Educ Asse Eval Acc (2018) 30:251–283
https://doi.org/10.1007/s11092-018-9280-x

Pointing teachers in the wrong direction: understanding Louisiana elementary teachers' use of *Compass* high-stakes teacher evaluation data

Timothy G. Ford[1]

Received: 3 October 2017 / Accepted: 26 June 2018 / Published online: 11 July 2018
© Springer Nature B.V. 2018

Abstract Spurred by *Race to the Top*, efforts to improve teacher evaluation systems have provided states with an opportunity to get teacher evaluation right. Despite the fact that a core reform area of *Race to the Top* was the use of teacher evaluation to provide on-going and meaningful feedback for instructional decision making, we still know relatively little about how states' responses in this area have led to changes in teachers' use of these sources of data for instructional improvement. Self-determination theory (SDT) and the concept of *functional significance* was utilized as a lens for understanding and explaining patterns of use (or non-use) of Compass-generated evaluation data by teachers over a period of 3 years in a diverse sample of Louisiana elementary schools. The analysis revealed that the majority of teachers exhibited either controlled or amotivated functional orientations to Compass-generated information, and this resulted in low or superficial use for improvement. Perceptions of the validity/utility of teacher evaluation data were critical determinants of use and were multifaceted: In some cases, teachers had concerns about how state and district assessments would harm vulnerable students, while some questioned the credibility and/or fairness of the feedback. These perceptions were compounded by (a) the lack of experience of evaluators in evaluating teachers with more specialized roles in the school, such as special education teachers; (b) a lack of support in terms of training on Compass and its processes; and (c) lack of teacher autonomy in selecting appropriate assessments and targets for Student Learning Target growth.

Keywords Data-driven decision-making · Data use · Teacher motivation · Self-determination theory · Teacher evaluation · Instructional improvement

✉ Timothy G. Ford
tgford@ou.edu

[1] Department of Educational Leadership and Policy Studies, Jeannine Rainbolt College of Education, University of Oklahoma, 4502 E. 41st Street, Building 4W, Suite 122, Tulsa, OK 74135-2553, USA

1 Introduction

Spurred by *Race to the Top* (RttT), efforts to improve teacher evaluation systems have provided states with an opportunity to get teacher evaluation right. Paramount in that improvement effort should be addressing the limitations of past teacher evaluation systems, limitations such as inadequate differentiation of teacher performance, inadequate evaluator training/expertise, lack of evaluation criteria focused on learning outcomes, and lack of meaningful feedback on performance for the purposes of improvement (Darling-Hammond 2013; Harris and Herrington 2015; Weisberg et al. 2009). Evidence of successes and remaining challenges in these areas is emerging, but there is still room for improvement (Ford et al. 2017; Hewitt and Amrein-Beardsley 2016; Kraft and Gilmour 2017).

Of the purposes for evaluating teachers, a basic distinction that can be drawn is between teacher evaluation for accountability (summative) versus professional development (formative) purposes. On one hand, the goal of summative teacher evaluation is to assess the teacher's performance or quality for the purposes of reaching a decision about whether to apply reward or sanction, or to otherwise inform personnel decisions (Organization for Economic Co-operation and Development [OECD], 2009). Formative evaluation, on the other hand, involves evaluation for the purposes of teacher support and professional development (Delvaux et al. 2013). Historically, teacher evaluation has been viewed as a facet of prescriptive instructional leadership associated with bureaucratic control and oversight (Blase and Blase 1999; Murphy et al. 2013); thus scholars use the term *supervision* to refer to evaluative activities related to teacher support and development (Hallinger et al. 2014).

In addition to its traditional role as an accountability tool, the *Race to the Top* competition also touted the potential of new teacher evaluation systems to provide ongoing and meaningful feedback for teacher development and instructional decision making (see, e.g., Great Teachers and Leaders subsection D, Part 5; U.S. Department of Education 2009). Many states, including Louisiana, have followed suit in also claiming this as a central purpose of their new teacher evaluation systems (Doherty and Jacobs 2015; Louisiana Department of Education 2012, 2013, 2014). However, an analysis of post-RttT data from the Teaching and Learning International Survey (TALIS) by Darling-Hammond (2014) revealed that American teachers are still less likely to believe that the information they receive from evaluation is useful for improving instruction. The fact is, we still know relatively little about how effective state teacher evaluation policies are in spurring teachers' use of these sources of data for instructional improvement (Ford et al. 2016; Datnow and Hubbard 2015; Sun et al. 2016).

Implementing new teacher evaluation systems comes at significant human and economic capital costs to states and districts, irrespective of the already substantial federal and philanthropic investments made since RttT. A recent study by RAND and the American Institutes for Research of three U.S. school districts found that startup and on-going expenditures for their new teacher evaluation system ranged from $6.4 million in the smallest district to $24.8 million in the largest over a 3-year period (Chambers et al. 2013). These estimates did not, however, include the human capital costs of teacher evaluation, namely, the time spent by personnel to observe and evaluate teachers. The number of formal observations conducted by principals and assistant

principals in the large districts was around 20,000 per year, with an average time spent on each observation of around 2–3 h (Chambers et al. 2013). A similar study of new teacher evaluation in New Jersey (TEACHNJ)—which required three teacher observations per year—estimated that the minimum amount of time needed to conduct classroom observations would increase by 35% under the new system (Larkin and Oluwole 2014). Even a conservative measure of average principal's salary puts the estimate of the yearly cost of *just observing* all 3.1 million U.S. K-12 public school teachers twice in a given year at around $700 million (Dynarski 2016).[1] Given the already demanding work lives of principals, finding the time to meet the demands of these new systems is also likely to be a challenge, requiring a significant reallocation of time and resources to be done well (Lavigne and Good 2015). Given these substantial investments, understanding if (and how) the information generated from new teacher evaluation systems is translating into improved teaching and learning seems particularly critical.

1.1 Current study

The use of various types of evidence to inform instructional improvement is a mainstay of good teaching practice. Recent definitions of data-informed decision-making (DIDM) in the literature have emphasized establishing structures, routines, and/or processes for the systematic use of data in creating actionable knowledge (Datnow and Park 2014; Ikemoto and Marsh 2007; Mandinach 2012; Mandinach et al. 2008; Marsh 2012; Marsh et al. 2006). This understanding of DIDM draws sharp delineations between the collection of *raw data*, the organization of these data into *information*, and the use of information combined with expertise to create *actionable knowledge* (Marsh 2012). Within the broader literature, we have learned much over the past decade about the types of data stakeholders use and the conditions that enhance or reduce DIDM (see, e.g., Datnow and Hubbard 2015; Farrell and Marsh 2016a; Marsh 2012; Schildkamp et al. 2017 for reviews of this literature). A few recent studies have found greater use of data for accountability than for instructional purposes, noting that higher accountability data use can often be associated with more compliance-oriented school cultures characterized by superficial responses to data, teaching to the test, and/or other gaming practices (Farrell and Marsh 2016b; Schildkamp et al. 2017). Yet others demonstrate that the closer the data are to what teachers do in the classroom, and the more autonomy they have over the data use process, the more likely they will use them and the more likely they are to spur changes in pedagogy (Farrell and Marsh 2016a; Huguet et al. 2017).

Teacher evaluation feedback can include commonly used sources of data in schools (student test scores, benchmark assessment performance, other student assessments), but also data sources more unique (but not necessarily exclusive) to the teacher evaluation context (lesson observations, student/parent surveys, self-assessments, etc.). We still know very little, however, about the conditions that surround stakeholders' use (or non-use) of teacher evaluation data for improvement. Because many of

[1] This estimate is based upon the product of Dynarski's (2016) estimate of a principal's salary of $45/h; the number of U.S. K-12 teachers (3.1 million); the average number of hours spent per evaluation; typical number of observations in a given year (2).

these sources of data are not unique to teacher evaluation, there is a high likelihood that those conditions that enhance or diminish the use of teacher evaluation feedback are similar to those found in the broader DIDM literature.

As scholars have noted, the literature on data use is largely atheoretical in nature (Farrell and Marsh 2016b). *Why* particular conditions motivate stakeholder data use while others do not is a question that has not been well addressed in the field. Undergirding the exploration of why is the discovery of a mechanism that explains teachers' disparate use of data generated by teacher evaluation—something arguably critical to designing and implementing appropriate and effective teacher evaluation policy. In this paper, I utilize Self-determination Theory (SDT) and the concept of *functional significance* as a lens for understanding and explaining patterns of use (or non-use) of Compass-generated evaluation data by teachers over a period of 3 years in a diverse sample of Louisiana elementary schools. The study was guided by the following research questions:

1. How did Louisiana teachers report using, (if at all), the information generated from Compass, Louisiana's high-stakes teacher evaluation system, to drive on-going instructional improvement?

2. Are there systematic patterns in Louisiana elementary teachers' use (or non-use) of Compass-generated performance information, and what, if anything, explains these patterns?

1.2 *Compass*: Louisiana's high-stakes teacher evaluation system

In 2010, Louisiana House Bill 1033 established Compass, a high-stakes system for the evaluation of teacher performance (Louisiana Department of Education 2012). According to Louisiana regulations, the assessment of school personnel is intended to fulfill both summative and formative purposes (paraphrased and condensed for brevity): (1) personnel retention, management, and decision-making; (2) improving teaching, leading, and learning; and (3) informing teachers' professional growth and development (Bulletin 130, LAC 28:147.103.A 2017). Starting in the 2012–2013 academic year, implementation of Compass went statewide, and schools fully transitioned to the Common Core State Standards (CCSS) during the following year (the 2013–14 academic year).

In the initial year, an individual teachers' yearly Compass score included both a lesson observation component and a student achievement growth component (grade 3–8 teachers received a Value-added Measure (VAM), all others a Student Learning Target (SLT) performance score). VAM scores were based on Louisiana Educational Assessment Program (LEAP) and iLEAP scores. In 2013–2014, Louisiana made the transition to the PARCC (Partnership for Assessment of Readiness for College and Careers) assessment by having teachers begin teaching to the standards, with a plan for the assessment to go into full force the following year. The same year, the use of VAMs for teacher evaluation purposes was eliminated and SLT performance became sole achievement growth measure for *all* Louisiana teachers.[2]

[2] All the descriptions of the Compass system discussed in this section are as they were during the study period of 2011–2015. Since this time, Compass has again changed to reflect adjustments to assessment policy as well as teacher evaluation policy.

