISSN 1068-2341

Evaluating the Validity Evidence Surrounding the Use of Value-Added Models to Evaluate Teachers: A Systematic Review

Audrey Amrein-Beardsley
Arizona State University
United States

Matthew Ryan Lavery
South Carolina Education Oversight Committee
United States

Jessica Holloway
Australian Catholic University
Austrailia

Margarita Pivovarova
Arizona State University
United States

&

Debbie L. Hahs-Vaughn
University of Central Florida
United States

Citation: Amrein-Beardsley, A., Lavery, M. R., Holloway, J., Pivovarova, M., & Hahs-Vaughn, D. L. (2023). Evaluating the validity evidence surrounding the use of value-added models to evaluate teachers: A systematic review. *Education Policy Analysis Archives, 31*(117). https://doi.org/10.14507/epaa.31.8021

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

Manuscript received: 22/5/2023
Revisions received: 2/9/2023
Accepted: 2/9/2023

Abstract: Local education agencies (LEAs) continue to use value-added models (VAMs) for teacher evaluation policies and purposes, often with consequences attached. Although the Every Student Succeeds Act (ESSA) provides more flexibility to LEAs, few have discontinued VAM use, suggesting they interpret VAMs as a valid measure of teacher effectiveness. In this systematic review, we used a framework built on the *Standards of Educational and Psychological Testing* (AERA et al., 2014) to examine validity evidence contained in 75 articles published in high-quality, peer-reviewed journals in which article authors supported or challenged user interpretations and uses of VAMs. Results with implications for educational policy are presented.
Keywords: validity/reliability; school/teacher effectiveness; teacher evaluation; educational policy; value-added models

Evaluación de la evidencia de validez sobre el uso de modelos de valor agregado para evaluar a los docentes: Una revisión sistemática

Resumen: Las agencias educativas locales (LEA) continúan utilizando modelos de valor agregado (VAM) para políticas y propósitos de evaluación docente, a menudo con consecuencias. Aunque la Ley Every Student Succeeds (ESSA) proporciona más flexibilidad a las LEA, pocas han descontinuado el uso de VAM, lo que sugiere que interpretan los VAM como una medida válida de la eficacia docente. En esta revisión sistemática, utilizamos un marco basado en los *Estándares de Pruebas Educativas y Psicológicas* (*Standards of Educational and Psychological Testing;* AERA et al., 2014) para examinar la evidencia de validez contenida en 75 artículos publicados en revistas revisadas por pares de alta calidad en los que los autores de los artículos apoyaron o cuestionó las interpretaciones de los usuarios y los usos de los VAM. Se presentan resultados con implicaciones para la política educativa.
Palabras-clave: validez/confiabilidad; eficacia de la escuela/docente; evaluación docente; política educativa; modelos de valor agregado

Avaliando as evidências de validade do uso de modelos de valor agregado para avaliar professores: Uma revisão sistemática

Resumo: As agências educativas locais (LEAs) continuam a utilizar modelos de valor acrescentado (VAM) para políticas e fins de avaliação de professores, muitas vezes com consequências associadas. Embora a Lei de Todos os Alunos com Sucesso (ESSA) proporcione mais flexibilidade aos LEAs, poucos descontinuaram o uso do VAM, sugerindo que interpretam os VAM como uma medida válida da eficácia dos professores. Nesta revisão sistemática, utilizamos uma estrutura baseada nos *Padrões de Testes Educacionais e Psicológicos* (AERA et al., 2014) para examinar as evidências de validade contidas em 75 artigos publicados em periódicos revisados por pares de alta qualidade, nos quais os autores dos artigos apoiaram ou desafiou as interpretações e usos dos VAMs pelos usuários. São apresentados resultados com implicações para a política educacional.
Palavras-chave: validade/confiabilidade; eficácia escola/professor; avaliação docente; política educacional; modelos de valor agregado

Evaluating the Validity Evidence Surrounding the Use of Value-Added Models to Evaluate Teachers: A Systematic Review

More than one century ago, Pittenger (1917) published an article about the "Problems of Teacher Measurement" in the *Journal of Educational Psychology*. He wrote about the professional imperative to ensure that throughout America's public schools' educational measurement systems were in place to ensure that the best teachers were teaching students using the best instructional methods possible. Pittenger (1917) described three possible construct domains that could be measured to indicate a teacher's instructional quality or effectiveness: "(1) the plane of results or of pupil achievement; (2) the plane of the teaching and learning process; and (3) the plane of the teacher's equipment for teaching, both native and acquired" (p. 104). These three planes represent the most common domains of what we would call "multiple measures" today (American Educational Research Association [AERA] et al., 2014; Bill & Melinda Gates Foundation, 2013; Fox, 2016). Pittenger's (1917) third plane included "all attainable facts as to her [sic] equipment" (p. 106), and pertained to teacher certification, professional advanced degrees earned, years of teacher experience, and the like, which is comparable to today's consideration of teachers' credentials. His second plane included "her [sic] classroom procedure" (p. 106), pertaining to teacher practice. Today, this is most often measured via observational rubrics and instruments meant to capture teacher practice based on a series of (un)planned classroom observations. Pittenger's (1917) first plane included "the results which she [sic] achieves" (p. 106), pertaining to the teacher's impact on student learning.

Today, this first plane is most commonly measured as the student achievement scores derived via large-scale standardized tests. Even after the passage of the Every Student Succeeds Act (ESSA; 2015), a federal policy which afforded states and districts more local control (i.e., less federal intervention) over their teacher evaluation and accountability systems, many states throughout the US (e.g., in Colorado, Florida, New Mexico, North Carolina, Maine, and Utah) are still using the test scores derived via the large-scale standardized tests mandated by No Child Left Behind Act of 2001 (NCLB, 2002) to hold teachers accountable for their measurable effects on student learning (Close et al., 2020).

Pittenger (1917) clearly saw the appeal of such measures, but he advised against their use. The plane of results would be the ideal plane upon which to build an estimate of a teacher's individual efficiency, if it were possible (1) to measure all of the results of teaching, and (2) to pick out from the body of measured results any single teacher's contribution to said results. At present these desiderata are impossible to attain (p. 107).

The first of these cautions, that academic achievement is not the sole product of effective teaching, has received less attention as per current teacher evaluation policies and efforts (Good, 2014). Attempts to isolate the effect of a single teacher on student learning have flourished, however, such that some states continue to require via state- and local educational policies that a measure of student learning gains or academic growth be included in teacher evaluation systems, with most of these states using value-added models (VAMs) to help produce these measures (Paufler & Amrein-Beardsley, 2014).

Both then and now, the plane of student results was/is the most controversial domain of this type of educational measurement as it suggests "a mathematical exactness of procedure which is clearly impossible in this field" (Pittenger, 1917, p. 103). Though much has changed over the last century, the question remains whether VAMs now provide the mathematical exactness necessary to measure the plane of student results.

The purpose of the present review, accordingly, was to understand the validity evidence presented in high-quality VAM literature published in the leading peer-reviewed journals in education and economics, and how the authors of the studies reviewed position the evidence presented to support or challenge the validity of the inferences about teacher effectiveness as based on VAM estimates. We used the lens of argument-based validation (AERA et al., 2014; Kane, 2004, 2013) to interrogate this literature and to assess whether the authors of the reviewed literature provided evidence that the criticisms Pittenger (1917) raised over a century ago have been met and resolved by contemporary VAMs.

Conceptual Framework

For the purposes of this review, VAMs are defined as complex regression models via which modelers use students' histories of scores on academic achievement tests to determine those students' expected scores or expected gains on current achievement tests, and thereby estimate the specific contributions of individual teachers (often referred to as "teacher effects") to observed gains (or losses) on those tests (AERA, 2015; American Statistical Association [ASA], 2014; Braun, 2005). Some VAMs include student attributes, such as socioeconomic status, disability, attendance, or English learner status, as covariates to adjust estimates of teacher effects, and some do not (Ballou et al., 2004; McCaffrey et al., 2004). Regardless of the specific procedure used, though, VAM scores clearly play an important role in many US (and international; Araujo et al., 2016; Levy et al., 2019; Sahlberg, 2011; Sørensen, 2016; Smith & Kubacka, 2017) teacher evaluation and accountability policies and systems. As well, they are often used to prompt many personnel actions (e.g., professional development, merit pay, and decisions about the hiring, tenure, and termination of teachers) and make causal inferences about teacher effectiveness, in that teachers are given direct responsibility for producing (or failing to produce) student learning, as measured by large-scale, standardized achievement tests and VAMs (Braun, 2012; Paufler & Amrein-Beardsley, 2014; Reardon & Raudenbush, 2009).

Validity Arguments

Tests and measures are neither valid nor invalid, and neither are the scores they produce; rather, it is the interpretations and uses derived from such scores that must be validated (AERA et al., 2014; Kane, 2013; Messick, 1995). Over the course of its historical development, the traditional validity literature has conceptualized validation as including evidence of different types of validity, such as criterion validity (Cronbach, 1971; Cureton, 1951; Moss, 1992, 1995), convergent, predictive, and discriminant validity (Messick, 1975, 1980, 1989, 1995), and consequential validity (Kane, 2013; Messick, 1980), all of which map onto construct validity (Cronbach, 1971; Cronbach & Meehl, 1955; Kane, 2013). The current edition of the *Standards of Educational and Psychological Testing* (AERA et al., 2014; henceforth referred to as the *Standards*) refer more contemporarily, instead, to different sources of validity evidence, stating that:

> These sources of [validity] evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use. (pp. 13-14)

Typically, test developers or researchers validate the proposed interpretations and uses of scores from individual tests. VAMs are not individual tests, but statistical procedures, which are used to analyze the scores of multiple tests within a specific context to produce scores for subsequent interpretation and use. If it is the proposed interpretation and uses of test scores, and not the tests or the scores themselves must be validated, then it is appropriate to subject the interpretations and uses of scores produced by some other means to the same rigorous validation process described in the *Standards* (AERA et al., 2014).

Likewise, the *Standards* describe validation as "a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (p. 11). Van Eemeren and Grootendorst (2010) define argumentation as "a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forth a constellation of propositions justifying or refuting . . . the standpoint" (p. 1). By positioning validation as a process of argumentation, the *Standards* (AERA et al., 2014) suggest that validation is a public, rather than solitary, process, and that validity is best appraised in the context of the arguments presented to, challenged, and supported by a community of reasonable critics.

In the context of the present review, the community of reasonable critics included the authors of the empirical papers and commentaries included in the review, the community of scholars who peer-reviewed these papers prior to publication, and scholars who subsequently read and cited (or did not cite) the reviewed papers post-publication. Thus, in the present study we reviewed this process of validation argumentation by examining the high-quality extant literature published in leading peer-reviewed journals to search for and make explicit the arguments that support or challenge the uses of VAM scores in teacher evaluation and accountability policies and systems. We also sought to understand how these critics apparently understood, positioned, and described the weight of the combined evidence that authors of this literature offered to support or challenge the validity of the inferences about teacher effectiveness as based on VAMs.

Note that it was impossible to separate our review of the ongoing scholarly arguments about the valid use of VAMs to evaluate teachers from the scholarly communities in which these arguments took place. In a recent survey of VAM scholars who have published articles about VAMs in high-quality journals (Lavery et al., 2020), we found that scholars in economics, education, and quantitative methods interpreted the evidence about the valid use of VAM for teacher evaluation policies and purposes differently. We did not seek to resolve these differences and identify the one correct way to interpret the reviewed validity evidence, nor did we seek to discredit one side of the argument and endorse the other. Although in this study we did synthesize and summarize many aspects of the literature reviewed, integrating this literature into a cohesive whole was also not the primary intent of this review. Readers interested in a synthesis of VAM literature written by an education scholar may wish to read Everson (2017), and readers interested in a similar synthesis written by economics scholars may wish to read Koedel et al. (2015). The present review, instead, represents an attempt to understand and examine the validity arguments and inferences surrounding the use of VAMs across both fields.

Framework for Validity Evidence

We developed the framework for validity evidence shown in Figure 1 as a visual representation of the validation process described in the *Standards* (AERA et al., 2014), and to help organize the validity evidence that we found in the literature reviewed. Note, however, that although we reference the most recent edition of the *Standards* (AERA et al., 2014), the previous edition of the *Standards* (AERA et al., 1999) is similar and supports the language and structure of this framework for validity evidence equally well. As a verbal process (van Eemeren & Grootendorst, 2010), a validation argument "logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (AERA et al., 2014, p. 11). Kane (2013) refers to this explicit statement as the interpretation/use argument (IUA; pictured as the topmost element in Figure 1). As per Kane (2013), the IUA "includes all of the claims based on the test scores (i.e., the network of inferences and assumptions inherent in the proposed interpretation and use)" (p. 2), and it must include a clear specification of the construct measured (AERA et al., 2014).

