Educational Assessment, Evaluation and Accountability (2020) 32:461–479
https://doi.org/10.1007/s11092-020-09342-8

School testing culture and teacher satisfaction

William C. Smith[1] · Jessica Holloway[2]

Received: 3 February 2020 / Accepted: 28 October 2020 / Published online: 5 November 2020

Abstract

Teachers, as frontline providers of education, are increasingly targets of accountability reforms. Such reforms often narrowly define 'teacher quality' around performative terms. Past research suggests holding teachers to account for student performance measures (i.e. test scores) damages their job satisfaction, including increasing stress and burnout. This article examines whether the relationship between test-based accountability and teacher satisfaction can be, in part, explained by the emphasis of student test scores in teacher appraisals. Although historically used for formative purposes, recent research demonstrates that across a large range of countries, nearly all teachers work in a system where their appraisal is based, in part, on students' test scores. Using data from the 2013 Teaching and Learning International Survey, we pool data from 33 countries to evaluate the direct and indirect effect of school testing culture on teacher satisfaction. Results suggest that there is a direct relationship between the intensity of the testing culture and the satisfaction of teachers, as well as an indirect relationship with test score emphasis in teacher appraisals suppressing potential positive effects of appraisals on teacher satisfaction.

Keywords Teacher satisfaction · Accountability · Testing culture · Teacher appraisal · TALIS

Recent decades have brought about a sharp increase in teacher-focused accountability policies and practices. This global phenomenon (Holloway et al. 2017; Verger and Parcerisa 2017) has relied heavily on the numerical measures of 'teacher quality', as various forms of standardised achievement tests grow in prominence. Large-scale international achievement tests, such the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA), as well as national (e.g. NAPLAN in Australia) and subnational tests (e.g. state-

✉ William C. Smith
w.smith@ed.ac.uk

[1] Moray House School of Education and Sport, University of Edinburgh, Edinburgh, UK

[2] Research for Educational Impact (REDI) Centre, Deakin University, Melbourne, Australia

level tests in the USA), have helped facilitate the incorporation of student test scores into accountability systems around the world. While most of these standardised tests were never designed or intended to be used for measuring teacher quality or effectiveness, it is becoming increasingly common for schools to incorporate student test scores in their teacher-level appraisal/evaluation systems. Indeed, Smith and Kubacka (2017) analysed international data from the OECD's Teaching and Learning International Survey (TALIS) and found that nearly all teachers (i.e. 97%) reported that their appraisals included some form of student test scores. While multiple measures of teacher performance are typically included in these systems, overall, teachers reported that student test scores have been increasingly prioritized in terms of appraisal focus, and thus supplanted other forms of more meaningful feedback (e.g. teacher portfolios, observations). According to leading organisations, such as the OECD (2014) and the American Education Research Association (AERA 2015), multiple measures of teacher performance are necessary for achieving the fair and valid appraisal/evaluation systems deemed so important.

Unfortunately, the disproportionate emphasis on student test scores has led to the production of what some have identified as a 'testing culture', where the professional identities and work of teachers are being fundamentally changed. This transformation is exacerbated when high stakes are attached to appraisal outcomes (Certo 2006; Larsen 2005), which is also increasingly common across most education systems worldwide (Smith and Kubacka 2017). Past literature using the lens of organizational theory and sociological institutionalism has highlighted the importance of environmental factors, such as occupational stress (Xiaofu and Qiwen 2007), work pressure, and practical support (Aldridge and Fraser 2016), on teacher satisfaction. Within the school climate, interpersonal relationships are also important (Price 2012). Grayson and Alvarez (2008) identified poor relationships between teachers and their principal as one of main factors predicting teachers feeling of depersonalization and cynicism in their job. Furthermore, expectations for teachers are shaped by emerging institutional norms (Booher-Jennings 2005; Smith 2016) that lay out appropriate scripts for behaviour. Increasingly, these require teachers to embrace the preparation, application, and interpretation of student test scores (Holloway 2019). Shaping the experiences of teachers, the school testing culture reinforces these emerging norms, while permeating the teaching and learning environment, and influencing interpersonal relationships.

In this paper, we seek to investigate the relationship between the school testing culture and teacher satisfaction, with a particular focus on how teacher appraisals may moderate the relationship. Specifically, we are investigating whether teacher appraisals can provide an explanation for the reported relationship between test-based accountability and teacher satisfaction. In the following section, we start with an overview of the literature, focusing on the varied ways that education systems have incorporated student test scores into teacher appraisal, and then we move towards a more specific focus on how these systems have affected teacher satisfaction.

1 Use of student test scores in teacher accountability

Test-based accountability, or testing for accountability (Smith 2014), is present when student test scores are used as one input to hold teachers or schools accountable. For

teachers, this often comes in the form of performance-based pay, where the results of student test scores influence whether the teacher continues in their current position and what their cumulative income equals. Although the current push originated in the USA and UK, this type of test-based accountability has expanded to countries around the globe (UNESCO 2017). For instance, in Portugal, salary scales were redesigned in 2007 to include student test scores, and in Chile up to 30% of teachers' salaries may be based on student test scores (Barnes et al. 2016).

The effects of test-based accountability have been studied in a variety of ways, most of which have focused on the general effects of high-stakes accountability, such as decreased teacher morale (Certo 2006; Larsen 2005), limited pedagogical approaches (e.g. narrowing of the curriculum, teaching to the test; Polesel et al. 2014; Warren and Ward 2018), and other intended and unintended consequences that have resulted from increased testing programmes (e.g. policy responses to PISA test results, see Breakspear 2014 for a review). Another large area of research focus has been on the measurement issues associated with using student test scores to measure teacher effectiveness (Amrein-Beardsley 2014; Hanushek and Rivkin 2010; Rothstein 2010). This body of research has primarily stemmed from the USA, as the USA has developed the most sophisticated method for directly linking student test scores to teacher effects via the student growth model (SGM) or value-added model (VAM). VAMs are statistical tools designed to capture and compare the predicted and real effects that individual teachers have had on their students' annual, standardized achievement tests. While there are technical differences between the SGM and VAM methodologies, these differences are irrelevant for the present study; therefore, for the sake of clarity, the term 'VAM' will be used throughout the rest of the paper to mean any form of student growth or value-added model.

