Educational Assessment, Evaluation and Accountability (2019) 31:257–287
https://doi.org/10.1007/s11092-019-09303-w

Methodological issues in value-added modeling: an international review from 26 countries

Jessica Levy[1] · Martin Brunner[2] · Ulrich Keller[1] · Antoine Fischbach[1]

Received: 23 October 2018 / Accepted: 18 July 2019 / Published online: 2 August 2019

Abstract

Value-added (VA) modeling can be used to quantify teacher and school effectiveness by estimating the effect of pedagogical actions on students' achievement. It is gaining increasing importance in educational evaluation, teacher accountability, and high-stakes decisions. We analyzed 370 empirical studies on VA modeling, focusing on modeling and methodological issues to identify key factors for improvement. The studies stemmed from 26 countries (68% from the USA). Most studies applied linear regression or multilevel models. Most studies (i.e., 85%) included prior achievement as a covariate, but only 2% included noncognitive predictors of achievement (e.g., personality or affective student variables). Fifty-five percent of the studies did not apply statistical adjustments (e.g., shrinkage) to increase precision in effectiveness estimates, and 88% included no model diagnostics. We conclude that research on VA modeling can be significantly enhanced regarding the inclusion of covariates, model adjustment and diagnostics, and the clarity and transparency of reporting.

Keywords Value-added modeling · Literature review · Primary and secondary education · Teacher effectiveness · School effectiveness

What is the added value from attending a certain school or being taught by a certain teacher? To answer this question, the value-added (VA) model was developed. In this model, the actual achievement attained by students attending a certain school or being

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11092-019-09303-w) contains supplementary material, which is available to authorized users.

✉ Jessica Levy
jessica.levy@uni.lu

[1] Luxembourg Centre for Educational Testing, Faculty of Language and Literature, Humanities, Arts and Education, University of Luxembourg, Campus Belval, Maison des Sciences Humaines, 11, Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg

[2] Department of Education, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

taught by a certain teacher is juxtaposed with the achievement that is expected for students with the same background characteristics (e.g., pretest scores). To this end, the VA model can be used to compute a VA score for each school or teacher, respectively. If actual achievement is better than expected achievement, there is a positive effect (i.e., a positive VA score) of attending a certain school or being taught by a certain teacher. In other words, VA models have been developed to "make fair comparisons of the academic progress of pupils in different settings" (Tymms 1999, p. 27). Their aim is to operationalize teacher or school effectiveness objectively. Specifically, VA models are often used for accountability purposes and high-stakes decisions (e.g., to allocate financial or personal resources to schools or even to decide which teachers should be promoted or discharged). Consequently, VA modeling is a highly political topic, especially in the USA, where many states have implemented VA or VA-based models for teacher evaluation (Amrein-Beardsley and Holloway 2017; Kurtz 2018). However, this use for high-stakes decisions is highly controversial and researchers seem to disagree concerning the question if VA scores should be used for decision-making (Goldhaber 2015). For a more exhaustive discussion of the use of VA models for accountability reasons, see, for example, Scherrer (2011).

Given the far-reaching impact of VA scores, it is surprising that there is scarcity of systematic reviews of how VA scores are computed, evaluated, and how this research is reported. To this end, we review 370 empirical studies from 26 countries to rigorously examine several key issues in VA modeling, involving (a) the statistical model (e.g., linear regression, multilevel model) that is used, (b) model diagnostics and reported statistical parameters that are used to evaluate the quality of the VA model, (c) the statistical adjustments that are made to overcome methodological challenges (e.g., measurement error of the outcome variables), and (d) the covariates (e.g., pretest scores, students' sociodemographic background) that are used when estimating expected achievement.

All this information is critical for meeting the transparency standards defined by the American Educational Research Association (AERA 2006). Transparency is vital for educational research in general and especially for highly consequential research, such as VA modeling. First, transparency is highly relevant for researchers. The clearer the description of the model, the easier it is to build upon the knowledge of previous research and to safeguard the potential for replicating previous results. Second, because decisions that are based on VA scores affect teachers' lives and schools' futures, not only educational agents but also the general public should be able to comprehend how these scores are calculated to allow for public scrutiny. Specifically, given that VA scores can have devastating consequences on teachers' lives and on the students they teach, transparency is particularly important to evaluate the chosen methodology to compute VA models for a certain purpose. Such evaluations are essential to answer the question to what extent the quality of VA scores allows to base far-reaching decisions on these scores for accountability purposes.

1 The historical background of VA

The idea of teachers and schools being held accountable for students' achievement is not new. For example, England has introduced "payment by results" in 1862 and teachers have been paid depending on their students' achievement in examinations.

Also in the USA, the idea of paying teachers depending on their students' achievement has already emerged in the early twentieth century (Harris 2007; Lavigne and Good 2013). The term "value-added" was first mentioned in an educational context by the economist Hanushek (1971). He described a model that could be used to analyze teacher effects, taking into account prior achievement. In the 1990s, test-based accountability started to play an important role in the USA and VA models became increasingly popular in teacher or school evaluations. Most notably, in the USA, the development of the "Tennessee Value-Added Assessment System" (TVAAS, Sanders and Horn 1994) helped popularize the use of VA modeling (e.g., Everson 2017; Hill et al. 2011). In the same year, the first VA models for school evaluation were also calculated in France ("Indicateurs de valeur ajoutée," Duclos and Murat 2014; MENDEP 1994). Another milestone was the "No Child Left Behind Act," which came into effect in 2002 and played an important role in educational evaluation (mostly teacher quality) and accountability in the educational context in the USA. With the enactment of the "Race to the Top Act" (2011), teachers were evaluated and held accountable increasingly by means of students' achievement gains rather than teacher observations (e.g., Lavigne and Good 2013). Built on the TVAAS, the "Educational Value-Added Assessment System" (EVAAS) has been developed and made VA models even more popular under the Race to the Top Act. The EVAAS is still the best known and probably most widely used VA model, even though it has been questioned by many researchers (for an overview, see, e.g., Amrein-Beardsley and Geiger 2017). In the UK, contextual VA (CVA) has been used for school monitoring (e.g., Bradbury 2011; Perry 2016). CVA takes into account sociodemographic student data such as gender, ethnicity, and eligibility for free school meals—an indicator of students' socioeconomic family background (SES)—in addition to prior achievement. In contrast to the teacher accountability measures in the USA, CVA measures in the UK are usually applied to evaluate school quality. Following these examples, it is clear that the targets and purposes of VA models differ depending on, among others, the countries in which they are applied.

2 Benefits and limitations of VA

The usefulness and limitations of VA scores are highly discussed topics. If VA scores are stable over time, they can make important contributions in the selection of effective teachers, leading to an improvement in students' achievement (as discussed in e.g., Goldhaber and Hansen 2013). However, if they are not stable (as discussed in e.g., Newton et al. 2010), they have to be used with caution, especially in a high-stakes context. More concretely, Nunnally and Bernstein (1994), as cited in Evers 2001) suggest a reliability of over .90 when a test is used for important decisions. While research on the stability of VA scores indicates low to moderate stability measures of VA scores (correlations between years ranging from .2 to .66; see, for example, Kersting et al. 2013), the authors also argue that the assumption that teachers do not change over time is unreasonable, indicating that even a benchmark of .8 would already be too high. For a more exhaustive discussion of limitations and their implications, see, for example, Everson (2017) or Perry (2016).

3 How VA scores are perceived by educational practitioners

VA scores are used for general evaluation purposes and to inform a broader public about teacher or school quality. Yet, VA modeling has the largest impact when it is applied to accountability and high-stakes decisions because financial consequences can be drawn for schools in the basis of their VA scores. Moreover, these scores may have a significant influence on important personnel decisions, such as teachers' promotion or dismissal.

Even though VA scores are used for such far-reaching decisions, neither principals nor teachers seem to perceive them as a trustworthy measure. First, in a survey of 764 school principals, Goldring et al. (2015) found that only 56% of the principals said that student growth measures are "valid to a large extent" (pp. 100–101). Furthermore, the area of VA measures was the one for which principals expressed the strongest desire for more support: Over 70% of the principals indicated that they wanted support in understanding VA measures. Second, teachers tend to voice similar concerns. In a survey of more than 24,000 teachers, Jiang et al. (2015) found that 65% of the teachers agreed that their evaluation relied too heavily on student growth. In addition, 50% of the teachers disagreed that the tests offered a fair assessment of their students' learning.

Taken together, these empirical results show that both school principals and teachers wish for more clarity and transparency in VA measures to help them better comprehend how the scores are calculated and what the scores really mean. In other words, these results empirically underscore the idea that the current practice of VA modeling does not fully meet the transparency principle as emphasized in the relevant reporting standards (AERA 2006), which stress that the analytical procedures and techniques should be described in a precise and transparent way to facilitate comprehension.

4 Who is the target of VA? What is the purpose of VA?

The two most common targets of VA models are teachers and schools, but VA can also be calculated at the levels of principals or classrooms/peers (e.g., Branch et al. 2012; Sund 2009). Closely linked to the question of the target is the question of the purpose of the use of VA models. There are two major applications of VA modeling: (a) evaluation of the VA target by means of VA scores (with or without consequences for the VA target) and (b) identification of effective teachers, schools, or pedagogical strategies (Blazar et al. 2016; Bonesrønning 2004; Merritt et al. 2017; Rutledge et al. 2015).