The *Standards* (AERA et al., 2014) describe five different sources of validity evidence which may be collected to support (or challenge) the arguments for (or against) the proposed IUA. These sources of validity evidence (specifically, evidence based on test content, response processes, internal structure, relations to other variables, and related consequences) form the foundation of the framework in Figure 1.

Figure 1

Framework for Validity Evidence. IUA=Interpretation Use Argument

Validation is not merely a process of accumulating some evidence in each of these five categories, however. As a rational process of argumentation, the collected evidence must directly support the propositions and arguments inherent in the IUA (AERA et al., 2014; Kane, 2004, 2013). The *Standards* suggest that "some types of evidence will be especially critical in a given case, whereas other types will be less useful" (AERA et al., 2014, p. 12), and Kane (2013) states that "the kinds of evidence required for validation are determined by the claims being made" (p. 3). Thus, some IUAs may require evidence from all sources, while others might require evidence from only a few key sources. It is the responsibility of the validator to clearly state how the reported evidence relates to the IUA under what conditions.

The central disc shown in Figure 1 represents the kinds of evidence that, although not derived from one of the five sources of validity evidence described in the *Standards* (AERA et al., 2014), are critically important in the validation of an IUA. The framework groups the three concerns of reliability/precision, methodological assumptions, and certain aspects of fairness under the umbrella of "Reliable, Precise, and Fair Methods & Measurement" (see Figure 1; henceforth referred to as *Methods and Measurement*). The authors of the *Standards* (AERA et al., 2014) point out that the measurement literature has historically used the term "reliability" in two distinct ways, using "reliability/precision to denote the more general notion of consistency of the scores across instances of the testing procedure" (p. 33). Though discussed "as an independent characteristic of test scores,…reliability/precision has implications for validity" (AERA et al., 2014, p. 34). Further, since the IUA under investigation involves the use of VAM scores, which are derived via complex statistical procedures, the reliability/precision of those scores may be subject to satisfaction of the methodological assumptions for the procedures used to generate them. In the graphical representation of the framework for validity evidence, the disc sits between the sources of validity evidence and the IUA to indicate that evidence of sound measurement and reliable, precise, fair scores are necessary but insufficient for validity (AERA et al., 2014; Messick, 1989). Without evidence that the scores are of sufficient quality, validity evidence from the five sources is insufficient to support the IUA as it falls through the gaps.

The final element captured in this portion of the framework, fairness, is a "fundamental validity issue and requires attention throughout all stages of test development and use" (AERA et al., 2014, p. 49). Fairness is a rather complex issue and a construct that manifests in several elements of the framework. Since the central disc in the framework addresses concerns affecting the overall quality of scores that inform the proposed IUA, evidence relating to measurement bias is included in this element. For a traditional test, such evidence might include tests of differential item functioning (DIF) or measurement invariance across subgroups of test takers. When applied to VAM scores, this component includes evidence that scores are free from construct irrelevant variance (CIV), or other indications of freedom from bias. CIV is a term coined by Messick (1989) to describe factors that falsely inflate or deflate the measurement of a variable and therefore distort its interpretation or validity. Note that other evidence relating to issues of fairness fall under the appropriate source of validity evidence (e.g., validity evidence based on related consequences might include differential outcomes for various subgroups resulting from the IUA). The instrument used to code the literature reviewed (described in a forthcoming section) includes operational definitions of the types of evidence captured in each portion of the framework (see our Review Instrument in the supplementary file titled "SupplementaryMaterials_ValidationReviewInstrument").

Methods

Kane (2004) argues that under ideal circumstances "the development of the test would follow or be concurrent with the initial development of the [IUA]" (p. 141). However, the present review occurred at a time when VAMs were still in widespread use throughout the US (and international) education system, and a mainstay of both district and state-level educational policies across the US, even though federal regulations no longer explicitly require them (Close et al., 2020). That is, after Race to the Top (American Recovery and Reinvestment Act of 2009, 2009) offered federal dollars for states and districts to include measures of student achievement gains in teacher evaluation policies and systems, and after the federal government required states to also adopt stronger teacher accountability policies and systems to secure waivers from meeting the NCLB goal of 100% of the student proficiency by the year 2014 (NCLB, 2002; see also Duncan, 2011; Layton, 2012), the use of VAMs in teacher evaluation policy dramatically increased (Braun, 2012; Paufler & Amrein-Beardsley, 2014). Now that ESSA (2015) has afforded states and districts more local control (i.e., less federal intervention) over their teacher evaluation and accountability-based policies and systems, the continued use of VAM by states and districts suggests that these stakeholders either believe VAM-based educational policies can support valid inferences about teacher effectiveness and quality, or that they do not see a suitable alternative to VAMs.

Since the *Standards* (AERA et al., 2014) identify validation as "the joint responsibility of the test developer[s] and the test user[s]" (p. 13), via the present review we sought to survey, analyze, and summarize the validation efforts of VAM makers, VAM users, and the communities of reasonable, critical stakeholders to answer the following research questions:

1. To what extent did the validity evidence presented in high quality, peer-reviewed journals support or challenge the use of VAMs in teacher evaluation?
2. To what extent did the validity evidence reported in each of the six areas of the framework (i.e., five sources of validity evidence, plus *Methods and Measurement*; see Figure 1) support or challenge the use of VAMs in teacher evaluation?

We used two distinct critical lenses to examine the literature reviewed. The first critical lens was the rigorous peer review process of the most reputable and respected journals and of the scientific community in their relevant disciplines. The second critical lens was the argument-based validation process described in the *Standards* (AERA et al., 2014) and other literature on validity and validation. Forthcoming sections describe both of these critical lenses.

Rigorous Peer Review

To investigate the compiled evidence of validity concerning the IUA under investigation, we included quality literature which had been (a) published in the most reputable, peer-reviewed journals in education, economics, and quantitative methods and (b) incorporated into the ongoing scholarly argument by the community of scholars in these same disciplines as indicated by sufficiently high annual citations in later papers. We consulted the *2018 Journal Citation Reports® Social Sciences Edition* (JCR; Clarivate Analytics, 2019) to select journals of sufficient scientific quality and rigor. The JCR provides several indicators to indicate journals' impacts in the field. We consulted the metrics indicating journals' Impact Factors, 5-Year Impact Factors, Eigenfactor® Scores, and Article Influence® Scores to select the most reputable journals for inclusion in the study.

We considered journals that JCR listed in "*Education & Educational Research*" (*n*=243), in "*Economics*" (*n*=363), in "*Psychology, Educational*" (*n*=59), or in "*Social Sciences, Mathematical Methods*" (*n*=49). As several journals appeared in more than one JCR category, the combined list contained 666 journals considered for inclusion. The goal of consulting the JCR metrics was to ensure a minimum level of quality and rigor, not to reduce the number of articles reviewed. As a result, we included any journal in the top quintile (i.e., 20%) of journals in its category, or in the top quintile of the combined list of considered journals, on any of the four key JCR metrics.

The final list contained 238 journals that qualified for inclusion in the study, representing 36% of all journals considered. Table 1 displays the journals from which at least one article was included in the present review (*n*=16; 7% of qualified journals; the other qualified journals had not published articles in which authors investigated the validity of VAM in teacher evaluation). Since the peer review process did not cease at the time of publication, we also considered the number of times per year the paper in question had been cited since publication as indicated by Google Scholar (https://scholar.google.com/), dropping articles in the lowest quintile of citations per year. This methodological decision reflected our assumption that, although publications in one of the high-quality, reputable journals included in this review would ensure that an article received both rigorous peer review and broad exposure to scholars in economics, education, and/or quantitative methods, its number of annual citations would help indicate the degree to which its contents were incorporated into the scholarly dialogue.

Table 1

Qualified Journals in the Fields of Economics and Education with 2018 Journal Citations Reports® Key Indicators and Number of Articles Included in the Review

	Key JCR Indicators					
Journal	JIF	5-Year JIF	Eigen-factor®	Article Influence®	Field	Articles Included
American Economic Review	4.097	7.048	0.1260	8.348	Econ	6
American Educational Research Journal	3.170	4.861	0.0060	2.177	Educ	5
American Journal of Education	1.316	2.071	0.0013	1.001	Educ	1
Econometrica	4.281	6.723	0.0500	12.055	Econ	1
Economics of Education Review	1.519	2.338	0.0066	1.262	Both	8
Education Finance and Policy	2.429	2.057	0.0020	1.460	Both	3
Educational Evaluation and Policy Analysis	3.127	3.567	0.0047	2.384	Educ	7
Educational Researcher	3.386	5.569	0.0063	2.274	Educ	13
Harvard Educational Review	2.190	4.151	0.0024	1.558	Educ	2
Journal of Economic Perspectives	6.451	9.932	0.0203	6.924	Econ	1
Journal of Educational and Behavioral Statistics	1.767	2.687	0.0029	1.919	Educ	13

	Key JCR Indicators					
Journal	JIF	5-Year JIF	Eigen-factor®	Article Influence®	Field	Articles Included
Journal of Educational Measurement	0.938	2.081	0.0019	1.143	Educ	1
Journal of Policy Analysis and Management	3.828	4.712	0.0063	3.048	Econ	3
Journal of Research on Educational Effectiveness	2.485	2.560	0.0027	1.662	Educ	3
Quarterly Journal of Economics	11.775	14.150	0.0553	20.930	Econ	1
Teachers College Record	0.910	1.727	0.0040	0.622	Educ	7

Note. JCR=*Journal Citations Reports*®, JIF=Journal Impact Factor. The primary field of the journal was determined by consulting the *2016 Journal Citation Reports® Social Sciences Edition* (JCR; Clarivate Analytics, 2017). Journals listed as "Econ" were listed in the "Economics" JCR category, journals listed as "Educ" were listed in the "Education & Educational Research" or the "Psychology, Educational" JCR categories, and journals listed as "Both" appeared in a JCR category from both fields.

Article Selection and Inclusion

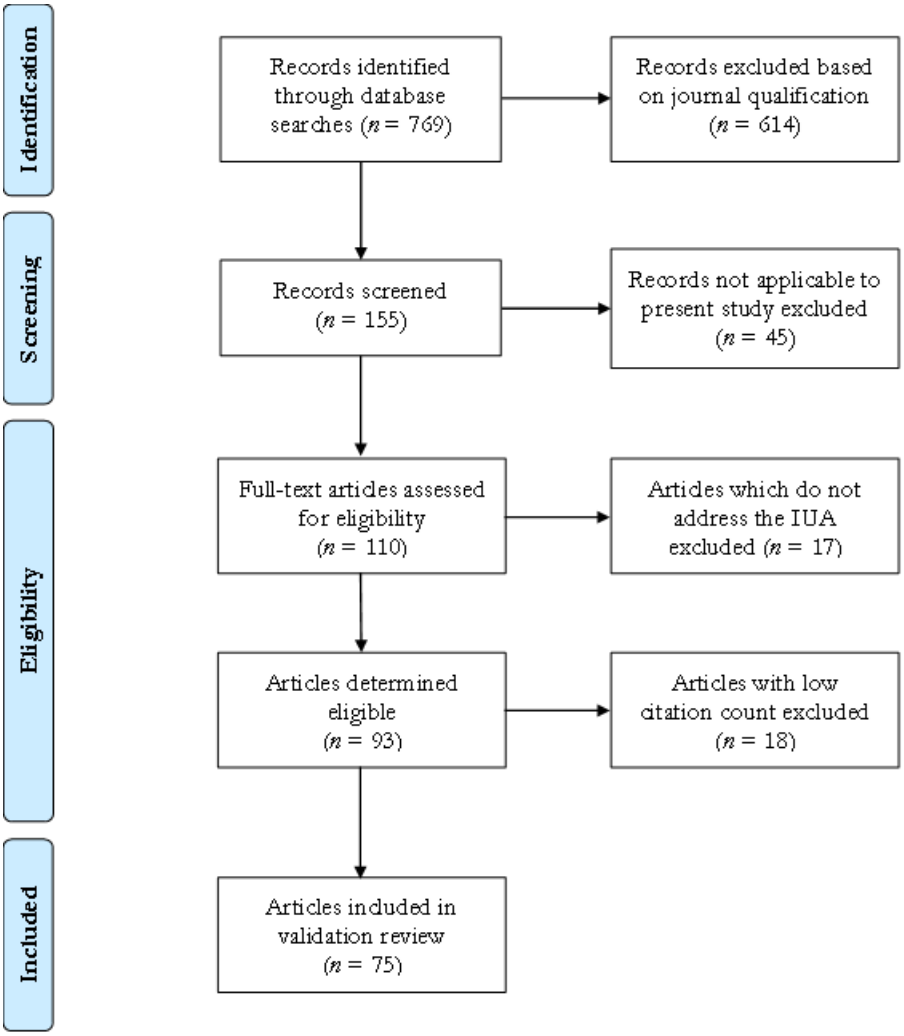
Researchers used the EBSCOhost online research database (EBSCO Industries, 2018) to search the Business Source Complete, the Education Full Text, the Education Research Complete, the Education Resources Information Center (ERIC), and the PsychINFO databases for articles containing variations of the term "value-added," mention of teachers, and some variation of the words "evaluation," "effective," or "quality." We selected articles published any time prior to the end of 2018, three years after ESSA (2015) was passed and states started to demonstrably decline in their use of VAMs, as well as curb state- and level-educational policies surrounding VAMs given ESSA (Close et al., 2020). Around this time, there was also a notable decline in the empirical and otherwise articles written about VAMs and VAM (mis)use. Related, our analyses also indicated that the number of qualifying articles peaked in 2014/2015.