These models have been extensively covered in the literature, with a particular focus on the methodological and logistical issues related to VAM measurement and use, such as studies on the validity, reliability, bias, and fairness of VAM-based evaluation and policy (Ballou and Springer 2015; Koedel and Betts 2011; Moore Johnson 2015; Rothstein 2010). There have been several reviews of these issues, from varied disciplines, see Amrein-Beardsley and Holloway (2017) for a review from an educational research perspective, Koedel et al. (2015) for an economic perspective, and Darling-Hammond (2015), Everson (2017), and the AERA official statement (AERA 2015) for general reviews of VAMs and VAM use. Together, these reviews demonstrate that VAMs are more sophisticated than previously-used status models—or models designed to measure student proficiency at a given time—as they measure growth in learning *over time* and are able to theoretically mitigate the effects of extraneous variables, such as socioeconomic status, English language status, and prior testing performance. However, the vast conditions that must first be met to guarantee that VAMs are valid, reliable, unbiased, and fair for teacher evaluation are nearly impossible in actual practice. As described in the AERA Statement (2015):

> Even if all of the technical requirements…are met, the validity of inferences from VAM scores depends on the ability to isolate the contributions of teachers and leaders to student learning from the contributions of other factors not under their control. This is very difficult, not only because of data limitations but also

> because of the highly nonrandom sorting of students and teachers into schools and classes within schools (p. 449).

Given these challenges, the general consensus is that VAMs should not be used for high-stakes purposes, though this warning has had little effect on most US states' adoption and use of VAMs in their high-stakes teacher evaluation systems (Collins and Amrein-Beardsley 2014; Close et al. 2019). While the global trend appears to be increased use of test-based accountability, caution in implementing VAMs for accountability purposes is still exercised in many countries. One notable exception is England, which has used various iterations of VAMs for school- and system-level accountability (Sørensen 2016).

This particular dimension of test-based teacher accountability is important for the current paper because it underlines the potential problems associated with the use of student test scores in teacher appraisals. In the USA specifically, and in the UK to some extent, the ways in which test scores have been used for high-stakes accountability purposes have led to problems with trust, utility, and satisfaction (Collins 2014; Garver 2019; Pizmony-Levy and Woolsey 2017). As mentioned previously, this has impacted teachers' interpersonal relationships (between colleagues and teachers and their supervisors), as well as made it difficult for teachers to see the value in the test scores for informing their instruction. As we will revisit throughout the paper, there are multiple ways that test scores can be used in appraisals, which might have bearing on whether teachers see such uses as beneficial or not. Worth considering here is how test-based accountability interacts with the 'global testing culture', which we describe next.

2 The global testing culture and its influence on teachers

The incorporation of test scores into teacher accountability schemes and the underlying belief that student test scores represent an objective, accurate account of student learning reflects a larger global testing culture (Smith 2016). The global testing culture is based on the assumptions of positivism and individualism. In agreeing with these assumptions, there is an almost unconscious belief that quantitative measures, such as test scores, represent the reality of the situation and that the outcomes of education are the result of individual actions and are not influenced by larger societal context or family circumstances. Based on sociological institutionalism, within this culture, behavioural expectations are laid out, including that teachers do everything in their power to help students succeed on the test.

The consistent pressure to improve test scores contributes to reshaping the 'possibilities by which the teaching profession, and teaching professionals, can be known and valued, and the ways that teachers can ultimately be and associate themselves in relation to their work' (Lewis and Holloway 2019, p. 48). Muller and Boutte (2019) further deconstruct the global testing culture by using the work of Paulo Friere to draw equivalences between standardized testing and the oppression of teachers. The divide and conquer dimension of oppression is clearly seen in past research that points to teachers blaming those in earlier grades for inadequately preparing students (Wiggins and Tymms 2000) and concerns that teachers will be stigmatized for not buying into the school's focus on student test scores (Booher-Jennings 2005).

A handful of studies (Holloway and Brass 2018; Perryman and Calvert 2019; Warren and Ward 2018) have explored how the cultural expectations of teachers, and the prevailing testing culture, are associated with an increase in teacher workload, and, consequently, work-related pressure, personal stress, and decreased job satisfaction. Perryman and Calvert (2019) have linked high-stakes accountability to excessive burnout and teacher turnover, arguing that their participants illustrated 'a discourse of disappointment, the reality of teaching being worse than expected, and the nature (rather than the quantity) of the workload, linked to notions of performativity and accountability, being a crucial factor' (p. 2) for why teachers were leaving the profession. Similarly, Garver (2019), who conducted an in-depth ethnographic study of a US middle school's use of a test-based teacher evaluation system, found that teachers experienced feelings of anxiety, distrust, and vulnerability. Wronowski and Urick (2019) found that although stress and worry were associated with the intent to leave their position, the factors only predicted actual departure for teachers frustrated by the accountability system.

Bringing together the statistical issues with using student test scores in teacher accountability, with the creeping pressures that are often associated with such systems, we argue that the testing culture is producing an environment where teacher satisfaction is potentially compromised. The relationship between satisfaction and appraisal has been studied in different contexts, and we see our study as extending this literature in important ways. First, though, we identify some of the studies that have explored similar questions.

3 Teacher appraisals and teacher satisfaction

Teacher appraisals have become the dominant tool for administering the accountability of teachers. Although initially separate from summative teacher evaluations, the inclusion of high-stakes and links to student test scores (Murphy et al. 2013; Xu et al. 2016) have made teacher evaluations and teacher appraisals practically indistinguishable (Smith and Kubacka 2017). Past research suggests that teacher appraisals, and how appraisals are experienced by teachers, are an artefact of the school climate and can impact individual job satisfaction. Past studies that have examined the role of teacher appraisals/evaluations on satisfaction have focused on general perceptions that the process was fair or inclusive (Brezicha et al. 2019). Ford et al. (2018) found that when the teachers viewed the evaluations as being part of a supportive process, and when the evaluations led to meaningful changes in their practice, teachers were more likely to report feelings of satisfaction. The authors emphasized that the utility of the evaluation was important for teachers to feel satisfied with their work and with their profession. In China, Liu et al. (2018) found that teachers who believed their evaluation to be inaccurate or subjective were more likely to have lower levels of teacher satisfaction.

What has received less empirical attention is the perspective of teachers on the use of student test scores in their appraisals/evaluations. In their international, large-scale study of TALIS data, Smith and Kubacka (2017) found that the overemphasis of student test scores in teacher appraisals was related to increased perceptions of the appraisal being an administrative task that carries little relevance for classroom practice. This result is similar to what other studies from the USA have found. Collins (2014) surveyed the teachers of Houston Independent School District, which is known for

having one of the strongest high-stakes teacher evaluation systems in the USA, about their experiences with their VAM-based teacher evaluation. One of the most prominent things she found was that teachers perceived little to no utility associated with their VAM scores or reports. The teachers claimed the reports to be too vague or unclear to produce any meaningful guidance for classroom practice. In fact, 'almost 60% of the teachers in this study reported that they do not use their SAS EVAAS® data for formative purposes whatsoever' (Collins 2014, p. 22). The participants also reported that VAM does not improve working conditions or enhance the school environment. Pizmony-Levy and Woolsey (2017) found similar results in their survey research with New Jersey teachers about their high-stakes teacher evaluation system. Their participants noted effects on classroom practice. They felt the emphasis on test scores forced them to teach to the test and remove non-tested content from their lessons. They also expressed concerns about the validity and fairness of evaluating teachers on student achievement scores.