As is the case for many states, SLTs (also known as Student Learning Objectives, SLOs) provide teachers with a framework to identify and track yearly performance goals for their students and utilize this information for instructional decision-making (Louisiana Department of Education 2015a). At the beginning of each school year, teachers fill out a "goal sheet" in which they are expected to specify (a) student learning goals, (b) the assessment instrument(s) that will be used to determine growth, (c) the indicator(s) for success in meeting the target, and (d) their plan for monitoring progress at three different "check points" throughout the school year (Louisiana Department of Education 2015a). While teachers were encouraged, through this process, to track student progress toward their identified learning goals, at the time of this study no formal process for providing teachers formative feedback on their SLT progress had been explicated. At the end of the year, the SLT process was to culminate in a reflection (either on one's own or with the building leader) on the successes and challenges of the year, as well as the leader-assigned SLT evaluation score from 1 to 4 which reflected varying degrees of attainment of the target (e.g., Insufficient, Partial, Full, or Exceptional). Depending on the policy of the district in which they work, control over the selection of the assessment instrument(s) for SLT use varied in our sample of schools—some districts required all teachers to use a common instrument while other districts allowed teachers more autonomy in choosing assessments aligned with their grade level and/or subject area.

In addition to SLT performance, all teachers were required to have a minimum of two lessons evaluated in a given year,[3] by means of a modified "Danielson" rubric.[4] These two scores (the growth component and the observations) each held equal weight (50/50) and were thus combined to determine the teacher's overall Compass performance score. In the event of multiple observations and/or SLTs for one teacher, these were also averaged to achieve a final score (for either the growth or observation component, or both). Teachers were required to be evaluated by either a principal, assistant principal, or other designated supervisor, and these individuals had to participate in ongoing trainings as a part of their role as Compass evaluator which included an assessment and activities to ensure inter-rater reliability (Bulletin 130, LAC 28:147.311 2017).[5] Observations originating from other professionals within the building such as instructional coaches and master/mentor teachers were considered "additional observers" and their evaluations could be used only to inform and/or assist assigned evaluators' decisions (Bulletin 130, LAC 28:147.311.B.1 2017). Compass is "high-stakes" in the sense that teachers who were rated "ineffective" in a given year were subject to a remediation plan and had 2 years from the start of this plan to achieve "effective" status; if

[3] Teachers who receive a "highly effective" rating in a given year are only required to have one formal observation the following year.

[4] The Compass teacher evaluation rubric utilizes only 5 of the 22 domains and 20 of the 76 elements of the full Danielson Framework for Teaching.

[5] While not clearly specified in the policy, in most cases in our sample the same evaluator observed both lessons conducted by the teacher. Assignment of evaluators was ultimately up to each building principal.

they did not, their teaching certificate would not be renewed, and they would be subject to disciplinary action up to termination (Louisiana House Bill 1033 2010).[6]

2 Theoretical framework: how the *functional significance* of performance information shapes teacher use

A key maxim of self-determination theory (SDT) is that of *dialectical integration*: the innate drive of the individual towards the resolution of the elements of the self, and the integration of this self to the social structures within which it is embedded. In other words, human beings have an intrinsic desire to engage in and interact with the world around them, exercise capacities, and pursue connectedness towards a more complex sense of self (Deci and Ryan 2000; Ryan and Deci 2017). SDT predicts that this drive will remain intact so long as certain key psychological conditions are met, namely the need for competence, autonomy, and relatedness. *Competence* refers to the individual drive to build upon existing skills and capacities in anticipation of future performance (Niemiec and Ryan 2009; Ryan and Deci 2002). *Autonomy* concerns action for which impetus derives not from the need to conform to external forces/expectations but rather from self-endorsed or determined values and beliefs (Ryan and Deci 2000). Finally, *relatedness* refers to the need to be cared for by colleagues as well as share a sense of belongingness to others in your community (Ryan and Deci 2002).

While SDT as a whole describes the various sets of conditions under which individuals are optimally motivated, it also provides a mechanism for understanding (and predicting) the extent to which individuals will integrate *new* events, experiences, or information to which they are exposed. The concept of *functional significance* states that the effects of external events on human motivation hinge on the psychological meaning they have for the recipient (Ryan and Weinstein 2009). For example, events have a positive effect on an individual's self-motivation when they have *informational* significance—that is, when they provide feedback that helps learners become more effective but without eclipsing autonomous action (Deci and Ryan 2000). In other words, the feedback "… points the way to being more effective in meeting challenges or becoming more competent, and does so without pressuring or controlling the individual" (Ryan and Brown 2005 p. 361).

On the other hand, events have *controlling* significance if they are experienced as pressure toward specific outcomes, or as an attempt to control the behaviors or efforts of individuals. When events or feedback attempt to control behavior, individuals often respond by exerting the least amount of effort needed to gain reward or avoid punishment (Ryan and Weinstein 2009). Thus, SDT predicts that such events might elicit short-term compliance but ultimately self-motivation for the activity will be undermined (Ryan and Deci 2017; Ryan and Weinstein 2009). Finally, events have *amotivating* significance when the arousal they engender is debilitating or when they

[6] While there is no available data on how many teachers have been dismissed during the Compass era due to ineffective ratings, aggregate results from the Louisiana Department of Education (2013, 2014, 2015b, 2016) report that around 4% of teachers were rated "ineffective" in 2012–2013, 2% in 2013–2014, and less than 1% in 2014–2015 and 2015–2016.

contain no inherent rationale for action. For example, events that are too challenging or feedback which is highly negative foster feelings of helplessness or incompetence (i.e., lack of self-efficacy), leading individuals to withdraw effort (Ryan and Brown 2005). Similarly, individuals will tend to exhibit amotivational responses to events (a) that are perceived to be irrelevant to their immediate tasks, (b) over which they feel no sense of control, and/or (c) for which no clear direction for action is provided (Deci and Ryan 2000; Ryan and Weinstein 2009).

The choice to use SDT instead of other potential theories of motivation to explain patterns of use (or non-use) of Compass-generated performance information was based on some important theoretical differences critical to the present study. Expectancy-Value Theory (EVT), for example, was considered as a potential framework for the study (Rice and Malen 2016). Under EVT, the strength of an individual's motivation toward a goal is the product of an individual's expectations of attaining the goal (i.e., self-efficacy and outcome expectancies) and the value the individual places on that attaining that goal (i.e., valence; Eccles et al. 1983; Wigfield and Eccles 1992). While they share considerable overlap, one important area where SDT and EVT diverge is in considering what shapes the value one places on attaining a goal and how this relates to overall motivation, behavior, and outcomes. EVT is generally agnostic about what lies behind the value one places on a goal—if motivation is strong, it should lead to positive outcomes. SDT, on the other hand, distinguishes between qualitatively different reasons for action, predicting that different motivational orientations (e.g., controlled versus autonomous) will lead to different outcomes, irrespective of the strength of one's motivation (Vansteenkiste et al. 2005). Empirical evidence in support of this claim has since been gathered through a series of carefully designed experiments (Vansteenkiste et al. 2006). Because the working hypothesis of this study was that distinctions in how individuals were oriented towards Compass performance information (i.e., the origins of their motivation to use it, either internal or external) would predict subsequent data use behaviors, it was necessary to draw on a theory that would be useful in helping to understand and interpret findings in this domain.

2.1 The role of high-stakes performance evaluation in shaping functional significance

Current accountability policies were designed as authority/incentive policy tools (Schneider and Ingram 1990). The underlying rationale of authority/incentive tools is that using rewards or punishment to induce desired behaviors is the most effective way to motivate individuals—a theory of motivation that owes its origins to classic operant theory (Ryan and Brown 2005). As such, accountability systems tend toward treating performance information as summative in nature, and this increases the likelihood of these data being perceived as controlling or amotivating (Curry et al. 2016; Ryan and Weinstein 2009; Ryan and Deci 2017). Extant evidence suggests that motivational strategies based on control/compliance, rewards, or threats of punishment are the poorest in producing long-term autonomous behavior precisely because they engender an externalized locus of causality over one which is more self-improvement oriented (Curry et al. 2016; Deci et al. 1999; Niemiec and Ryan 2009; Ryan and Deci 2017).

Schildkamp and her colleagues (2017) distinguish between data use for accountability, school development, or instruction, noting that it is often the case that these different purposes are under tension—emphasizing one typically carries the risk of marginalizing the others. The application of extrinsic tools to motivate improvement in the case of teacher evaluation frames it as a summative tool, and the pursuit of this goal will place a premium on information more likely to be of interest to stakeholders other than teachers (Adams et al. 2017)—often at the expense of information that teachers might find helpful or useful for improvement. For example, value-added (VAM) student growth measures might provide a school leader with information s/he can use for making personnel decisions, but the same performance information provides little to no information to a teacher trying to improve his/her instruction, other than that they are doing a "good" or "bad" job. Thus, one consequence of trying to use accountability performance information for improvement purposes is the paucity of information it provides school personnel on what is wrong and how to fix it (Adams et al. 2016, 2017; Beaver and Weinbaum 2015; Glover et al. 2016), and this can fundamentally undermine the functional significance of the information. Additionally, the pressure to perform absent corresponding information and support to make meaningful improvements further undermines functional significance by attenuating the perceived legitimacy of the indicators (Adams et al. 2016, 2017; Ford et al. 2016). This, in turn, increases the likelihood that school personnel will perceive them as controlling or amotivating.

Because evaluation for the purposes of teacher support and professional development (Delvaux et al. 2013) is generated specifically with teachers' improvement in mind, performance feedback from these systems is much more likely to be of functional significance to teachers. An analysis of teacher evaluation systems reveals a substantial (and growing) amount of variation in their design and implementation across states (Doherty and Jacobs 2015). While some states are becoming leaders in developing systems more supportive of teacher development and improvement (i.e., by incorporating multiple measures of performance and closely aligning these measures with classroom instruction and student learning, for example), other states' systems, while perhaps aspiring to these higher goals, remain primarily compliance/accountability-oriented (i.e., by using test scores as the preponderant criterion for evaluation and attaching dismissal, loss of tenure, teacher pay, and other extrinsic incentives to ineffective ratings). For those states who fall in the latter category, it remains to be seen whether or not data that are primarily accountability-oriented can be used effectively for teacher improvement (Schildkamp and Visscher 2010).

2.2 The dimensions of functional significance of performance information for teachers

Of the few research studies that have examined teachers' use of high-stakes teacher evaluation data in the classroom, the evidence demonstrates that teachers respond to controlling sources of performance information by externalizing their locus of control and disengaging in the use of such data for instructional improvement (Ford et al. 2017; Amrein-Beardsley and Collins 2012; Lavigne 2014). However, much remains to be understood about the conditions under which the functional significance of teacher performance data is either enhanced or compromised. If one important drawback of a

high-stakes approach to teacher evaluation is that it undermines the information in fundamental ways, an important question to consider is: What properties of teacher performance information and how it was generated enhance or diminish their functional significance for teachers? A systematic review of the broader literature on data use as well as teacher learning/evaluation reveals three broad and interrelated dimensions that evidence suggests likely affect the functional significance of performance information for teachers: validity, utility, and supportive conditions.

