Data Use For Instructional Improvement:

Tensions, Concerns, And Possibilities

For Supporting Ambitious And Equitable Instruction

By

Brette Garner

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Learning, Teaching & Diversity

May 11, 2018

Nashville, Tennessee

Approved:

Ilana S. Horn, Ph.D.

Barbara Stengel, Ph.D.

Rogers Hall, Ph.D.

Joanne Golann, Ph.D.

To educators, who strive to meet superhuman expectations

and

To students, whose brilliance cannot be captured by any single measure

ACKNOWLEDGEMENTS

The National Science Foundation and the Spencer Foundation provided financial support for this research. But this study would not be possible without the academic and emotional support of great teachers. I am forever grateful to Lani Horn, who is my advisor, cheerleader, coach, mentor, co-conspirator, and academic wingwoman. I owe so much to Miss Jackie Griffin; without her constant encouragement (and hassling), I would still be writing my MAP. Other Vanderbilt faculty — especially my committee (Barb Stengel, Rogers Hall, and Joanne Golann), as well as Peg Cagle, Nicole Joseph, Marcy Singer-Gabella, Teresa Dunleavy, Ebony McGee, and Rich Lehrer — have challenged my assumptions, pushed my thinking, and made me a better scholar. I also thank my teachers at The University of Texas, including Walter Stroup and Lupita Carmona for helping me critique standardized testing, Lisa Brown for exploring new possibilities in assessment, Uri Treisman for connecting policy and instruction, Cathy Seeley for empowering me and other young women in math education, Carolyn Eastman for making me argue a point, and Martha Norkunas for teaching me to honor others' stories. And I am especially thankful for my teachers at Mercer Elementary and Turpin High School — Jill Martin, Sylvia Elam, Barry Riehle, Ken Poe, and the late Trudi Boyd — who became my first role models of good teachers.

I am also indebted to my family members, who have made me who I am. The original Drs. Garner — my grandfather (Gary) and father (Brad) — encouraged me to follow in their academic footsteps. My mother (Linda) shows me how to be strong, ambitious, and balanced. My sister-friends, Brooke and Brie, are my fiercest supporters and strongest teammates. My aunts, uncles, and cousins are sources of love, encouragement, and support. And my late grandparents — Jim, Verdell, and Maryellen — continue to inspire me every day.

I am lucky to have a cadre of supportive friends and chosen family. My partner, Hope, has been my most trusted friend and confidante throughout the longest nights and darkest days of dissertating. My Peabody friends — Grace, Sammie, Nadav, Patty, Tess, Kate, Fai, Lara, Megan, Portia, Lauren, Danielle, Nick, Brooks, Dan, Ben, Amanda, Charlotte, Grant, Brtinie, Jonee, Jennifer, Amy, Robert, Leah, et al. — have built a caring and supportive grad student community. It is a joy to learn from and work with people whom I truly adore. My Nashville friends — including Abel, Graham, Jared, Steph, Meghsha, and the rest of the QDP, HotMess, DAM Band, and NFC communities — have kept me grounded and remind me to have as much fun as possible. And my Texas friends, who now live all across the country — Joe, Teater, Todd, Sarah, Stephanie, John, Melody and Sam, Mel and TJ, Emily, Amanda, Marcus and Chong-Hao, Pegah, Jenni and Bryan, Adam and Amelia, Shannah, Ingrid, and so many others — continue to enrich my life in more ways than I can count.

My scholarship is driven by a desire to improve math teaching and learning, and to re-humanize math classes for both teachers and students. This would not be possible without the teachers and coaches who participated in this study; they shared their time, energy, and expertise so that we can learn from them. And I would not be able to do this work without my former middle-school colleagues and students. My colleagues (especially Ginger, Rebecca, Eileen, Dane, and James) provided support and guidance as we navigated a flawed system, attempting to do our best for all of our students. In turn, my students taught me as much about pedagogy as I taught them about math. Their challenges, frustrations, and achievements expanded my understanding of what it means to do math, to be smart, and to be successful. I strive to honor them through this dissertation. It is, in many ways, an attempt to talk back to the system that failed my students, colleagues, and participants in order to imagine a more humane future.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

For nearly two decades, the U.S. education landscape has been dominated by test-based accountability policies, including the No Child Left Behind (NCLB) Act of 2001, the Race to the Top initiative of 2009 (RttT), and the Every Student Succeeds Act of 2016 (ESSA). In a purported effort to increase equity in students' educational outcomes, states hold teachers and schools accountable for student achievement on standardized assessments. Within this context, the U.S. Department of Education has promoted Data-Driven Decision-Making (DDDM) has emerged as a key strategy for supporting educational reform through test-based accountability (Duncan, 2009; Hamilton, Halverson, Jackson, Mandinach, & Supovitz, 2009; Means, Padilla, & Gallagher, 2010; National Forum on Education Statistics, 2011). The basic idea behind DDDM is that, by measuring student learning with a test, teachers can identify gaps in students' knowledge, and then make changes based on the data. Over time, this process will help students make progress toward their achievement goals, and schools will avoid the sanctions associated with a failure to make adequate progress.

But this process leaves a number of important questions unspecified: What counts as data? Which data are emphasized, and for which students? How do educators draw conclusions from data? What instructional responses are appropriate, based on the data? The answers to these questions have serious implications for instructional choices students' learning opportunities, and yet they vary widely across different contexts (Datnow, Park, & Kennedy-Lewis, 2012). In some cases, educators use data in ways that support instructional improvement and student learning; many successful school turnaround efforts cite data use as an important piece of their reforms (e.g., Schaffer, Reynolds, & Stringfield, 2012; Villavicencio & Grayman, 2012). Yet in many other settings, data use distorts teaching and learning, as teachers attempt to raise test scores by teaching to the test, narrowing the curriculum, and emphasizing the success of some students at the expense of others (Booher-Jennings, 2005; Jennings & Bearak, 2014; Lee, Louis, & Anderson, 2012). Rather than ameliorate educational inequities, such distortions often have the effect of exacerbating systemic racism and classism at the district and school levels (e.g., Horn, 2016; Khalifa, Jennings, Briscoe, Oleszweski, & Abdi, 2013).

The varied results of data use efforts point to the importance of understanding the details of educators' data use practices — the ways that they select, analyze, and make sense of data. Yet scholars have recently noted that the research base on educators' data use practices is thin (Coburn & Turner, 2011; Little, 2011). In this dissertation, I synthesize and build on this emerging literature in order to investigate the tensions between test-based accountability policies and instructional improvement projects. I focus the empirical portions of this study in the heavily-tested area of middle-school mathematics, where the effects of accountability policies are especially strong.

In Paper 1, I review the literatures on teachers' data use in practice and the implementation of test-based accountability policies in order to articulate the ways that test-based accountability policies distort the work of teaching and learning. I identify ten such distortions, which I call Distortive Data Use Practices in Education (DDUPEs). Even though data use has frequently been implemented in a distortive way, research on ambitious instruction (e.g., Lampert, Boerst, & Graziani, 2011) suggests that high-quality instruction requires that teachers use evidence of student learning to inform teaching practice. As an alternative to DDUPEs, I propose a set of Responsive Evidence Use Practices in Education (REUPEs), which are practices for using evidence to that are likely to support more ambitious and equitable instruction. This review informs the two empirical papers, which are case studies of teacher workgroups' data use in practice.

In Paper 2 (Garner, Thorne, Horn, 2017), we examined the ways in which the logic of test-based accountability policies influenced a middle-school mathematics teacher workgroup's data use practices. We found that three effects of accountability policies — namely, reducing complex constructs to quantitative variables, valuing remediation over instructional improvement, and enacting faith in instrument validity — distorted teachers' data use practices in ways that precluded the development of more ambitious and equitable instructional practices. This analysis bridges the policy-practice divide by analyzing the unintended consequences of test-based accountability policies, particularly with respect to the inequities that they seek to redress.

In Paper 3, I investigate the data use practices of two other middle-school mathematics teacher workgroups in order to gain deeper understanding of the ways that their epistemic stances on data shape their data use practices, evidence of student learning, and instructional responses. Data use is an inherently epistemic endeavor, as teachers negotiate what they can know, what is worth knowing, and what is possible to know about student learning. I juxtapose two similar workgroups — each in schools facing great accountability pressure, supported by accomplished instructional coaches, and engaging in concerted efforts to use data to inform instruction — that use different epistemic stances on data. Educators' notions about the ontological nature of data (i.e., what data represent) greatly shape their work, yet are typically engaged as tacit assumptions in workgroup conversations. I find that educators who use assessment data as an *indicator* of student learning are better positioned to use data for instructional improvement than those who use data as a *measurement* of student learning. This analysis highlights a foundational, yet little-researched, element of educators' data use practices.

Collectively, the papers that make up this dissertation add deeper nuance and understanding to issues that are particularly salient in the field of mathematics education, given the role of mathematics as a "gatekeeper" subject (Martin, Gholson, & Leonard, 2010) and its status as a frequently assessed content area under accountability policies. Data — particularly standardized test scores — have long been used as tools of white supremacy and to marginalize certain racial and ethnic groups (Gould, 1996). I am cognizant of this influence in current educational settings, and critique the system of test-based accountability policies accordingly (e.g., Garner et al., 2017).

An important implication of this series of analyses the development of more productive data use practices. Despite the various calls for data-driven decision-making, there are few efforts to prepare teachers to engage in nuanced discussions of data, or to prepare instructional coaches to facilitate those conversations (Mandinach & Gummer, 2013). By identifying potential pitfalls of data use efforts (e.g., DDUPEs) and articulating possibilities for data use that supports instructional improvement, this dissertation can inform future research into educators' development of productive data use practices.

References

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

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. Measurement: Interdisciplinary Research & Perspective, 9(4), 173-206

Datnow, A., Park, V., & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. Journal of Education for Students Placed at Risk (JESPAR), 17(4), 247-265.

Duncan, A. (2009, June). Robust Data Gives Us The Roadmap to Reform. Speech made at the Fourth Annual IES Research Conference, Washington, DC. Retrieved from https://www.ed.gov/news/speeches/robust-data-gives-us-roadmap-reform

Garner, B., Thorne, J. K., & Horn, I. S. (2017). Teachers interpreting data for instructional decisions: where does equity come in?. Journal of Educational Administration, 55(4), 407-426.

Gould, S. J. (1996). The mismeasure of man. WW Norton & Company.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., Pickens, C., Martin, E.S., & Steele, J. L. (2009). Using student achievement data to support instructional decision making.

Horn, I. S. (2016). Accountability as a design for teacher learning: Sensemaking about mathematics and equity in the NCLB era. Urban Education, 0042085916646625.

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. Educational Researcher, 43(8), 381-389.

Khalifa, M. A., Jennings, M. E., Briscoe, F., Oleszweski, A. M., & Abdi, N. (2013). Racism? Administrative and community perspectives in data-driven decision making: Systemic perspectives versus technical-rational perspectives. Urban Education, 0042085913475635.

Lampert, M., Boerst, T. A., & Graziani, F. (2011). Organizational resources in the service of schoolwide ambitious teaching practice. Teachers College Record, 113(7), 1361-1400.

Lee, M., Louis, K.S., & Anderson, S. (2012). Local education authorities and student learning: The effects of policies and practices. School Effectiveness and School Improvement, 23(2), 133-158.