Similarly, Hewitt (2015) looked at teachers' perceptions of a teacher evaluation system that included VAM scores in North Carolina. She found that, amongst other things, such systems had a profound impact on levels of stress, pressure, and anxiety. She also noted that a majority of teachers did not fully understand VAM or how to incorporate VAM outputs into their decisions about how to improve their practice. Overall, her participants reported feeling sceptical about the utility, fairness, or accuracy of VAM.

These issues point to potential problems principals must consider when thinking of incorporating and emphasizing student test scores in teacher appraisals. Broadly, when schools use test scores and appraisals in formative ways, there is a greater chance that teachers appreciate the feedback as a useful data point. Otherwise, when the scores are used in high-stakes and summative ways to label teacher quality and determine personnel decisions, teachers seem to feel greater pressure and more frustration. Building from these previous findings, we sought to more explicitly investigate the relationship between teacher satisfaction and the use of student test scores in teacher appraisal.

4 This study

Teacher appraisals/evaluations represent a relatively unexamined pathway that could help explain the relationship between test-based accountability and teacher satisfaction. Results from Burns and Darling-Hammond (2014), suggesting that low levels of feedback utility are associated with reduced teacher satisfaction, hint at this connection. In a rare study, Lacierno-Paquet and colleagues (Lacireno-Paquet et al. 2016) found that teachers in the USA were 2.5 times less likely to be satisfied with the evaluation process when it included student test scores.

This study further explores whether teacher appraisals are one potential path to explain the reported relationship between test-based accountability and teacher satisfaction. The two primary research questions include:

1. What is the relationship between school testing culture and teacher satisfaction?
2. Is this relationship mediated by feedback received on teacher appraisals?

5 Data and methods

Data from the 2013 Teaching and Learning International Survey (TALIS) was used in this study. TALIS is administered by the OECD and includes a cross-national survey of teachers and school environments, focusing on lower secondary education. As the largest international survey of teachers, TALIS has been used extensively to research factors associated with teacher satisfaction at the global (OECD 2016), regional (for Eastern Europe example see Smith and Persson 2016), and national level (for USA example, see Ford et al. 2018; for Spain, see Gil-Flores 2017). Essential for this study, TALIS teacher and principal questionnaires include information capturing the primary independent variable, school testing culture, the dependent variable, teacher satisfaction, and information on the proposed mediation path, teachers' perspectives on their appraisal. This study draws on information from the initial wave of participants from the 2013 TALIS, in which 33 countries or participating economies completed teacher and principal questionnaires. The stratified samples are nationally representative, with teachers nested in schools. Following Dicke et al. (2020) and Sun and Xia (2018), country surveys are combined into one pooled sample. Cases missing values on teacher satisfaction were dropped and missing data for the remaining analysis was dealt with through listwise deletion, producing a functional pooled sample for the final model of 66,592 teachers.

5.1 Dependent variable

The TALIS teacher questionnaire contains information on both general satisfaction with the profession and specific satisfaction with the school. Given that the school testing culture is unique to the school environment in which the teacher is employed, this study is limited to the latter. Following the approach of Smith and Persson (2016), teacher responses to three statements are included in the final job satisfaction variable: (1) 'I would like to change to another school if possible', (2) 'I enjoy working at this school', and (3) 'I would recommend my school as a good place to work'. Statements are reverse coded as needed so that a score of 1 indicates satisfaction with the current place of employment. The aggregated variable has a range of 0 (not satisfied at all) to 3 (satisfaction indicated in all three statements). For path analysis, the teacher satisfaction variable is standardized with a mean of zero and a standard deviation of one. Coefficients for each pathway are then interpreted as a one-unit increase in the corresponding variable is associated with a change in teacher satisfaction, relative to the standard deviation (i.e. a coefficient of 0.2 suggests a one-unit increase in the corresponding variables is associated with a 0.2 standard deviation increase in teacher satisfaction).

5.2 Predictor variables

Two variables are used to measure the presence of a school testing culture. First, a dichotomous variable captures whether student test scores are included in the teacher's appraisal (1 = yes; 0 = no). However, given recent research indicating that over 95% of teacher appraisals include student test scores (Smith and Kubacka 2017), a second measure is included for a more fine-grained analysis. To capture the extent to which teachers are held responsible for student test scores, principal responses to the statement

'I took actions to ensure that teachers feel responsible for their students' learning outcomes' are included. The teachers are responsible variable ranges from 0 (principal never or rarely took action) to 3 (principal very often took action). An independent samples *t* test identified a significant relationship between the two school testing culture variables ($t = -24.417$, $df = 91{,}088$, $p < .01$) in the expected direction, suggesting the two variables capture a similar construct. While many structural equation models include a large set of variables in their measurement models, given the strength of the relationship between the school testing culture variables and the suggestions by Hayduk and Littvay (2012) when considering whether many or few indicators should be included in structural equation modeling that 'using the few best indicators...encourages development of theoretically sophisticated models' (p. 1), we are confident that these variables capture, at a minimum, a key part of the pressure felt by teachers in test-based accountability systems.

The mediation pathway through teacher appraisal feedback consists of two variables. To capture whether student test scores are emphasized in appraisal feedback, this study follows that of Smith and Kubacka (2017). To identify which parts of teacher appraisal are emphasized, teachers are asked to evaluate eleven potential areas of feedback. Each area is coded on a Likert scale from 0 (not considered at all when feedback is received) to 3 (considered with high importance). The relative emphasis score is then calculated by taking the difference between the score related to student achievement and the mean score of the ten other potential areas of emphasis (see Eq. 1). Values over 0, therefore, indicate that student test scores were relatively more emphasized in teacher appraisal feedback, in comparison to the average score of other areas.

$$\text{Relative emphasis} = \text{Emphasis}_{\text{test score}} - \text{Mean}\left(\text{Emphasis}_{\text{all other factors}}\right) \quad (1)$$

The second variable in the teacher appraisal pathway captures the extent teachers feel the feedback they received had a direct, positive effect on their job satisfaction. Teacher responses ranged from 0 (feedback had no positive change on my job satisfaction) to 3 (feedback had a largely positive change on my job satisfaction).

5.3 Control variables

Four control variables are included at the teacher level: sex, age, years of education, and education level. Teacher's sex is coded 1 for female and 0 for male. Years of experience is a continuous variable that captures the years the teacher has spent at their current school. Age is a continuous variable that captures the age of the teacher. Education level is treated as an ordinal variable and coded from 1 for below ISCED level 5 (completion of secondary or below) to 4 for ISCED level 6 or above (completion of bachelors' degree or above).