5 Statistical models that can be used as VA models and their assumptions

Even though most VA models share the same goal—to estimate the effect of a certain teacher or a certain school on student achievement—the underlying statistical models differ. Common to all statistical models applied for VA modeling is the idea of comparing students' achievement with the achievement that is expected for students who have the same background characteristics that are relevant for student learning (e.g., pretest scores). There are several statistical models that allow users to calculate

this comparison (see, McCaffrey et al. 2004; Tekwe et al. 2004), but there are two statistical model types that have received the most attention in the methodological literature: (a) linear regression models and (b) multilevel models (see, e.g., Lopez-Martin et al. 2014, for other models that can be used for VA modeling; e.g., nonlinear models). In practice, the most frequently used models for the calculation of VA scores are ordinary linear regression models (Kurtz 2018).

5.1 Linear regression models

Linear regression models, including simple and multiple linear regression models, assume a linear relationship between the dependent and independent variable(s). Typically, student achievement serves as the dependent variable and the model contains prior achievement as an independent variable that serves as a covariate. An example of a model equation is as follows (see, e.g., Ray 2006):

$$A_{ijt} = \beta_0 + \beta_1 A_{ijt-1} + r_{ij}, \tag{1}$$

where A_{ijt} is the achievement of student i in group j (e.g., school j or class j) at time t; A_{ijt-1} is the prior achievement of student i in group j at $t-1$ (e.g., the previous school year); β_0 is the intercept term (i.e., the expected value for students who have a value of zero on the independent variables); and β_1 is the regression coefficient, indexing the relationship between achievement at time t and prior achievement at time $t-1$. r_{ij} is the residual error term for each student, representing the difference between the achievement score A_{ijt} that is predicted by the model and actual achievement A_{ijt}. The residual error term is assumed to be normally distributed (with a mean of 0 and variance σ^2_{rij}). Further, r_{ij} is assumed to be independent of all explanatory variables and to be homoscedastic (Hox 1995). Importantly, r_{ij} is used to compute the VA score:

$$\mathrm{VAj} = \bar{r}_j. \tag{2}$$

In Eq. 2, VA_j is defined as the average residual r_{ij} across all students for group j, for example, school j or class j taught by a certain teacher. The model from Eq. 1 can be extended by including student characteristics or context variables as further covariates.[1]

The method that is typically used to estimate regression coefficients (e.g., β_1) in multiple regression models is ordinary least squares (OLS). To apply OLS, it is important to check for the following assumptions (see, e.g., Cohen et al. 2003): the relationships between the dependent and independent variables are linear, residual error terms are independent, residuals are normal and homoscedastic, and all independent variables are measured without measurement error.

5.2 Multilevel models

In multilevel models, also known as hierarchical linear models, mixed models, or random coefficient models (McNeish et al. 2017), the nested structure of the

[1] Analogously, Eq. 1 can be calculated for each school, using data aggregated at school level. The residuals from each equation will be used as a VA score for each school.

data is taken into account. Such a nested structure is typical in the field of VA modeling (i.e., students nested in classes and schools). Multilevel models can be used to accommodate longitudinal and cross-sectional clustered data and to overcome some of the restrictions of linear regression models (Cohen et al. 2003; McNeish et al. 2017). Similar to the linear regression model, the dependent variable is usually an achievement measure in multilevel VA models, and the model generally contains prior achievement as a covariate (e.g., as applied in the Tennessee Value Added Assessment System, TVAAS, Sanders and Horn 1994). Other variables, such as student or context characteristics, can be included as additional covariates on the various levels. Equations 3 to 5 exemplify a VA model in terms of a two-level model (e.g., students nested in classrooms; see Dedrick et al. 2009; Hox 1995, 2013; McNeish et al. 2017):

$$\text{Level 1}: A_{ijt} = \beta_{0j} + \beta_{1j}A_{ijt-1} + e_{ij} \tag{3}$$

$$\text{Level 2}: \beta_{0j} = \gamma_{00} + \gamma_{01}C_j + \mu_{0j} \tag{4}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}C_j + \mu_{1j} \tag{5}$$

In Eq. 3, A_{ijt} is the achievement of student i in group j (e.g., school j or class j) at time t, A_{ijt-1} is the achievement of student i in group j in the prior year, β_{0j} is the intercept, β_{1j} is the regression coefficient linking prior achievement at time $t-1$ to achievement at time t, and e_{ij} is a residual error term (assumed to be normally distributed with a mean of 0 and a common variance of σ^2 for all teachers, classrooms, or schools). The largest difference from a linear regression model is that there is potentially a different intercept and a different slope coefficient for every level 2 variable, in this case, the classroom. In Eqs. 4 and 5, classroom variables are included to explain between-classroom differences in the intercept (β_{0j}) and slope (β_{1j}) in Eq. 3. For example, C_j is class size, and γ_{00}, γ_{01}, γ_{10}, and γ_{11} are the regression coefficients that link class size to the intercept and slope. Both γ_{00} and γ_{01} are assumed to be constant across all classes (fixed effects), and μ_{0j} and μ_{1j} are random residual error terms that can vary between classes. Usually, these error terms are assumed to be normally distributed and to have a mean of 0 and variances specified as $\sigma_{\mu 0}{}^2$ and $\sigma_{\mu 1}{}^2$, respectively (Hox 2013). The VA score of a school j or teacher j can be quantified in terms of an estimate of the residuals $\hat{\mu}_{0j}$ for this particular school or teacher at level 2 (e.g., the residual for a certain teacher or school; see Ferrão and Goldstein 2009).

Similar to linear regression models, the assumptions that should be checked for multilevel models are linearity, independence of residuals, normal distribution of residuals, and homoscedasticity (Snijders and Bosker 2012). In addition, a number of assumptions are usually made about the covariance structure in multilevel models (McNeish et al. 2017): that the covariance structures of the residual error terms and of random effects are properly specified, the residuals and the random effects do not covary, and the predictor variables do not covary with the residuals and the random effects. In addition, Snijders and Bosker (2012) recommended that the covariance between the random intercept and slope parameters should not be fixed to 0 but should rather to be freely estimated from the data.

To summarize both model types (linear regression and multilevel models), Eq. 1 shows a regression model that takes into account only student characteristics. However, students are clustered in classes, school, and districts. Even though multiple regression models can take into account the clustered structure of these variables with the use of dummy variables, multilevel models offer a more efficient method for two reasons. First, the dummy coding in multiple regression models does not allow generalization beyond the sample that is analyzed, and second, the evaluation of the extent to which the regression coefficients (e.g., β_1) vary across clusters becomes cumbersome when multiple regression is used (Cohen et al. 2003). In this respect, multilevel models represent a flexible and viable alternative statistical framework. As shown in Eq. 3, the first level of the multilevel model is similar to the equation representing the linear regression model (Eq. 1). However, in a multilevel context, at least one more level is used to estimate the intercept and slope from the first level (Eqs. 4 and 5). Also, the multilevel model contains more than one error term: the error term from level 1 (e_{ij}), the error term(s) from level 2 (e.g., μ_{0j} and μ_{1j}), and so forth.

5.3 Empirical studies on how VA scores depend on the VA model that is applied

The analysis and comparison of different types of VA models—sometimes also including or excluding certain covariates—have been the subject of several studies. For example, Wei et al. (2012) analyzed and compared five different VA models to calculate teachers' VA scores in a sample of 131 teachers. These models included, among others, a multiple linear regression model, a multilevel model, and an average score change model, indicating the average difference between students' test scores in year t and year $t-1$. The authors ranked teachers by their VA scores and found a lot of variability in these rankings for each teacher across the models that were applied to estimate the VA scores. They calculated correlations between these teacher rankings and found that the different models were only remotely to moderately related or even negatively related in some cases (correlations ranging from $-.22$ to .67).

Tekwe et al. (2004) compared VA scores obtained from applying different VA models. They used a sample of over 6000 students to compare different types of multilevel VA models, including the TVAAS (Sanders and Horn 1994) model. The correlations between the scores derived from these VA models ranged from .57 to .99, indicating that VA scores may differ considerably between models.

Evidence for this variation was also offered in a study by Newton et al. (2010) who compared five different VA models using linear regression and multilevel models. They analyzed how teacher rankings changed and found that the ranking of up to 14% of the teachers changed by three or more deciles across models. This means that, depending on which VA model has been calculated, teachers who actually have a middle or high VA score could be incorrectly assigned a low VA score, which would lead to these relatively effective teachers being sanctioned.

In sum, previous studies have found that VA scores can differ significantly depending on the applied model. For example, this means that, depending on the model that was used, a teacher can be ranked as one of the best teachers or as an average (perhaps even one of the worst) teachers. Consequently, a teacher who was able to keep his or

her job could have been dismissed if another model had been used. Most authors thus conclude that VA measures should be used with great caution, especially in accountability systems or high-stakes decisions.

6 Model diagnostics and statistical parameters

The AERA (2006) reporting standards emphasize that transparency in reporting is crucial in order to provide enough information to replicate, to evaluate, and to build upon the knowledge gained from a certain study. Transparency also involves reporting model diagnostics to check whether vital assumptions underlying the applied statistical models have been met as well as to provide information about key statistical parameters that describe important characteristics of the empirical data and that indicate how well the model approximates these data.