Little, J. W. (2011). Understanding data use practice among teachers: The contribution of micro-process studies. American Journal of Education, 118(2), 143-166

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. Educational Researcher, 42(1), 30-37.

Martin, D. B., Gholson, M. L., & Leonard, J. (2010). Mathematics as gatekeeper: Power and privilege in the production of knowledge. Journal of Urban Mathematics Education, 3(2), 12-24.

Means, B., Padilla, C., & Gallagher, L. (2010). Use of education data at the local level: From accountability to instructional improvement. US Department of Education.

National Forum on Education Statistics. (2011). Traveling Through Time: The Forum Guide to Longitudinal Data Systems. Book Four of Four: Advanced LDS Usage (NFES 2011-

802). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Schaffer, E., Reynolds, D., & Stringfield, S. (2012). Sustaining turnaround at the school and district Levels: The high reliability schools project at Sandfields Secondary School. Journal of Education for Students Placed at Risk (JESPAR), 17(1-2), 108-127.

Villavicencio, A., & Grayman, J. K. (2012). Learning from “turnaround” middle schools: Strategies for success. New York: Research Alliance for Ne*w York City Schools*.

CHAPTER II

DATA USE AND TEST-BASED ACCOUNTABILITY: CONTRADICTIONS, CONCERNS, AND POSSIBILITIES FOR INSTRUCTIONAL IMPROVEMENT

For decades, U.S. policymakers, educators, and the general public have expressed concern for the state of American education, particularly in STEM fields. Legislators and business leaders are concerned that American students are falling behind their international peers and therefore will not be competitive in the labor market or in the race for technological innovations. School reform advocates are concerned that deep inequities in our education system have resulted in “failing”[1] students, schools, and districts, particularly in underserved urban and rural communities. In response to these and other worries, various stakeholders have called for educators to make teaching and learning more rational and scientific by using data to tailor instruction and by using assessments to hold schools and teachers accountable to student performance. In theory, such data-driven decision-making (DDDM) efforts will result in higher student achievement and greater educational equity.

Yet the details of this process are left unspecified: There is no clear or straightforward path that leads from assessment data to student achievement. DDDM efforts are often modeled by generic processes that outline a sequence of steps: Educators collect, organize, analyze, and synthesize data, which point to a change that should be implemented (e.g., Mandinach, 2012).

[1] Even though policymakers, district leaders, and school leaders use test results to identify people and institutions as “failing,” I note that the failure is on the part of the education system, rather than on the students and teachers within the system. Namely, the U.S. education system has a long history of failure to provide funding and resources for schools with poor students and students of color; Ladson-Billings (2009) refers to this broadly as the “education debt.” Furthermore, as I articulate below, test-based accountability policies have failed to appropriately assess students from historically marginalized groups. These failures have reified and upheld the very systems of oppression that our education system purportedly seeks to redress.

Such representations collapse a set of complicated practices into neat and tidy boxes, thereby rendering invisible much of the interpretive work of using evidence to inform instruction (cf. Suchman, 1995). As a result, many aspects of educators' data use vary tremendously across contexts (Datnow & Hubbard, 2015): What counts as data? Which data are emphasized, and for which students? How do educators draw conclusions from data? What instructional responses are appropriate? Educators' answers to questions like these are consequential for instructional practice and students' learning opportunities.

Calls for DDDM have arisen in tandem with a movement toward test-based accountability policies. But the meaning of *data* has shifted as a result of accountability pressures. Within the assessment system that was institutionalized under the No Child Left Behind (NCLB) Act of 2001 — and that continues under the Every Student Succeeds Act (ESSA) of 2016 — the term *data* is typically used to refer to quantitative data (Marsh, Pane, & Hamilton, 2006). Standardized assessments (and tests meant to mimic standardized assessments) have become the primary metrics against which students, teachers, and schools are judged. These assessments — namely, state end-of-year tests and district interim benchmark assessments — are composed of primarily multiple-choice items and thus provide "objective" quantifiable results. This allows policymakers, district and school leaders, parents, and other stakeholders to compare and measure school success, teacher quality, and student learning. Within this system, other evidence of student learning — including qualitative information like student work or classroom conversations — is no longer ratified as *data*.[2]

[2] Many scholars and educators consider the term data to include qualitative and quantitative information about student learning. Indeed, I use the broader sense of data elsewhere (i.e., Chapters III and IV). But for the present analysis, I reserve the term data for quantitative information (most often, results from multiple-choice assessments) and I use the term evidence to refer to the more inclusive sense of data (e.g., student work as well as multiple-choice assessment results).

To further complicate the implementation of DDDM under test-based accountability policies, the narrow meaning of *data* is at odds with what is often considered good instructional practice. Using evidence of student learning to inform instruction is not a novel concept; thoughtful teachers have always considered students' prior knowledge and skills as they planned for instruction (Black & Wiliam, 1998). Throughout the school day, teachers constantly assess student learning in formal and informal ways (Jordan & Putz, 2004): They ask students questions during class discussions, examine students' work, observe students' public problem solving, and administer written assessments both large (e.g., unit tests) and small (e.g., daily exit tickets). These sources of information provide teachers with rich evidence about what students know and are able to do. Good teaching — what many scholars have come to call "ambitious" teaching — relies on teachers' ability to use such evidence to inform instruction (Lampert, Boerst, & Graziani, 2011).

But using a restrictive notion of *data*, these sources of evidence might not be considered sufficiently valid, reliable, or objective to use as part of DDDM. Adhering strictly to DDDM could lead teachers to disregard or overlook important evidence of student thinking that is not captured by a formal assessment. In this way, DDDM conflicts with ambitious instruction. This tension is important in that it points to a flaw in the enactment of test-based accountability policies: The purported goal of DDDM is to support student learning, and ambitious instruction is widely considered to be an effective pedagogy for supporting student learning. Yet DDDM (as enacted in the context of test-based accountability policies) and ambitious instruction rely on different meanings and uses of *data*, which differentially shape pedagogical choices and therefore students' learning opportunities. The result is that DDDM generally distorts teaching and learning in ways that are unlikely to prevent instructional improvement or deeper student learning.

In this paper, I review the literatures on teachers' data use in practice and the implementation of test-based accountability policies in order to examine these contradictions between DDDM (as enacted in the context of test-based accountability policies) and ambitious instruction. It is important that we recognize and acknowledge these points of tension between accountability policies and ambitious teaching in order to avoid them. To that end, I identify ten Distortive Data Use Practices in Education (DDUPEs), which are the ways that DDDM has distorted the work of teaching and learning. But from the literature, I also identify ways that educators can move beyond the distortions of test-based accountability policies to more productive uses of evidence. As an alternative to DDUPEs, I propose three Responsive Evidence Use Practices in Education (REUPEs), which are ways that educators can use data to respond to student thinking and support deeper learning. REUPEs are evidence use practices that have the potential to support instructional improvement. This analysis builds on the existing data use literature to highlight the ways that test-based accountability policies and associated DDDM efforts distort teaching and learning, as well as the ways that educators can resist such distortions by using evidence of student learning to support instructional improvement.

The Naive Optimism of Test-Based Accountability

Since the enactment of NCLB, states have been required to develop and administer annual standardized tests in order to assess student learning with respect to state standards. Many states began developing their own standards and tests prior to 2001, but these assessment systems were institutionalized through NCLB's mandate. Since then, the U.S. Department of Education has promoted DDDM as a key strategy for using assessments to drive educational reform (Duncan, 2009; Hamilton, Halverson, Jackson, Mandinach, & Supovitz, 2009; Means, Padilla, & Gallagher, 2010; National Forum on Education Statistics, 2011). District leaders, school leaders, and teachers are expected to use assessment data in their efforts for school improvement (Hamilton et al., 2009). The logic of these efforts is fairly straightforward (Figure II.1): If states measure students' learning with standardized assessments, they will be able to identify trouble spots (e.g., "failing" schools) and gaps in student outcomes (e.g., achievement differences between affluent students and poor students). Educators and district leaders can use the data from state tests to make changes that will gradually improve student learning. Schools will be invested in making such changes and improving student learning outcomes to avoid sanctions for failing to meet accountability goals.

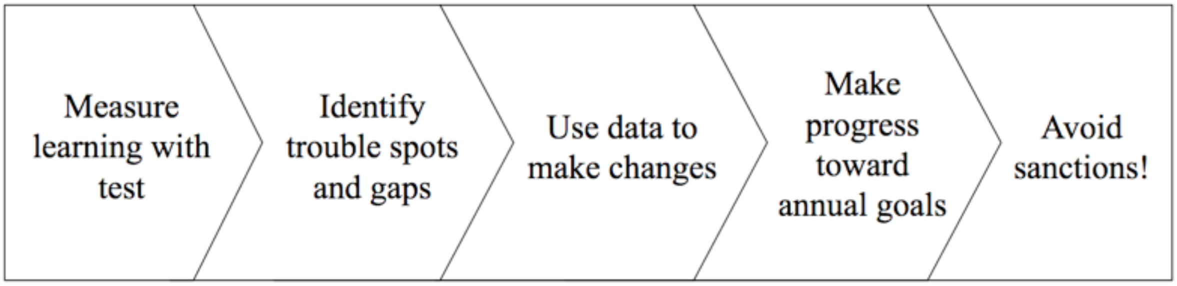

Measure learning with test
Identify trouble spots and gaps
Use data to make changes
Make progress toward annual goals
Avoid sanctions!

Figure II.1. Logic of test-based accountability policies.

Yet this is a naive and simplistic approach to assessment and educational reform that overlooks the myriad complicated reasons for educational inequities; simply using data does not always support increases in student achievement (e.g., West, Morton, & Herlihy, 2016). Even though data use is an important strategy for many successful reform efforts, it must be part of a coherent approach that also includes ambitious goals for instruction and teachers' professional development in order to support student learning (Ezzani, 2015; McNaughton, Lai, & Hsiao, 2012). Without accounting for the intricacies and nuances of teaching and learning, test-based accountability policies — and the resulting instantiation of DDDM — are unlikely to promote instructional improvement, student learning, or educational equity. Instead, increased accountability pressure has resulted in Distortive Data Use Practices in Education (DDUPEs), as educators attempt to increase student achievement results without improving instruction.

Table II.1. The logic of test-based accountability policies and the resulting DDUPEs.

Test-Based Accountability Logic	**Distortive Data Use Practices in Education (DDUPEs)**
Measure student learning with standardized test	1. Educators prioritize standardized test data over other evidence of student learning. 2. The tests are biased against students from historically marginalized groups
Identify trouble spots and gaps	3. Educators engage in educational triage by focusing instruction on frequently tested topics 4. Achievement gap rhetoric reinforces deficit orientations toward "failing" students and their schools
Use data to make changes	5. Data-use efforts focus on the logistics of data use, not educators' interpretations 6. Test-based accountability policies prioritize the outcomes of instruction, no matter what yields them
Make progress toward annual goals	7. Policymakers set goals to fit linear growth trajectories 8. Policies specify goals for student achievement based on narrow measures, ignoring other signs of progress
Avoid sanctions	9. Schools cultivate the appearance of achievement without increasing student learning 10. The public compares schools based on assessment results

In this section, I outline the ways in which the logic of test-based accountability policies results in DDUPEs (Table II.1). To identify the shortcomings of test-based accountability policies, I reviewed the literature on the implementation of accountability policies in schools, as well as the literature on educators' data use practices. Using the EBSCO databases, I searched for terms including "accountability policy," "data use," and "school turnaround." I also reviewed special issues of major education journals (e.g., *Teachers College Record* and the *American*

Education Research Association Journal) that focused on test-based accountability policies and educators' data use since the implementation of NCLB.