Validity In psychometrics, validity refers to "…the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, NCME, 2014, p. 11). Studies employing this approach to examining the validity of teacher evaluation performance information—in particular value-added measures (VAMs)—for accountability and improvement purposes abound in the literature (see, for example, Darling-Hammond et al. 2012; Haertel 2013; Herlihy et al. 2014; Papay 2011). Complementing more formal approaches to investigating validity, some studies of teacher evaluation have been concerned instead with understanding the perceptions and satisfaction of teachers regarding their evaluation system—sometimes referred to as "social validity" (Amrein-Beardsley and Collins 2012; Collins and Amrein-Beardsley 2014; Jiang et al. 2015; Reddy et al. 2018). Scholars investigating the social validity of policy recognize that success or failure often hinges on how teachers understand, interpret, and respond to policies that affect them and their work (Lipsky 2010; McLaughlin 1987). The success of teacher evaluation policy would seem to be particularly vulnerable to these types of perceptions not only because teachers themselves are the subject of the evaluation but also because teachers' use of these evaluation-generated data is largely contingent upon its perceived social validity (Reddy et al. 2018). The focus of this paper is primarily on more informal assessments of validity, as discussed in more detail below.

A review of the recent literature on teacher evaluation yielded three salient sub-dimensions of social validity around which the following discussion is organized: clarity, credibility, and fairness.[7] For this study, clarity refers to the degree to which the performance measures and the standards that comprise them are well defined and understandable to those being evaluated.[8] Performance information should be generated via a frequent, systematic evaluation process that yields understandable and actionable feedback on what to improve (Delvaux et al. 2013; Ford et al. 2016; Hallinger et al. 2014; TNTP 2010; Schildkamp and Visscher 2010; Tuytens and Devos 2011). These assessments should be based upon a set of high standards which reflect what is currently understood as good teaching practice (Darling-Hammond 2013; Lavigne and Good 2014). An evaluation approach based on the standards of good teaching is more supportive of improvement by establishing the clear expectations necessary to motivate

[7] It is important to mention that both the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014), and the Joint Committee on Standards for Educational Evaluation (JCSEE 2009) define and delineate issues of evaluation related to clarity (JCSEE), credibility (JCSEE), and fairness (AERA et al.). The operational definitions of these terms in this paper share some overlap but also differ somewhat from theirs, as will be discussed as each term is defined below.

[8] In the JCSEE, one aspect of clarity that aligns with the definition used in this paper is accuracy standard A2, "defined expectations." Another aspect, however, which was not a focus of our definition per se, is the necessity for clarity on how the assessments/evaluation tools are aligned with the expectations (JCSEE standard A1).

change (Kelly et al. 2008). Clear standards can also facilitate perceptions of the evaluation as an authentic and fair assessment of practice, and this can drive use of information from their evaluation to make changes in practice (Ford et al. 2016; Delvaux et al. 2013; Lavigne 2014).

Credibility refers to the degree to which those being evaluated believe that the performance measures are appropriate and/or legitimate for the intended use (Rice and Malen 2016).[9] Fairness is related to credibility and refers to the opportunity those being evaluated have to fully demonstrate their standing on the performance measure and for their scores to be interpreted with consideration for the various contexts under which it was assessed (AERA, APA, NCME, 2014).[10] A growing body of evidence suggests that an assessment of teacher performance should be based on a thorough assessment of teaching practice—no one measure of performance (whether student test scores or otherwise) is adequate to arrive at a determination of teacher effectiveness and construct a plan for change (Grissom and Youngs 2016; Lavigne and Good 2014; Master 2014; TNTP 2010). The prioritization (whether intentional or unintentional) of any one measure of teacher performance may undermine credibility and perceived fairness of the evaluation system if teachers do not feel that this measure truly reflects their classroom efforts (Lavigne 2014; Rice and Malen 2016). Such a narrow focus also ignores other valuable contributions of teachers to the development of the whole child (Grissom et al. 2016).

Finally, teachers' perceptions about what kinds of data are credible are driven by their own values, beliefs, and prior experiences with various sources of data (Farrell and Marsh 2016a; Ingram et al. 2004). Prior experiences, familiarity (or lack thereof) with particular types of data, and beliefs about what constitutes good teaching can all drive perceptions of credibility (Ingram et al. 2004). On the other hand, in the case of teachers exhibiting behaviors associated with controlling significance, credibility of the information takes a backseat to the consequences of non-compliance/poor performance. In these cases, responses to high-stakes data are likely to be superficial—avoiding sanction or maintaining the appearance of propriety can supersede efforts at genuine improvement (Booher-Jennings 2005; Farrell and Marsh 2016b).

Also important to teacher's perceptions of the credibility and fairness of evaluation information is the nature of the relationship between the evaluator and the teacher (Delvaux et al. 2013). If the teacher perceives the evaluator to be incompetent, or the feedback process is compromised by poor communication, this can have a detrimental effect on perceived usefulness of the process and the likelihood that a teacher will act on the evaluation results (Chow et al. 2002; Kelly et al. 2008). Perceptions of competence of the evaluator are related to the degree that the teacher feels the evaluator is qualified to rate their performance, and this is driven by knowledge of the evaluator's teaching experience, subject matter content/pedagogical knowledge, and depth of training in the evaluation process (Delvaux et al. 2013; Milanowski and Heneman 2001).

[9] The concept of credibility does not relate in any direct way to the JCSEE standards, but might nevertheless be an overall judgment rendered by an evaluatee of the process based on several of these standards. None of these standards are specifically referenced in this study.

[10] This aspect of fairness is only part of the *Standards for Educational and Psychological Testing* framework. Other aspects of fairness concern the degree of measurement bias as well as influences of test-taking contexts which were not as present in the literature on the topic of teacher evaluation.

Utility The other defining characteristic of the functional significance of performance information is its perceived usefulness to the end-user (Beaver and Weinbaum 2015; Farrell and Marsh 2016a; Ingram et al. 2004; Kerr et al. 2006; Marsh et al. 2006; Marsh 2012). A broad definition of utility refers to aspects of the evaluation process that enhance the practical value the performance information has for a wide range of stakeholders.[11] Utility is of course defined, in part, by the perceived validity of the information on the part of teachers. If performance data do not first meet other validity criteria, it will likely fail to meet the utility criterion; it stands to reason that individuals will not use data not first seen as valid. Additionally, we know that the type and/or nature of the data, its origins, its timeliness, and how these align with the past experiences of teachers also drive data perceived utility of the data (Datnow and Park 2014; Farrell and Marsh 2016a; Marsh 2012).

For teachers, the question of the utility of performance information could be summed up in the following general question: Will this information help me do something to improve things I care about? Because a majority of teachers enter the profession for altruistic reasons (Lortie 1975; Watt and Richardson 2014; Rosenholtz 1991), it is not surprising that a substantial portion of what motivates teachers to improve is the desire to see their students grow and thrive. Coupled with a desire for self-improvement, performance data which can aid in the pursuit of these "psychic rewards" is likely to be of interest to teachers (Ford et al. 2017). For instance, "soft" data generated from informal classroom assessments and student work that give teachers insights into how their students are thinking or what they know or can do from a developmental perspective are commonly used sources of data for teachers (Datnow and Hubbard 2015; Farrell and Marsh 2016a).

Of course, teachers' concerns about student thriving are not limited only to issues of student achievement and learning. A concern with performance information used for accountability purposes is that of risks to student equity in classrooms (Datnow and Park 2014; Datnow et al. 2017; Skrla et al. 2004). There are some exemplar studies of districts taking careful, proactive steps to make data useful in informing both equity issues and learning (Park et al. 2013), but scholars also identify this area of the literature as one in which further study is needed (Datnow et al. 2017).

Meeting teachers' psychological needs as a facilitating condition for data use Even if care is taken to ensure that performance data are valid, a lack of appropriate teacher preparation/training may prevent teachers from recognizing how data might be used for their own improvement (Datnow and Hubbard 2015). *Competence*, a key psychological need for teachers, is primarily manifested in the case of data use through teachers' feelings of preparation to use data in meaningful ways for decision making—as Mandinach (2012) has referred to it, the cultivation of *pedagogical data literacy.* School leaders can foster competence for data use by allowing for dedicated time and space for teacher collaboration (Farley-Ripple and Buttram 2014; Farrell 2015; Ingram et al. 2004; Little 2012; Mandinach 2012;

[11] Our use of the concept of utility refers most specifically to the JCSEE standards of utility related to evaluator qualifications and functional reporting (Standards U3 and U5). The other utility standards were not as salient in the teacher evaluation literature.

Marsh et al. 2006). They can also provide access to expertise through professional development, the hiring of instructional coaches, or other support staff (Honig and Venkateswaran 2012; Marsh et al. 2006, 2010). Leaders' efforts to build a strong school data culture, along with collegiality and trust, aids in the formation of human capital from social capital (Cosner 2011; Ikemoto and Marsh 2007; Little 2012; Young 2006). Perceptions of competence shape an individual's self-efficacy beliefs, and these too shape data-use behavior. As was mentioned earlier, SDT posits that competence beliefs alone without concomitant autonomous motivation will not necessarily lead to desired data use outcomes, and this makes the psychological needs of competence and autonomy a more proximal condition of data use than self-efficacy.

Once teachers have begun the process of building competence in data analysis and use, they must also be given the latitude to apply those learnings to practice in a way they find meaningful. *Autonomy*, also a key psychological need for teachers, allows teachers to experience success through their application of knowledge to problems of interest through self-determined action (Ford et al. 2016). Allowing for self-determined courses of action on the part of individuals ensures that interest in the feedback on successes and failures that emerge from the course of action will be optimized (Ryan and Deci 2000, 2017). Conversely, situations where the course of action is pre-determined and individuals are simply expected to carry it out, *ceterus paribus*, the interest in the information generated from the action is likely to decrease. Several studies of effective data use conditions have highlighted autonomy as a key condition which improved teachers' perceptions of the usefulness of the data for their work (Farrell and Marsh 2016a; Huguet et al. 2017). For example, Farrell and Marsh (2016a) found that, with respect to using common grade assessment data, teachers who reported finding them useful for instructional improvement did in part because they had developed the assessments themselves.

Finally, *relatedness* as it pertains to data use is the teacher psychological need that has the most empirical support to date. Indeed, numerous studies using wide-ranging samples have highlighted the importance of collaboration and/or the development of effective professional learning communities in the learning and use of data for instructional improvement (Cosner 2011; Huguet et al. 2017; Marsh and Farrell 2015; Sun et al. 2016; Van Gasse et al. 2017). Relatedness as a psychological need is critical particularly for data that are seen as "high stakes" in nature. Of concern is that data are "safe" enough to discuss with others, and for that, trust, collective responsibility, and norms of interaction and behavior are needed (Marsh 2012). Further, collaboration over time helps to form the common language, shared knowledge, understandings and routines, as well as the sense of community needed to support data use (Cosner 2011; Curry et al. 2016; Datnow and Park 2014).

The question of how teachers might use teacher evaluation data collaboratively for improvement is very much still in question, however. It is certainly possible that, under the right circumstances, teachers could be supported in working collaboratively with data from their individual evaluations, but the fact that evaluation data are information (often of a high-stakes nature) about an individual teacher's practice makes it much more likely for teachers to see themselves in competition with one another than partners in improving practice (Booher-Jennings 2005).