6.1 Model diagnostics

As noted earlier, for linear regression models, the assumptions that need to be checked include linearity of the relationships between the dependent and independent variables, homoscedasticity, normality of residuals, independence of residuals, and measurement without measurement error (see, e.g., Cohen et al. 2003). To avoid redundancy, we elaborate on measurement error in the section on statistical adjustments. The assumptions that accompany multilevel models are similar to the ones associated with linear regression models: linearity, homoscedasticity, normality of residuals, and independence of residuals, but these assumptions must be assessed for all levels involved in the multilevel model that was specified (Snijders and Bosker 2012).

The assumption of linearity implies that the relationships between the covariates and dependent variables are linear. A violation of the linearity assumption can lead to imprecise VA scores because this violation means that the model was not appropriate for approximating the relationship between achievement (the dependent variable) and the covariate(s). This assumption of linearity can be checked by plotting the dependent variable(s) against the independent variable(s) and checking to see if the relationships are adequately captured by a straight line (e.g., by means of a nonparametric smoothing function; see Cohen et al. 2003).

The assumption of homoscedasticity means that the variance of the residuals is constant and is not related to any of the independent variables or predicted value(s) (Cohen et al. 2003). To check this assumption, residuals can be plotted against predicted values (e.g., as done in Malacova 2007).

The assumption of normality of residuals means that the residuals are normally distributed around the regression line. If this assumption is violated, standard errors can be imprecise (Dedrick et al. 2009) and problems with significance tests and confidence intervals can occur, especially in small samples. Even though the violation of the normality assumption usually does not lead to large problems in large samples, nonnormality might be a sign of other problems with the model, for example, misspecification (Cohen et al. 2003).

The assumption of normality can be checked, for example, by analyzing a normal probability plot (e.g., as done in Malacova 2007).

Independence of residuals means that there is no relationship between the residuals (i.e., they are uncorrelated with one another). This assumption is met in a random sample. However, if data are clustered (i.e., collected from groups), and the clustering is not taken into account in the model, this assumption might be violated (Cohen et al. 2003). When calculating VA scores, this means, for example, that students within the same class tend to be more similar to each other than expected. The violation of the independence assumption can lead to imprecise residuals and can thus also affect the reliability of the VA scores.

6.2 Statistical parameters

What statistical parameters should be reported? First, in the amount of variance (R^2), the model explains on the various levels for which the model was specified should be reported. The amount of explained variance is an important measure of effect size, and it helps the reader evaluate and interpret the reported results as well as to put the reported results in the context of previous research. Second, the American Psychological Association (APA 2010) strongly recommends that researchers report the covariance structure for multilevel models because the covariance structure is critical for understanding the estimated model. In multilevel models, the covariance structure can be defined more flexibly than, for example, in linear regression models because multilevel models contain more error terms (Dedrick et al. 2009). Thus, studies with multilevel models should contain reports of the covariance structure.

Taken together, to safeguard the transparency of reporting, the covariance structure and amount of variance that is explained (at the various levels) are important pieces of information that can be used to evaluate the applied model and the obtained empirical results.

6.3 Empirical reviews of model diagnostics

Previous studies have analyzed model assumptions in the context of VA modeling or have argued for why certain assumptions should be checked. For example, Stacy et al. (2012) analyzed the sensitivity and stability of VA scores by considering heteroscedasticity. They reported evidence that the variability and stability of teacher effects (i.e., VA scores) depend on the characteristics of a teacher's class. Using simulations based on 5000 Monte Carlo repetitions, they found that teachers of students whose achievement scores were situated in the middle of the distribution tended to have greater stability in their VA scores than teachers whose students were at the bottom end of the achievement distribution. In addition, they found significant correlations between the squared residuals from the VA models and some of the covariates (e.g., limited English proficiency), thus potentially indicating heteroscedasticity. These findings mean that the teachers could have been subjected to positive or negative consequences because the VA model did not take into account the heteroscedasticity of their classrooms' characteristics.

Another example of assumptions that have been explored in studies of VA models is linearity. For example, Lopez-Martin et al. (2014) analyzed school VA using data from three cohorts (6755 students) at four different time points. They compared multilevel VA models with nonlinear (in this case quadratic) growth models, using different covariates (e.g., gender or socioeconomic status) in both model types. They found that nonlinear models fit better and that the inclusion of student- and family-level covariates provided results that were even more accurate. The latter is the reason why they emphasized the importance of knowing the characteristics of the analyzed school. The authors highlighted the importance of finding an appropriate model to determine schools' growth (i.e., VA scores).

To our knowledge, no empirical reviews of applications of VA models have analyzed how studies on VA models have reported their statistical parameters (covariance structure or explained variance).

7 Statistical adjustments: methodological challenges and empirical results

VA models can be used to estimate teacher or school effects on students' achievement with great precision. To achieve precise measures, it is required to find appropriate ways to resolve several methodological challenges by applying statistical adjustments. Before we will discuss the methodological challenges reviewed in the present article, we would like to indicate that we did not review the validity of test scores that underlie the calculation of VA scores. Of course, validity of the achievement measures is the fundamental prerequisite on which VA models build. In this regard, validity implies that theory and empirical evidence support the interpretations of test scores for proposed uses of tests (AERA, APA,, and National Council on Measurement in Education 2014), for example, the interpretation of achievement test scores as an assessment of student learning at school. Evidence for such interpretations should be manifold, including analyses of test content, response processes, internal structure, and relations to other variables (AERA et al. 2014). For example, validity of test scores is weak when conclusions on student learning vary widely depending on the applied scaling models (i.e., models on the internal structure of tests). Importantly, this variability in test scores also leads to variability in VA scores which in turn undermines their validity as a measure of teacher or school effectiveness (see, e.g., Amrein-Beardsley and Barnett 2012; Ng and Koretz 2015; Pham 2018). In the present article, we assumed that the underlying test scores used for calculating VA scores are valid measures of students' achievement.

7.1 Methodological challenges

Several methodological challenges need to be tackled to increase precision, involving (a) adjustment for measurement error, (b) treatment of missing data, and (c) shrinkage of VA score estimates.

First, when using standardized tests, measurement error will naturally occur. For example, a review of meta-analyses on score reliability showed that the average

coefficient alpha is .80 with a standard deviation of .09 and a range of .45 to .95 (Vacha-Haase and Thompson 2011). This implies that almost every measured variable in educational research has at least some degree of measurement error. This also includes achievement test scores. For example, a meta-analysis by Rodriguez and Maeda (2006) showed that the average reliability of a state-wide achievement test was about .92, which is high, but which also shows that (on average) about 8% of the total student-level variance is due to measurement error. Crucially, if there is measurement error in an independent variable, the regression coefficients (β, γ) and consequently the residual terms (r, e, μ) will be imprecise. Because the VA score is estimated using the residuals, the VA score will be imprecise, too. An important question is how to adjust for measurement error in order to get as close to the actual true value as possible. One possibility is to correct each correlation in a full correlation matrix for measurement error (Cohen et al. 2003).

One implication of measurement error is regression to the mean (e.g., Little 2013), which occurs when using repeated measures with an imperfect correlation between two estimates from different occasions. Regression to the mean influences regression coefficients and thus also VA scores. One way to adjust for regression to the mean is to use multiple baseline measurements to reduce measurement variability (Barnett et al. 2005).

Second, missing data usually occur in longitudinal studies. Different assumptions about the missing data can be made: missing at random, missing completely at random, or missing not at random (see Rubin 1976). If data are missing at random, this means that the probability of having missing data on a certain variable is unrelated to the value of the variable itself, after the other variables in the data set are controlled for. A special case of data that are missing at random is data that are missing completely at random. This means that the probability of having missing data on a certain variable is completely unrelated to the value of the variable itself and to the values of other variables in the data set. If the missing at random assumption is violated, data are missing not at random. This means that the missing values are statistically related to the reason for their missingness. Usually, data that are missing at random are also called ignorable, and data that are missing not at random are called non ignorable (Allison 2002; Schafer and Graham 2002). Missing data will decrease the precision of the variance of the residuals, whereas more complete data will increase the precision of the prediction of residuals and thus of the VA score (McCaffrey and Lockwood 2011). There are different methods for handling missing data (e.g., imputation or maximum likelihood; for more details, see, e.g., Allison 2002); their applicability depends on the assumptions behind the mechanism causing the missing data (e.g., multiple imputation is a method that can be applied to deal with data that are missing at random).

Third, shrinkage estimators, often also called empirical Bayes estimators, use estimates from the full sample to "shrink" the values of individuals or groups, bringing them closer to the population mean (Cohen et al. 2003). Shrinkage in VA modeling is used after the actual VA estimation to correct the teacher or school effects for the overrepresentation of high or low performance within the group of students who have one teacher or the students who attend one school (e.g., Johnson et al. 2012). When

using a shrinkage estimator in a multilevel model, the level 2 residual estimates $\hat{\mu}_{0j}$ are estimated using a constant shrinkage factor c_j (Ray 2006). This shrinkage factor has a value between 0 and 1. The larger the size of the level 2 units (e.g., class size), the closer this shrinkage factor is to 1, and the closer the multilevel residual estimates $\hat{\mu}_{0j}$ are to the true level 2 residuals, leading to more accurate VA scores.