Measure Student Learning with Standardized Assessments

Test-based accountability policies and DDDM arose from concerns that U.S. students were not learning enough in schools (Nichols & Berliner, 2007), extending the logic of the landmark document *A Nation at Risk* (1983). To oversee how much students were learning, states developed and administered end-of-year assessments to measure student learning. Thus this trend began long before the 2000s — and even has ties to the eugenics movement (Gould, 1996) — NCLB specifically required that all states administer high stakes assessments in grades 3-8 and at least once in high school. Much like a doctor might measure children's height, weight, or body temperature, state departments of education began to measure students' knowledge, particularly in mathematics and English language arts. Because annual testing does not provide educators feedback about their current students' learning, many districts also developed or purchased additional interim assessments to administer as a way to monitor students' progress toward end-of-year goals. These interim assessments, which mimic state tests in form and content, have become one of the primary sources of data that teachers, coaches, and principal use throughout the school year (Datnow & Hubbard, 2015).

Measuring student learning with high-stakes and interim assessments creates a DDUPE as *educators prioritize standardized test data over other evidence of student learning.* Though teachers have access to a variety of data on student learning, test-based accountability policies ratify students' performance on standardized tests as the only legitimate metrics of students' knowledge and skills. This distorts teachers' data use, because standardized test items — and the interim assessments meant to mimic them — typically emphasize procedural skills over conceptual understanding or mathematical practices (Yuan & Le, 2012). As a result, teachers tend to tailor their instruction to reflect the content and level of rigor that most often appears on high-stakes exams (Jennings & Bearak, 2014). This has effectively narrowed mathematics curriculum to emphasize procedural knowledge and students' test-taking abilities.

This practice creates further distortion as *the tests are biased against students from historically marginalized groups*, including Black, Latinx, and Indigenous students, low-income students, students with disabilities, and emergent bilingual students (Au, 2009; Pham, 2009). Mathematics assessments often include word problems, which require familiarity with the linguistic and contextual features of the items in addition to the mathematical content (Helms, 1992). Questions that seem innocuous from a White, middle-class perspective may subtly reinforce systemic racism and classism through colorblind ideology (Battey & Leyva, 2016). And even the "normalizing" process of psychometric validation can systematically select for assessment items that Black and Latinx students tend to answer incorrectly more often than White students (Kidder & Rosner, 2002). These factors combine to (re)produce a skewed understanding of student achievement; assessment results may more accurately reflect students' race and class than their mathematical knowledge.

Identify Trouble Spots and Gaps

One of the stated goals of NCLB was to end the "soft bigotry of low expectations" by identifying groups of students who are being underserved by schools and holding schools accountable for their success (Bush, 2000). To that end, standardized assessments have exposed "our nation's dreadful achievement gaps" (Duncan, 2009). Of particular concern is the relatively low performance of Black, Latinx, and Indigenous students, poor students, students with disabilities, and emergent bilingual students on these measures. Under NCLB, RttT, and ESSA, schools and districts are held accountable to ensuring sufficiently high passing rates for each of these subgroups of students. Ostensibly, this ensures that educators work to promote all students' learning, thereby promoting greater equity.

Yet these accountability criteria create a DDUPE as *educators engage in educational triage*, focusing on specific students at the expense of others; this dramatically reduces equity of learning opportunities. Many researchers have noted the emergent category of "bubble kids," or those just on the cusp of proficiency (e.g., Booher-Jennings, 2005). Rather than adjusting instruction to meet the needs of all students, some educators distort the purpose of assessments, using them as a magnifying glass to focus intense light on the students who will count the most for accountability purposes. Bubble kids often receive the bulk of this attention, since a bit of extra tutoring might push them over the bar thereby helping the school meet its performance goals. Some educators focus even more specifically on bubble kids in the specific subgroups that prevented their school from meeting accountability goals, like the low-income bubble kids or the Black bubble kids (Horn, 2016). Students who are deemed unimportant for accountability purposes — because they are very likely or very unlikely to pass, because they transferred to the school late in the school year, or because their subgroup tends to score well — are systematically left out of such learning opportunities.

Furthermore, the rhetoric of the achievement gap creates a DDUPE as it *reinforces deficit orientations toward "failing" students and their schools.* As Gloria Ladson-Billings (2006) and others have stated, the notion of the achievement gap attributes low scores to students and their teachers, rather than on the systemic and long-standing underfunding and under-resourcing of schools and communities with large populations of people of color. The gap, she argues, might more appropriately be referred to as an "education debt." There is also a selective attention to the gaps that fulfill long-standing societal narratives that position monolingual White students as smarter and more capable than emergent bilingual students, Black, Latinx, and Indigenous students. Administrators do not, for instance, fret about White students' underperformance in mathematics with respect to Asian-American students, or monolingual students' lack of language proficiency as compared to bilingual students (Gutiérrez, 2008). In these ways, test-based accountability policies encourage achievement gap rhetoric that supports racist and classist notions of academic success.

Use Data to Make Changes

In order to improve student outcomes, teachers are expected to use assessment data (particularly from state and benchmark assessments) to inform their instruction. DDDM has become a "mantra" for schools (Ikemoto & Marsh, 2007). For instance, the Regional Education Laboratory identified "ongoing data use for school improvement" as a key practice associated with high performance on state assessments (Weinstock, Yumoto, Abe, Meyers, & Wan, 2016). Indeed, those involved in successful turnaround efforts often cite teachers' use of data as a key strategy for increasing student achievement (e.g., Schaffer, Reynolds, & Stringfield, 2012; Villavicencio & Grayman, 2012). To that end, most schools receiving School Improvement Grants pushed substantial resources into supporting teachers' data use (Le Floch, O'Day, Birman, Hurlburt, Nayfack, Halloran, Boyle, Brown, Mercado-Garcia, Goff, Rosenberg, & Hulsey, 2016).

The taken-for-granted mantra of DDDM creates yet another DDUPE by *focusing on the logistics of data use, rather than the processes*. Data use is presented as a straightforward, rational — even scientific — endeavor: If teachers have enough data, a clear and appropriate response will reveal itself. But as studies across various contexts indicate, data use is a complicated interpretive endeavor that requires substantial pedagogical and professional judgment (Datnow, Park & Kennedy-Lewis, 2012; Goldstein & Hall, 2007; Wayman, Jimerson, & Cho, 2012). DDDM efforts leave many important questions unanswered: What data should be used and by whom? What kinds of data sets are adequate for what kinds of decisions? How should educators make sense of different data, and what conclusions should they draw? How should they respond instructionally? This lack of clarity is only worsened by educators' own lack of preparation for this task. Pre-service teachers generally receive little preparation for effective DDDM (Mandinach & Gummer, 2013). Many professional development efforts for in-service teachers focus on the logistics of accessing data, rather than the more complicated work of interpreting data (Jimerson & Wayman, 2015).

A corollary to this DDUPE is that *test-based accountability policies prioritize the outcomes of instruction*. Weiss (2012) notes that there are competing goals in using data for instructional improvement and using data for accountability purposes. Test-based accountability policies emphasize the latter, which leads to numerous unintended consequences. When faced with accountability pressure, educators often work to increase test scores rather than actually supporting student learning (Diamond & Cooper, 2007; Horn, Kane, & Wilson, 2015; Jennings, 2012; Lee et al., 2012). In the end, accountability policies do not support teacher learning; there are no provisions for professional development, instructional coaching, or any other resource that would support instructional improvement. This undermines the purported improvement goals of test-based accountability policies, as teachers turn to test preparation strategies and educational triage rather than instructional improvement.

Make Progress Toward Annual Goals

One of the key features of NCLB was the establishment of incremental achievement goals, known as Adequate Yearly Progress (AYP). Each year, schools' accountability targets increased slightly to promote gradual improvement, both for the overall student body and for each subgroup of students. Schools that failed to meet AYP targets in consecutive years were identified for interventions to help get them back on track, including school turnaround, transformation, or closure (Le Floch et al., 2016). The steady increase in AYP targets set a goal for all schools to achieve 100% proficiency by 2014 (NCLB, 2002).

A common critique of NCLB's AYP goals is that they are arbitrary and unrealistic (Darling-Hammond, 2007). This arbitrariness creates a DDUPE as *policymakers set goals to fit linear growth trajectories*, often requiring the schools with the lowest initial test scores to generate the greatest improvement. Yet this is not based in any understanding of human or organizational improvement; actual instructional improvement takes time and concerted effort, and it rarely happens in a linear fashion. It is more reasonable to expect large improvements as teachers and students adjust to new pedagogies, with relatively modest changes as they solidify their knowledge and skills (Elmore, 2004). Furthermore, successful instructional improvement efforts happen across multiple years, yet accountability policies demand significant growth each individual year.

Furthermore, test-based accountability policies distort data use as *they only specify goals for student achievement*, not for instructional quality or teacher learning. In an effort to avoid "micromanaging how schools run," NCLB set no vision for high-quality instruction or teacher learning (Bush, 2002). Policymakers established goals for student outcomes, but failed to account for teachers' learning needs in order to support those outcomes. Important case studies show that assessment and data use can be a powerful lever for improving student achievement when implemented in the context of effective instruction and professional learning (Ezzani, 2015; McNaughton et al., 2012). But this is not the norm: Test-based accountability policies offer no guidance on what effective instruction is, much less any supports for educators' professional learning.

Avoid Sanctions

As a way of incentivizing increased student achievement, test-based accountability policies threaten sanctions for schools that fail to meet annual goals. Ostensibly, this is to encourage educators to work harder on behalf of their students and to pay more attention to those who might otherwise slip through the cracks of the education system (Bush, 2002). Sanctions included intervention from the state Department of Education, loss of federal funding, or even school closure (Nichols & Berliner, 2007). NCLB also allowed for federal funding to be diverted to pay for students' transportation to higher-achieving schools in the same area (Au, 2009). These punitive measures were intended to encourage educators to work harder to meet accountability goals and to expand students' school choice if they attended a low-achieving school.

But test-based accountability policies do not account for the resources and support that schools need to improve instruction. Instead, the pressure to avoid sanctions creates a DDUPE as *schools cultivate the appearance of achievement without increasing student learning*. Nichols and Berliner (2007) detail the various ways educators inflate students' scores. Some district and school leaders invest heavily in for-profit test-prep tutoring services to help students perform better on assessments. Principals, teachers, and other school staff have engage in outright cheating by sharing test items ahead of time, helping students during the test, and even changing answers after the test. There are also accounts of educators discouraging students who are likely to fail from taking the test in the first place. These tactics artificially increase school performance, distorting the interpretation of the test results.