After removing duplicates, the initial search returned 769 records (111 from Business Source Complete, 285 from Education Full Text, 432 from Education Research Complete, 504 from ERIC, and 164 from PsychINFO). We then limited the results to articles published in one of the qualified peer-reviewed journals, which left 155 articles to be screened for possible inclusion in this review (see Figure 2 for the number of articles considered at each stage of the review; see also Moher et al., 2009).

Per the PRISMA guidelines (Moher et al., 2009), we next evaluated the titles and abstracts of the 155 articles to determine which ones met the inclusion criteria. We discussed specific inclusion and exclusion criteria in detail to ensure that all five members of the research team (all of whom have intermediate to extensive expertise in this area of research) applied the same criteria to the articles screened. Specifically, we operationally defined the terms relevant to the IUA under investigation: that VAMs, which use student scores on large-scale standardized tests to estimate teacher effects, support making valid inferences about teacher effectiveness. We defined "inferences about teacher effectiveness" as the use of VAM estimates to compare teachers or groups of teachers to rank-order or categorize them according to their respective contributions to student test scores.

Figure 2

Flow of Articles Through the Phases of the Systematic Review



Note: Model adapted from "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement," by Moher et al. (2009, p. 267).

After discussing the criteria and several discriminant examples, we felt comfortable with the criteria and divided up the 155 articles for screening, whereby each reviewer independently examined each article's title and abstract, pre-inclusion, at least once. If one reviewer was not completely confident about the inclusion or exclusion of an article, they flagged it for a second reviewer to screen. Two reviewers ultimately screened eight of the 155 articles (5%) twice, and we excluded a total of 45 of the 155 articles (29%) as not directly relevant to the IUA (see Figure 2). After screening, 110 of the original 155 articles (71%) remained for full-text review (see more forthcoming).

Review Instrument

To collect data on the validity evidence offered in each of the articles during full-text review, we created an online instrument to facilitate independent review of the 110 articles. We developed the review instrument to reflect the same body of literature that informed our framework for validity evidence (see Figure 1), again, as based on the *Standards* (AERA et al., 2014). Reviewers collaboratively developed the instrument, and piloted and revised it twice. After the second round of revisions, we agreed that no further revisions were necessary. Please see our final Review Instrument in the supplementary file titled "SupplementaryMaterials_ValidationReviewInstrument."

The instrument included space to enter information about the study reviewed, such as the title, authors, and the year of publication. We recorded the position that the authors of the article took on the IUA under investigation, as demonstrated within the text of the paper. These positions were coded using a Likert scale (Likert, 1932) ranging from 5-to-1, with 5=*supportive* (the authors of the publication proposed or supported the IUA); 4=*somewhat supportive* (the authors of the publication generally supported the IUA, though they may have advised some caution), 3=*neither supportive nor challenging* (the authors of the publication provided evidence or discussion in support of the IUA as well as evidence or discussion against it), 2=*somewhat challenging* (the authors of the publication critiqued, challenged, or offered caution about the IUA while acknowledging its use), and 1=*challenging* (the authors of the publication opposed or argued against the IUA). If the authors of the article did not take a position on the IUA and provided no relevant validity evidence, we coded the article as *not applicable* and excluded it from the review.

To capture feedback on each of the six areas of the framework, reviewers recorded whether each area of validity evidence was either *directly studied* (the authors of the article contributed original validity evidence related to the area), *mentioned with evidence* (the authors of the article did not contribute new validity evidence to the literature, but did mention the area while citing evidence provided in other papers), *mentioned* (the authors of the article mentioned the area, but provided no validity evidence), or *not addressed* (the authors did not mention the area of validity evidence). For those areas of the framework which were *directly studied* or *mentioned with evidence*, reviewers rated the degree to which the authors of the paper under review positioned the validity evidence provided as supportive or challenging the use of VAMs in teacher evaluation on a 5-point scale from 5=*supportive* to 1=*challenging* to (the same scale used to rate positions on the IUA). Note that reviewers coded the evidence strictly based on the content presented in the article reviewed, as positioned by its authors. As part of our reliance on rigorous peer review, we assumed that, if VAM scholars did not accept the evidence presented in an article, these scholars would indicate as much in later reviewed articles. Reviewers did not apply their own expertise to judge the quality of the evidence reviewed.

To facilitate the collection of relevant evidence, we broke the *Methods and Measurement* area of the review instrument into the subcomponents of *precision and stability*, *methodological assumptions*, and *construct irrelevant variance (CIV)/bias* (mentioned prior and described more in the forthcoming Results section). The review instrument included at least one text box for each area of the framework for validity evidence, including each subcomponent of *Methods and Measurement* in which reviewers were to insert notes, comments, or copy and paste entire passages directly from the article being reviewed (with page numbers intact) to allow for more qualitative evidence and textual feedback, per area and per article.

Review and Coding Procedures

Two members of the review team reviewed the full text of each article. Reviewer pairs reconciled discrepancies and all members of the team held multiple calibration meetings to discuss individual and collective understandings of the IUA, the framework for validity evidence, and the validation review instrument throughout the coding process. After successful calibration, coding commenced. Reviewers had 78% initial, exact agreement as based on all 19 categorical items for the 110 articles coded for full-text review. During reconciliation discussions, for all 110 articles read, researchers reached 100% agreement on final ratings. As shown in Figure 2, though, 17 articles were excluded as *not applicable* during full-text review, leaving 93 articles applicable to the current review. More specifically, authors of articles excluded as *not applicable* used teacher value-added scores as a covariate or predictor in a study in which they did not directly investigate teacher effectiveness.

As a final application of rigorous peer review as a critical lens for the present study, we recorded the number of times each of the 93 applicable papers had been cited in the literature tracked by Google Scholar (https://scholar.google.com/) as of December 2019. Since older articles had a greater opportunity to be cited than recent papers, we calculated the number of citations per year for each article since the year of publication through 2019 as an indicator of the article's influence in the field. The 93 eligible articles had been cited from 0.67 to 458 times per year (M=27.8, SD=56.7, Mdn=12.5). We dropped the 18 articles in the lowest quintile (i.e., 20%) of citations per year, leaving n=75 articles (marked with an asterisk in the reference list) for final inclusion in the systematic review reported here. See Table 2 for the authors' positions on the IUA, information about the validity evidence provided, and total citations since publication of the 75 articles included in the present review.

Table 2

Position on the IUA Investigated, Mean Ratings for Areas of the Framework for Validity Evidence, and Citation Counts for Each Article Included in the Review

Areas of Evidence				Citation Count	
Article Reviewed	Position	*n*	Mean	*N*	Per Year
AERA (2015)	1.5	4	2.00	83	27.7
Amrein-Beardsley (2008)	2.0	2	2.00	293	29.3
Backes, Cowan, Goldhaber, Koedel, Miller, & Xu (2018)	25.5	4	2.50	16	9.3
Ballou, Sanders & Wright (2004)	3.5	1	4.34	636	45.4
Ballou & Springer (2015)	2.0	2	1.00	101	33.7
Berliner (2014)	1.0	3	1.00	93	23.3
Blazar, Litke & Barmore (2016)	1.0	3	1.44	22	11.0
Braun (2015)	2.0	3	2.33	25	8.3
Buzick & Laitusis (2010)	1.5	1	1.33	66	8.3
Castellano, Rabe-Hesketh & Skrondal (2014)	2.5	1	2.00	22	5.5
Chetty, Friedman & Rockoff (2014a)	4.5	3	5.00	685	171.3
Chetty, Friedman & Rockoff (2014b)	5.0	2	5.00	770	192.5
Chetty, Friedman & Rockoff (2016)	4.0	1	2.33	18	9.0
Collins & Amrein-Beardsley (2014)	1.5	2	3.00	63	15.8

Areas of Evidence				Citation Count	
Article Reviewed	Position	*n*	Mean	*N*	Per Year
Condie, Lefgren & Sims (2014)	2.5	2	2.75	32	8.0
Darling-Hammond (2015)	1.5	4	1.00	137	45.7
Dieterle, Guarino, Reckase & Wooldridge (2015)	2.0	1	1.00	41	13.7
Everson, Feinauer & Sudweeks (2013)	2.0	2	1.00	32	6.4
Fox (2016)	2.0	3	2.33	11	5.5
Gabriel & Lester (2013)	1.5	1	1.00	31	7.8
Goldhaber (2015)	2.5	2	2.00	65	21.7
Goldhaber & Chaplin (2015)	3.0	1	4.00	65	21.7
Goldhaber, Cowan & Walch (2013)	3.5	3	2.44	38	7.6
Goldhaber, Goldschmidt & Tseng (2013)	2.0	2	1.67	47	9.4
Goldhaber & Hansen (2010)	4.5	2	5.00	141	17.6
Good (2014)	2.0	2	3.00	32	8.0
Grossman, Cohen, Ronfeldt & Brown (2014)	2.0	3	1.33	88	22.0
Guarino, Maxfield, Reckase, Thompson & Wooldridge (2015)	3.0	1	1.00	59	19.7
Hanushek & Rivkin (2010)	3.0	2	3.17	700	87.5
Harris (2009)	3.0	2	3.33	43	4.8
Harris, Ingle & Rutledge (2014)	3.5	5	1.47	145	36.3
Harris & Sass (2014)	2.5	4	2.50	62	15.5
Hill (2009)	1.5	3	1.00	56	6.2
Hill, Kapitula & Umland (2011)	2.0	3	1.33	293	41.9
Isenberg & Walsh (2015)	3.0	1	3.00	13	4.3
Johnson, Lipscomb & Gill (2015)	3.0	3	2.00	18	6.0
Jones, Buzick & Turkan (2013)	2.0	2	1.00	61	12.2
Kane & Staiger (2002)	2.0	2	1.00	601	37.6
Kennedy (2010)	2.0	1	1.00	296	37.0
Koedel (2009)	3.0	1	2.00	75	8.3
Koedel, Mihaly & Rockoff (2015)	5.0	3	4.33	132	44.0
Konstantopoulos (2014)	1.0	3	2.67	59	14.8
Kupermintz (2003)	1.5	3	2.00	344	22.9
Lavigne (2014)	3.0	3	1.67	61	15.3
Lee (2018)	2.5	1	2.50	13	4.3
Lefgren & Sims (2012)	4.5	1	4.00	41	6.8
Lockwood, Louis & McCaffrey (2002)	2.5	1	1.00	87	5.4
Lockwood & McCaffrey (2014)	3.0	1	2.00	44	11.0
Lockwood, McCaffrey, Hamilton, et al. (2007)	2.0	2	1.50	257	23.4
Lockwood, McCaffrey, Mariano & Setodji (2007)	3.0	1	3.00	94	8.6
Loeb, Soland & Fox (2014)	3.0	2	1.75	70	17.5
Mariano, McCaffrey & Lockwood (2010)	4.0	1	1.00	47	5.9