7.2 Empirical reviews of statistical adjustments

Previous studies have investigated different statistical adjustments in VA models. First, Koedel et al. (2012) analyzed how test measurement error could be accounted for with empirical and simulated data. In both data sets, they found that inferences from a VA model could be improved by adjusting for test measurement error. The improvements they found when they applied an adjustment for measurement error in an empirical sample were the same as the improvements that would be expected after increasing the sample sizes by 11 to 17%.

Second, most researchers seem to agree that violating the missing at random assumption can lead to a bias in VA scores (e.g., Karl et al. 2013; McCaffrey et al. 2003; McCaffrey and Lockwood 2011). For example, McCaffrey and Lockwood (2011) investigated VA models for which the missing data were assumed to be missing not at random. In their longitudinal sample of over 9000 students from grades 1 to 5, they found that only 21% of the students had observed scores on all measures and that students with fewer scores tended to have lower mean achievement scores. The correlations between VA scores from models allowing missing data to be missing not at random and models with the missing at random assumption were high (between .98 and 1 across grades). Even though allowing the data to be missing at random only had little impact on teacher VA scores, the authors point out the potential benefits of applying the missing not at random model, for example, in data sets where more students have missing data on the achievement test when they are taught by a certain target teacher.

Third, Herrmann et al. (2016) analyzed the use of shrinkage procedures when calculating VA models with data from over 17,000 students. They found that shrinkage improved the precision of the VA scores. However, differences between VA scores that were based on estimates that did and did not account for shrinkage were not large enough to lead to different conclusions concerning the evaluations of the teachers.

In sum, most of these studies investigating VA models with and without statistical adjustments found that a model with such adjustments led to greater precision in the VA score. Even though the improvement of accuracy after the use of statistical adjustments was not significant in every case, the improvement of precision could still be practically relevant for the future of some teachers or schools.

8 Student and context characteristics

8.1 Which covariates should be included in the VA models?

Recall that the VA score of a certain school or teacher juxtaposes the actual achievement attained by students attending a certain school or who have a certain teacher with

the achievement that is expected for students who have the same background characteristics. This comparison hinges on the rationale that these background characteristics affect students' achievement and that including these variables as covariates in the VA model should render VA scores as fair as possible. Yet, which variables should be included in the VA model? The answer to this question can be guided by models of school learning that emphasize that learning and achievement are influenced by various factors (e.g., Haertel et al. 1983; Wang et al. 1993). These factors can be grouped into student, family, and external context variables.

The most important student-level characteristic to explain student achievement seems to be prior achievement (e.g., Casillas et al. 2012). Other kinds of student-level characteristics which affect student achievement include sociodemographic variables (e.g., gender; Voyer and Voyer 2014): cognitive variables such as intelligence or memory (e.g., Baumert et al. 2009; Rohde and Thompson 2007) and motivational (e.g., Uguroglu and Walberg 1979), affective (e.g., attitude toward school subjects or anxiety, Hattie 2009; Ma 1999), and personality variables (e.g., Poropat 2009).

For the family background characteristics of the students, the socioeconomic status (SES) of the parents has been found to be positively associated with student achievement (e.g. Sirin 2005; White 1982). In addition, the language(s) spoken at home and migration background seem to affect student achievement (e.g., Genesee et al. 2005; Hopf 2005).

Educational context-level characteristics have been found to have an influence on the performance of students, too. They include teacher variables (e.g., experience; Harris and Sass 2011), classroom variables (e.g., class size; Hattie 2009), and school variables (e.g., school climate or location; Helmke 2003; Mahimuang 2005). These educational context characteristics constitute differences in learning environments that may affect students' development.

8.2 Empirical reviews of covariates used to compute VA scores

Several studies have sought to determine which variables should be included in or excluded from the VA model. Whereas there are studies that have identified prior achievement as the only efficient predictor of actual achievement (e.g., Mahimuang 2005), in other studies comparing models with and without covariates, the authors concluded that the inclusion of covariates could significantly improve the VA model. For example, Ferrão (2009) analyzed the inclusion or exclusion of SES variables in multilevel VA models. Using a longitudinal sample of about 1500 students in grades 1, 3, 5, 7, and 8, she found that SES was a significant predictor in almost all grades and that the impact of model choice (i.e., the choice of covariates) was greatest in primary school grades.

The inclusion or exclusion of SES and student demographics was also analyzed by Ballou et al. (2004). They analyzed the Tennessee Value-Added System (developed by Sanders and Horn 1994) and modified it by including student SES and demographic variables as covariates. They found that the inclusion of these variables had only a moderate impact on teacher effects (i.e., VA scores) such that teacher effects in the initial model were correlated with teacher effects after the variables were included

($r > .90$). This was even the case for teachers with classes comprised entirely of poor or minority students.

Johnson et al. (2015) also found correlations of above .90 between VA scores representing teacher effects calculated when including or excluding student- and peer-level background variables. However, they argued that these high correlations did not necessarily prevent teachers from being misclassified. They analyzed the teacher rankings that depended on the VA measure, and they found that 26% of the teachers in the bottom quintile ranked higher when the model used to estimate the VA scores included additional covariates.

9 Research objectives

VA modeling is used to identify highly effective teachers or schools as well as to evaluate educational systems and educational agents. Thus, VA scores contribute to the cumulative body of knowledge in educational research on teaching and school effectiveness, and even more importantly, they can affect teachers' lives and schools' futures. When a statistical method has such far-reaching implications, a high level of methodological rigor and transparency in reporting is essential for safeguarding valid knowledge and decision making and to allow the general public and the scientific community to scrutinize the methods. A clear description of the model is necessary so that the knowledge of previous research can be built upon, the chosen methodology can be evaluated, and the potential for replicating previous results can be guaranteed. However, large-scale surveys point to the fact that school principals and teachers seem to consider this transparency lacking (Goldring et al. 2015; Jiang et al. 2015). Furthermore, previous scientific reviews have highlighted that some studies have lacked methodological rigor (e.g., Koedel et al. 2015; McCaffrey et al. 2003). Even though some of the studies reviewed by McCaffrey et al. (2003) found that teachers' effects on students' achievement (i.e., their VA score) persisted over time, the authors suspected that the magnitudes of these effects were overstated. Because they identified different sources of potential errors, they recommended that any attempt to use VA measures for high-stakes decisions should be based on an understanding of these potential errors. Similarly, although Koedel et al. (2012) found consistent evidence for the benefits for students when using VA scores to inform decision making, they recommended additional exploration concerning the use of VA models to inform teacher assignments in order to improve instructional quality. The largest review to date by Everson (2017) was based on 99 studies, including both teacher and school VA models, and showed that many methodological challenges were not adequately addressed. For example, many of the reviewed studies did not justify their modeling choice. Although the excellent review by Everson provided important insights, these insights were limited to studies published between 2007 and 2015 in the English language with the majority of the studies coming from the USA.

The present international review was aimed at significantly contributing to the knowledge of how VA models are specified and communicated in the international educational research practice. Specifically, we focused on several vital methodological questions concerning VA modeling for various targets, involving teachers and schools: (a) Which statistical models were used to compute VA? (b) Which model diagnostics

were checked and which statistical parameters were reported? (c) Which statistical adjustments were made to tackle methodological challenges? (d) Which student and context characteristics were included as covariates in the VA models?

In addressing these research questions, we aimed to significantly extend knowledge of the application of VA models in educational practice. In particular, previous reviews have primarily tackled teacher VA scores (Koedel et al. 2015; McCaffrey et al. 2003) or have focused on methodological concerns in VA modeling for a selected sample of studies (Everson 2017). The present review addresses the generalizability of previous findings. Relative to previous reviews, we (a) included and systematically analyzed a considerably larger number of empirical studies (i.e., 370 studies), (b) covered a much longer time span (i.e., from 1971 to July 2017), and (c) covered a much larger number of countries (i.e., 26 countries) by including research reports written in the English, French, and German languages.

10 Method

10.1 Research process

The research process we applied followed Reed and Baxter's (2009) suggestions to use reference databases in research synthesis. First, we searched for articles on VA scores using "value added" or "added value" as search terms and specifications (if possible) for the research domains of psychology, education, or social sciences. We searched for articles in the ERIC,[2] Scopus,[3] PsycINFO,[4] and Psyndex[5] databases. Google Scholar[6] was used to get access to publications that were listed but not stored as full texts in the databases (i.e., only the title and abstract were available). The literature search was conducted between February 1 and July 15, 2017.

Studies were included in this review if they satisfied the following criteria:

- Research about VA modeling in primary or secondary education (not early childhood education, higher education, or adult education)
- Published as a scientific journal article, a conference proceeding, a report, a book, or a book chapter. Both peer-reviewed and nonpeer-reviewed studies were included because there are many reports that are officially not peer reviewed but nevertheless represent an important component of the VA literature
- Available as full text in English, German, or French
- Empirical application of VA models

Figure 1 presents a flow chart for study selection, based on the PRISMA flow diagram (Liberati et al. 2009). In the end, 370 studies met our criteria and were coded for 20 different categories. A full list with all the references included in this review is found in Table A1 (online resource).