5.4 Analytic strategy

Descriptive statistics were calculated to provide an initial illustration of all key variables. This was followed by a preliminary bivariate analysis to evaluate the initial association between independent and mediating variables with teacher satisfaction.

Independent *t* tests were performed to examine the mean difference in satisfaction by whether test scores were included as a component in the teacher appraisal. Pearson correlation coefficients are calculated to compare all continuous variables.

Multi-level structural equation modeling (SEM) is employed for the primary analysis. The approach in this study is similar to Sun and Xia (2018) who draw on a pooled sample of teachers across all participating countries in the 2013 TALIS and apply multi-level SEM to predict the relationship between distributed leadership and teacher job satisfaction. Multi-level SEM is appropriate for this analysis as it takes into consideration the nested characteristic of the data—with teachers nested in schools nested in countries (Hox 2013). Additionally, SEM allows us to distinguish between (a) the direct effect of school testing culture on teacher satisfaction and (b) the indirect effects of school testing culture on teacher satisfaction through teacher appraisal feedback (Schumacker and Lomax 2004). All results are presented graphically to ease interpretation (Hox 2010) and computed using the gsem option in Stata v14 (Huber 2013).

The full model (shown in Fig. 1) is completed through three additive steps. The baseline model predicts the direct effect of school testing culture on teacher satisfaction (paths A and B) and includes teacher-level control variables (path X). The second model adds the impact of teachers' positive perception of appraisal feedback (path G) to evaluate whether teacher appraisal feedback is a potential mediating mechanism. Model 3 completes the full multi-level model by adding test score emphasis in appraisal feedback and school and country-level error terms. To aid convergence, covariance of exogenous upper-level latent variables (school and country) are constrained (Huber 2013)[1] and not displayed in the results. Each model assumes exogenous variables are correlated. Error terms for each endogenous variable are included for each model and provided in the notes for each figure. The final total effect of school testing culture on teacher satisfaction is calculated as follows:

$$\begin{aligned} \text{Direct effect} &= \text{path A} + \text{path B} \\ \text{Indirect effect} &= (\text{path C}x\text{path E}x\text{path G}) + (\text{path D}x\text{path E}x\text{path G}) \\ &\quad + (\text{path C}x\text{path F}) + (\text{path D}x\text{path F}) \\ \text{Overall effect} &= \text{direct effect} + \text{indirect effect} \end{aligned} \tag{2}$$

5.5 Goodness of fit

As teacher satisfaction is the primary endogenous variable of interest, an r-squared or equation level goodness of fit was calculated to evaluate the precision of each model in predicting teacher satisfaction. While the gsem command in Stata provides flexibility in allowing the inclusion of multiple levels, it has limited options for model fit. Given the minimal differences in output using gsem and alternative approaches (see footnote 1), a single-level model was assumed to calculate the r-squared. Included in the notes for

[1] An alternative two-level approach using the vce (cluster) command at the school level to relax the assumption of independence of observations and adjust standard errors did not substantially change the value of coefficients or the level of significance of results.

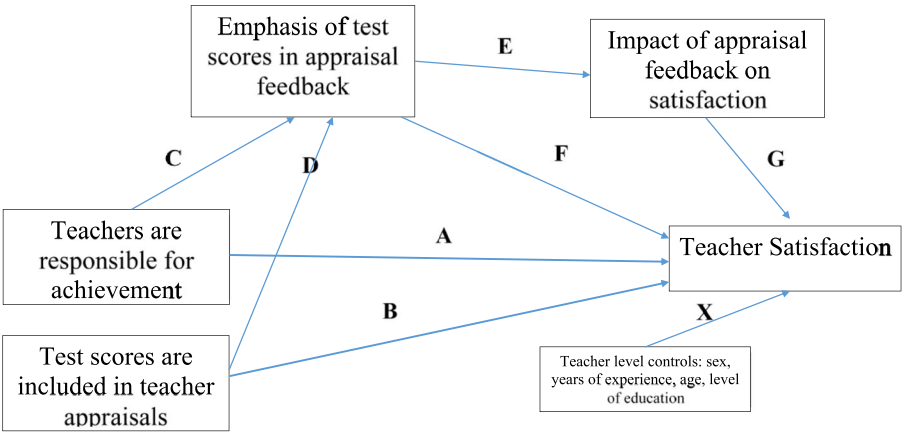

Fig. 1 Full path analysis

each figure, the r-squared illustrates the amount of variance in teacher satisfaction captured by each model.

6 Results

6.1 Preliminary analysis

Descriptive statistics are found in Table 1. In the pooled sample, teachers report relatively high levels of satisfaction with their current place of employment (mean = 2.528, SD = .849). Additionally, a fairly substantial school testing culture appears to be the norm. The mean value (1.976, SD = .719) suggests that the average principal often takes actions to ensure teachers know they are responsible for student outcomes and nearly 97% of teachers have student test scores incorporated into their teacher appraisal. Finally, for the teacher appraisal mediation pathway, the emphasis on student test scores is higher than the mean of other potential pieces of feedback (mean = .363, SD = .643) and teachers, on average, report feedback resulting in a small to moderate change in their satisfaction (mean = 1.796, SD = .987).

Bivariate analysis revealed few initial relationships between school testing culture variables or teacher appraisal feedback variables and teacher satisfaction. An independent samples t test found no significant relationship between including student test score in the appraisal and teacher satisfaction ($t = 1.454$, $df = 86{,}853$, $p = .93$). Amongst Pearson correlation coefficients, no relationships are significant and the only correlation above ± .2 is the correlation between whether the teacher believed the appraisal feedback had a positive impact on their satisfaction and their overall satisfaction level ($r = .229$).

6.2 Primary analysis

Our baseline path analysis (Fig. 2) reveals coefficients in the expected direction, but neither ensuring teachers know they are responsible for student achievement ($\beta =$

Table 1 Descriptive statistics

Variable	Observations	Mean	SD	Minimum	Maximum
Teacher satisfaction	101,775	2.528	.849	0	3
Teacher satisfaction (standardized)	101,775	0	1	−2.978	.557
Satisfaction in appraisal feedback	77,436	1.796	.987	0	3
Student test scores emphasized in appraisal feedback	83,515	.363	.643	−2.727	2.727
Student test scores included in teacher appraisal	91,601	96.8%			
Principal ensures teachers know they are responsible for student achievement	101,423	1.976	.719	0	3
Female	107,652	68.1%			
Years of experience	100,860	16.058	10.348	0	58
Age	107,563	42.530	10.520	18	76
Education level	106,176	2.907	.395	1	4

$-.003$, $p = .51$) or the inclusion of test scores in appraisals ($\beta = -.032$, $p = .11$) are significantly related to teacher satisfaction, after controlling for teacher demographic variables. Female ($\beta = .058$, $p < .01$) and older teachers ($\beta = .004$, $p < .01$) tend to be more satisfied. In addition, those with lower levels of education are more satisfied ($\beta = -.096$, $p < .01$). Teacher years of education are not significantly related to their satisfaction. Direction, magnitude, and significance levels of all control variables remained largely consistent across all models.