Areas of Evidence				Citation Count	
Article Reviewed	Position	*n*	Mean	*N*	Per Year
Martineau (2006)	1.5	3	1.00	168	14.0
McCaffrey, Lockwood, Koretz, Louis & Hamilton (2004)	4.0	1	1.00	663	47.4
Moore Johnson (2015)	2.0	3	1.00	52	17.3
Nye, Konstantopoulos & Hedges (2004)	4.0	3	3.67	1935	138.2
Papay (2011)	2.0	3	1.33	269	38.4
Papay (2012)	2.0	2	1.50	166	27.7
Paufler & Amrein-Beardsley (2014)	1.0	2	1.00	46	11.5
Polikoff (2015)	3.0	2	1.00	43	14.3
Polikoff & Porter (2014)	1.5	5	1.20	108	27.0
Raudenbush (2004)	1.5	1	1.00	299	21.4
Raudenbush (2015)	3.0	1	3.00	21	7.0
Reardon & Raudenbush (2009)	2.0	1	1.00	127	14.1
Reckase (2004)	2.0	1	1.00	66	4.7
Rivkin, Hanushek & Kain (2005)	3.0	1	3.00	5736	441.2
Rockoff & Speroni (2010)	3.0	2	2.50	138	17.3
Rothstein (2009)	2.0	1	1.00	402	44.7
Rothstein (2010)	3.0	2	1.75	968	121.0
Rowan, Correnti & Miller (2002)	2.5	3	2.00	1134	70.8
Rubin, Stuart & Zanutto (2004)	2.0	2	1.00	320	22.8
Schochet & Chiang (2013)	1.5	2	1.50	58	11.6
Stacy, Guarino, & Wooldridge (2018)	2.0	2	1.50	19	3.2
Winters & Cowen (2013)	4.0	3	3.00	64	4.5
Winters, Dixon & Greene (2012)	4.0	3	3.00	29	4.8

Note. Author(s)' positions on and evidence for the IUA under investigation are rated on a scale from 1 to 5, where 1 = *challenging*, 2 = *somewhat challenging*, 3 = *neither supportive nor challenging*, 4 = *somewhat supportive*, and 5 = *supportive*. Counts (*n*) and means reported under "Areas of Evidence" indicate the number of areas of the framework (see Figure 1) for which authors of the indicated article reported validity evidence relevant to the use of VAMs in teacher evaluation and the mean rating of its support for the IUA rated on the same five-point scale used for the authors' positions. Number of citations (*n*) reported under "Citation Count" were taken from Google Scholar (https://scholar.google.com/).

Results

To gauge the findings of the high-quality literature from the most reputable peer-reviewed journals regarding the IUA under investigation, the central goal was to synthesize the evidence found within the articles reviewed, and to determine the magnitude and direction of its weight as positioned by the authors of the studies reviewed. Researchers present the findings of this review by examining the validity evidence itself, as framed by the *Standards* (AERA et al., 2014) next.

Overall Support or Challenge

On average, authors provided evidence for two areas of the framework for validity evidence (M=2.1, SD=1.0). To determine the weight and direction of the evidence across all articles reviewed, we calculated both an unweighted mean (in which each article is considered equivalent to all other articles) and a weighted mean (in which the ratings are weighted by the number of citations per year for that article). We recognize that our validation review instrument had not itself been sufficiently validated for the scores derived from it to be interpreted as a direct measure of validity. However, we calculated these scores as a form of quantitizing (Sandelowski et al., 2009) to aid in our own analyses, and present them to the reader to support interpretation the qualitative syntheses of the reviewed validity evidence. Across all articles we included, the validity evidence from all six areas of the framework was somewhat challenging to the use of VAMs in teacher evaluation with an unweighted mean rating of 2.08 (SD=1.15). When weighted by annual citation count, the articles reviewed had a mean rating of 2.68 (SD=1.38), which was neither supportive nor challenging. The weighted mean challenge rating was significantly different than the unweighted mean, t (2616)=3.76, $p < .001$, $\overline{D}$=0.60, 95% CI [0.34, 0.87], d=0.44, which yielded a small-to-medium effect (Cohen, 1992). The difference between the weighted and unweighted means suggests that the reviewed papers whose authors were more supportive of VAMs had been cited more frequently than the papers whose authors were more critical of VAMs. It is unknown whether the higher citation count suggests that authors of these papers were more accepted or influential among VAM scholars, but the papers were generally more prominent in the scholarly debate. Thus, in answer to the first research question (and absent a discussion of the specific validity evidence which is given in the forthcoming sections), these results suggest that the accumulated validity evidence did not support (when weighted), or somewhat challenged (when unweighted), the use of VAMs in teacher evaluation.

Validity Evidence by Area

The second research question we used to ask whether evidence provided in each of the areas of the framework (see Figure 1) supported or challenged the use of VAMs in teacher evaluation. Table 3, forthcoming, displays mean support ratings (both unweighted and weighted by citations per year), capturing how authors of the reviewed literature positioned the evidence provided across all 75 articles for each area of validity evidence. These findings also suggest that the accumulated validity evidence did not fully support the use of VAMs in teacher evaluation, regardless of the source of validity evidence. However, in some cases when annual citation counts were taken into consideration, the validity evidence seemed less challenging of the use of VAMs in teacher evaluation. Each subarea of validity evidence is discussed in more detail next.

Methods and Measurement

Although it is true that neither the test nor its scores are valid or invalid, the IUA does rely upon the quality of the scores that inform it. The *Standards* (AERA, et al., 2014) assert, "reliability/precision of measurement is always important. However, the need for precision increases as the consequences of decisions and interpretations grow in importance" (p. 33). This is important, here, in that performance evaluations are both important and consequential. Beyond reliability, any source of random or systemic error evident in the scores that inform the IUA may also compromise the decisions made upon their basis. Thus, researchers looked for possible violation or satisfaction of methodological assumptions. We also included CIV as a source of bias and as a source of validity evidence, as well.

More authors reported validity evidence in the *Methods and Measurement* area than any of the other areas in the framework for validity evidence (n=74; 99%), reporting evidence that somewhat challenged the IUA when unweighted (M=1.95, SD=1.18) but was significantly less challenging when weighted by citation count (M=2.54, SD=1.40), t (2609)=3.57, $p < .001$, $\overline{D}$=0.59, 95% CI [0.31, 0.86], d=0.42, which yielded a small-to-medium effect (Cohen, 1992). This finding, again, suggested that papers in which authors challenged VAMs and VAM use were cited less frequently in the VAM literature than more supportive papers. This area may have received so much attention in the literature because of the "mathematical exactness of procedure" it requires, as Pittenger (1917, p. 103) reported over a century ago. Since the validation review instrument also contained three subcomponents for this area of validity evidence, each of which reviewers used to measure different yet related aspects of reliable, precise, and fair methods and measurements supporting the IUA, we also examined each subcomponent separately (see Table 3).

Table 3

Number of Articles with Validity Evidence by Subcomponent of Methods and Measurement with Unweighted and Weighted Mean Rating of Evidence

			Unweighted		Weighted [b]			
Area of Validity Evidence	*n*	(%) [a]	*M*	(*SD*)	*M*	(*SD*)	*t*	*d*
Overall (any of the subcomponents)	74	(99%)	1.95	(1.18)	2.54	(1.40)	3.76***	0.44
Precision and stability	49	(65%)	1.59	(1.19)	2.12	(1.55)	2.37*	0.34
Methodological assumptions	47	(63%)	1.83	(1.20)	1.96	(1.09)	0.79	0.12
Construct irrelevant variance / bias	56	(75%)	2.21	(1.37)	2.75	(1.43)	2.77**	0.37

Notes. For each area of the framework for validity evidence, evidence is rated on a scale from 1 to 5, where 1=*challenging*, 2=*somewhat challenging*, 3=*neither supportive nor challenging*, 4=*somewhat supportive*, and 5=*supportive*. The *t*-test and Cohen's *d* reported represent a test of mean differences between the weighted and unweighted means.

[a] Percent of the 75 articles included in the review

[b] Mean weighted by article citations per year (see Table 2)

* $p < .05$. ** $p < .01$. *** $p < .001$.

Precision and Stability. The authors of 49 reviewed articles in this area provided evidence related to the precision and stability of teacher effectiveness scores derived from students' large-scale standardized test scores. The unweighted mean rating of the evidence provided by the authors of these articles was somewhat challenging to the IUA investigated (M=1.59, SD=1.19), while the weighted mean was a significantly less challenging (M=2.12, SD=1.55), t (1425)=2.37, p=.018, $\overline{D}$=0.53, 95% CI [0.19, 0.87], d=0.34, which yielded a small effect (Cohen, 1992). As we examined the comments and quotations entered into the review instrument related to this subcomponent, a pattern emerged in which the authors of reviewed articles largely agreed on the stability of VAM scores, but they disagreed on the implications of that stability.

Across the board, there was consistent agreement with the evidence that Koedel et al. (2015) found that "the year-to-year correlation in estimated teacher value-added . . . range[s] from 0.18 to 0.64" (p. 186; see also Amrein-Beardsley, 2014; Backes et al., 2018; Goldhaber & Hansen, 2010; Harris, 2009; Hill, 2009; Jones et al., 2013; Kennedy, 2010; Lockwood et al., 2002; Nye et al., 2004;

Papay, 2011, 2012). However, some scholars positioned this as adequate for evaluative purposes (e.g., Goldhaber, 2015; Goldhaber & Hansen, 2010), and others positioned it as inadequate to justify the use of students' large-scale test scores for teacher evaluation, especially when consequences were attached, (e.g., Gabriel & Lester, 2013; Hill, 2009; Jones et al., 2013; Papay, 2012). Konstantopoulos (2014), for example, found the stability of VAM estimates of teacher effects like "estimates used for high-stakes decisions in other fields such as health care, real estate, and sports" (p. 12), but he argued against the use of VAMs for high-stakes decisions in their current form. Goldhaber and Hansen (2010), on the other hand, found that VAM measures are stable enough to predict student achievement three years later, arguing that "this finding lends credence to the notion that these implicit measures of teacher quality are a reasonable metric to use as a factor in making substantive personnel decisions" (p. 254).

The conversations surrounding precision and stability were also split in terms of the implications of such estimates of stability and precision. Some authors posited that there were ways of increasing stability (e.g., including multiple years of data or optimally weighting subject area tests; Hanushek & Rivkin, 2010; Lefgren & Sims, 2012), and others were more concerned with the mislabeling of teachers (Berliner, 2014; Hill, 2009) and what doing so might entail in terms of the personnel decisions attached to such outcomes (e.g., teacher improvement plans, merit pay, tenure decisions, teacher termination). Blazar et al. (2016) found that context matters, and that VAM categorizations "were sensitive to within- versus across-district comparisons and the specific set of teachers to whom an individual teacher was compared" (p. 337).

We acknowledge that, anytime a method is used to put people in categories, there is likely to be some misclassification, particularly around the cut points. Perhaps that is why AERA (2015) includes, as one of its eight technical requirements for the use of VAM, that statistical estimates of error must be included in classifications and reports based on VAMs. They also include strong recommendations related to maximizing precision, writing that:

> VAM scores should not be used unless they are derived from data obtained from sufficient numbers of students over multiple years. VAM scores should always be accompanied by estimates of uncertainty to guard against over interpretation of differences. Further, care must be taken to address estimate instability that results from teacher mobility across schools, grades, and subjects. (AERA, 2015, p. 450)

There were also researchers who cautioned about conditions that impact stability. For example, Goldhaber, Goldschmidt et al. (2013) warned that specifying the correct format, selecting the right parameters, and modeling their relationships to one another is very important in classifying teachers, stating that "there is a considerable number of teachers who end up switching quintiles based solely on model specification" (p. 230). In their comparison of model specifications, Johnson et al. (2015) found that "26% of teachers who are ranked in the bottom quintile under one specification would be ranked above this quintile under the alternative specification" (pp. 63-64). Similarly, Guarino et al. (2015) found in their investigation of different VAM estimators that 15% of above average teachers were misclassified as below average, regardless of estimator, and that "estimators also misclassif[ied] 28% of the teachers who should [have been] classified in the bottom quintile" (p. 209). Schochet and Chiang (2013) provided evidence that, not only are three years of data needed for estimates to be considered reliable, but also that "type I and II error rates for teacher-level analyses will be about 26% if 3 [sic] years of data are used for estimation" (p. 166). Finally, McCaffrey et al. (2004) found that "posterior precision would need to be 2 to 4 [sic] times greater to provide meaningful estimates of ranks and accurate identification of teachers in [the] extremes of the distribution" (p. 96). Overall,

authors of the literature reviewed generally agreed about the numbers and the coefficients surrounding the precision and stability of teacher effectiveness estimates, though disagreement remained on whether VAM-based estimates are precise or stable *enough* for use in teacher evaluations.