3 Method

The data presented in this investigation were collected over a 3-year period from 2012 to 2015 as part of a study of the effects of the *Compass* and *Common Core State Standards* initiatives on the work of early childhood and elementary level teachers and administrators in Louisiana. The overall study was designed as an embedded, multiple-case study. Embedded case studies allow for sub-units of analysis to be examined (in this case teachers) alongside the larger organizations or units within which they are embedded. Multiple case studies allow for both cross-case and replication approaches to be used in understanding phenomena under study (Yin 2017). The rationale behind this choice of design was to allow flexibility in answering a wide range of research questions that emerged over time as a part of the larger study of the Compass and CCSS policy initiatives. In order to address the research purposes and questions in this present study, for example, this design provided a means to leverage the individual cases for the purpose of uncovering theorized patterns about data use among teachers across sites (i.e., investigate the degree of replication of findings across sites). Thus, the current study purpose placed more emphasis on the embedded units (teachers) than the individual cases (schools) themselves, and this was in part because this particular investigation, as will be elucidated later, revealed little evidence that the cases and their respective contexts shaped in a substantial way the phenomenon in question (data use) as experienced by teachers.

3.1 Case and participant selection

In order to select the participating districts/schools for the multiple case study design, we employed *maximum variation sampling*. This is a type of purposive sampling which provides for a limited but wide range of cases (in this case schools) to be selected along a list of identified dimensions (Miles et al. 2014). Using this sampling strategy allows researchers to explore both unique aspects of the phenomenon under study within each site context but also to examine cross-cutting themes. This site sampling strategy considered the following school-level factors in the selection: school performance score (SPS), percent poverty and/or minority, urbanicity, and school size in order to maximize heterogeneity among this vector of factors.

A total of seven schools within five school districts (parishes) across Louisiana—each in a distinct geographical area of the state. As can be seen in Table 1, these schools represent a wide and varied group with respect to the displayed school-level factors. Concomitantly, within each case, maximum variation sampling was also utilized to examine our embedded units of interest within the study (in this case teachers), while considering teacher-level factors such as teacher experience, elementary grade level (K-5), and subject/content matter expertise (including special education). This resulted in a final sample of 37 teachers at the start of the study. Over the three study waves (discussed in more detail below), each teacher was interviewed once, and each interview was around 30–45 min in duration. Over the course of the 3-year interview process, 5 of the 37 total teachers in the sample were lost due to attrition or illness, and two of these were not available for re-interview after the first wave of data collection. Thus, a total of 104 teacher interviews were available for analysis ((37 + 32 + 32) +

Table 1 Breakdown of maximum variation sample of schools by demographic factors

	Emerson	Fisher	Sampson	Bennett	Carver	Wayne	Palmer
Urbanicity	Suburban	Suburban	Urban	Rural	Urban	Rural	Rural
SPS (Grade)	A	A	C	D	C	A	C
% Poverty	~ 71%	< 50%	~ 71%	> 91%	> 91%	< 50%	> 91%
% Minority	~ 55%	< 25%	> 80%	> 80%	> 80%	< 25%	> 80%
#Teachers	5	5	7	5	6	5	4

School names are all pseudonyms. Data from the 2012–2013AY on school demographics acquired from the Common Core Data and the Louisiana Department of Education. To protect the identities of schools participating in the study, the above percentages represent cutoffs for one standard deviation above (designated by the > symbol), below (designated by the < symbol), and the Louisiana state average (designated by the ~ symbol) for % poverty (as measured by % free and reduced lunch) and % minority

SPS, School Performance Score, *#Teachers* number of teachers in the final sample from each school

3) = 104.[12] Table 2 displays some descriptive information about the final sample of teachers chosen to participate in the study over the three waves of data collection.

3.2 Data collection and analysis

From within each case, the primary approach for collecting and analyzing the data was phenomenological. Critical to an understanding of how high-stakes teacher evaluation data was used by teachers and the various contexts that might shape their motivation toward data use, it was necessary to capture their perceptions and lived experiences with it. To this end, qualitative data were gathered primarily by means of semi-structured interviews (see Appendix 1 for each wave's interview protocol). Data for the over-arching study were collected in three waves, once per school year for 3 years. The first wave of data collection focused on establishing a baseline for the implementation of the high-stakes teacher evaluation system *Compass* and teachers' feelings regarding their performance prior to having received their final *Compass* scores for the 2012–2013 school year. The second wave of data collection in the late fall of the 2013–2014 academic year sought to accomplish two tasks: one, to follow up on teachers experiences regarding *Compass* after having received their final scores (which were announced at the end of June 2013), and also their feelings of and support for the CCSS initiative which was new to all Louisiana school districts in the 2013–2014 academic year. The third and final wave was undertaken in the late fall/early spring of the 2014–2015 school year for the purposes of following up with teachers regarding their Compass performance and to probe their explicit use of these data for improvement, as well as follow up on the continued transition to the CCSS. Principals at each of the schools in the study were also interviewed once per wave, and instructional coaches (in

[12] The three added interviews in the equation ((37 + 32 + 32) + 3) = 104 refer to three of the five teachers that were lost after the first wave that we were able to track down and interview one final time. Our main purpose in interviewing them was to get a sense of why they left. This is why they were not included in the second wave teacher sample numbers, but added their interviews separately to the total.

Table 2 Breakdown of teacher sample by data collection wave

Years experience	Wave I	Wave 2	Wave 3
0–5	6	5	5
6–10	13	10	10
11–20	9	8	8
21+	9	9	9
Grade level			
Pre-K/K	6	6	6
1	5	4	4
2	5	4	4
3	6	3	3
4	5	5	5
5	4	5	5
Special education	6	5	5
Totals	37	32	32

the cases where a school employed one) were interviewed during the third wave.[13] For the current study, these sources of interview data were used only for the purposes of data triangulation, which is explained in more detail below.

Post-interview, each member of the research team completed a contact summary form as a part of the reflective process, and these were included in the analysis. Prior to collecting data, members of the research team as well as other interested students and faculty participated in two half-day trainings (one of each prior to the first and second waves of data collection) in order to review the main theoretical constructs governing the research, as well as to review and practice administering the research protocols for the purposes of pilot-testing as well as to ensure consistency of administration across sites. Development of the research protocols and coding approach were carried out in regularly scheduled closed research team meetings, and in the meetings that followed data collection, a second reader from within the research team was used in the coding and analysis process to ensure the trustworthiness of the conclusions drawn from the data.

To establish inter-rater reliability, all transcripts were coded by two researchers, and the coding results of all interviews between the two coders were reviewed and compared for the purposes of addressing/negotiating intercoder agreement for each wave of data collection. A randomly selected subset of these interviews was used to calculate the indices of inter-rater reliability.[14] Disagreements were categorized both by

[13] The final sample of principal interviews was 20, and there were two instructional coaches interviewed in the third wave.

[14] Thirty randomly selected interview transcripts across the three waves (about one third of the total) were selected for the purpose of checking inter-rater reliability.

code and by unitization (Campbell et al. 2013). The simplicity of our coding scheme led to high rater discrimination in coding and thus disagreements among the raters with respect to which codes applied to the text were fewer—reliability in these cases was 76.66%—which is considered "acceptable" (Marques and McCall 2005). Disagreements about unitization, or to which segment of the text the code applied, while still infrequent, occurred more frequently than coding disagreements—reliability in these cases was 68%, also considered "acceptable."[15] Using "negotiated agreement" (Campbell et al. 2013), when discrepancies between specific rater's codings arose, they were then discussed with respect to the code's intended use and then, when possible, changed to reflect the consensus between both researchers.

For data analysis purposes, interviews were transcribed and coded using ATLAS.ti data analysis software. The data analysis strategy across all three waves of collection was essentially the same: first cycle descriptive codes were created according to constructs identified in the theoretical framework and corresponding protocol(s). Some example code categories were types of data (SLT/VAM/Lesson Observation/Other Data Use), supportive conditions for use (including competence, autonomy, or relatedness), District/School differences in use, Data equity issues, and responses to data (informational, controlled, or amotivated). Subcodes and magnitude codes were also used (where applicable) to highlight distinctions in perceived functional significance of the data, supportive structures, and to render salient changes in attitudes or behaviors that occurred across the data collection waves (Miles et al., 2013). During second and third wave coding and data analysis, we continued to add to our first cycle code list, but we also incorporated emotion and "in vivo" coding, which reflected our evolving theoretical framework to capture differential responses to high stakes data. Once coding was complete, data arrays (matrices) were constructed to display data by focal codes and subcodes and also by teacher, school, grade level, and experience. Matrices of data use perceptions were also constructed by teacher over the three waves in order to elucidate any marked changes in these perceptions over time. It is worth noting that we collected the Compass performance scores for each teacher (where they were willing to divulge it to us) and readers will note that all exemplar quotes below also indicated the teachers' final rating (ineffective, emerging, proficient, and highly effective, where available) at the time of the second (or third) interview.

From within these arrays, data were triangulated using primarily the *data source* approach (Denzin 2001), which attempted to corroborate the feelings and perceptions of teachers within and across schools and also across school role (administrators, teachers, coaches, etc.). For example, we looked at reasonable consistency among teachers within a school regarding the use of high-stakes teacher evaluation data as important evidence of corroboration. This strength of corroborative evidence was extended further when we saw reasonable alignment of principals' reports to those of teachers within the same school.

[15] Inter-rater reliability was calculated via the proportion agreement method (Campbell et al. 2013), which takes the sum of the number of coding agreements and disagreements for a given code divided by the total number of codings of the lowest submitter (the coder with the fewest instances of the code).

4 Findings and analysis

4.1 Summary/overview

According to Self-determination Theory, a primary determinant of the effects of feedback on an individual's motivation to use it resides in the psychological meaning, or *functional significance*, the information has for the recipient (Ryan and Brown 2005). An analysis of interview data revealed that Compass data use was low across all school sites, and this will be evidenced throughout the following presentation of the findings. As teachers reflected on their responses, they often discussed having made small adjustments to their practice to address particular criteria rated as deficient on their evaluation. Absent were discussions of using the results to reflect more deeply and globally about instructional changes and the ramifications these changes would have for classroom practice. As I probed deeper, it was clear that, for various reasons, perceptions of the functional significance of the Compass-generated information were undermined for Louisiana elementary teachers. In describing their experiences with Compass information, they often revealed controlled or amotivated orientations to these data sources, and these orientations were linked to the three broad categories outlined in the theoretical framework: (a) the degree to which support for teachers' psychological needs for competence, autonomy, and relatedness accompanied the feedback; (b) the perceived validity of the information as a representation of practice as well as for enhancing student performance and/or equity; and (c) the perceived degree of utility the information was expected to have for improvement.