In the second model, we add part of the teacher appraisal feedback pathway to examine the potential benefits of feedback on teacher satisfaction. The results (Fig. 3) illustrate that teachers that view their feedback as positively impacting their satisfaction are more likely to report higher levels of overall satisfaction ($\beta = .235$, $p < .01$). The increased magnitude of school testing culture coefficients and change from non-significant to significantly related to teacher satisfaction suggest that one avenue the overall school climate is influencing teacher satisfaction is through their individual interaction with appraisals and appraisal feedback.

The full model (Fig. 4) completes the hypothesized mediation pathway by including whether student test scores are emphasized in appraisal feedback and provides, marginally, the best fit for predicting teacher satisfaction (r-squared = .060). Here, it is clear that school testing culture has both direct effects on teacher satisfaction and indirect effects on teacher satisfaction through the teacher appraisal feedback pathway. In the full model, which controls for teacher demographics, the inclusion of test scores in teacher appraisals is directly related to a .103 ($p < .01$) standard deviation reduction in teacher satisfaction. Furthermore, a one-unit increase in principals ensuring teachers are responsible for student outcomes is associated with a .023 ($p < .01$) standard deviation decrease in teacher satisfaction.

Teachers in school testing cultures are more likely to have student test scores emphasized in their appraisal feedback (ensuring teachers are responsible, $\beta = .071$, $p < .01$; test scores in appraisal, $\beta = .100$, $p < .01$). Emphasizing test scores above other areas in teacher appraisal feedback is associated with a .010 ($p < .10$) standard deviation decrease in satisfaction and reduces the likelihood that the teacher would state their

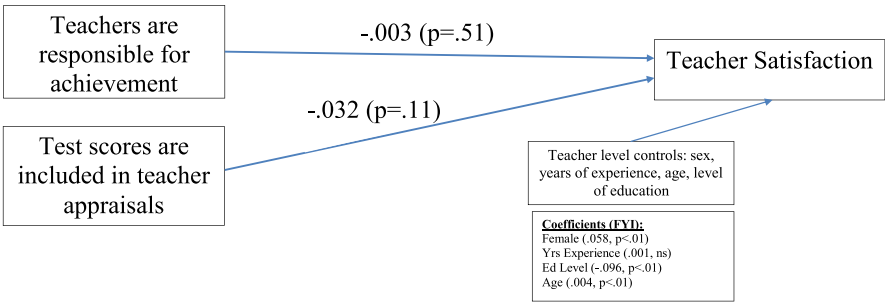

Fig. 2 Baseline analysis—Direct effect of school testing culture on teacher satisfaction ($n = 81{,}361$). *Notes*: r-squared for teacher satisfaction = .004. Measurement error for teacher satisfaction ($\beta = 1.010$)

feedback positively impacts their satisfaction ($\beta = -.183$, $p < .01$) reducing any potential benefits from the teacher appraisal pathway.

Figure 5 provides the total unstandardized effect of school testing culture on teacher satisfaction by intensity of testing culture. Total effects are calculated using Eq. 2 (see above). The figure predicts teacher satisfaction by setting all control variables to the mean and assuming the teacher's sex is female. The first bar indicates that a female teacher of average age, years of experience, and education level would have a satisfaction score of 2.16. As the school testing culture intensifies the predicted satisfaction score decreases from 2.16 to 1.87. Of the net drop of 0.29 points, the direct effect of school testing culture accounts for 59% of the reduction while the indirect effect through teacher appraisal feedback accounts for approximately 41%.

In terms of standard deviation, the difference between no school testing culture and the most intense testing culture is 0.35. Although an effect size of 0.35 standard deviations would be considered between small and medium by Cohen (1969), it would be considered 'substantively important' (SD > .25) by the What Works Clearinghouse (WWC 2014, p. 23). While teacher satisfaction and student achievement are distinct dependent variables making it difficult to compare, the total effect size is also larger

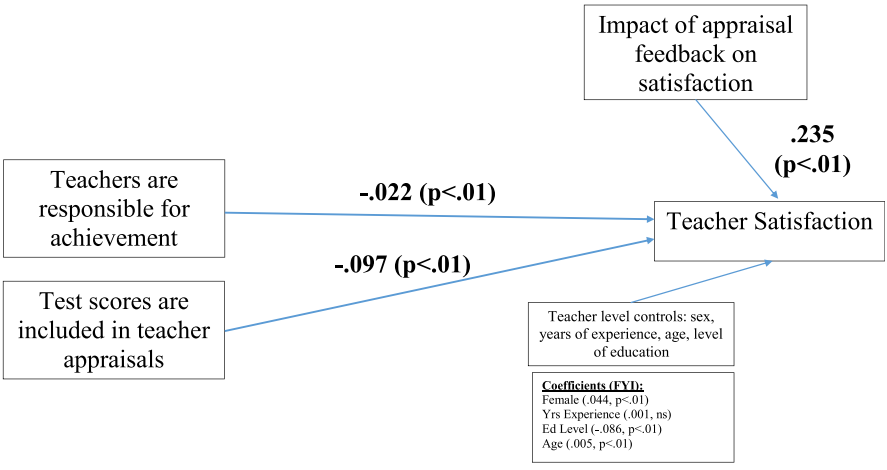

Fig. 3 Model 2—Exploring the Potential benefits of appraisal feedback ($n = 70{,}613$). *Notes*: r-squared for teacher satisfaction = .059. Measurement error for teacher satisfaction ($\beta = .921$)

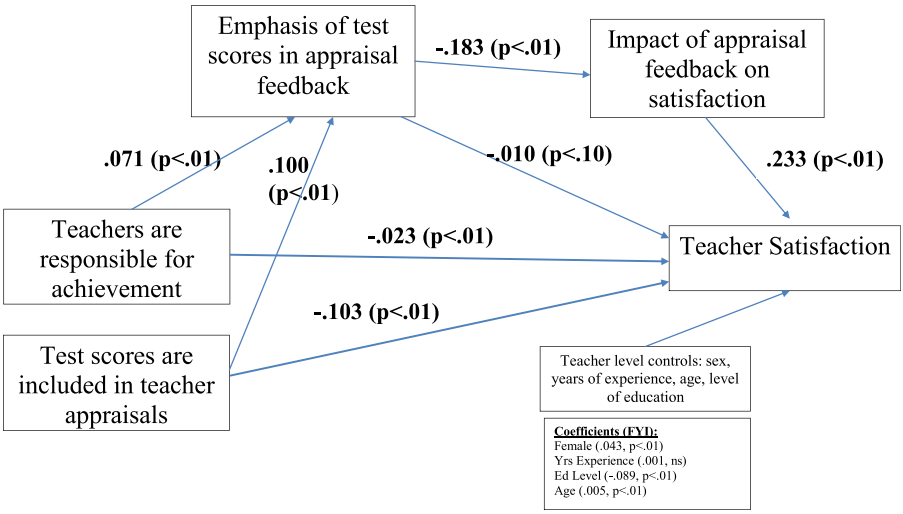

Fig. 4 Full model—Direct and indirect effect of school testing culture on teacher satisfaction ($n = 66{,}592$). *Notes*: Full model controls for school- and country-level effects. Error terms for both levels are regressed on teacher satisfaction and constrained to 1 (see "Data and Methods" section for more information). R-squared for teacher satisfaction = .060. Measurement error for teacher satisfaction ($\beta = .924$). Measurement error for emphasis on test scores ($\beta = .409$). Measurement error for the impact of appraisal on satisfaction ($\beta = .958$)