Methodological Assumptions. Authors of 47 of the reviewed articles (63%) provided evidence related to the methodological assumptions of VAMs. There was no difference in the mean rating of the evidence provided by the authors of these articles to the IUA investigated whether unweighted (M=1.83, SD=1.20) or weighted by citation count (M=1.96, SD=1.09), t (1257)=0.79, p=.429, $\overline{D}$=0.13, 95% CI [-0.22, 0.48], d=0.12, which yielded a small effect (Cohen, 1992). Reardon and Raudenbush (2009) identified six assumptions associated with drawing causal inferences from VAMs. Though the authors discussed the estimation of school effects, they pointed out that "the basic logic of [their] argument would [remain] unchanged if [they] considered the estimation of teacher effects instead" (p. 498-499). They asserted that VAMs assume (a) that student assignment practices are manipulable (i.e., any given student could theoretically be assigned to any of the teachers evaluated), (b) that there is no interference between units (i.e., that a student's outcomes with a given teacher does not depend on the other students assigned to the same teacher), (c) that test scores are measured on an interval scale that supports meaningful comparisons between the teachers evaluated, (d) that teacher effects are homogeneous and stable (i.e., that all teachers teach all students equally well and consistently over time), (e) that the assignment of students to teachers is strongly ignorable, or as good as random, and (f) that the measurement model can estimate the effect of all teachers for all types of students, even if that teacher is never assigned students of certain types. Reardon and Raudenbush (2009) concluded that few, if any, of these assumptions are plausible, however. Nonetheless, the Reardon and Raudenbush (2009) list of assumptions effectively captured some of the major themes in the literature reviewed regarding methodological assumptions.

One theme, which also has implications for the precision and stability of the effectiveness estimates discussed prior, was the assumption that teacher effects are homogeneous and stable over time (i.e., assumption d). Some reviewed authors reminded us that "we have known for some time that teachers' effects on student achievement are highly unstable from year to year" (Good, 2014, p. 8; see also Kennedy, 2010). Condie et al. (2014) found evidence of differential teacher effects both by subject and by student factors. For Condie et al. (2014), the observed differential effects offered caution in the use of VAMs in teacher evaluation, but they still recommended their use stating that "under any reasonable conditions, commonly suggested personnel policies based on value-added [would] raise student achievement" (p. 89). Fox (2016) also found teacher effects to be heterogeneous, but they argued this as a strength of VAMs, suggesting that human resource policies play to teachers' strengths by assigning them to teach the students with whom they are most effective.

The second assumption asserted by Reardon and Raudenbush (2009) pertains to the presence of peer effects (i.e., assumption b). Berliner (2014) included several pages of evidence from studies in which authors challenged this assumption, while Kennedy (2010) discussed the contribution of classmates as part of co-constructed learning environments, adding that the "notion that we should hold teachers accountable for student outcomes presumes that those outcomes are largely in teachers' hands and overlooks the role of the textbook, the physical space, and other resources" (p. 592; see also Schochet & Chiang, 2013). Isenberg and Walsh (2015) explored another source of interference between units, namely co-teaching, which allows both teachers to directly

impact student learning. The authors tested several different methods of accounting for co-teaching, finding each method equally acceptable (Isenberg & Walsh, 2015).

The third assumption asserted by Reardon and Raudenbush (2009) is that student scores are measured on an interval scale that supports comparisons (i.e., assumption c). This was examined by authors of five papers in the present review (Buzick & Laitusis, 2010; Harris, 2009; Mariano et al., 2010; Martineau, 2006; Reckase, 2004). The papers of Buzick and Laitusis (2010) and Harris (2009) included discussions of the potential problems caused when estimates of teacher effectiveness are drawn from tests that are not properly vertically aligned. Buzick and Laitusis (2010) examined several methods of vertical scaling, finding that it can help, "but only when the variability in student achievement on a proposed vertical scale decreases substantially from grade to grade relative to the pattern that would have been observed in an alternative scaling" (p. 552). Martineau (2006) simulated the problem of construct shift, common when complex multidimensional constructs are treated as unidimensional, finding that current teacher effects can be corrupted by past teachers who taught different dimensions of the measured construct. Mariano et al. (2010) offered a solution by offering a method of estimating teacher effects that purports not to depend on vertical scaling, and Reckase (2004) discussed several scaling problems and alluded to multidimensional item response theory as holding some promise.

Another theme that emerged from discussions of methodological assumptions in the literature reviewed was the nonrandom assignment of students and teachers (i.e., Reardon & Raudenbush's assumption e). Authors of reviewed papers discussed nonrandom assignment of students and teachers as concerning, particularly in terms of biasing (upwards or downwards) teachers' estimated effects (Condie et al., 2014; Everson et al., 2013; Hill, 2009; Kupermintz, 2003; Paufler & Amrein-Beardsley, 2014; Rothstein, 2009). Inversely, Rockoff (2004) tested his dataset to determine whether classroom assignment was random, ultimately asserting that student "assignment appear[ed] similar to random assignment" within the districts he studied, and bias was, thus, "unlikely to affect [his] results" (p. 248). Johnson et al. (2015) also acknowledged that although a small number of authors of studies who used random assignment boded well for VAMs, authors of these studies "examined only a few of the many possible VAM variations" (p. 61). Blazar et al. (2016) argued that nonrandom assignment of students to teachers may not reduce the accuracy of VAMs, as much as the nonrandom assignment of both teachers and resources to districts and schools. In two recent papers that were not included in this review, Yeh (2019, 2020) argued that the methods of random assignment used in the studies that Johnson et al. (2015) found promising were insufficient to counteract the kinds of nonrandom assignment that Blazar et al. (2016) discussed. Though not an assumption identified by Reardon and Raudenbush (2009), Amrein-Beardsley (2008) noted that many VAMs assume non-informative missing data (see also McCaffrey et al., 2004; Raudenbush, 2004), or data that are missing at random, stating that "this assumption is extremely problematic because it is well known that disproportionate numbers of students who do not participate in large-scale tests are low performing" (p. 69).

CIV and Bias. The authors of more reviewed articles examined evidence of bias (n=56; 75% of the articles reviewed) than the other two subcomponents of the *Measurement and Methodology* area. There was a difference in the rating of CIV and bias evidence provided by these authors between the unweighted mean (M=2.21, SD=1.37) and the mean weighted by annual citation count (M=2.75, SD=1.43), t (2284)=2.77, p=.006, $\overline{D}$=0.53, 95% CI [0.17, 0.90], d=0.37, which yielded a small-to-medium effect (Cohen, 1992). The significant difference between the weighted and unweighted means, again, suggested that papers in which authors challenged the valid use of VAM

in teacher evaluation was not promoted by VAM scholars as broadly as papers which supported the general use of VAMs. Specifically, authors of eight of the articles directly studying this issue found evidence suggesting freedom from bias, thus supporting the use of VAMs in teacher evaluation (Backes et al., 2018; Ballou et al., 2004; Chetty et al., 2014a, 2014b; Goldhaber & Chaplin, 2015; Goldhaber, Cowan et al., 2013; Lockwood, McCaffrey, Hamilton, et al., 2007; Loeb et al., 2014), while double that number, or authors of 16 articles found evidence of bias, thus challenging the IUA (Ballou & Springer, 2015; Castellano et al., 2014; Chetty et al., 2016; Dieterle et al., 2015; Goldhaber, Goldschmidt et al., 2013; Guarino et al., 2015; Harris et al., 2014; Johnson et al., 2015; Kupermintz, 2003; Lockwood & McCaffrey, 2014; Martineau, 2006; McCaffrey et al., 2004; Polikoff & Porter, 2014; Rothstein, 2009, 2010; Stacy et al., 2018).

At first glance, it may seem difficult to reconcile these two lists of references, especially when the names of multiple researchers, such as Chetty et al. (2014a, 2014b) appear in both. However, each of the authors who directly studied bias used carefully qualified language to emphasize that, though the measurement models they examined appeared free from bias under certain conditions, they may still demonstrate vulnerability to bias under different conditions. As Lockwood, McCaffrey, Mariano et al. (2007) wrote, as based on their analyses:

> The effects of interest are not necessarily causal effects or intrinsic characteristics of teachers. Rather, they account for unexplained heterogeneity at the classroom level. Ideally, they provide information about teacher performance, but there might be many sources of this heterogeneity, including omitted student characteristics. (p. 127)

As a further illustration, authors of another seven papers found evidence of bias under some conditions and absence of bias under other conditions (Fox, 2016; Harris et al., 2014; Koedel, 2009; Koedel et al., 2015; Nye et al., 2004; Rowan et al., 2002; Rivkin et al., 2005). These findings seem to suggest, just as Kane (2013) and the authors of the *Standards* (AERA et al., 2014) did, that each specific interpretation and use of test scores should be independently validated, as a score which is suitable in one context may be unsuitable in others. Further, the evidence found in the present review did not suggest that eliminating bias is impossible, but that doing so is very challenging.

Validity Evidence Based on Test Content

The first source of validity evidence discussed in the *Standards* (AERA et al., 2014) is validity evidence based on the content of the test(s) used to support the IUA. Originally discussed in the literature as content-validity, this source of evidence is primarily concerned with the relationship between the content of the tests and the constructs the tests are purported to measure (Cronbach, 1971; Cureton, 1951; Kane, 2013). Validity evidence of this type may include but not be limited to analyses of the content domain as aligned to the tested domain, discussions of the relevance or sufficiency of the tested content domains as pertinent to the proposed interpretations, or recommendations and testimonials taken from content-area experts (AERA et al., 2014; Kane, 2013).

There are times, however, when the inferences made about a construct are not directly measured by a test. Perhaps the most salient example was the evaluation of teacher effectiveness in the plane of results discussed prior by Pittenger (1917). The logic of the IUA under investigation was that, since teachers are responsible for promoting student learning, and since standardized tests are chiefly designed to measure student learning, then student growth on standardized tests can be used to evaluate teachers. Since the standardized tests used in VAMs are not typically designed to measure effective teaching, we included articles in this section in

which authors examined how closely the standardized tests administered to students aligned with the instruction they purportedly received from the teachers evaluated. Authors of 13 of the 75 reviewed articles (17%) discussed validity evidence related to test content, providing evidence that somewhat challenged the IUA that was no different whether the mean rating was unweighted (M=1.92, SD=1.11) or weighted by citation count (M=2.25, SD=0.98), t (438)=1.18, p=.239, $\overline{D}$=0.33, 95% CI [-0.29, 0.94], d=0.33, which yielded a small effect (Cohen, 1992).

Authors of the articles recorded in this area provided evidence supporting the general statement made by Nye et al. (2004), "that the effects of . . . teacher effectiveness are expected to be largest when the content covered during instruction is closely aligned with school outcomes such as student achievement measures" (p. 253). Lockwood, McCaffrey, D. F., Hamilton et al. (2007) investigated the differences in teacher VAM scores generated by different subscales of the same mathematics assessment, while Grossman et al. (2014) and Papay (2011) directly investigated the sensitivity of VAM estimates to different student achievement tests used in support of the IUA. Findings from both of sets of authors' studies indicated that different measures yield different results. Papay (2011) concluded that "using different achievement tests produces substantially different estimates of individual teacher effectiveness. The variation in teacher value-added estimates that arises from using different outcomes far exceeds the variation introduced by implementing different model specifications" (p. 3).

These general assertions, as also related to Martineau's (2006) investigation of construct drift, suggest that teacher effectiveness estimates may be sensitive to the specific knowledge and skills measured on such tests. Polikoff and Porter (2014) studied instructional alignment as a predictor for VAM scores using data from the Bill & Melinda Gates Foundation's (2013) Measures of Effective Teaching (MET) studies, finding weak correlations between measures of instructional alignment and teacher value-added. If, as the evidence reviewed seemed to suggest, teacher effectiveness scores are very sensitive to the specific achievement test administered to students, it might make sense that correlations between measures of instructional alignment and value-added scores are stronger than those that Polikoff and Porter (2014) observed. The authors of the study themselves stated that "the correlations [they] found were much smaller than [they] expected. The design anticipated an increase in R^2 of .10, suggesting a correlation of greater than .30. In fact, there were few if any correlations that large" (Polikoff & Porter, 2014, p. 411).