The expansion of school choice creates further pressure on schools to succeed, creating a DDUPE as *the public compares schools based on assessment results*. This impacts mobility patterns (and property values), as parents seek out the "best" school for their children. But high-achieving schools are not always available in certain neighborhoods or districts; families are not always able to take advantage of the public school choice provision within NCLB (Jin, 2016). Predictably, this this disproportionately affects families in urban and rural areas and families from historically marginalized communities (Jin, 2016; Zhang & Cowen, 2009). But perhaps most dangerously, comparing schools based primarily on test scores obscures other characteristics of effective schools and teachers, such as relational trust and culturally responsive pedagogy (Bryk & Schneider, 2002; Ladson-Billings, 2009).

The predictability of DDUPEs: Barriers to instructional improvement

Even though policymakers espoused goals of high achievement for all students through optimistically-titled policies like "No Child Left Behind" and the "Every Student Succeeds Act," they organized assessment systems that are unlikely to support instructional improvement. In many ways, accountability systems make DDUPEs predictable, if not inevitable, thereby establishing barriers for the development of more ambitious instructional practices. Though a complete redesign of current assessment systems is outside the scope of this paper, I note that a dramatic systemic shift is necessary to repair the distortions caused by test-based accountability policies.

Bureaucratic controls generally have a poor track record for supporting meaningful and sustainable educational reform: Some, like value-added measures, have had little effect on instructional practice (e.g., Hill, Kapitula, & Umland, 2011), while others, like test-based accountability policies, have had a detrimental effect on instruction. One explanation for this phenomenon is that the education profession lacks a coherent epistemic community (Glazer & Peurach, 2015). That is, educators are not organized around common tools or theories that support the dissemination of knowledge or the development of shared professional practices. Developing a coherent approach to generating, organizing, and sharing knowledge is difficult, as teachers have relatively few opportunities to collaborate around problems of practice (Horn, Garner, Kane, & Brasel, 2017; Little, 1990; Lortie, 1974). This impedes instructional improvement in general, but it makes educators especially susceptible to distortions from bureaucratic policies.

Without a firm epistemic grounding to fall back on, teachers respond to bureaucratic controls in unintended ways. The ways that individuals respond to top-down policies are mediated by their collegial community as well as their own knowledge, beliefs, and experiences (Diamond, 2007; Spillane, Reiser, & Reimer, 2002). Since most teachers are not part of an epistemic community that has a firm commitment to ambitious instruction, it is unreasonable to expect accountability policies or DDDM to inspire a shift toward instructional improvement. Rather, it is more likely that teachers will enact new assessments in ways that encourage more straightforward (and often distortive) uses of data — for instance, by teaching to the test (Jordan & Putz, 2004).

The design of the assessments opens up opportunities for further distortions. The assessment systems developed under NCLB are designed to emphasize student achievement, rather than student learning. Ideally, these should be equivalent — students with higher test scores should be those who have learned more, and students who have learned more should receive higher scores. Fredericksen and Collins (1989) would describe such an assessment as having high *systemic validity*. But to assess student learning in an objective, efficient, and cost-effective manner, assessment items are designed to be indirect: That is, to measure discrete skills rather than students' performance on authentic tasks. In high-pressure settings, indirect assessments are especially susceptible to distortion: By emphasizing test-preparation strategies and focusing resources on "bubble kids," schools can artificially inflate test scores without supporting student learning. This short-circuits the connection been student learning and assessment scores, thereby undermining the systemic validity of the assessment.

Considering these systemic forces shaping test-based accountability and associated data use efforts, it may be unsurprising that policies like NCLB have not met their goals of high achievement for all students. Indeed, in order to move beyond DDUPEs, the education system will need to undergo dramatic transformations: 1) developing responsive practices for using evidence to inform instruction, 2) supporting an epistemic community that can sustain and proliferate such practices, and 3) instituting assessments with higher systemic validity. The latter two changes are outside the scope of this paper, but I take a step in that direction by attending to the first change: developing responsive evidence use practices.

Using Evidence for Instructional Improvement

Despite the seemingly inevitable flaws of DDDM and test-based accountability policies, the appropriate response is not to discard with data use entirely. The underlying principle of DDDM is that teachers should systematically and routinely collect evidence that allows them to determine what students understand, and use that evidence to design an instructional response. Using evidence of student learning in this way is crucial for high-quality instruction. Indeed, it is one of the cornerstones of ambitious instruction (Lampert et al., 2011). DDUPEs creep in to teachers' practice when 1) accountability policies place undue emphasis on assessments that provide thin evidence of student learning, 2) educators lack preparation to analyze and interpret data in productive ways, and 3) educators receive little support for designing instructional responses. More humane data use practices are possible, but they must address these concerns.

REUPEs: Responsive Evidence Use Practices in Education

A humanizing vision of data use for instructional improvement fundamentally rests on a vision of high-quality instruction that I have alluded to as ambitious. Ambitious instruction seeks to support all students in learning academic content and be able to apply it in authentic contexts (Lampert et al., 2011). In addition to being able execute skills, like reading passages aloud and performing calculations, students should be able to engage in disciplinary practices, like developing arguments and solving complex problems. The goal of ambitious instruction, then, is not memorize facts and procedures, but to understand the underlying disciplinary concepts (Lampert et al., 2100). This is aligned with the goals for student learning that professional organizations have been advocating for decades, as well as with the practice standards that have more recently been adopted in the Common Core and the Next Generation Science Standards. In order to support students' development of conceptual understanding and disciplinary practices, students need opportunities to engage in rich tasks that develop problem-solving skills (Lampert et al, 2011; Stein, Grover, & Henningsen, 1996). Rich tasks can simultaneously support students in developing disciplinary practices and understandings while also making some of their thinking visible to teachers. Through collaboration and discussion, students can develop both academic and social skills that support their involvement in an intellectual community (Horn, 2012; Lotan, 2003; cf. Cobb & Yackel, 1996).

As its name suggests, ambitious instruction is not easy. Ambitious instruction requires teachers to be able to constantly elicit, interpret, and build on student thinking in order to support deeper student learning. This introduces a great deal of uncertainty, as teachers decide what evidence to seek out, how to interpret the evidence, and how to respond to student thinking within the context of a complex classroom setting. In this section, I propose three Responsive Evidence Use Practices in Education (REUPEs) to guide teachers' use of evidence to support more ambitious instruction.

Coordinating multiple sources of evidence. All evidence of student learning highlight some attributes of a phenomenon and downplay others; part of using evidence ethically and responsibly to gauge their progress and learning requires that educators recognize this limitation (NFES, 2010). Any assessment can elicit particular elements of a student's knowledge of a topic — perhaps their procedural knowledge for calculating a unit rate, or their ability to apply algebraic principles to a specific context — but no assessment item can reveal the entirety of students' understanding. To that end, collecting multiple pieces of evidence, from various sources, is an important element of responsive evidence use (Hamilton et al., 2009). As teachers coordinate multiple sources of evidence — e.g., student work, observations of students during class, students' responses to multiple-choice questions, etc. — they can develop richer, more nuanced understandings of what students know and are able to do.

Of course, coordinating multiple evidence sources makes the interpretive process more complicated: There is more evidence to process, and more possibilities to consider. Ikemoto and Marsh (2007) note the increased complexity of using additional types of evidence, but also conceptualize complex evidence as a key feature of *inquiry-focused* data use, which they argue is likely to support instructional improvement. One way to deal with the increased complexity —

examining a reasonable amount of evidence, while maintaining sufficient detail to pick up on student thinking — is to purposively sample evidence for closer analysis. Garner and Horn (2018) share the evidence use practices of a mathematics teacher workgroup that did just that: They used quantitative benchmark data to identify students who were on the cusp of mastery, and then closely examined work from those students on an open-ended assessment. This approach allowed them to find a balance between breadth and depth of their evidence use.

Examining student thinking. By coordinating multiple sources of evidence, teachers can develop more nuanced understandings of student thinking. This is crucial for developing ambitious instruction that builds on students' knowledge and supports students' development of key disciplinary ideas and practices. As teachers examine the range and degree of student understandings, they are better positioned to improve their instruction (Bocala & Boudett, 2015; Horn et al., 2015; Nelson, Slavit, & Deuel, 2012). Of course, this also places a great demand on teachers' pedagogical content knowledge (Shulman, 1986). In order to deeply examine students' thinking, teachers must understand content well enough to be able to make sense of students' ideas and know how they relate to larger learning goals. For instance, if a student solves a math problem using a unique algorithm, the teacher would need to make sense of the algorithm and determine the conditions under which the algorithm works. Interpreting student thinking can be a challenging endeavor.

Notably, many scholars have found that multiple-choice assessments, like interim benchmark assessments, do not provide enough information for teachers to interpret student thinking (Farrell & Marsh, 2016; Garner, Thorne, & Horn, 2017; Olàh, Lawrence, & Riggan, 2010). Instead, benchmark assessments might be best used for identifying areas for deeper inquiry, augmented by other evidence such as classroom discourse or student work (Garner, &

Horn, 2018). This qualitative evidence can give a fuller picture and thus clearer insight into student thinking (Farrell & Marsh, 2016).

Using evidence to reflect on and adjust instruction. After examining student thinking, teachers can reflect on previous instruction and adjust future instruction. Reflecting on instruction allows teachers to identify the elements of their practices that support (and do not support) student learning (Blanc, Christman, Liu, Mitchell, Travers, & Bulkley, 2010). Adjusting instruction to respond to students' learning needs is a critical piece of using evidence for instructional improvement (Horn et al., 2015; Nelson, et al., 2012). As teachers take a critical look at their professional work, they can deepen their pedagogical knowledge, skill, and effectiveness.

Yet it is also imperative that teachers maintain an asset orientation about students as they engage in this work. Bertrand and Marsh (2015), for instance, note that teachers who attributed student difficulties to the efficacy of instruction are best positioned to improve their pedagogical practices. But teachers who simultaneously invoked concerns about student characteristics, like special education status or English proficiency, actually reinforced deficit notions of students' abilities: They tended to lower expectations of students' capabilities, rather than supporting all students to engage in ambitious instruction.

Particularly when integrated with other instructional improvement efforts, this vision of evidence use can support the development of ambitious instruction. These principles underscore the importance of using evidence to launch inquiries into student understanding so that teachers can build on student thinking during instruction. Taking an inquiry-oriented approach to evidence use also helps educators attend to students' conceptual understanding, supporting deeper learning for all students.

Discussion

The pressures of test-based accountability policies create perverse incentives that distort educators' use of data. Most DDDM efforts are agnostic with respect to instructional quality; they encourage educators merely to use data, but rarely specify what data should be used, how it should be interpreted, or the instructional goals of data use. When combined with intense pressure to raise student achievement, this agnosticism creates distortions that warp the meaning and use of assessment data. Rather than supporting aspirations for instructional improvement or deeper student learning, test-based accountability policies encourage DDUPEs that work at cross-purposes with calls for ambitious instruction.

Despite these tensions, it is possible to use evidence of student learning toward a humanizing vision of instructional improvement. By coordinating multiple sources of evidence, examining student thinking, and using evidence to reflect on and adjust instruction, educators can support more ambitious instruction. Using evidence in these ways requires supports for teacher learning and collaboration. Collaboration with colleagues and instructional leaders can foster rich conversations around evidence and, when necessary, insulate against accountability pressures.