[2] Educational Resources Information Center, produced by the Institute of Education Sciences
[3] Produced by Elsevier B.V.
[4] Produced by the American Psychological Association (APA)
[5] Produced by the Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID)
[6] Produced by Google LLC

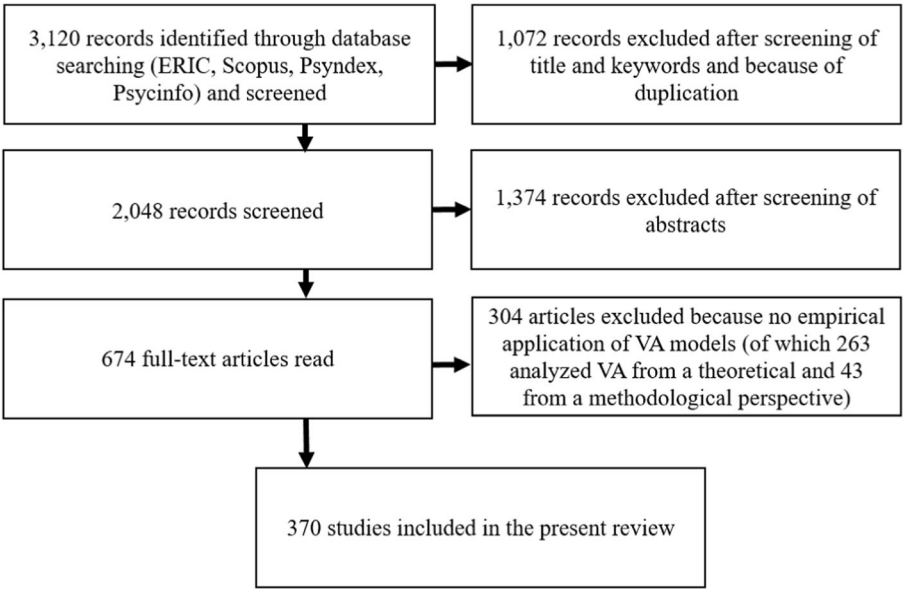

Fig. 1 Flowchart of the research process, structured after the PRISMA flow diagram (Liberati et al. 2009)

10.2 Coding

A coding manual and coding forms were used in the coding process, as recommended by Cooper et al. (2009). The coding manual is found in Table A2 (online resource). Categories for coding have been chosen based on previous research on VA modeling (e.g., Everson 2017) theories on learning and previous research on student achievement (e.g., Haertel et al. 1983) and statistical models that can be used for the prediction of student achievement (e.g., Dedrick et al. 2009). The initial coding was done by the first author of this study. Forty randomly chosen studies were double coded. The second coder was trained by the first coder to work with the coding manual and subsequently took part in a practice coding phase that comprised regular meetings to discuss and adapt the coding process. The average agreement for both ratings across 20 categories was 90% (ranging from 78 to 100%). When accounting for chance, interrater reliability was still substantial with an average $\kappa = .75$ (ranging from .43 to 1). A table with the interrater-reliability for every category is found in Table A3 (online resource).

11 Results

11.1 Description of publications

We analyzed a total of 370 publications in which VA modeling had been applied. They were published in 26 different countries; of these studies, 253 studies were conducted in the USA and 117 studies in other countries. A list of all the countries is found in Table A4 (online resource). Additionally to the results presented below, we have calculated all the results divided by country (the USA vs. not the USA, Tables A5–

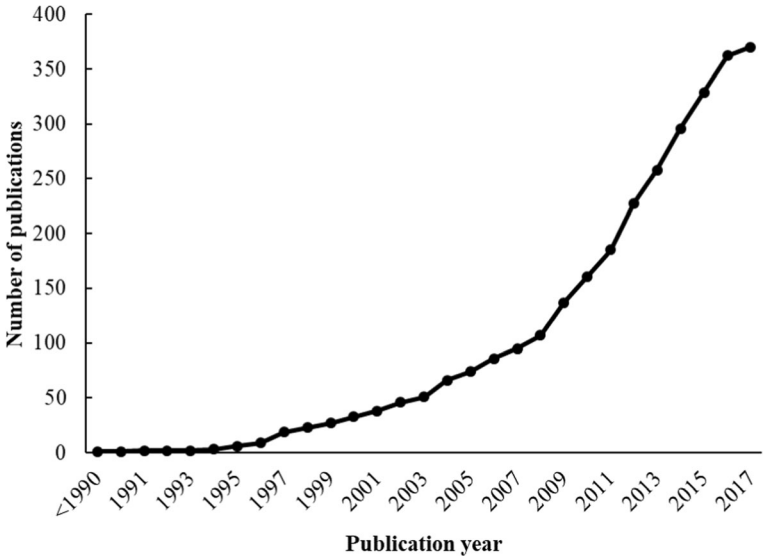

Fig. 2 Accumulated number of publications with empirical application(s) of value-added models

A9 [online resource]) and by year (before 2000, from 2000 to 2009, and after 2010, Tables A10–A14 [online resource]). Unless otherwise stated, the tendencies that can be observed in these tables are the same as in the general tables.

As is seen in Fig. 2, the total number of empirical publications on VA models has increased in recent years. The increase gets steeper from 2002 onward. The samples in the studies consisted of primary school students (130 studies, 35%), secondary school students (118 studies, 32%), and both (120 studies, 32%). Two studies (1%) did not specify the educational level.

As shown in Table 1, teachers were the VA model target in 159 studies (43%). The other studies analyzed VA models applied at the school, classroom, or principal levels or a combination of some of these variables (including a more general level, e.g., district level). Most of the studies analyzing teacher VA scores were conducted in the USA (153 studies vs. six studies from the "rest of the world"). Most studies from countries other than the USA analyzed VA models at the school level (89 studies vs. 55 studies from the USA, see Table A5 [online resource] for further detail).

Table 1 Frequencies of value-added targets in publications

Value-added target	Frequencies	Percent
Teacher	159	43
School	144	39
Principal	8	2
Class	1	0
Combination	58	16
Total	370	100

Table 2 Frequencies of value-added models used in publications

Value-added model	Studies with teacher as VA target	Studies with school as VA target	Studies with other VA target	All studies
Linear regression VA model	93 (58%)	35 (24%)	34 (51%)	162 (44%)
Multilevel VA model				
TVAAS /EVAAS[a]	7 (4%)	5 (3%)	2 (3%)	14 (4%)
Other multilevel	28 (18%)	53 (37%)	15 (22%)	96 (26%)
Other VA model	4 (3%)	11 (8%)	1 (1%)	16 (4%)
More than one VA model[b]	13 (8%)	12 (8%)	8 (12%)	33 (9%)
VA model not specified	14 (9%)	28 (19%)	7 (10%)	49 (13%)
Total number of studies	159	144	67	370

Sum of percentages is not always 100 because of rounding

[a] Tennessee Value-Added Assessment System/Education Value-Added Assessment System

[b] More than one model type was used, most frequently a linear regression model and a multilevel model

11.2 Which statistical models were used to compute VA scores?

Out of a total of 370 studies, 228 studies (62%) described the applied statistical model and included a formula, 108 studies (29%) gave only a verbal description, and 34 studies (9%) gave no description at all.

Table 2 shows the frequencies of the different models that have been used: Most studies (162; 44%, of which 141 from the USA; see Table A6 [online resource]) calculated linear regression models; 110 studies (30%) used multilevel models, of which 14 studies (4%; all from the USA; Table A6 [online resource]) applied the TVAAS or the EVAAS model. The remaining 49 studies (13%) used either a different model type (16 studies, 4%), mainly a gain score model, or more than one model type (33 studies, 9%), mainly to compare different models, such as linear regression and multilevel models. The other 49 studies (13%) did not specify any model.

11.3 Which model diagnostics were checked and which statistical parameters were reported?

As shown in Table 3, most studies did not report key statistical parameters (i.e., the covariance structure of a multilevel model or the amount of explained variance), indicated with a "Yes" in Table 3. Moreover, most studies did not report any model diagnostics (i.e., linearity, homoscedasticity, normality, independence of residuals). In particular, 88% of the studies (321 studies) did not include any model diagnostic check; 51% of the multilevel studies (71 out of 140 studies) reported neither the covariance structure nor the explained variance.