4.2 Superficial uses of evaluation results

While Compass-generated data use was low across all school sites, when teachers reported using it for instructional purposes, their intention in using it was to target a specific evaluation criterion for improvement in lieu of broader, deeper changes to their practice—reflecting a more controlled response to the results. When asked how they used the data from the teacher evaluation process in their teaching, they often mentioned focusing on discrete items which would help boost their scores in particular areas of the evaluation rubric rated as deficient. Conspicuously absent in their discussions of how they used these data for improvement was any evidence of deep reflection on the implications their Compass results had for their practice, or the changes that they had made as a result. And while this orientation was particularly prevalent with respect to teachers rated as "highly effective," the mindset was not exclusive to this group. One example of a controlled orientation to improvement information was expressed by a 4th grade teacher at Fisher Elementary:

> P: The scores…ok. My first one [lesson observation], at the beginning of the year, I think it was like in September, I made all 3's. So I want to know what things I could improve in. So I looked at some of the easier ones for me to increase my score. [Effective proficient rating].

When asked how they were using the results of their evaluation to improve their instruction, a typical area of improvement that many teachers indicated (particularly

lower elementary teachers) was in the area of student–student interaction—an important effectiveness indicator for Common Core.

> "…right now I'm working on letting the students lead more which is tough for first grade." [Highly effective rating, 1st grade teacher at Wayne Elementary]

What seemed salient about these responses among lower elementary teachers is that they all indicated that the expectation of student–student interaction was somewhat unrealistic for the younger grades. While this is not necessarily an unreasonable expectation in reality, it nevertheless further suggests that theirs is a controlled response to the feedback rather than motivated by genuine desire to improve. When asked about the appropriateness of this expectation, a first grade teacher at Bennett commented: "Um, you know…we pretty much do what's on the rubric every day. But as far as the student led activities, it takes a lot to train these children."

4.3 Conditions which modulated Compass data use

The role of autonomy in shaping motivation for data use At the outset of Compass implementation, decision-making about the assessments and benchmarks upon which SLT data were based was left largely in the hands of teachers. In this district policy context, teachers' reported use of their SLT results suggests that it truly enhanced their ability make on-going decisions about practice. By the third year of Compass, however, districts' and schools' response to the challenge of teacher capacity for data was instead to centralize the selection of assessments for SLTs, thus exerting control over the process. The centralization of control over SLT assessment selection led to a steep decline in the perceived utility (and thus use of) SLT-generated data for teachers in both tested and non-tested grades. A teacher from Sampson Elementary reflects on the utility of the EAGLE test, an interim assessment system for the state test:

> No, we were told that K, 1st and 2nd grade were to use DIBELS…for 3rd, 4th and 5th, we were told we had to use EAGLE. Now, the year before, we were allowed to choose. I think I may have used STAR math, at that point [Interviewer: somebody else said that was good for that purpose…] Um-hm, it was very good, um-hum oh-yeah, because it told you…I mean EAGLE doesn't tell you much. I think the 5th grade test may have had 17 questions on it. How can you get a good view of what the kids knew out of 17 questions?" [4–5th grade teacher, Sampson Elementary, Wave 3].

The transformation of a Wayne Kindergarten teacher (and her Kinder colleague) from Wave 1 of data collection (when she was given full autonomy over assessment choices) and Wave 3 was stark, as seen in the two quotes below:

> Well, first of all we used the DRA [Directed Reading Activity] rather than DIBELS and we found that to be a better tool, more appropriate. So, we were good with that and we used the ISTEEP [the Integrated System to Enhance Educational Performance] for math because we weren't real sure about what to use for math. As far as SLT's I mean, we are very pleased with our results…our

> reading levels improved tremendously from in the past and we have kept a record of it. So what it forced us to do was less phonics instruction, less drill of phonics and more writing. And we found that the more writing we did the better readers we had. So we had a really good year with it [Wave 1, Wayne Kindergarten Teacher].

> [Discussing the district SLT assessment] I don't think the people who make this stuff up understand that you have a small percentage of children who are ready for those types of assessments. I don't think it's effective so we have our own things we look at every day [Wave 3, Wayne Kindergarten Teacher].

Evident in the first year of SLT assessment choice, these Kindergarten teachers (and many of our other teacher participants) took full advantage of their autonomy by carefully thinking through their assessment choices so that they would yield the best information for improving instruction and student learning. Because they took issue with the appropriateness of the district-chosen SLT assessment, they had adopted by the third year a controlled or amotivated orientation to the assessment and instead continued to use information from the original assessment sources they had chosen to drive instruction and student learning decisions.

The role of competence support Yet, because the process of generating SLTs was new to them, many teachers made mistakes. Many of our teacher participants expressed throughout the study the need for time, practice, and support in developing the data skills needed to learn from the early mistakes they made in setting their SLTs. In the beginning of Compass, teachers as well as administrators were, generally speaking, not well equipped by the program to provide the technical support needed for the careful crafting of SLTs. As a result, many teachers struggled to write reasonable ones—those which set realistic yet challenging goals for their yearly performance. Some districts tried to take the "guesswork" out of the process, by centralizing the selection of the assessment, or, in many cases, even setting the SLT performance targets for them, but this guidance ultimately exacerbated the support problem.

Assistance from school leadership, while helpful in one sense, also created problems when teachers were expected to do them on their own. One teacher showed us her goal sheet as she explained her confusion with the initial guidance on how to craft appropriate SLTs from leadership:

> I, you know, I had—I tried. We [s/he and her principal] did what we thought was right, and ended up not being, and because of my inexperience with it, I'm probably not gonna score what I should have scored…So, now, we're running around going, "Will I get a three?" You know, and if I'd have known, I would have not set my goal as high, you know, that wasn't explained enough. [3rd grade teacher, Sampson Elementary]

A Special education teacher from Emerson elementary also commented on the issue of support for writing SLTs:

> "…that was an area I had trouble finding support…So, people didn't really know—you know they gave me some ideas of different assessments that I could use. I ended up going with the district created, for k—since I have kids across grade level. I did two like, bulk SLTs, one was for first and second grade and I based it off the district [pre-test/post-test].

When asked about how more support for SLT writing from others would have benefitted her, a teacher from Carver remarked on how her confusion about the process wasted the "best year" she has had in terms of student performance:

> The only thing we knew about it [the SLT goal setting process] was what our principal was told, and then it was almost like she wasn't really told anything because she didn't really know. And then the assistant principal was like, 'we'll try this' and so you word one thing wrong and it's just…It's a whole year gone. Like the best year I've had ever. [Effective proficient 1st grade teacher from Carver Elementary]

4.4 Perceived validity of Compass feedback for student/teacher outcomes

Knowing what drives teachers when it comes to teaching is critical to understanding what kinds of feedback are likely be perceived by teachers as informational as opposed to controlling or amotivating. The analysis of teacher interview data revealed that perceptions of the validity of feedback in the case of Compass derived from multiple sources: student performance/equity concerns, credibility/fairness, who was evaluating, and the perceived utility of the information for improvement.

Student performance/equity issues Replete in our interview data are concerns about how particular groups of students would be harmed by district benchmark and state testing—the growth measure component of the teacher evaluation system. The perceived consequences of these tests for their students led many teachers to express frustration about assessments that did not advance positive learning outcomes for their students. In particular, SPED and early primary teachers expressed concerns about the appropriateness of the district-mandated SLT assessments and the effects they were having on their more vulnerable students. A SPED teacher at Sampson Elementary remarked:

> My SLTs are connected to how they perform on standardized tests. My students are considered LAA2 [Louisiana Alternative Assessment]—not as grueling as LEAP or iLEAP [state test], but we don't have a lot of information on what alternative assessment we will have in PARCC. With these new assessments, students seem lower and lower every year and I'm worried. I've got to meet their needs. At some point it has to be about life skills instead of can they pass some test?" [SPED teacher 4–5 grade, Sampson Elementary, high performing].

When asked how she then used her SLT assessment data [STAR test in Reading] in improvement she flatly stated:

> Um…I used it for my SLTs [long pause]…It was interesting to see their growth. I would do that whether I had the SLTs or not because it is interesting to see…"

Using the assessment data only for accountability reporting, as this teacher indicates, is the very definition of a controlled response. At the same time, she does recognize the value of the data, but it remains unclear from our conversation, however, how the information she derived from this data source (growth/no growth) would be used to adjust her classroom practice.

Teachers in other districts which had moved to a mandated SLT performance assessment expressed their concerns over the "developmental appropriateness" of the assessment for lower-elementary students. The comments of a kindergarten teacher at Wayne demonstrated just how her perceptions of the issues with the district-mandated SLT assessment drove how she responded to the assessment itself:

> If SLTs were something we created ourselves it would be okay, but because they are on the computer and the district picks the test company, they are so developmentally inappropriate it's just scary. It has brought to light some things [some student issues/needs], but basically we are just teaching to the test on this right before they go in there." [Wave 3, Wayne Kinder Teacher 2, highly effective]

Credibility/fairness One critical source of validity is whether or not teachers believe that the evaluation is a true representation of practice. Perceptions of authenticity led many teachers to dismiss their lesson observations as important sources of information to improve their practice. Many teachers referred to the process as a "dog and pony show," referring to the fact that the announced nature made the lesson performance seem inauthentic and more for the observer's benefit than to get useful feedback on their teaching. Skepticism stemmed from various perceptions of the process itself, but also in lack of fairness of the evaluation, and this led to doubts about its credibility as a measure of teaching quality/effort. A 3rd grade, highly effective teacher from Bennett elementary quipped: "First of all, I think…if you follow the rubric verbatim it is not… that is not a real world classroom at all, it just isn't." Other teachers made assessments of the fairness of the process based on conversations with colleagues about their experiences/results. A highly effective 5th grade teacher from Sampson confided:

> People talked…people shared their scores and how they did, and I'm thinking, I know for a fact that she threw that together the night before, and she made the same thing I did…it's like putting on a dog-and-pony show…you know it's coming. You have your little folder ready with your dog-and-pony show, as I call it, and you just pull it out.

A Kindergarten teacher at Wayne elementary expressed wanting to see more fairness in terms of adjustments to the indicators of "standard met" on the observation rubric based on grade level. She provided a specific example of this "dog and pony show" in action, when she described how she addressed student-student questioning/interaction in her lessons:

P: Right, right. When I did my first one, my first evaluation, I got 3's there [on student–student interaction], and I am thinking, 'well we were building words, onset and rhyme, and what was the word…um…I can't even remember but let's just say it was 'clack' and one of the children says, "Well what is clack?" and another child said, "Well that's such and such".

I: Perfect example.

P: Was that not higher level questioning in Kindergarten?…Children asking each other and the other child responded…and that happened several times in the first lesson?

I: But for you that was not acknowledged?

P: No and for me that was higher level questioning.

I: Absolutely, I agree.