than the average effect size (SD = .28) on student achievement across 124 random trials (Lipsey et al. 2012) or reported effects of individualized tutoring (SD = .23, Cook et al. 2015) or universal free school lunch (SD = .09, Frisvold 2015) on math test score.

7 Concluding discussion

The global increase in the use of student test scores to hold teachers' accountability has seen a rush of scholars working to understand how the trend has impacted policy and practice. There has been a great deal of empirical studies that have looked specifically at the effects of teachers, ranging from large-scale survey studies (e.g. Collins 2014; Pizmony-Levy and Woolsey 2017) to small-scale qualitative studies (e.g. Garver 2019; Hardy 2018; Perryman 2009). There have also been a number of studies that have looked at measurement issues related to using student test scores to measure teacher quality (see Darling-Hammond 2015 for a review). While these studies focus on a wide range of topics and contexts, what most of them have in common is that their findings and conclusions indicate a troubled relationship between the use of student test scores in teacher accountability and how teachers feel about their practice and work-place conditions. This is particularly pronounced in systems where the stakes for teachers are high, such as in the USA and the UK. Not only are researchers finding that teachers are significantly modifying their practice in response to these sorts of accountability systems, which has been a long-standing concern about testing more generally (Amrein and Berliner 2002; Nichols and Berliner 2007; Ravitch 2016), but it has also begun to influence the way teachers feel about their work and their professional identity (Brass and Holloway 2019; Garver 2019; Perryman 2009).

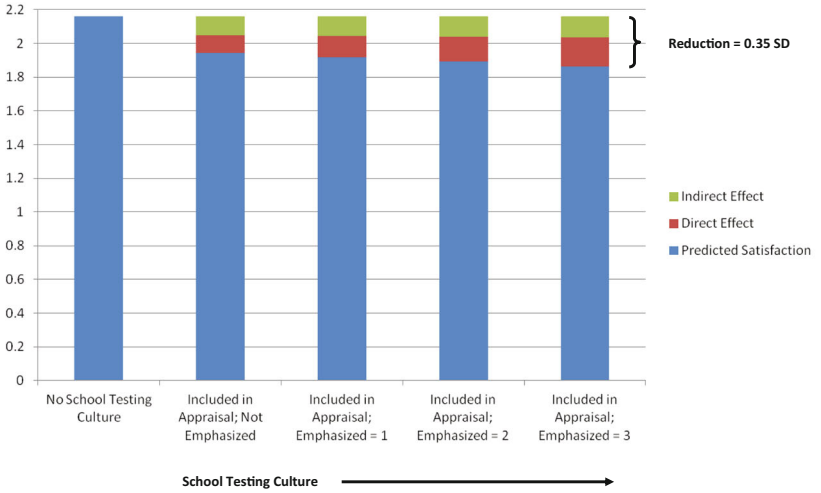

Fig. 5 Total unstandardized effect of school testing culture on teacher satisfaction

Our research extends past studies, providing a more nuanced view of the relationship between test-based accountability and teacher satisfaction. First, our results support past research that draws a direct line between an increased focus on student testing and decreased teacher satisfaction. Second, it is clear that the use and perception of teacher appraisals have an important role to play regarding teacher satisfaction. Our findings suggest that teacher appraisals are not pre-destined to have a negative impact on overall teacher satisfaction. In fact, when not emphasizing student test scores, teacher appraisals can boost teacher satisfaction. This is an important finding because it aligns with what many scholars and education leaders have argued for a while now—that appraisal and accountability are not in and of themselves *bad* for teachers and schooling (Darling-Hammond 2014; Darling-Hammond et al. 2012). However, what many of these researchers and their more critical colleagues (e.g. Perryman 2009; Perryman and Calvert 2019; Holloway and Brass 2018; Hursh 2007; Lipman 2004) have argued is that it is the pervasiveness of the testing culture, and the overemphasis on student test scores in teacher appraisals, that is having a profoundly negative effect on teachers and their practice. Our current study provides another layer to our understanding about this phenomenon. Once established, the school testing culture appears to both directly relate to teacher dissatisfaction and reduce the potential benefits of teacher appraisals by indirectly and negatively influencing teacher satisfaction by warping the teacher appraisal process. This is similar to what Ford et al. (2018) found in their study—that the degree to which teachers view the utility of appraisal and feedback is closely related to their satisfaction. Ultimately, as the testing culture intensifies, teachers' overall satisfaction decreases.

These findings are not without limitations. First, while the results make clear that the common global trends associated with the increased emphasis on student test scores appear to be reflected in the school testing culture, potentially harming teacher satisfaction, the school testing culture and school appraisal mediation pathway only capture a limited level of variance in teacher satisfaction (r-squared = .060). While school testing culture may play a small, but important role, many more factors such as self-

efficacy (Kasalak and Dağyar 2020), teachers involvement in decision-making (Smith and Persson 2016) and distributed leadership within the school (Sun and Xia 2018) should be considered to get a full understanding of teacher satisfaction. Additionally, TALIS data includes information from both the principal (i.e. school testing variables) and teacher (i.e. appraisal and teacher satisfaction variables). Past research has suggested that principals and teachers have different perceptions related to school climate (Casteel 1994), including whether or not teachers are satisfied (Dicke et al. 2020). Teacher self-reported satisfaction is used in this study and we believe appropriately captures the affective relationship between the teacher and the school. Still, the perceptions of teachers may not represent those of other actors within the school.

Given our results, we strongly urge school leaders to consider carefully the ways they use student test scores, as well as appraisals more broadly. Situating student test scores amongst multiple indicators can partially mitigate, but is unlikely to remove, the pressure felt by the school testing culture. Even appraisals that include multiple metrics, appearing more holistic, often end with principals emphasizing student test scores above other components (Smith and Kubacka 2017). School leaders need to be cautious when including student test scores. If used, they need to be treated as a source of formative feedback, rather than as a summative judgment about the teacher's quality or ability. This has serious implications for policy and practice, which are described next.