11.4 Which statistical adjustments were made to tackle methodological challenges?

In Table 4, studies were coded as "Yes" if they made a certain adjustment or if they explained why they did not use it. Most studies did not include statistical adjustments

Table 3 Model diagnostics and statistical parameters reported in the publications

	Studies with teacher as VA target		Studies with school as VA target		Studies with other VA target		All studies	
Statistical parameters and model diagnostics	Yes	No	Yes	No	Yes	No	Yes	No
Statistical parameters								
Explained variance								
Studies with multilevel model	14 (29%)	34 (71%)	36 (36%)	33 (33%)	11 (48%)	12 (52%)	61 (44%)	79 (56%)
Studies without multilevel model	30 (27%)	81 (73%)	12 (16%)	63 (84%)	15 (34%)	29 (66%)	57 (25%)	173 (75%)
All studies	44 (28%)	115 (72%)	48 (33%)	96 (67%)	26 (39%)	41 (61%)	118 (32%)	252 (68%)
Covariance structure[a]								
Studies with multilevel model	6 (13%)	42 (88%)	6 (9%)	63 (91%)	2 (9%)	21 (91%)	14 (10%)	126 (90%)
Studies without multilevel model	–	–	–	–	–	–	–	–
Model diagnostics								
Linearity	13 (8%)	146 (92%)	7 (5%)	137 (95%)	10 (15%)	57 (85%)	30 (8%)	340 (92%)
Homoscedasticity	12 (8%)	147 (92%)	2 (1%)	142 (99%)	7 (10%)	60 (90%)	21 (6%)	349 (94%)
Normality	2 (1%)	157 (99%)	1 (1%)	143 (99%)	1 (1%)	66 (99%)	4 (1%)	366 (99%)
Independence	6 (4%)	153 (96%)	0 (0%)	144 (99%)	1 (1%)	66 (99%)	7 (2%)	363 (98%)
Total number of studies	159		144		67		370	

Sum of percentages is not always 100 because of rounding

[a] We analyzed the reporting of the covariance structure for the 140 out of 370 studies in which multilevel models were applied, including studies applying more than one model type of which one of the model types was a multilevel model ($n = 48$ for teacher VA; $n = 69$ for school VA; $n = 23$ for other VA target)

Table 4 Statistical adjustments in the value-added models

	Studies with teacher as VA target		Studies with school as VA target		Studies with other VA target		All studies	
Statistical adjustments	Yes	No	Yes	No	Yes	No	Yes	No
Measurement error	48 (30%)	111 (70%)	15 (10%)	129 (90%)	19 (28%)	48 (72%)	82 (22%)	288 (78%)
Missing data	31 (19%)	128 (81%)	24 (17%)	120 (83%)	18 (27%)	49 (73%)	73 (20%)	297 (80%)
Shrinkage	58 (36%)	101 (64%)	19 (13%)	125 (87%)	22 (33%)	45 (67%)	99 (27%)	271 (73%)
Total number of studies	159		144		67		370	

(measurement error, missing data, shrinkage). In particular, 55% of the studies did not apply any of the proposed adjustments, and 25% of the studies applied only one. The number of studies which included adjustments for measurement error increased over the years (0% before 2000; 21% between 2000 and 2009; 25% after 2009) and stayed more constant for adjustments for missing data and shrinkage (see Table A13 [online resource]) .

11.5 Which student and context characteristics were included as covariates in the VA models?

Table 5 presents the kinds of covariates that were included in the VA models. The numbers in the "Yes" column indicate the number of studies that included a certain covariate. Most studies included prior achievement in their VA models. Sociodemographic and socioeconomic background variables were included in almost two thirds of the studies (63% and 64%, respectively). By contrast, other cognitive (e.g., intelligence), motivational, or affective student variables were included in only eight studies (2%, all from other countries than the USA; see Table A9 [online resource]). Language variables and educational context characteristics (class, teacher, and school) were included in 16 to 35% of the studies. The number of studies from countries other than the USA which included language variables (3%) and educational context characteristics (12%, 4%, and 42%, respectively) was respectively lower than for the studies conducted in the USA (see Table A9 [online resource]). Additionally, more studies included class and teacher variables over time (from 19% and 7% before 2000 to 33% and 17% after 2009), whereas less studies included school variables (41% before 2000 and 32% after 2009; see Table A14 [online resource]).

12 Discussion

VA modeling is used to identify highly effective teachers or schools as well as to evaluate educational systems and educational agents. Thus, VA scores contribute to the cumulative body of knowledge in educational research on teaching and school effectiveness, and even more importantly, they can affect teachers' lives and schools' futures. Despite the far-reaching impact of VA scores, and although consequential decisions are made on the basis of results from VA models, there is a critical lack of knowledge regarding the estimation and use of VA scores. A real-life example of the implications for teachers' lives can be seen in a recent court case: A teacher went to court because her VA score dropped from 14/20 to only 1/20 in 1 year, identifying her as "ineffective" (Cimarusti 2016). The teacher's arguments (e.g., that the VA model was not transparent and the calculations of her VA score were not made available to her) were convincing to the court, and she won the case.

The objective of the present review was to provide an integrative summary of the use of VA models in empirical applications, focusing on methodological questions. An analysis of 370 empirical studies about VA modeling from 26 different countries shows the growing presence and importance of the use of VA models. The increase in publications in recent years shows that the relevance of the topic is present not only in practice and at a political level for accountability purposes but also in research. Since

Table 5 Student and context characteristics included as covariates in the value-added models

	Studies with teacher as VA target		Studies with school as VA target		Studies with other VA target		All studies	
Covariates	Yes	No	Yes	No	Yes	No	Yes	No
Student characteristics								
Prior achievement	147 (92%)	12 (8%)	108 (75%)	36 (25%)	59 (88%)	8 (12%)	314 (85%)	56 (15%)
Sociodemographic variables	114 (72%)	45 (28%)	78 (54%)	66 (46%)	42 (63%)	25 (37%)	234 (63%)	136 (37%)
Cognitive student variables	1 (1%)	158 (99%)	4 (3%)	140 (97%)	3 (4%)	64 (96%)	8 (2%)	362 (98%)
Motivational/affective student variables	0 (0%)	159 (100%)	2 (1%)	142 (99%)	6 (9%)	61 (91%)	8 (2%)	362 (98%)
Family background characteristics								
Socioeconomic variables	113 (71%)	46 (29%)	78 (54%)	66 (46%)	45 (67%)	22 (33%)	236 (64%)	134 (36%)
Language variables	79 (50%)	80 (50%)	25 (17%)	119 (83%)	26 (39%)	41 (61%)	130 (35%)	240 (65%)
Educational context characteristics								
Class variables	86 (54%)	73 (46%)	5 (3%)	139 (97%)	17 (25%)	50 (75%)	108 (29%)	262 (71%)
Teacher variables	50 (31%)	109 (69%)	2 (1%)	142 (99%)	7 (10%)	60 (90%)	59 (16%)	311 (84%)
School variables	47 (30%)	112 (70%)	55 (62%)	89 (33%)	22 (33%)	45 (67%)	124 (34%)	246 (66%)
Total number of studies	159		144		67		370	

Sum of percentages is not always 100 because of rounding

2002, the year in which the No Child Left Behind Act was enacted, the number of publications seems to be increasing even faster, with consequences for accountability systems in education and the more frequent use of VA models for evaluating teacher and school effectiveness.

In line with Everson (2017), we found that most publications using VA models have stemmed from the USA, even after considering publications in German and French. This might be explained by the importance of VA scores in the USA for educational decision making. For half of the studies, we found that the target of the VA models was the teacher and that most of these studies were conducted in the USA. This is probably because VA scores are used for accountability purposes and high-stakes decisions in assessments of teacher quality in the USA, especially since the enactment of the No Child Left Behind Act (2002). In the remaining 25 countries included in this review, the VA target was predominantly the school or sometimes a combination of the teacher and school. This could be a result of the use of VA scores for school monitoring purposes in the UK or for school rankings in France. In addition, VA scores usually do not play a role in high-stakes decisions in countries outside the USA. Rather, in these countries, the goal is to use VA scores to identify highly effective teachers or schools to learn about factors that significantly promote students' achievement.

12.1 Which statistical models are used to compute VA scores?

Various statistical models have been proposed to compute VA scores. Further, transparency standards (AERA 2006) require researchers to specify which models they used. In the present study, we found that 9% of the studies provided no description of the model they used, which means that neither the statistical model nor the covariates were described. A larger percentage of studies (13%) were classified as not having specified the statistical model. The difference between these numbers can be accounted for by the fact that some studies described parts of the model (e.g., the covariates that were used) but gave no exact specification of the statistical model that was used. If this information is not included, it is impossible to replicate the study or evaluate the results.

When the model was specified, we found (in line with Everson 2017) that most studies used linear models (i.e., linear regression or multilevel models). However, there seems to be no consensus about which statistical model should be used. This lack of consensus is crucial when considering the fact that VA scores resulting from different model types are used to draw the same conclusions (e.g., for teacher or school accountability). Not only several studies found differences in teacher rankings that depended on the model that was used (e.g., Goldhaber et al. 2013; Tekwe et al. 2004), but also, even when the same VA model was used across several years, the rankings were not necessarily stable. For example, Newton et al. (2010) found that the rankings of 19 to 41% of the teachers changed by three or more deciles even when the same model was applied across several years. These findings emphasize the importance of model choice in order to obtain a VA model that estimates VA scores as accurately and with as much stability as possible.

The fact that most studies have used either linear regression or multilevel models seems to be an indicator of a conflict between accurate analyses and the aim to keep the model as simple and understandable as possible. One argument for keeping VA models simple is to ensure that teachers and parents can understand them. However, a higher

level of complexity will lead to greater accuracy in VA scores (e.g., Kelly and Downey 2010; Wei et al. 2012). Considering the importance of the consequences of decisions made by using information from VA models, we recommend that researchers always aim to arrive at a model that is as accurate as possible. Even if the model might be more difficult to understand, if it is explained well, educational agents will probably appreciate the additional accuracy and fairness of the model. This means that if a choice has to be made between linear regression and multilevel models, multilevel models should be preferred in order to respect the nested structure of students within classes and schools. In addition, further investigation into nonlinear model types (e.g., Lopez-Martin et al. 2014) or other statistical models (e.g., propensity score matching; Everson et al. 2013) could be a promising avenue for future VA research.