P: It wasn't acknowledged in the first one but the second time whenever I said to the children, "Okay, after we have read the first two pages then we are going to turn to our partner and ask each other questions" then it was acknowledged [by the evaluator]…

Similar perceptions of the process from other teachers also prompted an amotivated orientation to the information generated from the observations. When asked about how s/he used the information from hers, another teacher remarked: "It's a dog and pony show, so I don't worry about it" [Effective proficient 3rd grade teacher at Emerson Elementary]. Another expressed a similar sentiment: "I don't worry about it because it's just so subjective. I don't see the benefit nor do I take anything away from it." [Highly effective rating, Kindergarten teacher at Wayne Elementary].

Who is evaluating? Certainly the sporadic nature of lesson observations and the fact that at least one of these was "announced" shaped perceptions of lesson observations as a "dog and pony show." Yet, analysis of the data revealed that validity of the results from the standpoint of fairness was also shaped by the perceived credibility of evaluator and fairness of the evaluation itself. When asked about how s/he did on her first evaluation, a Pre-K teacher at Bennett elementary remarked: "Well I got a really good score. I think that all depends on who is evaluating you. Which I always try to do my best. I always try to do everything I'm supposed to do."

In almost all cases, the principal was the primary evaluator for all teachers within the school. Within our sample of teachers, deviations from this pattern often (but not always) led to more favorable views of the lesson observation process. A teacher from Sampson elementary reflected on an important change to her evaluator in her first year (outside evaluator) and her second year (the principal):

Last year, I had a different evaluator. She was extremely tough, but she was extremely fair…She would point out, OK, this is what we need to do, this is what

> we want to do, so I learned from that. Because I had a different evaluator this year, that I knew was not difficult at all, [in planning/changing for this year] I still held myself to that same standard. So when I write lesson plans now, I'll look at them differently because of my previous observer. [Effective proficient rating last year, 5th grade teacher at Sampson Elementary].

In some ways, the response of this teacher to the perceived loosening of expectations on the part of the new evaluator was to maintain a prior standard of rigor to ensure that the observations still had improvement value for her.

Special education teachers were one group in particular that expressed frustration at evaluators' lack of understanding about their instructional environments. A SPED teacher from Sampson lamented that her first evaluator did not take the time to see her classroom before coming to evaluate. It was only when her new evaluator took the time to visit and see her classroom prior to evaluation that she received more empathy and consideration for how challenging her task was, but also how unfair the "one size fits all" lesson observation rubric was for her particular teaching context. She confided in her desire to be evaluated by an expert in her field: "That would be the icing on the cake…I would love to have feedback from an expert in my area."

Utility: formative data for improvement The ostensible purpose of generating Student Learning Targets under Compass has always been as a goal setting tool. In fact, while teachers in tested grades were, under the first iteration of Compass, not subject to evaluation based on SLTs but rather VAMs, Compass policy required all teachers to engage in SLT development, recognizing that goal setting is a cornerstone of good teaching practice. Though SLT performance does now constitute part of Compass evaluation score for teachers in tested grades, evidence suggests that the recognition of SLTs as sources of information about performance are losing out to the primacy of state testing and accountability data.

When asked about how they use their SLT data, the responses of teachers in tested grades largely reflected mindsets or behaviors consistent with controlled significance. Unlike teachers in non-tested grades, however, the recognition of the potential informational significance of SLT data were reduced due to the salience of state PARCC (Partnership for Assessment of Readiness for College and Careers) testing results. A third grade teacher from Fisher, a high-performing school, put it best:

> [Because of access to PARCC results] We don't focus on the SLTs as much and I'm not as familiar as I should be because I don't see how it will come into play—they just aren't necessary in a testing grade [Wave 3, 3rd Grade, Fisher Elementary].

When asked how s/he uses the information generated by the SLT process, another teacher's response was also indicative of a controlled orientation to SLT data: "I don't use it [SLT data]—you do it, put it in the computer and then at the end of the year you look at it" [Wave 3, 3rd Grade, Emerson Elementary (high performing)]

Taking this orientation a step further, there were several teachers who exhibited more of an amotivated orientation to the SLT performance information. One teacher from

Fisher elementary admitted that she is not even informed as to the assessment upon which part of her evaluation is based. For her, this was borne out of a frustration with her lack of a choice of SLT assessment:

> We are using a district test for our SLT's, but I have not seen it. You should not be told what it's going to be…because it's your kids [Wave 3, 3rd Grade, Fisher Elementary].

When asked if she used her lesson observation feedback to inform her practice, another teacher, this time at Carver Elementary confided: "No, not really, I just do the same thing you know, I do what I do and it may or may not be on the rubric." [Effective proficient, Kindergarten teacher, Carver Elementary].

5 Discussion

At the time of this study, the Louisiana Department of Education (2012) described the Compass evaluation process as one with a central formative component: It was purported to give teachers the support and feedback they need to improve their practice and, as a consequence, student learning. However, the responses of this sample of Louisiana teachers to the Compass system suggest that they perceived it as much more summative than formative. In every case, it seems as though the framing of the Compass system from within the larger context of state and district mandates undermined teachers' ability to view this feedback as informational in nature. Instead, with few exceptions, teachers exhibited Compass data use orientations which reflected a strong controlled or amotivated response; they did what was needed to "check the box" and move on. In cases where teachers perceived the data from feedback to be informational, one would expect to see evidence in their accounts of use which show thought, intention, and intricacy of use for improvement. Instead, teachers' response to the question of how do you use these data was often different shades of: "I use these data for this purpose and no other."

Even in cases where other high-stakes accountability data were more salient, as in the case of the PARCC test data, teachers struggled to recognize the value and/or usefulness of the SLT data for driving instruction. Their responses to these data were again controlled and/or amotivated in nature—some teachers going so far as to admit that they had not even seen the assessment that was used for their SLT growth measure. In cases where teachers did mention using/responding to the data in some way, their responses were often superficial in nature and did not reflect the responses of teachers who were using the data to deeply change their practice. Teachers often mentioned trying to work on discrete items from their evaluation—things that could be easily changed without much upheaval to their overall instructional practice. Items like increasing student-student interaction—an aspect of the observation rubric that most teachers, in the early years of Compass and Common Core, struggled to do well—represented easy ways to boost one's score with minimal disruption to other more "high scoring" areas.

Support for teachers' psychological needs in the areas of competence and autonomy in particular also undermined teachers' use of evaluation feedback to

improve their practice. Analysis revealed that district and school assessment policy played a particularly critical role, and this has been found to be the case in previous studies as well (Honig and Venkateswaran 2012; Young 2006). In the early years of Compass, teachers struggled to write SLTs which would give them a fighting chance for success, some overshooting or undershooting their mark by a considerable margin. They received little training on how to do this, and oftentimes district and school leaders were as in the dark as teachers on how to do it well, so they were ill-equipped to provide guidance. As a result, many teachers lamented that their final SLT score did not reflect the year student learning-wise that they had.

Instead of investing in competence-building training around the SLT process, many (but not all of) the studied districts instead centralized the assessment to be used for all teachers in attempt to exert some semblance of control over what must have seemed like disparate experiences across schools and their teachers. Unfortunately, the choice of several districts to centralize the decision over SLT assessments for Compass led to a sharp decline in the functional significance of the information it generated. As has been found in previous studies, autonomy and/or self-determination in the types/sources of data to be used for decision making enhances teachers' perceptions of the usefulness of the data for their work (Farrell and Marsh 2016a; Huguet et al. 2017). Teachers did not create (or choose) the district mandated SLT assessment, and thus it comes as no surprise that their orientation to these data sources was necessarily controlled or amotivated.

While competence and autonomy support were critical, prevalent issues for these Louisiana teachers, it is important to mention that perceptions of the role of relatedness support were conspicuously absent across our sample. At first glance, such a uniform absence of discussion from teachers on this point was somewhat surprising, but when taking into account the aggregate response to teacher evaluation data exhibited, it was less so. After all, much of the current evidence in this area highlights the ways in which collaboration can enhance or improve *on-going* data use efforts (Cosner 2011; Huguet et al. 2017; Marsh and Farrell 2015; Sun et al. 2016; Van Gasse et al. 2017). In this case, no forward momentum with respect to Compass-generated data use had materialized and that rendered any discussion or action around collaboration for data use moot. It is highly likely that—and there were hints of this in some of the data presented here—the high-stakes, competitive nature of the Compass data and process led some to be more guarded about their results, and thus likely precluded the emergence of teacher collaboration around using these data sources for improvement. This being said, there was substantial evidence of collaboration around DIDM with respect to the Common Core initiative, particularly in the later data collection waves, but this did not involve any of the Compass data sources.

A review of the extant literature suggested several different criteria that shape the functional significance of feedback for teachers: validity and utility of the data, as well as conditions that are supportive of competence, autonomy, and relatedness for teachers. Like many other data use studies, perceptions of the validity/utility of teacher evaluation data (or lack thereof) (Amrein-Beardsley and Collins 2012; Hewitt 2015; Jiang et al. 2015; Longo-Schmid 2016; Reddy et al. 2018; Rice and Malen 2016) were critical drivers of Louisiana teachers' responses to

their evaluation data. Perceptions of validity were multifaceted in our sample. In some cases, teachers had concerns about how state and district assessments would harm vulnerable students, while some questioned the credibility and/or fairness of the feedback. Drawing on their expertise and experience, special education and early childhood teachers in our sample questioned the appropriateness of the chosen assessments used for their SLTs, citing concerns about the long-term harm of such assessments on their students. These concerns necessarily undermined their perceived validity and thus their use for meaningful improvement.

This was compounded, in some cases, by the lack of experience of evaluators in evaluating teachers with more specialized roles in the school, such as special education teachers. Jones (2016) reported on the ways in which one-size-fits-all evaluation approaches fail to accurately capture the contributions of special education teachers and their teaching practice. Special education teachers in our sample too expressed frustration at having been evaluated by individuals who often were unfamiliar with their classrooms and/or their students and longed for an evaluator who was a colleague/mentor in their field.

6 Conclusion

As other scholars have noted, the literature on data use is largely atheoretical in nature (Farrell and Marsh 2016b). This study, it is hoped, sheds light on the ways in which theory (in this case self-determination theory) can be used as a lens for understanding why teachers respond in unintended ways to data that is generated, in large part, for their benefit. Interpreting behavior without the benefit of theory is ultimately harmful because it invites speculation as to why things are or are not working. This can lead to false assumptions about the root of the problem and ultimately misguided solutions. Indeed, it would be a mistake to assume that teachers responded how they did to their Compass evaluations because they have negative attitudes about evaluation or because they do not care about improving. Evidence from our exploration of the issue of teacher evaluation in Louisiana—as has been the case in other states/contexts—leads me to conclude that these assumptions could not be farther from the truth.