7.1 Implications for policy, practice, and future research

While this study adds to our overall understanding about the impact of appraisal on teacher satisfaction, it also prompts further questions about the utility of evaluation and the use of student test scores in holding teachers accountable. This is especially important if we consider this in line with current trends that prioritize numerical data for making sense of school and teacher quality more broadly. In this way, we argue that there is a critical need for school leaders to grapple with the various approaches to appraisal, as well as how appraisals and student test scores might be used in more formative ways. One way this could be achieved is through training and ongoing professional development for principals and other school leaders. Training on topics such as data literacy, assessment, and accountability might incorporate sections on how to use such techniques in ways that support teacher development. This might help leaders navigate the complicated relationship between being not only evidence-driven but also supportive of teacher wellbeing and growth.

However, training and professional development can only achieve so much if policies continue to prioritize high-stakes testing as a means for identifying school and teacher quality. Principals are left with little discretionary space if there are policies that require them to use test scores and appraisals for making personnel decisions (e.g. promotion, performance-based pay). This is where countries like the USA and the UK might benefit from considering how other countries are taking a more holistic and formative approach to test score use.

We acknowledge that 'satisfaction' is a difficult construct to measure, and knowing specifically how satisfaction might affect teacher practice or subsequent decisions about whether to remain in the classroom is hard to say at this time. There is a growing criticism of the testing culture, coming from a variety of perspectives (e.g. from

governments to teacher organizations; see Strauss 2012a, b), with a particular warning about how these conditions are creating a dire environment for teachers (Perryman and Calvert 2019). For example, concerns about teacher shortages, decreased interest in becoming teachers amongst young people, and teachers' personal and professional wellbeing, have all been highlighted in calls for reduction to the widespread testing culture. Therefore, we need more research about how the testing culture is changing the make-up of the profession (e.g. are teacher shortages related to increased accountability and testing?), especially with regard to student test scores. We add the findings of the current study to these considerations by urging school leaders and policymakers to weigh critically the purpose and consequences of test-based appraisals. As we have shown, it is possible for teachers to have high levels of satisfaction within schools that use teacher appraisals. However, this relationship changes as the intensity of the testing culture increases, which signals time for reflection on how the pervasiveness of the testing culture can be challenged.

References

Aldridge, J. M., & Fraser, B. J. (2016). Teachers' views of their school climate and its relationship with teacher self-efficacy and job satisfaction. *Learning Environments Research, 19*, 291–307.

American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448–452.

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

Amrein-Beardsley, A., & Holloway, J. (2017). Value-added models for teacher evaluation and accountability: commonsense assumptions. *Educational Policy, 33*(3), 516–542.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education policy analysis archives, 10*, 18.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77–86.

Barnes, S.-A., Lyonette, C., Atfield, G., & Owen, D. (2016). *Teachers' pay and equality: a literature review – longitudinal research into the impact of changes to teachers' pay on equality in schools in England*. Warwickshire: Warwick Institute for Employment Research.

Booher-Jennings, J. (2005). Below the bubble: 'Educational triage' and the Texas accountability system. *American Education Research Journal, 42*(2), 231–268.

Brass, J., & Holloway, J. (2019). Re-professionalizing teaching: the new professionalism in the United States. *Critical Studies in Education*, 1–18.

Breakspear, S. (2014). How does PISA shape education policy making. In *Why how we measure learning determines what counts in education, CSE Seminar series* (Vol. 240). Melbourne: Centre for Strategic Education.

Brezicha, K. F., Ikoma, S., Park, H., & LeTendre, G. K. (2019). The ownership perception gap: exploring teacher satisfaction and its relationship to teachers' and principals' perception of decision-making opportunities. *International Journal of Leadership in Education*, 1–29.

Burns, D., & Darling-Hammond, L. (2014). *Teaching around the world: what can TALIS tell us*. Stanford: Stanford Center for Opportunity Policy in Education.

Casteel, D. B. (1994). Principal and teacher perceptions of school climate related to value-added assessment and selected school contextual effects in the First Tennessee District. PhD Dissertation. East Tennessee State University.

Certo, J. L. (2006). Beginning teacher concerns in an accountability-based testing environment. *Journal of Research in Childhood Education, 20*(4), 331–349.

Close, K., Amrein-Beardsley, A., & Collins, C. (2019). Mapping America's teacher evaluation plans under ESSA. *Phi Delta Kappan, 101*(2), 22–26.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). New York: Academic Press.

Collins, C. (2014). Houston, we have a problem: teachers find no value in the SAS education value-added assessment system (EVAAS®). *Education Policy Analysis Archives, 22*, 98.

Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: a national overview. *Teachers College Record, 116*(1), 1–32.

Cook, P. J., Dodge, K., Farkas, G., Fryer, R. G., Guryan, J., Ludwig, J., & Mayer, S. (2015). Not too late: improving academic outcomes for disadvantaged youth. Working paper WP-15-01. Northwestern University: Institute for Policy Research.

Darling-Hammond, L. (2014). One piece of the whole: teacher evaluation as part of a comprehensive system for teaching and learning. *American Educator, 38*(1), 4.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher, 44*(2), 132–137.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.

Dicke, T., Marsh, H. W., Parker, P. D., Guo, J., Riley, P., & Waldeyer, J. (2020). Job satisfaction of teachers and their principals in relation to climate and student achievement. *Journal of Educational Psychology, 112*(5), 1061–1073.

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

Ford, T. G., Urick, A., & Wilson, A. S. (2018). Exploring the effect of supportive teacher evaluation experiences on US teachers' job satisfaction. *Education Policy Analysis Archives, 26*, 59.

Frisvold, D. E. (2015). Nutrition and cognitive achievement: an evaluation of the school breakfast program. *Journal of Public Economics, 124*, 91–104.

Garver, R. (2019). Evaluative relationships: teacher accountability and professional culture. *Journal of Education Policy*, 1–25.

Gil-Flores, J. (2017). The role of personal characteristics and school characteristics in explaining teacher job satisfaction. *Revista de Psicodidáctica/Journal of Psychodidactics, 22*(1), 16–22.

Grayson, J. L., & Alvarez, H. K. (2008). School climate factors related to teacher burnout: a mediator model. *Teaching and Teacher Education, 24*, 1349–1363.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267–271.

Hardy, I. (2018). Governing teacher learning: understanding teachers' compliance with and critique of standardization. *Journal of Education Policy, 33*(1), 1–22.

Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology, 12*, 159.

Hewitt, K. K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: undermined intentions and exacerbated inequities. *Education Policy Analysis Archives, 23*(76).

Holloway, J. (2019). Teacher evaluation as an onto-epistemic framework. *British Journal of Sociology of Education, 40*(2), 174–189.