12.2 Which model diagnostics are checked and which statistical parameters are reported?

Even though the relevant reporting standards (AERA 2006) imply a high level of transparency in research on VA models, and many journals aim to publish papers that are intelligible to a broad audience, educational practitioners (e.g., teachers and principals) still perceive that VA scores are not transparent, and they question the validity of VA scores. The present results provide empirical support for the perceptions of these teachers and school principals. For example, 51% of the 140 studies that applied a multilevel model did not report the covariance structure or the amount of explained variance. In addition, most of the 370 studies (88%) did not report any model diagnostics. More concretely, only 8% of the studies checked the assumption of linearity. However, the relationship between prior achievement and actual achievement does not appear to be linear (Lopez-Martin et al. 2014). Even if studies do not employ a nonlinear model, it would be important at least to check for whether the assumption of linearity might be violated in order to guarantee the accuracy of VA scores. The same picture was observed for the other model assumptions of homoscedasticity, normality of residuals, and independence of residuals (checked in 6%, 1%, and 2% of the studies, respectively) and statistical parameters (covariance structure reported in 10% of the studies using multilevel models, explained variance in 32% of all the studies). This lack of reporting of statistical parameters and assumptions is not necessarily limited to VA studies. For example, in a review of 99 studies applying multilevel models, Dedrick et al. (2009) found that 58% of the studies did not discuss the covariance structure of their multilevel models. The quality of reporting in studies could be improved by adding statistical parameters (e.g., the covariance structure of multilevel models) or other methodological information to the reports of results, thus allowing readers to interpret the results and conduct replication studies. This is the case not only for studies in general but also for publications on VA modeling.

12.3 Which statistical adjustments are made to tackle methodological challenges?

Several statistical adjustments have been proposed to increase the precision of VA scores: adjustments for measurement error, methods for dealing with missing data, and a shrinkage procedure. Even though statistical adjustments seem to increase precision in VA scores (Herrmann et al. 2016; Koedel et al. 2012; McCaffrey and Lockwood

2011), more than half of the studies (55%) analyzed in the present review did not use any of the proposed statistical adjustments. This shows a serious lack of rigor at the methodological level of the studies. Not using statistical adjustments when calculating VA models may lead to imprecise VA scores because the regression coefficients or standard errors may be imprecise. Decisions based on these imprecise VA scores (concerning the future of schools or teachers' careers) could thus be faulty, and teaching as well as school processes that are supposed to be effective might in fact not be. For example, given that most studies use error-prone achievement data to compute VA scores, it is obvious that these VA scores are imprecise because they contain measurement error. Consequently, statistical adjustments may improve the quality of such studies and their calculation of VA scores and increase precision in VA scores. Thus, the integrity of decisions may increase because the decisions will be based on more reliable VA scores. One issue that has not been explicitly reviewed in the present paper was the validity of the test scores used to calculate VA scores. We have assumed that the underlying test scores are valid, but we are aware that validity is a central question, which will profit from further research.

12.4 Which student and context characteristics are included as covariates in VA models?

Even though the inclusion or exclusion of certain covariates has been the subject of several studies on VA models (Ballou et al. 2004; Ferrão 2009; Johnson et al. 2015), to date, there is no consensus or consistency regarding which covariates should be included in VA models. In the present review, we found that most studies included prior achievement in the models. Prior achievement can usually explain the largest amount of variance. However, models of school learning predict, and empirical results show, that adding additional covariates (e.g., students' personality, motivation, intelligence, or SES) to the model helps to improve the prediction of future achievement as an outcome variable (e.g., Ferrão 2009; Johnson et al. 2015). Thus, including these covariates helps to improve the fairness of VA scores because a larger number of background characteristics that are relevant for students' future achievement are taken into account. We found that sociodemographic and socioeconomic variables were included in the majority of studies, but a remarkable number of studies lacked these variables. Importantly, cognitive and motivational or affective variables were included in only 2% of the studies. However, this pattern of results might change in the near future. With the enactment of the Every Student Succeeds Act (2015), noncognitive measures could become more important in accountability systems, and it is interesting to follow the development of VA models and the integration of these student characteristics as covariates (and outcomes).

13 Limitations

Although the present paper is the largest review of methodological issues in VA modeling to date, it is possible that we missed some pertinent studies because we considered only research that was published in English, French, or German as well as properly referenced (see Sect. 2). Given that VA modeling is part of various

governmental accountability systems, it is reasonable to assume that there is potentially valuable literature on applied VA modeling that is not available in the public domain. On the other hand, given that the number of referenced publications keeps growing, the role of research is most likely becoming more important for practitioners, too. In addition, it could be argued that quality standards are better met in a research context than in practice because of, for example, financial resources or time constraints. Thus, the open question is: How far and in what direction does the use of VA in practice differ from the results of the present review?

Another limitation of the present study is that we cannot exclude the possibility that the authors of the reviewed publications included a certain variable, statistical adjustment, or parameter without explicitly specifying it and that we thus missed some information in the classification process. If this is the case, the results reported in the present review might be too pessimistic. However, given that reporting standards exist, we believe that authors most likely respected them, especially concerning important information such as statistical adjustments.

In the present study, we found that most studies applied some kind of regression-based VA modeling. However, other types of models can also be used to calculate VA scores (e.g., propensity score matching). Given that we encountered such alternative models only in a very limited subset of studies, we coded them under "other VA models" but we did not discuss or analyze them in detail.

Finally, the present paper addressed only methodological questions pertaining to VA modeling. As pointed out by two anonymous reviewers of our paper, other essential aspects of VA models and their application require further investigation (see, for example, the works by Amrein-Beardsley and Barnett 2012; Everson 2017). This involves the following questions: In which countries are VA scores used to make high-stakes decisions (e.g., personnel decisions)? What difference does a high versus low VA make on student achievement outcomes? Has the use of VA in school or teacher evaluations improved schools or instructional quality? Is VA continuing to gain momentum or has there been push-back (from practice and research)? Should VA scores used to make high-stakes decisions?

14 Conclusion

Given the high relevance and political importance of VA models, VA scores should be as accurate and transparent as possible. However, the present literature review revealed a lack of transparency, rigor, and consistency in a large number of studies on VA modeling. The quality and clarity of studies on VA modeling could—and should—be improved by respecting relevant reporting standards (AERA 2006) when reporting methods and results. In addition, the understanding and acceptance of teachers and principals could be improved by describing, specifying, and explaining the model in considerably greater detail. Our recommendation is thus to promote Darling-Hammond's (2015) idea that more detailed information on the quality of VA scores would allow policymakers to acknowledge the limitations of VA models and enable a more thoughtful development of VA models, with the goal to improve the quality of teaching rather than only to rank teachers and dismiss the ones located in

the bottom part of the distribution. Specifically, to improve the quality of studies on VA modeling, reviewers and editors should require VA researchers to submit thorough descriptions of the models they used and the covariates they included. Furthermore, we suggest that future studies on VA modeling use appropriate statistical adjustments or that they explain why they decided not to use adjustments. In addition, checking the violation of model assumptions and reporting statistical parameters is crucial for ensuring that results are comprehensible and for allowing for public and scientific scrutiny.

Acknowledgements The authors would like to thank Isabelle Klee for her rigorous work in double-coding.

Funding The present research was supported by a PRIDE grant (no. 10921377) of the Luxembourg National Research Fund (FNR).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Allison, P. D. (2002). Missing data: quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology, 55*, 193–196. https://doi.org/10.4135/9780857020994.n4.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40. https://doi.org/10.3102/0013189X035006033.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Amrein-Beardsley, A., & Barnett, J. H. (2012). Working with error and uncertainty to increase measurement validity. *Educational Assessment, Evaluation and Accountability, 24*(4), 369–379. https://doi.org/10.1007/s11092-012-9146-6.

Amrein-Beardsley, A., & Geiger, T. (2017). All sizzle and no steak: value-added model doesn't add value in Houston. *Phi Delta Kappan, 99*(2), 53–59.

Amrein-Beardsley, A., & Holloway, J. (2017). Value-added models for teacher evaluation and accountability: commonsense assumptions. *Educational Policy*. https://doi.org/10.1177/0895904817719519.

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

Barnett, A. G., Van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology, 34*, 215–220. https://doi.org/10.1093/ije/dyh299.

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*, 165–176. https://doi.org/10.1016/j.edurev.2009.04.002.

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*, 324–359. https://doi.org/10.3102/0002831216630407.

Bonesrønning, H. (2004). Can effective teacher behavior be identified? *Economics of Education Review, 23*, 237–247. https://doi.org/10.1016/j.econedurev.2003.07.002.

Bradbury, A. (2011). Equity, ethnicity and the hidden dangers of 'contextual' measures of school performance. *Race Ethnicity and Education, 14*, 277–291. https://doi.org/10.1080/13613324.2010.543388.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2012). *Estimating the effect of leaders on public sector productivity: the case of school principals.* (working paper no. 66). National Center for Analysis of Longitudinal Data in Education Research. https://doi.org/10.3386/w17803.

Casillas, A., Robbins, S., Allen, J., Kuo, Y.-L., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology, 104*(2), 407–420. https://doi.org/10.1037/a0027180.

Cimarusti, D. (2016). *NY teacher victorious in having court throw out VAM score*. Retrieved April 27, 2018, from https://networkforpubliceducation.org/2016/05/ny-teacher-victorious-court-throw-vam-score/. Accessed 27 April 2018.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah: Lawrence Erlbaum Associates, Inc..