Instead, the case of Compass and the information that is generated as a result of its high-stakes approach to evaluation should be viewed as a cautionary tale of the subtle but powerful ways that the significance of information which is intended to help teachers improve can be undermined. In light of the findings here, it is certainly no wonder why there is skepticism in the educational community regarding the effectiveness of teacher evaluation to improve teacher capacity and student learning (Hallinger et al. 2014). Here again, policy and practice which is not mindful of the basic conditions under which individuals are optimally motivated to pay attention to and use performance feedback will fail to effect change in teaching and learning, and ultimately lead to wasted resources and stakeholder resentment and frustration. The good news is that, like all policies, improvements can be made, and, with some important changes, it may still be possible to recalibrate *Compass* and get teachers headed in the right direction.

Teacher Interview Protocols

Data Collection Wave I

1. TELL ME ABOUT YOURSELF: Could you describe your background in education and your current responsibilities? **Probes: Years of experience teaching, certification, grade level(s) taught, content area expertise, outside school-related involvement, school committees.**
2. **WHAT HAS IT BEEN LIKE** TRANSITIONING TO THE *CCSS* AT YOUR SITE? Tell me a little bit about the transition of your school/district to the *Common Core State Standards (CCSS).* What has been good about the transition? Where have the challenges been? **Probes: Opinions/orientations to the idea of *CCSS* (centralization of curricula across states); impact on students; developmentally appropriate practices; training/preparation/PD**
3. HOW **PREPARED** ARE YOU TO IMPLEMENT THE *CCSS*? What is the aspect of the *CCSS* that you feel most equipped to deal with/handle in the classroom? The least? Do you feel as though there is a system in place to support you in addressing this area of concern? **Probe the nature of their feelings of support more deeply if necessary.**
4. HOW DO YOU **FEEL** ABOUT Compass? Moving from *CCSS* to Compass, the new teacher evaluation system, what have been your initial impressions of this evaluation tool? (Reword: What do you see as the benefits of this evaluation system? What are the drawbacks of such an evaluation system in your opinion?) **Probes: Student Learning Targets; Value-Added Measures; observation rubric and scaling (modified Danielson rubric); training and support.**
5. Can you give an example of when you feel the most control over your teaching in terms of who you are as an individual and a professional? Can you give an example of when you feel the least control over your teaching?
6. Since the beginning of these two initiatives, have you ever found yourself questioning your identity either professionally or personally? Would you mind describing a good example of one these instances?

Data Collection Wave II

1) (Overall Background) Tell me a little bit about how this school year is going for you right now. What are some successes? Some challenges? **(Verify same teaching position as last year/same responsibilities. Probe any new responsibilities.)**
2) (District/School Context) Now that the *Common Core State Standards* are being fully implemented across the state, what has your school/district done differently this year to incorporate these standards into your curriculum? (Probes: adopt an already-developed system (e.g., Engage NY) or develop your own).
3) (*CCSS* Feelings activity) Have your feelings changed towards *CCSS* now that you have had some time to work with the standards? How would you describe your feelings about the *CCSS* using an image or a metaphor? Would you mind drawing writing your feelings for me? (Give them a separate sheet of paper with space to dra*w*/write).

4) (*CCSS* Support activity) Directions: Provide participant with the Support activity instrument. In the center of the circle the participant is to label "self" or "teacher." In each of the boxes with arrows pointing toward the center, participants are to write ***one type of support they are receiving in implementing the CCSS.*** In boxes that have arrows pointing away from the circle, participants are to write in ***feelings*** about the support they are receiving. Finally, interviewer is to ask the teacher to fill in the blank/answer the question "What do you feel is missing in this picture of support for *CCSS*?"
5) (Compass Follow-up) Can you tell me about your initial reaction to your *overall* Compass score? How are you dealing with your rating personally? How well do you feel your Compass evaluation and SLT/VAM score reflect your teaching? Will you share your overall rating with me? **(Probes: Student** Learning Targets; Value-Added Measures; observation rubric and scaling (modified Danielson rubric); training and support. Collect their COMPASS evaluation score from last year (ineffective, effective emerging, etc.) both overall and the subscores, if teacher willing to share).

Data Collection Wave III

1.) (Overall Background) Tell me a little bit about how this school year is going for you right now. What are some successes? Some challenges? **(Verify same teaching position as last year/same responsibilities. Probe any new responsibilities.)**
2.) (District/School Context) How has your district/school responded to the PARCC test from a curriculum standpoint? (**Probes:** Are new programs being implemented to better align with PARCC and/or are curriculum materials the same from last year?) Do you feel the conversation this year in your district/school has moved toward PARCC preparation or continued with curriculum alignment to the CCSS?
3.) (CCSS Feelings) Last time we met we discussed your feelings about the CCSS. Let us look at what you said then **(present written metaphor activity to teacher and use the following as guiding questions for unpacking the image)**. Are you still feeling this way? (**Probes:** Has the easing of VAM-based accountability changed your attitude toward the standards this year or the way you are teaching? Has the PARCC test added or taken away from your feelings (positive or negative) towards the CCSS?)
4.) (CCSS Support activity) Last time we met we also discussed the type of support you are getting from your school/district. (**Present prior written support activity image to teacher, and use the following as guiding questions for unpacking the image)**. How would you modify this image to add or take away support items? Have you received more/less support this year? Has the support focused changed as a result of the new testing guidelines?
5.) (COMPASS Follow-up) Would you discuss with me your informal evaluation from the fall? If you would, share with me an example of how you have used the information from your COMPASS evaluation last year to prepare for your evaluation/teaching this year. **(Probes: feelings about using data for improvement; training/support for improvement, collaboration around using data)**

(Also collect their COMPASS evaluation score from last year (ineffective, effective emerging, etc.) both overall and the subscores, if teacher willing to share.)

6.) (SLT Follow-up) As they have been emphasized at the state level this year, would you talk a little bit about your experience now with SLTs? How has this increased focus shaped your preparation and teaching this year? **(Probe: Have your feelings towards your SLT's (if applicable) changed? In what ways?; professional development/support; follow-up on results).**

References

Adams, C. M., Forsyth, P. B., Ware, J. K., & Mwavita, M. (2016). The informational significance of A-F school accountability grades. *Teachers College Record, 118*(7), 1–31 Retrieved from: http://www.tcrecord.org/Content.asp?contentid=20925. Accessed 15 Oct 2017.

Adams, C. M., Ford, T. G., Forsyth, P. B., Ware, J. K., Barnes, L. B., Khojasteh, J., Mwavita, M., Olsen, J. J., & Lepine, J. A. (2017). *Next generation school accountability: A vision for improvement under ESSA*. Palo Alto, CA: Learning Policy Institute.

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

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS-EVAAS) in the Houston independent School District (HISD): Intended and unintended consequences. *Educational Policy Analysis Archives, 20*(12) Retrieved from: http://epaa.asu.edu/ojs/article/view/1096.

Beaver, J. K., & Weinbaum, E. H. (2015). State test data and school improvement efforts. *Educational Policy, 29*(3), 478–503.

Blase, J., & Blase, J. (1999). Principals' instructional leadership and teacher development: Teachers' perspectives. *Educational Administration Quarterly, 35*(3), 349–378.

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

Bulletin 130. La. Admin. Code. tit. 28, pt. 147, §103 (2017). Retrieved from: http://www.doa.la.gov/osr/lac/28v147/28v147.doc

Bulletin 130. La. Admin. Code. tit. 28, pt. 147, §311 (2017). Retrieved from: http://www.doa.la.gov/osr/lac/28v147/28v147.doc

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294–320.

Chambers, J., de los Reyes, I. B., O'Neil, C. (2013). How much are districts spending to implement teacher evaluation systems? Case studies of Hillsborough County Public Schools, Memphis City Schools, and Pittsburgh Public Schools. (RAND working paper # WR-989-BMGF). Retrieved from https://www.rand.org/pubs/working_papers/WR989.html

Chow, A. P. Y., Wong, E. K. P., Yeung, A. S., & Mo, K. W. (2002). Teachers' perceptions of appraiser–appraisee relationships. *Journal of Personnel Evaluation in Education, 16*(2), 85–101.

Collins, C., & Amrein-Beardsley A (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1). Retrieved from https://www.tcrecord.org/Content.asp?ContentId=17291. Accessed 15 Oct 2017.

Cosner, S. (2011). Teacher learning, instructional considerations and principal communication: Lessons from a longitudinal study of collaborative data use by teachers. *Educational Management Administration & Leadership, 39*(5), 568–589.

Curry, K. A., Mwavita, M., Holter, A., & Harris, E. (2016). Getting assessment right at the classroom level: Using formative assessment for decision making. *Educational Assessment, Evaluation and Accountability, 28*(1), 89–104.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York, NY: Teachers College Press.

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), 4–13.

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

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record, 117*(4).

Datnow, A., & Park, V. (2014). *Data-driven leadership*. San Francisco: Jossey-Bass.

Datnow, A., Greene, J. C., & Gannon-Slater, N. (2017). Data use for equity: Implications for teaching, leadership, and policy. *Journal of Educational Administration, 55*(4), 354–360.

Deci, E. L., & Ryan, R. M. (2000). The "what" and the "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627–668.

Delvaux, E., Vanhoof, J., Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluation have an impact on professional development? A multilevel analysis. *Teaching and Teacher Education, 36*, 1–11.

Denzin, N. K. (2001). *Interpretive interactionism* (2nd ed.). Thousand Oaks, CA: Sage.

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading, and learning*. Washington, DC: National Council on Teacher Quality.

Dynarski, M. (2016, December 8). Teacher observations have been a waste of time and money. The Brookings Institution. Retrieved from https://www.brookings.edu/research/teacher-observations-have-been-a-waste-of-time-and-money/

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Farley-Ripple, E. N., & Buttram, J. L. (2014). Developing collaborative data use through professional learning communities: Early lessons from Delaware. *Studies in Educational Evaluation, 42*, 41–53.

Farrell, C. C. (2015). Designing school systems to encourage data use and instructional improvement: A comparison of school districts and charter management organizations. *Educational Administration Quarterly, 51*(3), 438–471.

Farrell, C. C., & Marsh, J. A. (2016a). Metrics matter: How properties and perceptions of data shape teachers' instructional responses. *Educational Administration Quarterly, 52*(3), 423–462.

Farrell, C. C., & Marsh, J. A. (2016b). Contributing conditions: A qualitative comparative analysis of teachers' instructional responses to data. *Teaching and Teacher Education, 60*, 398–412.

Ford, T. G., Van Sickle, M. E., & Fazio-Brunson, M. (2016). The role of "informational significance" in shaping Louisiana elementary teachers' use of high-stakes teacher evaluation data for instructional decision making. In K. K. Hewitt & A. Amrein-Beardsley (Eds.), *Student growth measures in policy and practice: Intended and unintended consequences of high-stakes teacher evaluations* (pp. 117–135). New York: Palgrave Macmillan.

Ford, T. G., Van Sickle, M. E., Clark, L. V., Fazio-Brunson, M., & Schween, D. C. (2017). Teacher self-efficacy, professional commitment and high-stakes teacher evaluation (HSTE) policy in Louisiana. *Educational Policy, 31*(2), 202–248.