Repairing the distortions of DDDM and test-based accountability policies will require addressing large systemic issues. Developing assessments with greater systemic validity is a first step, but Fredericksen and Collins (1989) note that this will require a large investment to implement on a large scale, due to the resources required to design and evaluate assessments that more accurately capture students' understanding — which will almost certainly be open-ended or performance-based. Fredericksen and Collins also propose decentralizing the assessment system so that teachers can be charge of assessing their own kids. In order for such changes to be effective, they would need to come in the context of a larger effort for supporting teacher learning. This could come in the form of a push toward an epistemic community (Glazer & Peurach, 2015) that develops practices and tools for generating knowledge about student understanding.

Though the field's understanding of teachers' data use has grown over the last five years, there are still areas that are under-researched. In particular, there are few studies that investigate teachers' conceptions of and assumptions about data. Because it is a fundamentally interpretive activity, data use rests on educators' epistemic assumptions In a sense, it engages perennial philosophical questions: How do teachers know what students know? What is worth knowing, and what can be known, about student understanding? Few researchers have tackled this question (a notable exception is Slavit, Nelson, & Deuel, 2012), yet these epistemic assumptions remain integral to teachers' sensemaking with assessment data.

Furthermore, there is relatively little research in supporting teachers' development of more productive evidence use practices. There is substantial evidence that teachers' school contexts shape their evidence use; their collaborations and relationships with colleagues, instructional coaches, and principals are particularly important (e.g., Farrell, 2014; Marsh, Farrell, & Bertrand, 2016). Yet there are relatively few studies of professional learning about productive evidence use practices. Most efforts to support teachers' data use address the logistics of data use, helping teachers gain familiarity with data management software or measuring the frequency (but not the details) of teachers' data use. I anticipate that this analysis of productive data use practices will inform the work of researchers, teacher educators, and instructional leaders who seek to support teachers' development of more productive data use practices.

References

Aguirre, J., Mayfield-Ingram, K., & Martin, D. (2013). *The impact of identity in K-8 mathematics: Rethinking equity-based practices*. The National Council of Teachers of Mathematics.

Aguirre, J. M., & del Rosario Zavala, M. (2013). Making culturally responsive mathematics teaching explicit: A lesson analysis tool. *Pedagogies: An international journal*, *8*(2), 163-190.

Au, W. (2009). Unequal by design. *High-stakes testing and the standardization of inequality. London: Routledge*.

Battey, D., & Leyva, L. A. (2016). A framework for understanding whiteness in mathematics education. *Journal of Urban Mathematics Education*, *9*(2).

Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, *20*(10), 1-33.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: principles, policy & practice, 5(1), 7-74.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, *85*(2), 205-225.

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. *Teachers College Record*, *110*(3), 608-645.

Bocala, C., & Boudett, K. P. (2015). Teaching educators habits of mind for using data wisely. *Teachers College Record*, *117*(4).

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

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. Russell Sage Foundation.

Bush, G. W. (2000). President's remarks to NAACP. Retrieved April 4, 2015, from http://www.washingtonpost.com/wp-srv/onpolitics/elections/bushtext071000.htm

Bush, G. W. (2002). President Signs Landmark No Child Left Behind Education Bill. Remarks made at Hamiton High School, Hamilton OH. Retrieved from https://georgewbush-whitehouse.archives.gov/news/releases/2002/01/20020108-1.html

Cobb, P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational psychologist*, *31*(3-4), 175-190.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173-206

Common Core State Standards Initiative. (2010). Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of 'No Child Left Behind'. *Race Ethnicity and Education*, *10*(3), 245-260.

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., & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. Journal of Education for Students Placed at Risk (JESPAR), 17(4), 247-265

Diamond, J. B., & Cooper, K. (2007). The uses of testing data in urban elementary schools: Some lessons from chicago. In Yearbook of the national society for the study of education (Vol. 106, pp. 241-263). Wiley Online Library.

Duncan, A. (2009, June). Robust Data Gives Us The Roadmap to Reform. Speech made at the Fourth Annual IES Research Conference, Washington, DC. Retrieved from https://www.ed.gov/news/speeches/robust-data-gives-us-roadmap-reform

Duncan, A. (2010, July). Unleashing the Power of Data for School Reform. Keynote address at the Educate with Data: Stats-DC 2010 Conference, Bethesda, MD. Retrieved from https://www.ed.gov/news/speeches/unleashing-power-data-school-reform-secretary-arne-duncans-remarks-stats-dc-2010-data-

Elmore, R. F. (2004). School reform from the inside out: Policy, practice, and performance. Harvard Educational Pub Group.

Ezzani, M. (2015). Coherent district reform: A case study of two California school districts. Cogent Education, 2(1).

Farrell, C. C. (2014). Designing school systems to encourage data use and instructional improvement. Educational Administration Quarterly, 51(3), 438-471

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

Garner, B. & Horn, I.S. (2018) Using Standardized Test Data as a Starting Point for Inquiry: A Case for Thoughtful Compliance. In Barnes, N. & Fives, H.R. (eds.), Teachers' Data Use: Cases of Promising Practice. Routledge, New York City.

Garner, B., Thorne, J. K., & Horn, I. S. (2017). Teachers interpreting data for instructional decisions: where does equity come in?. Journal of Educational Administration, 55(4), 407-426.

Glazer, J. L., & Peurach, D. J. (2015). Occupational control in education: The logic and leverage of epistemic communities. Harvard Educational Review, 85(2), 172-202.

Goldstein, B. E., & Hall, R. (2007). Modeling without end: Conflict across organizational and disciplinary boundaries in habitat conservation planning. In Kaput, E. Hamilton, S. Zawojewski, & R. Lesh (Eds.), Foundations for the future.

Gould, S. J. (1996). The mismeasure of man. WW Norton & Company.

Gutiérrez, R. (2008). A" gap-gazing" fetish in mathematics education? Problematizing research on the achievement gap. Journal for Research in Mathematics Education, 357-364.

Gutiérrez, R. (2012). Context matters: How should we conceptualize equity in mathematics education?. In Equity in discourse for mathematics education (pp.17-33). Springer Netherlands.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., Pickens, C., Martin, E.S., & Steele, J. L. (2009). Using student achievement data to support instructional decision making.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing?. American Psychologist, 47(9), 1083.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics, 65-97.

Horn, I. S. (2008). Turnaround students in high school mathematics: Constructing identities of competence through mathematical worlds. Mathematical Thinking and Learning, 10(3), 201-239.

Horn, I. S. (2012). Strength in numbers: Collaborative learning in secondary mathematics. National Council of Teachers of Mathematics.

Horn, I. S. (2016). Accountability as a design for teacher learning: Sensemaking about mathematics and equity in the NCLB era. Urban Education, 0042085916646625.

Horn, I. S., Kane, B. D., & Wilson, J. (2015). Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. American Educational Research Journal, 52(2), 208-242

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

Jennings, J. (2012). The effects of accountability system design on teachers' use of test score data. Teachers College Record, 114(11).

Jennings, J. L., & Bearak, J. M. (2014). “Teaching to the test” in the NCLB era: How test predictability affects our understanding of student performance. Educational Researcher, 43(8), 381-389.

Jimerson, J. B. & Wayman, J. C. (2015). Professional learning for using data: Examining teacher needs and supports. Teachers College Record, 117(4).

Kazemi, E., Franke, M., & Lampert, M. (2009, July). Developing pedagogies in teacher education to support novice teachers’ ability to enact ambitious instruction. In Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia (Vol. 1, pp. 12-30). Adelaide, SA: MERGA.

Kazemi, E., & Stipek, D. (2001). Promoting conceptual thinking in four upper-elementary mathematics classrooms. The Elementary School Journal, 102(1), 59-80.

Kidder, W. C., & Rosner, J. (2002). How the SAT creates built-in-headwinds: An Educational and legal analysis of disparate impact. Santa Clara L. Rev., 43, 131.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in US schools. Educational researcher, 35(7), 3-12.

Ladson-Billings, G. (2009). The dreamkeepers: Successful teachers of African American children. John Wiley & Sons.

Ladson-Billings, G. (2014). Culturally relevant pedagogy 2.0: aka the remix. Harvard Educational Review, 84(1), 74-84.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. American educational research journal, 27(1), 29-63.

Lampert, M., Boerst, T. A., & Graziani, F. (2011). Organizational resources in the service of schoolwide ambitious teaching practice. Teachers College Record, 113(7), 1361-1400.

Lee, J. (2016). Paying for School Choice: Availability Differences among Local Education Markets. International Journal of Education Policy and Leadership, 11(5).

Lee, M., Louis, K.S., & Anderson, S. (2012). Local education authorities and student learning: The effects of policies and practices. School Effectiveness and School Improvement, 23(2), 133-158.

Le Floch, K.C., O’Day, J., Birman, B., Hurlburt, S., Nayfack, M., Halloran, C., Boyle, A., Brown, S., Mercado-Garcia, D., Goff, R., Rosenberg, L., and Hulsey, L. (2016). Case Studies of Schools Receiving School Improvement Grants: Final Report (NCEE 2016-4002). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Little, J. W. (2011). Understanding data use practice among teachers: The contribution of micro-process studies. American Journal of Education, 118(2), 143-166

Lotan, R. A. (2003). Group-worthy tasks. Educational Leadership, 60(6), 72-75.

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. Educational Researcher, 42(1), 30-37.

Marsh, J. A., Farrell, C. C., & Bertrand, M. (2016). Trickle-down accountability: How middle school teachers engage students in data use. Educational Policy, 30(2), 243-280

McNaughton, S., Lai, M. K., & Hsiao, S. (2012). Testing the effectiveness of an intervention model based on data use: A replication series across clusters of schools. School Effectiveness and School Improvement, 23(2), 203-228.

Means, B., Padilla, C., & Gallagher, L. (2010). Use of education data at the local level: From accountability to instructional improvement. US Department of Education.

Moll, L. C., & González, N. (2004). Engaging life: A funds of knowledge approach to multicultural education. Handbook of research on multicultural education, 2, 699-715.

National Council of Teachers of Mathematics (2000). Principles and standards for school mathematics

National Council of Teachers of Mathematics (NCTM). (2014). Principles to actions: Ensuring mathematical success for all.

National Forum on Education Statistics. (2011). Traveling Through Time: The Forum Guide to Longitudinal Data Systems. Book Four of Four: Advanced LDS Usage (NFES 2011-802). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

National Forum on Education Statistics. (2010). Forum Guide to Data Ethics (NFES 2010-801). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Nelson, T. H., Slavit, D., & Deuel, A. (2012). Two dimensions of an inquiry stance toward student learning data. Teachers College Record, 114(8), 1-42.

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

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. Peabody Journal of Education, 85(2), 226-245.

Pham, V.H. (2009) “Computer modeling of the instructionally insensitive nature of the Texas Assessment of Knowledge and Skills (TAKS) exam”, Unpublished dissertation, University of Texas at Austin.

Schaffer, E., Reynolds, D., & Stringfield, S. (2012). Sustaining turnaround at the school and district Levels: The high reliability schools project at Sandfields Secondary School. Journal of Education for Students Placed at Risk (JESPAR), 17(1-2), 108-127.

Schoenfeld, A. H. (1988). When good teaching leads to bad results: The disasters of'well-taught' mathematics courses. Educational psychologist, 23(2), 145-166.