Braun (2015) discussed the problem of comparing teachers for whom VAM estimates are available to the more than 70% of teachers who teach in grade levels or content areas which prevent them from receiving a VAM score. AERA (2015) suggests that "VAM scores must not be calculated in grades or for subjects where there are not standardized assessments that are accompanied by evidence of their reliability and validity" (p. 450). Teachers who teach in a tested content area and grade level may face different measurement challenges, however. Darling-Hammond (2015) explained that annual state achievement tests were required by NCLB (2002) policy mandates to measure only the appropriate grade-level content standards, which prevented these tests from accurately measuring the achievement of students who performed above or below grade level, creating both floor and ceiling effects (Darling-Hammond, 2015).

Accordingly, reviewed literature suggested that teacher effectiveness estimates vary substantively based on different tests within the same content area (Lockwood, McCaffrey, Hamilton, et al., 2007; Papay, 2011), even when the same sample of students taking such tests and the time of administration were the same or held constant. More recent analyses by Backes

et al. (2018) suggested that VAM estimates are robust to changes in the achievement test administered, however. As per Grossman et al. (2014) "researchers and policymakers need to pay careful attention to the [tests] used to measure student achievement in designing teacher evaluation systems as these decisions will yield different results" (p. 301).

Validity Evidence Based on Response Processes

The *Response Processes* source of evidence captures how well the construct or performance of the target domain matches the responses or performances measured to inform the IUA (AERA et al., 2014). When the responses or performances measured by a test are substantially different from that of the target domain, irrelevant method variance may limit the validity of proposed interpretations of the test (Messick, 1989, 1995). For individual tests, evidence in this area may include but not be limited to investigations into how test takers interpret and respond to the tests used, logical or empirical analyses of the match between response processes and the target domains, or studies of raters and graders and the processes that they employ (AERA et al., 2014). Although there is not a clear and straightforward correlation between evidence based on response processes for individual tests and for VAMs, we examined this source of evidence by asking, "In what way does the response process measured by VAM (i.e., raising students' scores on large-scale standardized tests) match or differ from effective teaching?" In the *Response Processes* area, reviewers coded articles in which the authors examined whether VAMs, which are meant to capture the teacher's contribution to student achievement gains, are sufficient to measure the construct of teacher effectiveness. In this area, we considered Kane's (2013) suggestion that, "if the [measured] tasks seem to involve the same processes as most tasks in the target domain, extrapolation is likely to seem reasonable" (p. 28).

Authors of eight reviewed articles (11%) provided validity evidence related to response processes, which somewhat challenged the IUA whether the mean was unweighted (M=1.63, SD=0.92) or weighted by annual citation count (M=1.54, SD=0.76), t (173)=-0.29, p=.770, $\overline{D}$=-0.08, 95% CI [-0.73, 0.57], d=-0.11, which yielded a very small effect (Cohen, 1992). The concern for this source of validity evidence was whether VAMs are sufficient to capture the construct of teaching effectiveness. Standardized tests may simply not capture important components of teaching (see also Good, 2014; Konstantopoulos, 2014). Notably, in the second most regularly cited paper reviewed, Chetty et al. (2014b) examined VAMs as a predictor of some other possible outcomes of teaching measured in adulthood. The Chetty et al. (2014b) paper is discussed more in the forthcoming section on validity evidence related to other variables.

When comparing confidential supervisor evaluations to VAM results, Harris and colleagues (Harris et al., 2014; Harris & Sass, 2014) found that principals considered criteria beyond teachers' contributions to student test scores when they evaluated teacher quality, which they distinguished from teacher effectiveness (as measured by VAMs; see also Backes et al., 2023). This suggests that there may be more to the construct than can measured by student test scores alone. However, as Moore Johnson (2015) discussed, "principals who are not instructional experts [could] be left to interpret discrepancies between what they see in the classroom and what they read on a VAM score sheet" and suggest that "value-added scores may unduly influence their decisions" (p. 121). Thus, authors of the literature reviewed seemed to collectively suggest that teacher quality may be too complex and multidimensional to be measured solely by student test scores.

Validity Evidence Based on Internal Structure

In the historical validity literature, evidence based on internal structure is related to construct validation (Cronbach & Meehl, 1955) and concerns the degree to which the test or analytical process that supports the IUA reflects the theory on which it is based. This includes evidence of how faithfully the sub-scales, sub-tests, or various components of the testing procedure follow their hypothesized relationships. Across the literature reviewed, however, authors of very few studies probed the internal structure of measurements of teacher effectiveness based on VAMs. Indeed, authors of only three of the 75 articles reviewed (4%) reported validity evidence related to internal structure which challenged the use of VAMs in teacher evaluation, whether examining the weighted mean (M=1.30, SD=0.46) or the unweighted mean (M=1.33, SD=0.58), t (97)=-0.13, p=.894, $\overline{D}$=-0.04, 95% CI [-0.70, 0.63], d=-0.08, which yielded a very small effect (Cohen, 1992). Harris et al. (2014) performed an exploratory factor analysis (EFA) to identify latent constructs measured by principal evaluations and VAM-based measures. Though in their results they identified several latent constructs that were significantly related to principal evaluations, only *technical skill* was significantly related to teachers' value-added scores in both mathematics and English/language arts (ELA). In their analyses of data from the MET Project (Bill & Melinda Gates Foundation, 2013), Polikoff and Porter (2014) assigned several different weights when calculating an overall effectiveness score using value-added measures, observations, and student surveys. The authors found little difference between weights, noting that the composite scores that they tried had "weak reliability, because of low correlations among the components—internal consistency reliability [r =].40 in mathematics and [r =] .30 in ELA" (Polikoff & Porter, 2014, p. 410). Related to these investigations, AERA (2015) added as one of its eight technical requirements for the use of VAM that "VAM scores must never be used alone or in isolation in educator or program evaluation systems" (p. 450), but that they should be used as but one of multiple measures that is "integrated into judgments about overall teacher effectiveness" (p. 450).

Validity Evidence Based on Relations to Other Variables

Relationships between the measurement of interest and other variables have been historically discussed as a primary aspect of criterion validity (Cronbach, 1971; Kane, 2013; Moss, 1992, 1995). Such evidence may be concurrent or convergent, demonstrating the degree to which two measures of constructs that theoretically should be related are, in fact, related; discriminant, demonstrating the degree to which concepts or measurements that are supposed to be unrelated are, in fact, unrelated; or predictive, concerning the degree to which measurement output can be used to predict other outcomes, which are typically assessed or observed at some later point in time. In all cases, researchers typically seek evidence related to other variables to demonstrate that the IUA is consistent with conclusions based on other related measures, observations, or outcomes. Authors of 23 reviewed articles (31% of the articles reviewed) discussed validity evidence based on such relations to other variables. There was a difference in the weight of accumulated validity evidence in this area between when treating reviewed articles equally (M=2.65, SD=1.64), and when weighting articles by citation count (M=3.42, SD=1.70), t (832)=2.14, p=.033, $\overline{D}$=0.77, 95% CI [0.09, 1.45], d=0.45 (a small-to-medium effect; Cohen, 1992) which, again, suggested that papers in which authors challenged the use of VAMs to evaluate teachers were promoted in the VAM literature more than papers which supported VAMs and VAM use.

More specifically, the concurrent validity evidence (i.e., collected at the same time) in this area was quite consistent in that teacher effects estimated from students' large-scale test scores did not demonstrate strong relationships to other variables that should, theoretically, also measure teacher effectiveness. For example, Grossman et al. (2014) investigated the relationships between the Protocol for Language Arts Teaching Observation (PLATO) and value-added scores, analyzing data from the MET Project (Bill & Melinda Gates Foundation, 2013). They found that PLATO scores were weakly related to VAM scores based on the SAT-9, but PLATO scores were even more weakly related to VAM scores based on the state tests. They concluded that "PLATO measures designed specifically to identify ambitious instructional practices are especially sensitive to which test is used to construct value-added scores" (Grossman et al, 2014, p. 7). However, in his analysis of the MET Project data for relationships between teacher VAM estimates and other measures of effectiveness, Raudenbush (2015) found that "the relationships, though not strong, were statistically reliable and large enough to be of practical significance" (p. 138). Authors (2014) wrote that research on the topic suggests that value-added scores are mildly related to observational scores but even less related to teachers' portfolio scores, and Authors (2014) identified correlations between multiple teacher evaluation measures as unacceptably low. In their review of the VAM literature, Koedel et al. (2015) found VAM scores to be positively but imperfectly correlated across subjects, across different tests in the same subject, and across other teacher evaluation metrics (see also Blazar et al., 2016). Harris et al. (2014) also found low correlations between teacher observations and VAM scores, but while similar in size, they called for a new way to conceptualize this discrepancy, arguing that the problem was that teacher effectiveness and teacher quality constructs are too often considered synonymous. Rather, they positioned the two constructs as distinct from one another; hence, they suggested they should be measured differently, and they should not necessarily correlate highly. Fox (2016) acknowledged a similar point but positioned it as a weakness of VAM-use in teacher assignment decisions. She argued that if aspects of effective teaching that support non-tested but positive outcomes are not measured by VAMs, then assigning either teachers or students on the basis of VAM may lead to higher test scores in the short term at the expense of valued long-term outcomes.

The predictive validity evidence (i.e., collected one in advance of the other) offered by authors varied, not only in terms of what was predicted, but also in terms of the degree of support or challenge also offered to the IUA. Hill et al. (2011) provided evidence that VAM scores are the strongest predictors of later VAM scores. Rockoff and Speroni (2010) found that a teacher's prior-year VAM scores predicted the achievement of his or her current students, while Rothstein (2010) found that fifth grade teacher VAM scores predicted students' fourth grade achievement just as well as they predicted current achievement, suggesting that VAM scores may measure contextual effects much more strongly than teacher effects. Chetty et al. (2014a, 2014b) used a longitudinal data set of achievement data and tax records to demonstrate that students of high value-added teachers were less likely to become teenage parents, and more likely to graduate, attend college, and earn higher income as adults. Polikoff and Porter (2014) also tested whether instructional alignment had a relationship with VAM outcomes, yet they found almost no relationship, raising the question "If VAMs are not meaningfully associated with either the content or quality of instruction, what are they measuring?" (p. 414).

Validity Evidence Based on Related Consequences

If the inferences and decisions that the scores support must be validated, rather than the scores themselves, then it is appropriate to also evaluate the consequences of those score uses as part of the validity argument. Whether the IUA produces its (positive or negative) intended outcomes is an important consideration, as is whether the IUA yields (positive or negative) unintended outcomes, as well as, or in lieu of, those which are intended. Kane (2013) wrote that "a decision rule that achieves its goals at an acceptable cost and with acceptable consequences is considered a success. A decision rule that does not achieve its goals or has unacceptable consequences is considered a failure" (p. 47). Note, however, that not all consequences might be immediately apparent and that, particularly in the case of unintended consequences, an IUA might be in use for some time before the full extent of its consequences are known and understood. Subsequently, validity evidence based on related consequences is the element of the framework most likely to reopen a scholarly discussion that might have previously been thought settled or resolved. The authors of the 38 reviewed articles herein discussed consequential evidence (51% of all papers reviewed). The mean rating of the evidence provided by the authors of these articles was more challenging to the IUA when unweighted (M=2.23, SD=1.44), than when weighted by citation count (M=2.91, SD=1.75), t (1373)=2.36, p=.018, $\overline{D}$=0.68, 95% CI [0.21, 1.14], d=0.39, which yielded a small-to-medium effect (Cohen, 1992).

Notably, the authors of the reviewed articles discussed consequential validity evidence more frequently than any other area of evidence, besides *Methods and Measurement* (discussed prior). The authors of 25 articles discussed evidence related to whether the intended consequences of the IUA were being met (33%), and authors of five of these articles presented evidence that supported the IUA (Chetty et al., 2014a, 2014b; Goldhaber, 2015; Goldhaber & Hansen, 2010; Hanushek & Rivkin, 2010; Koedel et al., 2015). Hanushek and Rivkin (2010) evidenced that replacing the worst 6-10% of teachers with average teachers would hypothetically raise overall student achievement (i.e., a positive consequence). Chetty et al. (2014a, 2014b) tested a series of potential positive consequences and suggested several benefits to students who were taught by a high value-added teacher. Likewise, Lee (2018) found that students having a series of high-VAM teachers was associated with an increased likelihood of students attaining an undergraduate degree, positioning this as evidence of VAMs having their desired effect. Extending upon the recommendation of Hanushek and Rivkin (2010), Chetty et al. (2014a, 2014b) also claimed that "replacing a teacher whose current VA is in the bottom 5 percent with an average teacher would increase the mean present value of students' lifetime income by $250,000 per classroom over a teacher's career" (p. 2635). Berliner (2014) stood out as one of the reviewed authors who explicitly rivaled this line of thinking, stating that a "recent reanalysis of the data [on the economic benefit of VAMs] suggested that it was not just a small effect that was found, but actually none at all" (pp. 12-13).