Glover, T. A., Reddy, L. A., Kettler, R. J., Kurz, A., & Lekwa, A. J. (2016). Improving high-stakes decisions via formative assessment, professional development, and comprehensive educator evaluation: The school system improvement project. *Teachers College Record, 118*(14), 1–26.

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

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores*. Princeton, NJ: Educational Testing Service Retrieved from http://www.ets.org/Media/Research/pdf/PICANG14.pdf.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28.

Harris, D. N., & Herrington, C. D. (Eds.). (2015). Value added meets the schools: The effects of using test-based teacher evaluation on the work of teachers and leaders [special issue]. *Educational Research, 44*(2), 71–141.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1) Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17292. Accessed 15 Oct 2017.

Hewitt, K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: Undermined intentions and exacerbated inequities. *Education Policy Analysis Archives, 23*(76). Retrieved from). https://doi.org/10.14507/epaa.v23.1968.

Hewitt, K., & Amrein-Beardsley, A. (2016). Introduction: The use of growth measures for educator accountability at the intersection of policy and practice. In K. Hewitt & A. Amrein-Beardsley (Eds.), *Student growth measures in policy and practice: Intended and unintended consequences of high-stakes teacher evaluations* (pp. 1–25). New York: Palgrave Macmillan.

Honig, M. I., & Venkateswaran, N. (2012). School–central office relationships in evidence use: Understanding evidence use as a systems problem. *American Journal of Education, 118*(2), 199–222.

Huguet, A., Farrell, C. C., & Marsh, J. A. (2017). Light touch, heavy hand: Principals and data-use PLCs. *Journal of Educational Administration, 55*(4), 376–389.

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the "data-driven" mantra: Different conceptions of data-driven decision making. *Yearbook of the National Society for the Study of Education, 106*(1), 105–131.

Ingram, D., Louis, K. S., & Schroeder, R. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record, 106*, 1258–1287. Retrieved from: https://www.tcrecord.org/content.asp?contentid=11573. Accessed 15 Oct 2017.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher, 44*, 105–116.

Jones, N. D. (2016). Special education teacher evaluation: An examination of critical issues and recommendations for practice. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 63–76). New York: Teachers College Press.

Kelly, K. O., Ang, S. Y. A., Chong, W. L., & Hu, W. S. (2008). Teacher appraisal and its outcomes in Singapore primary schools. *Journal of Educational Administration, 46*(1), 39–54.

Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes and lessons from three urban districts. *American Journal of Education, 112*, 496–520.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234–249.

Larkin, D., & Oluwole, J. O. (2014, March). The opportunity costs of teacher evaluation: A labor and equity analysis of the TEACHNJ legislation. New Brunswick, NJ: New Jersey educational policy Forum. Retrieved from https://njedpolicy.files.wordpress.com/2014/03/douglarkinjosepholuwole-opportunitycostpolicybrief.pdf

Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*(1). Retrieved from https://www.tcrecord.org/Content.asp?ContentId=17294. Accessed 15 Oct 2017.

Lavigne, A. L., & Good, T. L. (2014). *Teacher and student evaluation: Moving beyond the failure of school reform*. New York: Routledge.

Lavigne, A. L., & Good, T. L. (2015). *Improving teaching through observation and feedback: Beyond state and federal mandates*. New York: Routledge.

Lipsky, M. (2010). *Street-level bureaucracy: Dilemmas of the individual in public service* (2nd ed.). Thousand Oaks, CA: Russell Sage Foundation.

Little, J. W. (2012). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education, 118*(2), 143–166.

Longo-Schmid, J. (2016). Teachers' voices: Where policy meets practice. In K. Kappler Hewitt & A. Amrein-Beardsley (Eds.), *Student growth measures in policy and practice* (pp. 49–71). New York: Palgrave Macmillan.

Louisiana Department of Education. (2012). Compass: Louisiana's path to excellence—Teacher evaluation guidebook. Baton Rouge, LA: Author.

Louisiana Department of Education (2013). 2013 Compass final report. Baton Rouge, LA: Author. Retrieved from: https://www.louisianabelieves.com/resources/library/compass. Accessed 30 April 2018.

Louisiana Department of Education (2014). 2013–2014 Compass annual report. Baton Rouge, LA: Author. Retrieved from: https://www.louisianabelieves.com/resources/library/compass. Accessed 30 April 2018.

Louisiana Department of Education (2015a). Teacher student learning targets. Retrieved from: https://www.louisianabelieves.com/resources/classroom-support-toolbox/teacher-support-toolbox/student-learning-targets. Accessed 30 April 2018.

Louisiana Department of Education (2015b). 2014–2015 Compass teacher results by LEA. Retrieved from: https://www.louisianabelieves.com/resources/library/compass. Accessed 30 April 2018.

Louisiana Department of Education (2016). 2015–2016 Compass teacher results by district. Retrieved from: https://www.louisianabelieves.com/resources/library/compass. Accessed 30 April 2018.
Louisiana House Bill 1033. (2010). Evaluation and Assessment Programs.
Lortie, D. (1975). *Schoolteacher: A sociological analysis*. Chicago: University of Chicago Press.
Mandinach, E. B. (2012). A perfect time for data-use: Using data-driven decision making to inform practice. *Educational Psychologist, 47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064.
Mandinach, E. B., Honey, M., Light, D., & Brunner, C. (2008). A conceptual framework for data driven decision making. In E. B. Mandinach & M. Honey (Eds.), *Data-driven school improvement: Linking data and learning* (pp. 13–31). New York: Teachers College Press.
Marques, J. F., & McCall, C. (2005). The application of interrater reliability as a solidification instrument in a phenomenological study. *The Qualitative Report, 10*(3), 439–462.
Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 114*(11), 1–48.
Marsh, J. A., & Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making: A framework for understanding capacity building. *Educational Management Administration & Leadership, 43*(2), 269–289.
Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). Making sense of data-driven decision making in education (RAND occasional paper #OP-170-EDU). Santa Monica, CA: RAND. Retrieved from: http://www.rand.org/pubs/occasional_papers/OP170.html
Marsh, J. A., McCombs, J. S., & Martorell, F. (2010). How instructional coaches support data-driven decision making: Policy implementation and effects in Florida middle schools. *Educational Policy, 24*, 872–907. https://doi.org/10.1177/0895904809341467.
Master, B. (2014). Staffing for success: Linking teacher evaluation and school personnel management in practice. *Educational Evaluation and Policy Analysis, 36*(2), 207–227. https://doi.org/10.3102/0162373713506552.
McLaughlin, M. W. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis, 9*, 171–178.
Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level: From accountability to instructional improvement*. Washington, DC: U.S. department of Education Retrieved from: https://www2.ed.gov/rschstat/eval/tech/use-of-education-data/use-of-education-data.pdf
Milanowski, A. T., & Heneman, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education, 15*(3), 193–212.
Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: Sage.
Murphy, J., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation: The case of the missing clothes? *Educational Researcher, 42*, 349–354. https://doi.org/10.3102/0013189X13499625.
Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education, 7*, 133–144.
Organization for Economic Co-Operation and Development. (2009). Teacher evaluation. A conceptual framework and examples of country practices. Retrieved from: http://www.oecd.org/edu/school/44568106.pdf
Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*, 163–193.
Park, V., Daly, A. J., & Guerra, A. W. (2013). Strategic framing: How leaders craft the meaning of data use for equity and learning. *Educational Policy, 27*(4), 645–675.
Reddy, L. A., Dudek, C. M., Peters, S., Alperin, A., Kettler, R. J., Kurz, A. (2018). Teachers' and school administrators' attitudes and beliefs of teacher evaluation: A preliminary investigation of high poverty school districts. *Educational Assessment, Evaluation, and Accountability, 30*, 47–70.
Rice, J. K., & Malen, B. (2016). When theoretical models meet school realities: Educator responses to student growth measures in an incentive pay program. In K. Kappler Hewitt & A. Amrein-Beardsley (Eds.), *Student growth measures in policy and practice* (pp. 29–47). Palgrave Macmillan US.
Rosenholtz, S. J. (1991). *Teachers' workplace: The social organization of schools*. New York: Teachers College Press.
Ryan, R. M., & Brown, K. W. (2005). Legislating competence: The motivational impact of high-stakes testing as an educational reform. In C. Dweck & A. Elliot (Eds.), *Handbook of competence and motivation* (pp. 354–372). New York: Guilford Press.
Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory: An organismic dialectical perspective. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3–33). Rochester, NY: University of Rochester Press.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. New York: Guilford Press.

Ryan, R. M., & Weinstein, N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory and Research in Education, 7*(2), 224–233. https://doi.org/10.1177/1477878509104327.

Schildkamp, K., & Visscher, A. (2010). The use of performance feedback in school improvement in Louisiana. *Teaching and Teacher Education, 26*(7), 1389–1403.

Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement, 28*(2), 242–258.

Schneider, A., & Ingram, H. (1990). Behavioral assumptions of policy tools. *The Journal of Politics, 52*(2), 510–529.

Skrla, L., Scheurich, J. J., Garcia, J., & Nolly, G. (2004). Equity audits: A practical leadership tool for developing equitable and excellent schools. *Educational Administration Quarterly, 40*(1), 133–161.

Sun, M., Mutcheson, R. B., & Kim, J. (2016). Teachers' use of evaluation for instructional improvement and school supports for such use. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 169–183). New York: Teachers College Press.

The Joint Committee on Standards for Educational Evaluation [JCSEE]. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators*. Thousand Oaks, CA: Corwin.

The New Teacher Project. (2010). Teacher evaluation 2.0. New York: Author.

Tuytens, M., & Devos, G. (2011). Stimulating professional learning through teacher evaluation: An impossible task for the school leader? *Teaching and Teacher Education, 27*(5), 891–899.

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

Van Gasse, R., Vanlommel, K., Vanhoof, J., & Van Petegem, P. (2017). The impact of collaboration on teachers' individual data use. *School Effectiveness and School Improvement, 28*, 1–16.

Vansteenkiste, M., Lens, W., De Witte, H., & Feather, N. T. (2005). Understanding unemployed people's job search behavior, unemployment experience and well-being. A comparison of expectancy-value theory and self-determination theory. *British Journal of Social Psychology, 44*, 269–287.

Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist, 41*(1), 19–31.

Watt, H. M. G., & Richardson, P. W. (2014). Why people choose teaching as a career: An expectancy-value approach to understanding teacher motivation. In P. W. Richardson, S. A. Karabenick, & H. M. G. Watt (Eds.), *Teacher motivation: theory and practice* (pp. 3–19). London: Routledge.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). In) (Ed.), *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

Wigfield, A., & Eccles, J. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12*, 265–310.

Yin, R. K. (2017). *Case study research and applications: Design and methods* (5th ed.). Thousand Oaks, CA: SAGE publications.

Young, V. M. (2006). Teachers' use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education, 112*, 521–548.