Slavit, D., Nelson, T. H., & Deuel, A. (2012). Teacher groups' conceptions and uses of student-learning data. Journal of Teacher Education, 64(1), 8-21

Smith, M. S., & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. Mathematics teaching in the middle school, 3(5), 344-50.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. American educational research journal, 33(2), 455-488.

Stigler, J. W., & Hiebert, J. (2009). The teaching gap: Best ideas from the world's teachers for improving education in the classroom. Simon and Schuster.

Villavicencio, A., & Grayman, J. K. (2012). Learning from “turnaround” middle schools: Strategies for success. New York: Research Alliance for New York City Schools.

Wayman, J. C., Jimerson, J. B., & Cho, V. (2012). Organizational considerations in establishing the data-informed district. School Effectiveness and School Improvement, 23(2), 159-178.

Weinstock, P., Yumoto, F., Abe, Y., Meyers, C., & Wan, Y. (2016). How to use the school survey of practices associated with high performance. REL 2016-162. Regional Educational Laboratory Midwest.

Weiss, J. A. (2012). Data for improvement, data for accountability. Teachers College Record, 114(11)

West, M. R., Morton, B. A., & Herlihy, C. M. (2016). Achievement Network’s Investing in Innovation Expansion: Impacts on Educator Practice and Student Achievement.

Yuan, K., & Le, V. (2012). Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests. Santa Monica, CA: RAND Corporation. Retrieved from the RAND website: http://www.rand.org/content/dam/rand/pubs/working_papers/2012/RAND_WR967.pdf

Zhang, H. C., & Cowen, D. J. (2009). Mapping academic achievement and public school choice under the No Child Left Behind legislation. Southeastern Geographer, 49(1), 24-40.

CHAPTER III

TEACHERS INTERPRETING DATA FOR INSTRUCTIONAL DECISIONS: WHERE DOES EQUITY COME IN?

Test-based accountability policies place pressure on teachers and schools to increase test scores, particularly for students from historically marginalized groups. Educators and policymakers use results from standardized assessments to identify potential problem areas – content that students have not mastered, students who are underperforming, schools and teachers deemed ineffective, and so forth. In theory, the purpose of these policies is to detect such problems and find solutions, thus creating more equitable outcomes in schools.

However, policymakers underspecify the details of data-use processes. Despite accountability policies' stated intention to reduce educational inequity and improve the academic standing of students of color, emergent bilinguals[1], and students from low-income families, the underlying theory of action treats inequality as a technical problem rather than a political one: Data will point educators toward ameliorative actions without forcing them to confront systemic inequities that contribute to achievement disparities. To imagine another tack, efforts advocating for increased data use could instead address how teachers recognize and respond to the racial ideologies and injustices that operate both in the data and within teachers' own data literacy practices (Philip and Garcia, 2013, 2015; Philip et al., 2016). Using this counterfactual as a point of contrast, this paper identifies key problems with the techno-rational logic of accountability policies and reflects on the ways in which they influence teachers' data-use practices. To

[1] This paper uses the term "emergent bilingual" instead of labels such as "Limited English Proficient" to highlight students' existing linguistic competencies and to decenter English as the dominant language.

illustrate this argument, the analysis critically considers the data-use practices of a group of middle school mathematics educators examining the results of a district benchmark assessment to plan instruction in the weeks leading up to the high stakes, end-of-year state test. The analysis seeks to reveal the ways in which issues of equity intersect with their work as they interpret student assessment data for instructional decision-making.

A critical take on test-based accountability policies

Data-driven decision-making (DDDM) is a common strategy for school- and district-level improvement in many countries (Ah-Teck and Starr, 2014; Datnow and Hubbard, 2015; Lynch et al., 2016). In the USA, DDDM is shaped largely by test-based accountability policies. Beginning with the No Child Left Behind (NCLB) Act of 2001, US states have implemented annual high-stakes standardized assessments in mathematics and reading; similar policies have continued through the Race to the Top initiative of 2009 and the Every Student Succeeds Act (ESSA) of 2015. Under these policies, schools and districts are held accountable to students' performance on end-of-year exams that primarily feature multiple-choice questions. Students' scores are disaggregated by subpopulations (including categories for race, ethnicity, poverty, language, and special education status). This allows policymakers, educators, and the general public to identify and monitor differences in performance of subpopulations ("achievement gaps"). To avoid sanctions, schools must demonstrate sufficiently high passing rates for each subpopulation as well as the overall student body. By specifying achievement goals for students from various groups, accountability policies seek to encourage "continuous and substantial academic improvement for all students" (NCLB, 2002).

These accountability policies share the following underlying theory of action: By shining a light on students' performance, examining differences across groups, and maintaining high expectations for all, schools can provide more equitable outcomes for their students. Considered in the context of social stratification and historical disenfranchisement, the logic of this theory quickly unravels. As critical scholars have noted, policies that emphasize the test scores of historically marginalized subpopulations often reinscribe existing power structures by reinforcing deficit-oriented perspectives toward non-dominant communities (Milner, 2013). Framing differences among groups of students as an achievement gap – instead of as an education debt (Ladson-Billings, 2006) – pathologizes students who fail state tests, without acknowledging or redressing underlying reasons for performance differentials. As Ladson-Billings (2006) notes, the USA has a well-established history of limiting educational opportunities for students in marginalized communities through policies that supported segregation, unequal school funding, differential school staffing patterns, and related differential distributions of resources. Accordingly, describing the academic underperformance of marginalized students as an "achievement gap" highlights students' failure to learn grade-level material rather than society's ongoing failure to provide adequate resources, opportunities, and civil liberties for students and their families.

Through this techno-rational logic, test-based accountability policies focus on the present-tense outcomes, rather than historically rooted causes, of educational inequities. Student learning is measured by standardized tests, which hold all students to the same metric, regardless of the educational debt owed to them and their families. Scholars have found that schools facing higher pressure from accountability policies – often, schools with higher populations of students of color and students from low-income families – turn to more intensive test preparation strategies (Diamond and Cooper, 2007; Horn, 2016). The most charitable reading of this trend is to view it as an attempt to prepare students to engage in the "culture of power" that governs schooling (Delpit, 1988); that is, by explicitly preparing historically disenfranchised students for success on consequential metrics, teachers give them greater access to the knowledge and skills that will allow them to acquire academic and social power. However, this view is likely overly hopeful: In reality, standardized tests typically assess a narrow range of mathematical skills (Gutiérrez, 2009), and aiming instruction toward these goals limits students' access to richer mathematical understandings that would sustain their future learning. Indeed, critical scholars have argued that in order to access the academic and cultural power of mathematics, students need opportunities to engage in intellectually rich mathematics practices like argumentation and communication (Aguirre and Zavala, 2013; Gutiérrez, 2008), forms of understanding that are seldom captured in standardized tests.

In sum, accountability policies like NCLB seek to change educational outcomes for marginalized students without addressing systemic oppression. Social and historical inequity is bracketed off, with a close and incomplete focus on performance at one point in time on a narrow measure. As a consequence, students' academic trajectories are reduced to a series of test scores, enabling teachers, administrators, and the public to reinstate deficit narratives about marginalized subpopulations as they are measured against the White, middle-class norm. This functions to reify the racism, colonialist values, and White supremacy and privilege that are ingrained in the fabric of US society (Rollock and Gillborn, 2011).

Data use in practice

Under NCLB and ESSA, schools are required to administer state assessments at the end of each school year. Results from these tests often arrive over the Summer or even at the beginning of the following school year; they are therefore not useful for informing routine instruction of teachers' current students. To monitor students' progress toward end-of- year assessment goals, interim benchmark assessments have become a common practice in US schools (Datnow and Hubbard, 2015; Jennings, 2012). Districts develop or purchase assessments meant to mimic state tests in format and content; teachers administer them periodically throughout the school year. Whether locally developed or purchased from vendors, these assessments are problematic in different ways: Locally developed benchmark assessments lack the (costly) psychometric validation of published tests, while purchased benchmark assessments are often poorly aligned to local curricula. Nonetheless, the quantitative results they produce – the data referred to in this paper – become consequential in teachers' and students' lives, almost to the exclusion of other forms of data (e.g. student work or classroom talk). Despite the exhortations to do so, educators have few guidelines for using these data to inform their instruction. Accountability policies do not, for instance, specify how one might organize, synthesize, and summarize multiple data points, either within or across students. For this reason, there has been much variation in how teachers and administrators use assessment data in schools (Datnow and Hubbard, 2015; Jennings, 2012; Marsh, 2012).

Scholars studying educators' data use have identified characteristics of productive data-use practices – that is, ways of using data that attend to goals for deeper student learning. Horn et al. (2015) draw a distinction between two orientations for data use: one they call an instructional improvement orientation and another they refer to as an instructional management orientation.

Educators who take an instructional improvement approach coordinate multiple sources of data to reflect on instruction, invest in professional learning, and develop deeper pedagogical skill. They tend to attribute areas of poor performance to flaws in instruction rather than flaws in students (Bertrand and Marsh, 2015). They are thus positioned to respond to students' learning needs and support higher achievement.

In contrast, educators taking an instructional management approach are primarily concerned with organizing instructional work to maximize test scores, often without fundamentally changing their teaching practice. They tend to emphasize test preparation and student triage strategies that may result in higher scores without supporting students' understanding. In bypassing reflection on instructional practices and responses to students' learning needs, the instructional management approach is unlikely to support long-term instructional improvement or deeper student learning. Schools facing higher accountability pressures are more likely to use data for instructional management than schools that are already successful on standardized test measures (Diamond and Cooper, 2007). As a result, paradoxically, the schools that accountability policies squarely target are prone to use data in the least productive ways.

The misalignment between test-based accountability and equitable teaching

While various scholars have examined educators' data-use practices, few have done so from a critical perspective. Critical analysts accept race and power as part of the fabric of life in modern society (Milner, 2013) and work to challenge and change racist policies that disenfranchise certain groups in an effort to maintain the status quo (Tate, 1997). This paper employs such a stance, recognizing that intersections of race, class, teaching, and learning should be considered in examining teachers' interpretations of standardized testing, the resulting data, and the broader data-driven educational climate. This paper thus considers the broader social context and longer histories that are captured in assessment data and in educators' use of these data under accountability policies. This represents a move toward reconciling educators' data use with calls to develop culturally responsive pedagogies, which seek to redress broader inequities by building on students' cultural strengths and resources. By applying a critical lens to data-use practices, this paper articulates the misalignment of the techno-rational vision of educational equity and the culturally responsive vision.

The techno-rational view of equity is clear in the public conversations and media coverage of student achievement and accountability that embrace neoliberal notions of meritocracy (Solomon et al., 2005). Standardized assessments are unproblematically taken as measures of intellectual and academic merit, as though they are neutral to students' race, class, gender, and disability status. Students who perform poorly – and their schools and teachers – are presented as deficient or defective, without acknowledging systems of oppression that may have limited their learning opportunities, hampered their performance, or created assessment bias. In contrast, culturally responsive views posit that for teachers to be effective with non-dominant populations, they need to treat students as agents with funds of knowledge to bring to bear on their learning (Gay, 2010; Ladson-Billings, 2009; Moll and González, 2004) and to subvert oppressive systems as they manifest in schools (Gutiérrez, 2016). While standardized tests focus on the binary of what students know and do not know according to one metric, more open, asset-oriented stances might also apply when analyzing data to inform instructional decisions.