Authors of eight articles provided evidence that educational policies designed to evaluate teachers based on student test scores did not accomplish their intended consequences (Berliner, 2014; Hill et al., 2011; Kupermintz, 2003; Lavigne, 2014; Martineau, 2006; Moore Johnson, 2015; Polikoff, 2015; Rothstein, 2010). Kupermintz (2003) presented evidence supporting several alternative explanations of student gains, warning that "policy makers and administrators who wish to use [VAMs] must consider these alternative explanations when contemplating the likely consequences, intended and unintended, of any policy move" (p. 289). Rothstein (2010) also warned that educational policies based on VAMs would "reward or punish teachers who do

not deserve it and fail to reward or punish teachers who do" (p. 211). Similarly, Hill et al. (2011) stated that their case study findings "lead [them] to conclude that value-added scores, at least in [the] district [of study] and using these not-uncommon models, [were] not sufficient to identify problematic and excellent teachers accurately" (p. 825). Finally, an article written by Lavigne (2014) was entirely devoted to validity evidence based on related consequences, in which she concluded that high-stakes accountability systems based on student test scores failed to achieve their intended consequences, at a cost of several negative unintended consequences instead.

Authors of more than half of the articles who discussed related consequences provided evidence of unintended consequences. Everson et al. (2013) warned that evaluation systems that compare teachers with one another promoted competition over collaboration. Kane and Staiger (2002) documented the "natural fluctuation in noisy test score measures" and warned that such volatility could "wreak havoc if policymakers [drew] inferences from short-term fluctuations" (p. 102). Related, Lavigne (2014) warned that "policymakers should be concerned about how high-stakes teacher evaluation and related firing policies may be counter-productive and harm student achievement rather than help it" (p. 19). Winters and colleagues (Winters & Cowen, 2013; Winters et al., 2012) found that students assigned to teachers who would have been dismissed under VAM-driven personnel policies had smaller observed learning gains, suggesting that such policies would have been beneficial, but also found some evidence that higher-VAM teachers are more likely to leave the classroom under such policies. Papay (2011) explored the financial impact of using different tests in salary decisions, warning that "the average teacher in the district would see his or her pay changed by $2,178 simply by switching outcome measures [i.e., the tests]" (p. 181).

The three most common unintended consequences discussed were that teacher evaluations based on student standardized test scores via VAMs could (a) create a disincentive to teach the highest need students (Everson et al., 2013; Hanushek & Rivkin, 2010; Kupermintz, 2003; Moore Johnson, 2015; Rubin et al., 2004), (b) lead to an unwanted narrowing of the curriculum and undesirable degree of teaching to the test (Chetty et al., 2014a; Fox, 2016; Goldhaber, Cowan et al., 2013; T. J. Kane & Staiger, 2002; Papay, 2012), and (c) encourage cheating or other morally questionable practices designed to game the system (Ballou & Springer, 2015; Chetty et al., 2014a; Goldhaber, Cowan et al., 2013; Rubin et al., 2004). In sum, the validity evidence based on the related intended and unintended consequences did not conclusively demonstrate that the IUA meets its intended consequences, nor did it provide sufficient, conclusive evidence of negative unintended consequences. In addition, nowhere did authors present in any discreet, benefit-to-cost terms, whether any set of (positive or negative) intended consequences duly outweighed any set of (positive or negative) unintended consequences, or vice versa.

Discussion with Implications

It was our intent in this study, as external reviewers of one set of high-quality literature about VAMs as used for teacher evaluation purposes, to make explicit the arguments supporting the use of VAMs to evaluate teacher effectiveness (i.e., the IUA) and to interrogate the validity evidence advanced and positioned by the authors of the articles reviewed to support or challenge the same IUA. Across research questions, we found that the evidence provided by authors of the literature reviewed did not appear to support (and in some cases neither supported nor challenged) the valid use of VAMs to evaluate teachers. Alternatively, we could say that the reviewed literature contained about as much evidence to support the use of VAM in teacher evaluation as it contained evidence to

challenge it. This was consistently true for all areas of the framework for validity evidence, whether considering unweighted means, or by weighting articles by annual citation count when calculating mean support ratings. Qualitative syntheses of the validity evidence reviewed suggested that the weight of evidence provided by authors of these articles, again, as systematically and hence, arguably representative of the highest quality, most prominent papers on this topic, was consistent with our quantitative ratings. That is, there appeared to be a disconnect between the lukewarm and often mixed support for the use of VAMs in teacher evaluation and their widespread use for those purposes even after implementation of ESSA provided more flexibility to LEAs, especially in VAM-based educational policy regards. Several areas of disagreement about the evidence remain, however. A prominent example of this disagreement can be seen in the scholarly debate that transpired between Rothstein (2017) and Chetty et al. ((2017).

Some possibilities exist for this apparent disconnect between the validity evidence and educational policy and practice, as such. First, we consider that the present study is limited by the scope of its investigation. To identify, analyze, and synthesize relevant, high-quality, peer-reviewed literature, we investigated an IUA that was necessarily broad in scope. Via this review, accordingly, we investigated the use of VAMs in teacher evaluation, rather than investigating a *specific* VAM, based on scores from a *specific* set of tests, to evaluate *specific* teachers. As the authors of the *Standards* (AERA et al., 2014) recommend, each intended use of scores must be separately validated. Although via this study we yielded insufficient evidence to validate the broad use of VAMs in teacher evaluation, that does not preclude the validation of some specific application of VAMs in teacher evaluation or teacher evaluation policy in the future. It is also possible that some of the most challenging evidence that we found came from investigations of particularly ill-advised applications of VAMs in teacher evaluation, and that VAMs which support more valid inferences about specific teachers may already be in use in some contexts.

We also acknowledge that scholars from the fields included in this review (namely economics, education, and quantitative methods) may think about the IUA under investigation in different ways as colored by the lenses of their specific discipline. As was discussed in the papers we reviewed, we noticed authors of the articles published in economics journals seemed to position and capture VAMs differently than articles published in education journals. Although a thorough exploration of these differences is beyond the scope of our current paper, it appears that the authors of the 12 papers published in economics journals reported more favorable evidence (M=2.8, SD=1.5) than did the authors of the 52 papers published in education journals (M=1.8, SD=0.9; see Table 1 for identification of journal field). Papers in economics journals were also cited more per year (M=96.5, SD=133.7, Mdn=28.3) than papers in education journals (M=22.6, SD=22.8, Mdn=17.2), suggesting that economics papers may also more prominently frame the scholarly debate surrounding the use of VAMs in teacher evaluation and teacher evaluation policy.

Note also that we chose to use annual citation counts in this study as an indicator of a paper's influence within the scholarly debate. We dropped the articles with the lowest annual citation count prior to analyzing these papers, not because we believed that the infrequently cited papers were flawed, but because we wanted to understand the scholarly debate about the validity evidence surrounding the use of VAM to make valid (or invalid) inferences about teacher effectiveness and quality. We recognize that we may have dropped important articles that report high-quality studies or papers which may one-day be highly regarded and well-cited.

We also find it interesting that for some areas of our framework the unweighted mean rating of the evidence was significantly more challenging to the valid use of VAM in teacher evaluation

than weighting the mean by citation count. This suggests that authors of the papers who supported the use of VAM in teacher evaluation published papers that were being discussed in the scholarly literature more often than the papers in which authors challenge VAMs. Related, we acknowledge that highly cited papers could be oft-lauded or oft-criticized and those citations would still be counted the same way. It was also beyond the scope of this paper to analyze this caveat.

Related, we concede that as new evidence was published and scholarly arguments were developed over the course our of review period, this evidence was not included. Likewise, the authors of the literature reviewed in this study may have changed their positions on the IUA since the initial publication of their papers reviewed. Since these reviewed authors at least theoretically represent the most knowledgeable experts and scholars on VAMs, though, it is important to understand their current positions on the use of VAMs to evaluate teachers. This was the purpose of our survey of VAM scholars, which was also informed by the framework for validity evidence and the findings of this review; that was, to understand their current thoughts on VAMs, VAM use, and VAM-based educational policies (Lavery et al., 2020).

Finally, the findings of this review suggest that, even though the publication rate for articles in which authors investigated the validity of VAMs in teacher evaluation seem to have fallen off sharply since 2018, VAM scholars have not yet found consensus on this topic. This finding has scholarly significance as it suggests that further research is needed into specific VAMs used in specific contexts and the specific areas of validity evidence that remain unresolved. Considering our findings, school and district administrators should not make high-stakes personnel decisions based exclusively on VAMs until stronger validity evidence more consistently indicates that whatever VAMs might be used for whatever purposes are valid for each said purpose. Just as federal educational policy provides more flexibility to states and districts on the teacher evaluation methods that they use (ESSA, 2015), the lukewarm and somewhat mixed findings of this review suggest that it may still be inappropriate for state or local policymakers to continue to require the use of VAMs in teacher evaluation, especially when prompting high-stakes, consequential, personnel decisions.

Conclusions

In sum, and at present, as per the findings of this systematic review, it seems that Pittenger's (1917) warnings about the challenges of measuring the plane of results and his assertion that the measurement of teacher effectiveness may never be "more than a carefully controlled process of estimating a teacher's individual efficiency" (p. 103) stand true. Given the evidence presented in this review, these issues have not been fully resolved in the scholarly debate surrounding VAMs.

References

Note: References with an * indicate the 75 articles that researchers included in this review.

*American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448-452. https://doi.org/10.3102/0013189X15618385

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5, 123 § 115 (2009). https://www.congress.gov/bill/111th-congress/house-bill/1

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

*Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher, 37*(2), 65-75. https://doi.org/10.3102/0013189X08316420

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

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in Kindergarten. *The Quarterly Journal of Economics*, 1415–1453. doi:10.1093/qje/qjw016 Retrieved from http://qje.oxfordjournals.org/content/131/3/1415.abstract

*Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review, 62*, 48-65. https://doi.org/10.1016/j.econedurev.2017.10.004

Backes, B., Cowan, J., Goldhaber, D., & Theobold, R. (2023). *How to measure a teacher: The influence of test and nontest value-added on long-run student outcomes*. Center for Analysis of Longitudinal Data in Education Research (CALDER), American Institutes for Research. https://caldercenter.org/sites/default/files/CALDER%20WP%20270-0423-2.pdf

*Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65. https://doi.org/10.3102/10769986029001037

*Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77-86. https://doi.org/10.3102/0013189X15574904

*Berliner, D. C. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116*(1). Retrieved from http://www.tcrecord.org/Content.asp?ContentID=17293

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study.* http://metproject.org/reports.php

*Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal, 53*(2), 324-359. https://doi.org/10.3102/0002831216630407

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models. Policy information perspective*. Educational Testing Service. http://www.ets.org/research/policy_research_reports/publications/report/2005/cxje

Braun, H. (2012). Conceptions of validity: The private and the public. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 46-49. https://doi.org/10.1080/15366367.2012.679159

*Braun, H. (2015). The value in value added depends on the ecology. *Educational Researcher, 44*(2), 127-131. https://doi.org/10.3102/0013189X15576341

*Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39*(7), 537-544. https://doi.org/10.3102/0013189X10383560

*Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics, 39*(5), 333-367. https://doi.org/10.3102/1076998614547576

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

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

*Chetty, R., Friedman, J. N., & Rockoff, J. E. (2016). Using lagged outcomes to evaluate bias in value-added models. *American Economic Review, 106*(5), 393-399. http://dx.doi.org/10.1257/aer.p20161081

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2017). Measuring the impacts of teachers: Reply. *American Economic Review, 107*(6), 1685. http://dx.doi.org/10.1257/aer.20170108

Clarivate Analytics. (2017). 2016 Journal Citation Reports® Social Sciences Edition. http://jcr.incites.thomsonreuters.com

Close, K., Amrein-Beardsley, A., & Collins, C. (2020). Putting teacher evaluation systems on the map: An overview of states' teacher evaluation systems post-Every Student Succeeds Act. *Education Policy Analysis Archives, 28*(1), 1-58. https://doi.org/10.14507/epaa.28.5252

*Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1). http://www.tcrecord.org/Content.asp?ContentId=17291

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

*Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review, 40*, 76-92. https://doi.org/10.1016/j.econedurev.2013.11.009

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

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

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (2nd ed., pp. 621-694). American Council on Education.

*Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher, 44*(2), 132-137. https://doi.org/10.3102/0013189X15575346

*Dieterle, S., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added. *Journal of Policy Analysis and Management, 34*(1), 32-58. https://doi.org/10.1002/pam.21781

Duncan, A. (2011). *Winning the future with education: Responsibility, reform and results.* Testimony given to the U.S. Congress. U.S. Department of Education. https://www.ed.gov/news/speeches/winning-future-education-responsibility-reform-and-results

EBSCO Industries. (2018). EBSCOhost. http://web.a.ebscohost.com

Everson, K. C. (2017). Value-added modeling and educational accountability: Are we answering the real questions? *Review of Educational Research*, *87*(1), 35–70. https://doi.org/10.3102/0034654316637199

*Everson, K. C., Feinauer, E., & Sudweeks, R. R. (2013). Rethinking teacher evaluation: A conversation about statistical inferences and value-added models. *Harvard Educational Review, 83*(2), 349-370. https://doi.org/10.17763/haer.83.2.m32hk8q851u752h0

Every Student Succeeds Act of 2015, Pub. L. No. 114-95, § 114 Stat. 1177. (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

*Fox, L. (2016). Playing to teachers' strengths: Using multiple measures of teacher effectiveness to improve teacher assignments. *Education Finance and Policy, 11*(1), 70-96. https://doi.org/10.1162/EDFP_a_00176

*Gabriel, R., & Lester, J. N. (2013). The romance quest of education reform: A discourse analysis of the Los Angeles Times' reports on value-added measurement teacher effectiveness. *Teachers College Record, 115*(12). http://www.tcrecord.org/content.asp?contentid=17252

*Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher, 44*(2), 87-95. https://doi.org/10.3102/0013189X15574905

*Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein falsification test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness, 8*(1), 8-34. https://doi.org/10.1080/19345747.2014.978059

*Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review, 36*, 216-228. https://doi.org/10.1016/j.econedurev.2013.06.010

*Goldhaber, D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis, 35*(2), 220-236. https://doi.org/10.3102/0162373712466938

*Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review, 100*(2), 250-255. https://doi.org/10.1257/aer.100.2.250

*Good, T. L. (2014). What do we know about how teachers influence student performance on standardized tests: And why do we know so little about other student outcomes? *Teachers College Record, 116*(1), 1-22. http://www.tcrecord.org/content.asp?contentid=17289

*Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293-303. https://doi.org/10.3102/0013189X14544542

*Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of empirical Bayes's estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics, 40*(2), 190-222. https://doi.org/10.3102/1076998615574771

*Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271. https://doi.org/10.1257/aer.100.2.267

*Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*(4), 693-699. https://doi.org/10.1002/pam.20464

*Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73-112. https://doi.org/10.3102/0002831213517130

*Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review, 40*, 183-204. https://doi.org/10.1016/j.econedurev.2014.03.002

*Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*(4), 700-709. https://doi.org/10.1002/pam.20463

*Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. doi:10.2307/27975308

*Isenberg, E., & Walsh, E. (2015). Accounting for co-teaching: A guide for policymakers and developers of value-added models. *Journal of Research on Educational Effectiveness, 8*(1), 112-119. doi:10.1080/19345747.2014.974232

*Johnson, M. T., Lipscomb, S., & Gill, B. (2015). Sensitivity of teacher value-added estimates to student and peer control variables. *Journal of Research on Educational Effectiveness, 8*(1), 60-83. https://doi.org/10.1080/19345747.2014.967898

*Jones, N. D., Buzick, H. M., & Turkan, S. (2013). Including students with disabilities and english learners in measures of educator effectiveness. *Educational Researcher, 42*(4), 234-241. https://doi.org/10.3102/0013189X12468211

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135-170. https://doi.org/10.1207/s15366359mea0203_1

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

*Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(4), 91-114. https://doi.org/10.1257/089533002320950993

*Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598. https://doi.org/10.3102/0013189X10390804

*Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review, 28*(6), 682-692. https://doi.org/10.1016/j.econedurev.2009.02.003

*Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180-195. https://doi.org/10.1016/j.econedurev.2015.01.006

*Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record, 116*(1). http://www.tcrecord.org/content.asp?contentid=17290

*Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287-298. https://doi.org/10.3102/01623737025003287

Lavery, M. R., Amrein-Beardsley, A., *Geiger, T., & Pivovarova, M. (2020). Value-added model (VAM) scholars on using VAMs for teacher evaluation, after the passage of the Every Student Succeeds Act (ESSA). *Teachers College Record, 122(8).* https://www.tcrecord.org/content.asp?contentid=23320

*Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*(1). http://www.tcrecord.org/content.asp?contentid=17294

Layton, L. (2012). Rethinking the classroom: Obama's overhaul of public education. *The Washington Post*. http://www.washingtonpost.com/local/education/rethinking-the-classroom-obamas-overhaul-of-public-education/2012/09/20/a5459346-e171-11e1-ae7f-d2a13e249eb2_print.html

*Lee, S. W. (2018). Pulling back the curtain: Revealing the cumulative importance of high-performing, highly qualified teachers on students' educational outcome. *Educational Evaluation and Policy Analysis, 40*(3), 359-381. https://doi.org/10.3102/0162373718769379

*Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis, 34*(1), 109-121. https://doi.org/10.3102/0162373711422377

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2019). Methodological issues in value-added modeling: An international review from 26 countries. *Educational Assessment, Evaluation and Accountability 31*, 257–287. https://link.springer.com/article/10.1007/s11092-019-09303-w

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5-55.

*Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255-270. https://doi.org/10.3102/10769986027003255

*Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22-52. https://doi.org/10.3102/1076998613509405

*Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67. https://doi.org/10.1111/j.1745-3984.2007.00026.x

*Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125-150. https://doi.org/10.3102/1076998606298039

*Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis, 36*(4), 457-475. https://doi.org/10.3102/0162373714527788

*Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics, 35*(3), 253-279. https://doi.org/10.3102/1076998609346967

*Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35-62. https://doi.org/10.3102/10769986031001035

*McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101. https://doi.org/10.3102/10769986029001067

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*(10), 955-66. https://doi.org/10.1037/0003-066X.30.10.955

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan Publishing Co, Inc.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *Annals of Internal Medicine, 151*(4), 264-269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

*Moore Johnson, S. (2015). Will VAMs reinforce the walls of the egg-crate school? *Educational Researcher, 44*(2), 117-126. https://doi.org/10.3102/0013189X15573351

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*, 229–258. https://doi.org/10.3102/00346543062003229

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14*(2), 5-13. https://doi.org/10.1111/j.1745-3992.1995.tb00854.x

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425. (2002). https://www.congress.gov/bill/107th-congress/house-bill/1

*Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257. https://doi.org/10.3102/01623737026003237

*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. https://doi.org/10.3102/0002831210362589

*Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123-141. https://doi.org/10.17763/haer.82.1.v40p0833345w6384

*Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal, 51*(2), 328-362. https://doi.org/10.3102/0002831213508299

Pittenger, B. F. (1917). Problems of teacher measurement. *Journal of Educational Psychology, 8*(2), 103-110. https://doi.org/10.1037/h0072599

*Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183-212. https://doi.org/10.1086/679390

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

*Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121-129. https://doi.org/10.3102/10769986029001121

*Raudenbush, S. W. (2015). Value added: A case study in the mismatch between education research and policy. *Educational Researcher, 44*(2), 138-141. https://doi.org/10.3102/0013189X15575345

*Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519. https://doi.org/10.1162/edfp.2009.4.4.492

*Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*(1), 117-120. https://doi.org/10.3102/10769986029001117

*Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

*Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*(5), 261-266. https://doi.org/10.1016/j.labeco.2011.02.004

*Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571. https://doi.org/10.1162/edfp.2009.4.4.537

*Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*(1), 175-214. https://doi.org/10.1162/qjec.2010.125.1.175

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review, 107*(6), 1656-1684. https://doi.org/10.1257/aer.20141440

*Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the "Prospects" study of elementary schools. *Teachers College Record, 104*(8), 1525-1567. http://www.tcrecord.org/content.asp?contentid=11041

*Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116. https://doi.org/10.3102/10769986029001103

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* Teachers College Press.

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research, 3*(3), 208-222. https://doi.org/10.1177/1558689809334210

Smith, W. C., & Kubacka, K. (2017). The emphasis of student test scores in teacher appraisal systems. *Education Policy Analysis Archives, 25*(86). https://doi.org/10.14507/epaa.25.2889

Sørensen, T. B. (2016). Value-added measurement or modelling (VAM). Education International. http://download.ei-ie.org/Docs/WebDepot/2016_EI_VAM_EN_final_Web.pdf

*Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*(2), 142-171. https://doi.org/10.3102/1076998611432174

*Stacy, B., Guarino, C., & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve? *Economics of Education Review, 64*, 50-74. https://doi.org/10.1016/j.econedurev.2018.04.001

van Eemeren, F. H., & Grootendorst, R. (2010). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

*Winters, M. A., & Cowen, J. M. (2013). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher, 42*, 330-337. https://doi.org/10.3102/0013189X13496145

*Winters, M. A., Dixon, B. L., & Greene, J. P. (2012). Observed characteristics and teacher quality: Impacts of sample selection on a value added model. *Economics of Education Review, 31*(6), 19-32. https://doi.org/10.1016/j.econedurev.2011.07.014

Yeh, S.S. (2019). Origin of the achievement gap: Evidence from a league table analysis of 23 interventions. *Journal of Education Finance 45*(1), 23-41. https://www.muse.jhu.edu/article/747803

Yeh, S. S. (2020). Educational Accountability, Value-Added Modeling, and the Origin of the Achievement Gap. *Education and Urban Society*, *52*(8), 1181-1203. https://doi.org/10.1177/0013124519896823

About the Authors

Audrey Amrein-Beardsley
Mary Lou Fulton Teachers College, Arizona State University
audrey.beardsley@asu.edu
https://orcid.org/0000-0001-6924-3025
Audrey Amrein-Beardsley, Ph.D., is a Professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research focuses on the use of value-added models (VAMs) in and across states before and since the passage of the Every Student Succeeds Act (ESSA). More specifically, she is conducting validation studies on multiple system components, as well as serving as an expert witness in many legal cases surrounding the (mis)use of VAM-based output.

Matthew Ryan Lavery
South Carolina Education Oversight Committee
mlavery@eoc.sc.gov
https://orcid.org/0000-0002-4208-7277
Matthew Ryan Lavery, Ph.D., serves as the Deputy Director for the South Carolina Education Oversight Committee. Prior to the EOC, he taught at the secondary level for eight years before earning a Ph.D. in Educational Research Methodology, Measurement, and Analysis from the University of Central Florida. His research focuses on the valid use of educational data to inform the improvement of educational leadership, policy, programs, and student outcomes.

Jessica Holloway
Institute for Learning Sciences and Teacher Education, Australian Catholic University
jessica.holloway@acu.edu.au
https://orcid.org/0000-0001-9267-3197
Jessica Holloway, Ph.D., is a Senior Research Fellow within the Institute for Learning Sciences and Teacher Education (ILSTE) at the Australian Catholic University. Her research draws on political theory and policy sociology to ask how metrics, data, and digital tools produce new conditions, practices, and subjectivities, especially in relation to teachers and schools. Her recent books include *Expertise* (2023, with Jessica Gerrard) and *Metrics, Standards and Alignment in Teacher Policy: Critiquing Fundamentalism and Imagining Pluralism* (2021).

Margarita Pivovarova
Mary Lou Fulton Teachers College, Arizona State University
margarita.pivovarova@asu.edu
https://orcid.org/0000-0002-2965-7423
Margarita Pivovarova, Ph.D., is an Associate Professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research focuses on the relationship between student achievement, teacher mobility, and school contextual factors. More specifically, she explores the factors associated with teacher attrition from public and charter schools, and immigrant student achievement in schools with varied student demographics.

Debbie L. Hahs-Vaughn
Department of Learning Sciences & Educational Research, University of Central Florida
Debbie.Hahs-Vaughn@ucf.edu
https://orcid.org/0000-0002-1217-5161
Debbie L. Hahs-Vaughn, Ph.D., is a Professor in Methodology, Measurement, and Analysis at the University of Central Florida. Her primary research relates to methodological issues associated with applying quantitative statistical methods to survey data obtained under complex sampling designs and using complex survey data to answer substantive research questions. She is the author of six quantitative statistics textbooks and over 60 articles in professional outlets.

education policy analysis archives

Volume 31 Number 117 October 24, 2023 ISSN 1068-2341