Holloway, J., & Brass, J. (2018). Making accountable teachers: the terrors and pleasures of performativity. *Journal of Education Policy, 33*(3), 361–382.

Holloway, J., Sørensen, T. B., & Verger, A. (2017). Global perspectives on high-stakes teacher accountability policies: an introduction. *Education Policy Analysis Archives, 25*(85), 1–18.

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. *The Oxford handbook of quantitative methods, 2*(1), 281–294.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (Second ed.). New York: Routledge.

Huber, C. (2013). Generalized structure equation modelling using Stata. *Presentation at Italian Stata Users Group Meeting (Florence, Italy)*, November 14–15, 2013.

Hursh, D. (2007). Assessing no child left behind and the rise of neoliberal education policies. *American Educational Research Journal, 44*(3), 493–518.

Johnson, S. M. (2015). Will VAMS reinforce the walls of the egg-crate school? *Educational Researcher, 44*(2), 117–126.

Kasalak, G., & Dağyar, M. (2020). The relationship between teacher self-efficacy and teacher job satisfaction: a meta-analysis of the Teaching and Learning International Survey (TALIS). *Educational Sciences: Theory and Practice, 20*(3), 16–33.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and policy, 6*(1), 18–42.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195.

Lacireno-Paquet, N., Bocala, C., & Bailey, J. (2016). *Relationship between school professional climate and teachers' satisfaction with the evaluation process. (REL 2016–133)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands.

Larsen, M. A. (2005). A critical analysis of teacher evaluation policy trends. *Australian Journal of Education, 49*(3), 292–305.

Lewis, S., & Holloway, J. (2019). Datafying the teaching 'profession': remaking the professional teacher in the image of data. *Cambridge Journal of Education, 49*(1), 35–51.

Lipman, P. (2004). *High stakes education: inequality, globalization, and urban school reform*. London: Routledge.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., et al. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, D.C.: National Center for Special Education Research.

Liu, S., Xu, X., & Stronge, J. (2018). The influences of teachers' perceptions of using student achievement data in evaluation and their self-efficacy on job satisfaction: evidence from China. *Asia Pacific Education Review, 19*, 493–509.

Muller, M., & Boutte, G. S. (2019). A framework for helping teachers interrupt oppression in their classrooms. *Journal for Multicultural Education., 13*, 94–105. https://doi.org/10.1108/JME-09-2017-0052.

Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation. *Educational Researcher, 42*, 349–354.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: how high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

OECD. (2014). *TALIS 2013 results: an international perspective on teaching and learning*. Paris: Organisation for Economic Cooperation and Development.

Perryman, J., & Calvert, G. (2019). What motivates people to teach, and why do they leave? Accountability, performativity and teacher retention. *British Journal of Educational Studies*, 68(1), 3–23.

OECD. (2016). *Supporting teacher professionalism*. Paris: OECD.

Perryman, J. (2009). Inspection and the fabrication of professional and performative processes. *Journal of Education Policy, 24*(5), 611–631.

Pizmony-Levy, O., & Woolsey, A. (2017). Politics of education and teachers' support for high-stakes teacher accountability policies. *Education Policy Analysis Archives, 25*, 87.

Polesel, J., Rice, S., & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: a teacher perspective from Australia. *Journal of Education Policy, 29*(5), 640–657.

Price, H. E. (2012). Principal-teacher interactions: how affective relationships shape principal and teacher attitudes. *Educational Administration Quarterly, 48*(1), 39–85.

Ravitch, D. (2016). *The death and life of the great American school system: how testing and choice are undermining education*. New York: Basic Books.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*(1), 175–214.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. London: Psychology Press.

Smith, W. C. (2014). The global transformation toward testing for accountability. *Education Policy Analysis Archives, 22*(116).

Smith, W. C. (2016). *The global testing culture: shaping education policy, perceptions, and practice*. Oxford: Symposium Books.

Smith, W. C., & Kubacka, K. (2017). The emphasis of student test scores in teacher appraisal systems. *Education Policy Analysis Archives, 25*(86).

Smith, W. C., & Persson, A. M. (2016). Teacher satisfaction in high poverty schools: searching for policy relevant interventions in Estonia, Georgia, and Latvia. *Educational Studies Moscow, 2*, 146–182.

Sørensen, T.B. (2016). Value-added measurement or modelling (VAM). Education international discussion paper. Available at: https://worldsofeducation.org/en/woe_homepage/woe_detail/14860/discussion-paper-value-added-measurement-or-modelling-vam

Strauss, V. (2012a). Moco schools chief calls for three-year moratorium on standardized testing. Washington Post. Available at: https://www.washingtonpost.com/news/answer-sheet/wp/2012/12/10/moco-schools-chief-calls-for-three-year-moratorium-on-standardized-testing/. Accessed 23 Mar 2020.

Strauss, V. (2012b). Texas schools chief calls testing obsession a 'perversion'. Washington Post. Available at: https://www.washingtonpost.com/blogs/answer-sheet/post/texas-schools-chief-calls-testing-obsession-a-perversion/2012/02/05/gIQA5FUWvQ_blog.html. Accessed 23 Mar 2020.

Sun, A., & Xia, J. (2018). Teacher-perceived distributed leadership, teacher self-efficacy and job satisfaction: a multilevel SEM approach using TALIS 2013 data. *International Journal of Educational Research, 92*, 86–97.

UNESCO. (2017). *Accountability in education: meeting our commitments*. Paris: UNESCO.

Verger, A., & Parcerisa, L. (2017). A difficult relationship: accountability policies and teachers—International Evidence and Premises for Future Research. In Akiba, M. & LeTendre, G. K. (eds.), *International handbook of teacher quality and policy* (pp. 241–254). London: Routledge.

Warren, A. N., & Ward, N. A. (2018). 'This is my new normal': teachers' accounts of evaluation policy at local school board meetings. *Journal of Education Policy, 33*(6), 840–860.

WWC (What Works Clearinghouse). (2014). *WWC procedures and standards handbook (Version 3.0)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

Wiggins, A. & Tymms, P. (2000). Dysfunctional effects of public performance indicator systems: a comparison between English and Scottish primary schools. *Paper presented at the European Conference on Educational Research (Edinburgh, UK)*, 20-23 September, 2000.

Wronowski, M. L., & Urick, A. (2019). Examining the relationship of teacher perception of accountability and assessment policies on teacher turnover during NCLB. *Education Policy Analysis Archives, 27*(86).

Xiaofu, P., & Qiwen, Q. (2007). An analysis of the relation between secondary school organizational climate and teacher job satisfaction. *Chinese Education & Society, 40*(5), 65–77.

Xu, X., Grant, L. W., & Ward, T. J. (2016). Validation of a statewide teacher evaluation system. *NASSP Bulletin, 100*(4), 203–222.