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

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

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: a review of methodological issues and applications. *Review of Educational Research, 79*, 69–102. https://doi.org/10.3102/0034654308325581.

Duclos, M., & Murat, F. (2014). Comment évaluer la performance des lycées ? Un point sur la méthodologie des IVAL (Indicateurs de valeur ajoutée des lycées). *Éducation & Formations, 85*, 73–84.

Evers, A. (2001). Improving test quality in the Netherlands: results of 18 years of test ratings. *International Journal of Testing, 1*(2), 137–153.

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

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

Every Student Succeeds Act, Pub. L. No. 114–95, 129 Stat. 1802 (2015).

Ferrão, M. E. (2009). Sensivity of value added model specifications: measuring socio-economic status. *Revista de Educacin, 348*, 137–152.

Ferrão, M. E., & Goldstein, H. (2009). Adjusting for measurement error in the value added model: evidence from Portugal. *Quality & Quantity, 43*, 951–963. https://doi.org/10.1007/s11135-008-9171-1.

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in US schools: an overview of research findings. *Journal of Education for Students Placed at Risk, 10*, 363–385. https://doi.org/10.1207/s15327671espr1004_2.

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., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319), 589–612. https://doi.org/10.1111/ecca.12002.

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

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*, 96–104. https://doi.org/10.3102/0013189X15575031.

Haertel, G. D., Walberg, H. J., & Weinstein, T. (1983). Psychological models of educational performance: a theoretical synthesis of constructs. *Review of Educational Research, 53*, 75–91. https://doi.org/10.2307/1170327.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: estimation using micro data. *The American Economic Review, 61*(2), 280–288.

Harris, D. C. (2007). *The promises and pitfalls of alternative teacher compensation approaches [policy brief]*. Madison: Wisconsin Center for Education Research, University of Wisconsin-Madison.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*, 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009.

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Helmke, A. (2003). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze: Kallmeyersche Verlagsbuchhandlung.

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy, 3*(1), 1–10. https://doi.org/10.1080/2330443X.2016.1182878.

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. https://doi.org/10.3102/0002831210387916.

Hopf, D. (2005). Zweisprachigkeit und Schulleistung bei Migrantenkindern. *Zeitschrift für Pädagogik, 51*, 236–251.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-publikaties.

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In *The Oxford handbook of quantitative methods in psychology* (Vol. 2: Statistical Analysis). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0014.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's reach students. *Educational Researcher, 44*, 105–116. https://doi.org/10.3102/0013189X15575517.

Johnson, M. T., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012). *Value-added models for the Pittsburgh public schools*. Mathematica Policy Research, Inc.

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*, 60–83. https://doi.org/10.1080/19345747.2014.967898.

Karl, A. T., Yang, Y., & Lohr, S. L. (2013). A correlated random effects model for nonignorable missing data in value-added assessment of teacher effects. *Journal of Educational and Behavioral Statistics, 38*, 577–603. https://doi.org/10.3102/1076998613494819.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: looking inside the 'black box' of complex metrics. *Educational Assessment, Evaluation and Accountability, 22*, 181–198. https://doi.org/10.1007/s11092-010-9100-4.

Kersting, N. B., Chen, M.-K., & Stigler, J. W. (2013). Value-added teacher estimates as part of teacher evaluations: exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archives*, *21* (special issue on Value-added: what America's policymakers need to know and understand). https://doi.org/10.14507/epaa.v21n7.2013

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test measurement error and inference from value-added models. *The B.E Journal of Economic Analysis & Policy, 12*(1), 1–37. https://doi.org/10.1515/1935-1682.3314.

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.

Kurtz, M. D. (2018). Value-added and student growth percentile models: what drives differences in estimated classroom effects? *Statistics and Public Policy, 5*(1), 1–8. https://doi.org/10.1080/2330443X.2018.1438938.

Lavigne, A. L., & Good, T. L. (2013). *Teacher and student evaluation: moving beyond the failure of school reform*. New York: Routledge. https://doi.org/10.4324/9780203070901.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ, 339*, b2700. https://doi.org/10.1136/bmj.b2700.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: The Guilford Press.

Lopez-Martin, E., Kuosmanen, T., & Gaviria, J. L. (2014). Linear and nonlinear growth models for value-added assessment: an application to Spanish primary and secondary schools' progress in reading comprehension. *Educational Assessment, Evaluation and Accountability, 26*(4), 361–391. https://doi.org/10.1007/s11092-014-9194-1.

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education, 30*, 520–540. https://doi.org/10.2307/749772.

Mahimuang, S. (2005). Factors influencing academic achievement and improvement: a value-added approach. *Educational Research for Policy and Practice, 4*, 13–26. https://doi.org/10.1007/s10671-005-0677-1.

Malacova, E. (2007). Effect of single-sex education on progress in GCSE. *Oxford Review of Education, 33*, 233–259. https://doi.org/10.1080/03054980701324610.

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics, 5*, 773–797. https://doi.org/10.1214/10-AOAS405.

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

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*, 67–101. https://doi.org/10.3102/10769986029001067.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*, 114–140. https://doi.org/10.1037/met0000078.

MEN-DEP. (1994). *Trois indicateurs de performances des lycées (Les dossiers d'éducation et formations)*. Paris: Ministère de l'Éducation nationale et Direction de l'Évaluation et de la Prospective.

Merritt, E. G., Palacios, N., Banse, H., Rimm-Kaufman, S. E., & Leis, M. (2017). Teaching practices in grade 5 mathematics classrooms with high-achieving English learner students. *The Journal of Educational Research, 110*, 17–31. https://doi.org/10.1080/00220671.2015.1034352.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: an exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23), 1–24. https://doi.org/10.14507/epaa.v18n23.2010.

Ng, H. L., & Koretz, D. (2015). Sensitivity of school-performance ratings to scaling decisions. *Applied Measurement in Education, 28*, 330–349. https://doi.org/10.1080/08957347.2015.1062764.

No Child Left Behind Act, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Perry, T. (2016). English value-added measures: examining the limitations of school performance measurement. *British Educational Research Journal, 42*, 1056–1080. https://doi.org/10.1002/berj.3247.

Pham, G. (2018). *Complex interplay effects of classroom instructional and context factors on student growth in English as a foreign language in Vietnam: the case of nonlinear relationship, regularized regression and the relevance of the scaling model* (dissertation). Universität Koblenz-Landau, Campus Landau.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*, 322–338. https://doi.org/10.1037/a0014996.

Race to the Top Act of 2011, S.844-112th Congress. (2011). Retrieved from www.govtrack.us/congress/bills/112/s844.

Ray, A. (2006). *School value added measures in England* (a paper for the OECD Project on the development of value-added models in education systems).

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 73–101). New York: Russell Sage Foundation.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306–322. https://doi.org/10.1037/1082-989x.11.3.306.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83–92. https://doi.org/10.1016/j.intell.2006.05.004.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Rutledge, S. A., Cohen-Vogel, L., Osborne-Lampkin, L., & Roberts, R. L. (2015). Understanding effective high schools: evidence for personalization for academic and social emotional learning. *American Educational Research Journal, 52*, 1060–1092. https://doi.org/10.3102/0002831215602328.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311. https://doi.org/10.1007/BF00973726.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*, 147–177. https://doi.org/10.1037//1082-989X.7.2.147.

Scherrer, J. (2011). Measuring teaching using value-added modeling: the imperfect panacea. *NASSP Bulletin, 95*(2), 122–140. https://doi.org/10.1177/0192636511410052.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: a meta-analytic review of research. *Review of Educational Research, 75*, 417–453. https://doi.org/10.3102/00346543075003417.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and applied multilevel analysis*. Thousand Oaks: SAGE Publications.

Stacy, B., Guarino, C. M., Reckase, M. D., & Wooldridge, J. (2012). *Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve?* (discussion paper no.

7676). Institute for the Study of Labor (IZA). Retrieved from https://www.econstor.eu/bitstream/10419/89833/1/dp7676.pdf. Accessed 6 Dec 2018.

Sund, K. (2009). Estimating peer effects in Swedish high school using school, teacher, and student fixed effects. *Economics of Education Review, 28*, 329–336. https://doi.org/10.1016/j.econedurev.2008.04.003.

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*, 11–36. https://doi.org/10.3102/10769986029001011.

Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. *Journal of Research in Reading, 22*, 27–36. https://doi.org/10.1111/1467-9817.00066.

Uguroglu, M. E., & Walberg, H. J. (1979). Motivation and achievement: a quantitative synthesis. *American Educational Research Journal, 16*, 375–389. https://doi.org/10.2307/1162831.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: a retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*, 159–168. https://doi.org/10.1177/0748175611409845.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological Bulletin, 140*, 1174–1204. https://doi.org/10.1037/a0036620.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*, 249–294. https://doi.org/10.2307/1170546.

Wei, H., Hembry, T., Murphy, D. L., & McBride, Y. (2012). *Value-added models in the evaluation of teacher effectiveness: a comparison of models and outcomes* (Pearson's Research Reports). Retrieved from http://images.pearsonassessments.com/images/tmrs/ComparisonofValue-AddedModelsandOutcomes.pdf. Accessed 6 Dec 2018.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461–481. https://doi.org/10.1037/0033-2909.91.3.461.