Because data use has been imagined as a learning opportunity for teachers, the contribution of these conversations to teachers' professional development is consequential. This analysis builds on previous studies that examine teachers' data use from the perspective of teachers' learning opportunities and professional development (Horn et al., 2015; Jackson et al., 2014) by looking for moments in data-use conversations where educators consider issues of race and ethnicity. In an examination of all 25 data-use conversations collected for this study, the authors found that concerns about diversity and equity were rarely mentioned, except insofar as they are represented by categories of students and associated descriptions of underachievement based in deficit-oriented perspectives. This paper produces a response to this finding by putting forth a proposal for how teachers might engage with student data in ways that support culturally responsive instruction that leverages students' assets, strengths, and agency.

This paper focuses on educators' use of benchmark assessment data – rather than other forms of data – not because it is affords the greatest opportunities for instructional improvement or discussions of equity, but because it is a common form of data use across US schools, and such assessments are consequential for teachers' work. Indeed, participants in our study (including teachers, instructional coaches, principals, and district leaders across two large urban districts) described their data use largely in reference to the quantitative data from benchmark assessments. They also described using such data to change instructional plans, sort students for remediation activities, and assess teachers' efficacy. This aligns with Datnow and Hubbard's (2015) findings that, in response to test-based accountability policy, benchmark assessments have become a common strategy for districts' improvement efforts. Given the prevalence of benchmark assessments as teachers' primary source of data and the purported equity goals of test-based accountability policies, it is important to investigate the ways in which the use of benchmark assessment data do (or do not) support considerations of equity.

As a starting place for this discussion, this paper presents an analysis of a data-use meeting. Though the meeting participants engaged in a very common activity (analyzing benchmark data), they had additional supports that could lead to productive data use. By examining how these educators persisted in colorblind, techno-rational discourses within their data-use conversation, this analysis seeds a critique of accountability logics and related data-use practices that work against deeper, culturally respectful visions of equitable education.

Methods

Research context

The data for this analysis come from a larger study of instructional improvement in middle school mathematics. Starting in 2007, the Middle-School Mathematics and the Institutional Setting of Teaching project investigated large-scale support of mathematics teachers' development of ambitious and equitable instruction. From a representative sample of schools in two large, urban districts, focal teacher workgroups were purposively sampled to over-represent "successful" cases of teacher learning through data use. Key district informants were asked to identify "strong" workgroups. Researchers then followed up to select strong workgroups that had potential catalysts for teachers' learning. Examples of catalysts included the presence of knowledgeable instructional coaches, accomplished teachers, or unusual supports of time and professional development. Approximately eight groups were selected each year for three years. Though the research team sought out longitudinal cases, new sites were selected each year due to high rates of teacher turnover and other sources of institutional churn.

To understand teachers' data-use practices in workgroup conversations, the primary corpus of data consists of approximately 25 video- and audio-recorded data-use meetings (typically 45-60 minutes in length). For closer analysis, this paper presents a case study (Yin, 2009) of one teacher workgroup's data use during a full-day data analysis session. This investigation is also informed by secondary data sources from the larger study, including semi-structured interviews about teachers' instructional views and workplace experiences, surveys that provide supplementary information about the school contexts and participants' backgrounds, and copies of assessments used by teachers in these conversations.

Analytic approach: Understanding data use as workplace learning

This analysis aligns with studies of professionals at work in the sociocultural historical tradition (e.g. Goldstein and Hall, 2007; Hall and Horn, 2012). These studies often use video data to examine joint participation in problem-solving and analytic activities, paying close attention to talk, gesture, and tool use in these contexts. By emphasizing video data, this analysis addresses the persistent methodological "say-do" problem: That is, often what participants actually do differs from what they describe themselves doing, and the latter may more closely resemble authorities' expectations than reality. Because of the pressure of test-based accountability (as one teacher said at the start of the meeting, "I am so stressed out. Like, this is my job!"), analyzing what educators actually do with data adds needed nuance to the field's understanding of accountability policy.

Unit of analysis. Horn and colleagues analyze learning opportunities in teachers' conversations by parsing out episodes of pedagogical reasoning (EPRs; Horn, 2007; Horn et al., 2015). Horn defines EPRs as "moments in teachers' interaction in which they describe issues in or raise questions about teaching practice that are accompanied by some elaboration of reasons, explanations, or justifications" (2007, p. 46). To analyze learning opportunities in teachers' data-use conversations, this paper adapts the unit of analysis by parsing the conversation into episodes of data reasoning (EDRs). A new unit of analysis was necessary because data-use conversations do not always include considerations of pedagogy. Instead, in EDRs educators describe issues and raise questions about data and make decisions that they back up with reasons, explanations, and justifications, which may or may not point to issues of pedagogy and instruction. An important finding in this work is that EPRs and EDRs are frequently separate: That is, teachers often discuss data without considering pedagogy.

Like EPRs, EDRs are identified within conversations through topic shifts. For instance, an EDR might begin when a teacher asks, "What did students do on question no. 27?" and end when the discussion moves to another test item or another topic. Over the course of the six-hour data analysis session, the focal workgroup engaged in 18 EDRs, ranging in length from 40 seconds to 13 minutes. The average EDR length was approximately four minutes. Additional time during the session was spent discussing logistical concerns (e.g. district-wide meetings, schedules for state testing), socializing, and working independently.

Understanding teachers' sensemaking. Interaction analysis methods (Derry et al., 2010; Jordan and Henderson, 1995) were used to interpret videos records of EDRs. Within EDRs, the analysis attended to what the focal educators considered relevant for sensemaking of the data, treating their ways of participation as members' phenomena and as situated in social contexts (Lave and Wenger, 1991; Sacks, 1967/1992; Stevens, 2010). Gesture and sequential turns at talk were taken as important for understanding participants' sense-making processes (Sacks et al., 1974; Schegloff, 1992). Within utterances, noun and verb predication were examined to understand who or what (e.g. students, content standards, test items) was animated and given agency and authority in teachers' conversations (Goldstein and Hall, 2007; Ochs et al., 1996). The focus of teachers' attention, the resources they mobilized, and the actors they emphasized in their discourse and gesture were taken as indicators of how the teachers' data-use practices aligned with test-based accountability policies.

Researcher positionality. It is important to note that this analysis is not a critique of the teachers' data-use practices but rather of the accountability policy. As White women examining a workgroup made up primarily of teachers of color, we are mindful of our own privileged social position in relation to our participants. The goal of this analysis is to illuminate the systemic forces that work against deeper engagement with equitable pedagogies, not to indict a group of caring and hardworking teachers. In this vein, the analysis invokes a teacher solidarity lens (Philip et al., 2016), taking into account the district, state, and national policy contexts in which the focal teachers worked. The teachers' approach is a logical and reasonable response to the constraints and incentives of the system – the district's requirement to administer benchmark assessments, the principals' requirement that teachers use benchmark data to inform instruction, and the state's accountability measures that held teachers accountable for their students' performance. These pressures shaped the teachers' work in ways that make their data-use practices sensible. It is also important to acknowledge that this analysis captures a subset of the teachers' work; there may be other avenues through which they strived for equity that are not captured by the study. The present critique is of test-based accountability policies and the ways that they are presented as solutions to educational inequity (while simultaneously reinforcing it); it is not a critique of educators' responses to a flawed system.

Case selection

The focal group from Riverview Middle School (all proper names are pseudonyms) includes three sixth-grade mathematics teachers who analyze benchmark data during a full-day session known locally as a “Data Day.” Following the logic of the larger study, this workgroup was selected as a “best case.” Like many workgroups in the study, the Riverview teachers analyzed data from a district benchmark assessment. But they also had additional resources that mitigated common barriers to productive data use and attention to equity: The teachers had an unusual amount of time to devote to data analysis (Riverview was the only school in the study to organize Data Days) and did so under the guidance of a principal with high mathematics and pedagogical expertise. Furthermore, three of the four educators were Black, potentially attuning them to the ways in which racism is enacted in schools (Villegas and Irvine, 2010).

Beyond these resources of time and insider cultural knowledge, the Riverview workgroup had access to an accomplished instructional leader. The Riverview Principal, Vera Cardwell, emphasized the importance of using data to inform instruction. Prior to her work at Riverview, Ms. Cardwell was a middle school mathematics teacher, instructional coach, and assistant principal who successfully supported teachers in using data to improve students’ scores on state assessments. In interviews, Ms. Cardwell reported that the strategic use of student performance data was one of the main strategies to which she attributed her success as an instructional leader. In addition to having experience as a mathematics teacher and administrator, Ms. Cardwell described a sophisticated vision of mathematics instruction (Munter, 2014), emphasizing students’ engagement with high-level tasks and discussions to support their conceptual understanding. Ms. Cardwell identified as Black and spoke passionately about the importance of educational equity. For these reasons, Ms. Cardwell was an instructional leader with above-average mathematics instructional expertise and a commitment to educational equity, giving her atypically strong potential to support rich learning opportunities for the teachers whom she supervised (Horn et al., 2015).

Aside from her expertise, Ms. Cardwell also believed in the potential of data to improve student learning outcomes. Out of this conviction, Ms. Cardwell sought a grant from the district to fund full-day Data Days for mathematics and language arts teachers to be held after each of the four district benchmark assessments. During Data Days, teachers were given a full work day to examine students' performance and to use the data to plan for future instruction. At other schools in the larger study, teachers typically analyzed assessment data during weekly hour-long workgroup meetings; the Data Days afforded significantly more time for the Riverview teachers to analyze benchmark data. Data Days were critical events (Emerson et al., 2011) for the Riverview teachers' work because of the amount of time and resources devoted to the day, the importance of the benchmark assessments in the school and district, and Ms. Cardwell's emphasis on using data in teachers' instructional decision-making.

From the math teacher workgroups at Riverview, Ms. Cardwell recommended the sixth-grade team as a group that worked especially well together. They were also particularly invested in using data to inform instruction. During semi-structured interviews, they each reported the importance of using data as part of their professional practice. Crystal, for instance, said that she was "data heavy," meaning that she focused on collecting and analyzing assessment data to inform her instruction. Devon also described a special "ceremony" that the sixth-grade teachers conducted to share data with students and celebrate success on benchmark assessments. There may be a performative aspect to this emphasis, where the teachers (perhaps unintentionally) overstated the importance of data use for Ms. Cardwell's benefit. Yet the consistency and earnestness with which they reported using data suggest that their data-use practices are a central part of their professional work. Their commitments to use data to inform instruction further establish the Riverview sixth- grade team as a "best case" for investigating the implications of accountability policies.

Riverview's diverse student population, with many students from historically marginalized groups, also made the school an appropriate site to study the ways in which accountability policies – particularly the equity-oriented goals of NCLB – were reflected in teachers' data use. During the 2013-2014 school year, the 700 students at Riverview were approximately 45 percent Latinx[2], 30 percent Black, and 20 percent White, with smaller percentages of students identified as Indigenous or Asian/Pacific Islander. Approximately 10 percent of students were emergent bilinguals, and 10 percent of students received special education services. Nearly 80 percent of the students at Riverview qualified for free and reduced-price lunch, indicating that many Riverview students came from low- income families.

Overview of the Data Day

Participants and organization. The session in this analysis is the final Data Day for the school year. Since this Data Day occurred approximately one month before the state test, accountability pressure was at its peak. The sixth-grade team consisted of Rachel (White, 17 years of teaching experience), Crystal (Black, 12 years of teaching experience), and Devon (Black, three years of teaching experience), as shown in Figure III.1. The teachers analyzed student assessment data to make preliminary plans for instruction for the weeks remaining before

[2] This paper uses "Latinx" rather than "Latino" to decenter the patriarchal nature of the Spanish language (Gutiérrez, 2013). The "-x" suffix further decenters the gender binary and is inclusive of trans and genderqueer individuals (Monzó, 2016).

the state test. Their goal for the day was to identify questions and standards that students struggled with and discuss potential instructional responses. Ms. Cardwell occasionally participated in the teachers' conversations; her time was split among groups from other grade levels and content areas that were meeting at the same time.

Material resources. Ms. Cardwell organized a variety of documents for teachers to use during the Data Day. Teachers had the list of test items by percentage correct, showing how many sixth-grade students answered each question correctly (Figure III.2). Teachers also received copies of the benchmark test and an item analysis document showing the distribution of student responses. Items were indexed to state standards, which were grouped thematically into larger reporting categories. As Ms. Cardwell set teachers to work at the beginning of the Data Day, she suggested that teachers look for patterns in student performance across items within standards and reporting categories.

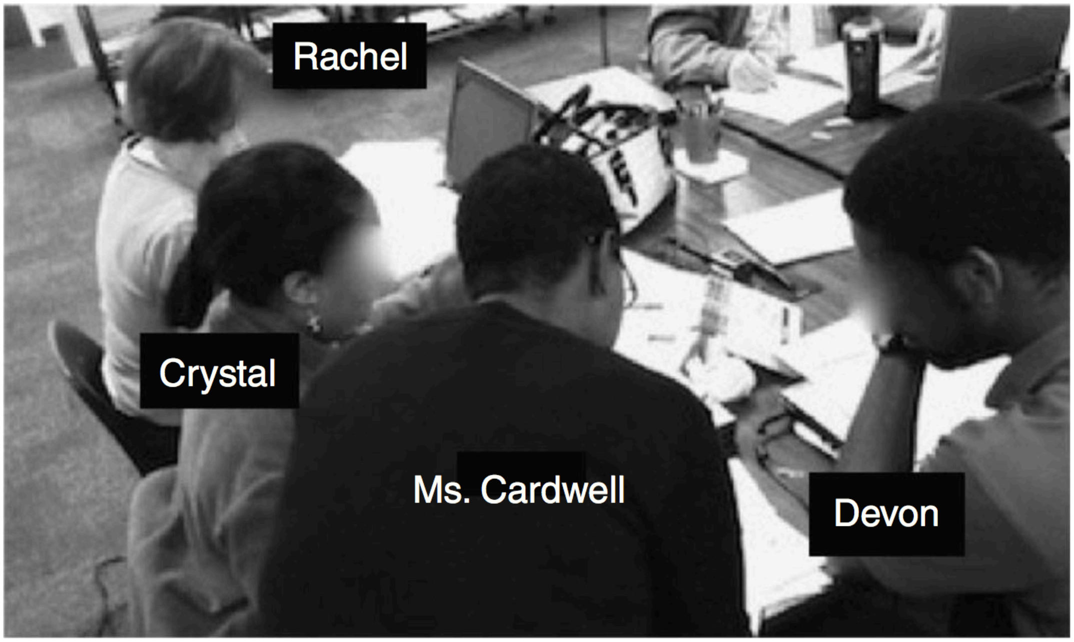


Figure III.1. Sixth-grade math teachers Rachel, Crystal, and Devon are pictured talking with Ms. Cardwell, the Principal, to analyze data from the most recent district benchmark assessment.

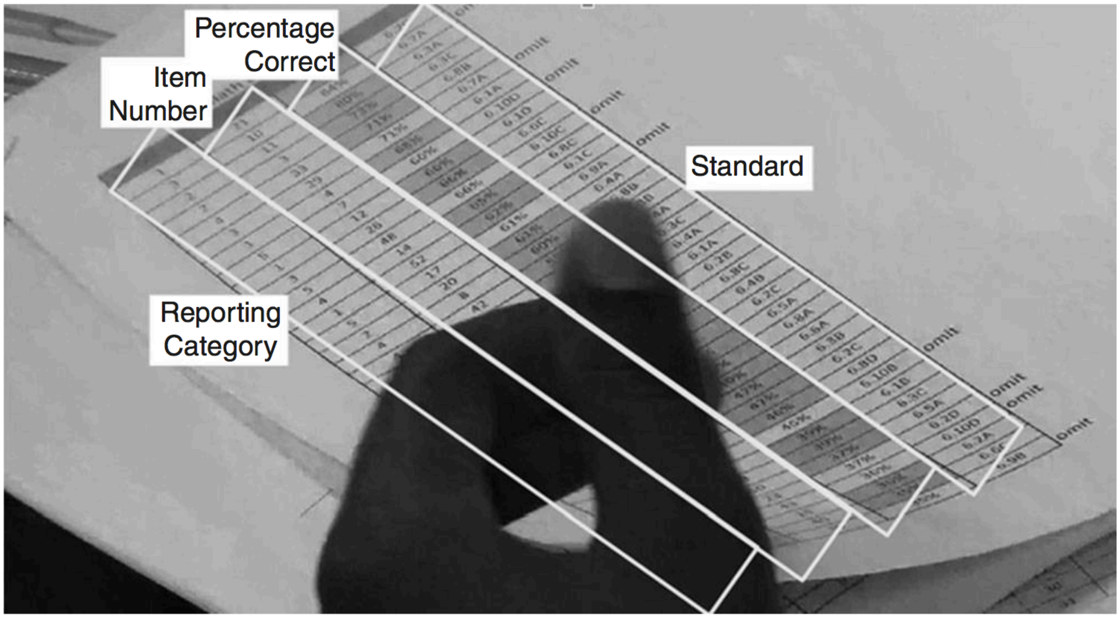


Figure III.2. List of benchmark test items by percentage correct, with standards and reporting categories listed, as provided to teachers during the Data Day.

Findings: Influence of accountability policy in data use for teaching and learning

The analysis revealed three primary ways that techno-rational accountability policy logics were reflected in the Riverview teachers' data-use practices:

(1) *reduction of complex constructs*: complicated constructs, like students' mathematical knowledge, were over-simplified and represented as quantitative variables;

(2) *remediation over instructional improvement*: instructional choices were aimed at remediating students' performance on specific questions, rather than supporting their mathematical understanding, reflecting a perverse incentive; and

(3) *enacted faith in instrument validity*: teachers acted as though they accepted questions at face value, without considering other factors that may reduce the validity of assessment items (e.g. students' comprehension of the wording of the items).

In the following sections, each of these techno-rational logics is explained, and then its reflection in practice is illustrated through an excerpt from the Data Day. Each illustrative excerpt is representative of the Riverview sixth-grade math teachers' work throughout the Data Day.

Reduction of complex constructs to quantitative variables

Under test-based accountability systems, complex constructs – including students' identities and mathematical ability – are reduced to quantitative variables. NCLB required that states disaggregate assessment data by subpopulations of students (including racial and ethnic groups, emergent bilingual students, students receiving special education services, and students from low-income families); this requirement continues under ESSA. Labeling students with discrete categories, whether by race, English proficiency, economic status, or special education qualification, collapses complicated identities and varied experiences into coarse measurement categories.

Students' mathematical knowledge is similarly reduced to a single score, which states use to categorize students based on their ability (e.g. as novice, proficient, or advanced). Using students' performance on a single assessment as a proxy for their mathematical knowledge and understanding is a gross oversimplification that fails to capture students' relative strengths and weaknesses across topics. Furthermore, assessments composed primarily of multiple-choice items that measure only a small piece of students' mathematical knowledge and skills, as they fail to capture more complex practices like proof or argumentation. Since such assessments are not sensitive to non-dominant forms of knowledge (Gutiérrez, 2008), students from historically marginalized groups are likely to be placed at a further disadvantage in this process.

Reduction of complex constructs in practice. During the Data Day, the Riverview teachers interpreted students' performance on individual items as representative of their mastery of the standard assessed by the item. A conversational routine (Horn and Little, 2010) emerged: The teachers typically began by identifying a particular question or standard that students had trouble with, taking a single number (e.g. 29 percent of students answered a question correctly) as a measurement of a complicated construct (e.g. students' understanding of the order of operations). In many EDRs, the teachers discussed strategies for reteaching the content, framing students as passive recipients of procedures instead of investigating how instruction might need to change to better address students' understandings. The following episode (Table III.1)[3] is representative of the teachers' data-use practices throughout the Data Day.

Table III.1. Episode of Data Reasoning I: Discussing instructional strategies for the order of operations.

1.	Rachel:	Let's talk about what we're gonna do with Standard 3E. Are we going to do the thing where they number each part?
2.	Devon:	Where, what's 3E again? The adding fractions?
3.	Rachel:	3E is order of operations
4.	Devon:	I'd say a checklist – like you go down a list and you check off answers
5.	Rachel:	You do what now?
6.	Devon:	((*Gesturing in the air as he speaks*)) 3E, order of operations? I'd make a PEMDAS[4] checklist, and you check it as you, you know, do each one. Does it have this? No. Does it have this? Yes, okay, then they do it line-by-line. Maybe we should have them do it line-by-line. Maybe we should emphasize doing it line-by-line, just so [unintel] the problem each time, so that it makes a little V shape
((*9 turns omitted: teachers discuss getting a snack*))		
15.	Rachel:	((*Laughs*)) Okay. So. You said you go line-by-line. Apparently what we've

[3] Turns at talk are numbered for identified speakers. Continuous speech at turn boundaries is shown with – long dashes. EMPHATIC talk is shown in caps, and elong:::ated enunciation is shown with repeated colons. ((*Activity descriptions*)) appear within double parens and in italics. Speech that was not captured clearly on the recording is noted with [Unintel].

[4] PEMDAS is a mnemonic for the order of operations: parentheses, exponents, multiplication, division, addition, and subtraction.

been doing doesn't work, but ((*laughs*))

16. Devon: [Unintel] redo it. When I tried just doing it the way I did, [unintel]

17. Rachel: They didn't get it

18. Devon: When you do one problem each time, and draw it line-by-line, then students follow it

19. Rachel: ((*Nodding*)) Mmhmm, that's good. I'm trying to remember, Ms. Jone – Ms. Jones showed me a way – it had to do with sticky notes. I think what she did was like, pretty much what you're saying. She – they had sticky notes for each part of the order of operations, and they would put the sticky note. Oh, they would move over the sticky note that's in this problem. K, in this problem, there's no grouping, there's no – so I'm just going to move over the ones that are in this problem ((*gesturing moving sticky notes from right to left side of workspace*)). And then like I have them stuck on my desk and I move over the ones ((*gesturing moving sticky notes from left to right side of workspace*)). Now, I have to put these that I moved over in order, according to the order of operations, and I look at that while I solve the problem. That was her idea, I think that's pretty good.

What is examined. In this EDR, the teachers interpreted students' low score on an item as indicative of students' understanding of the mathematical content. Students' understanding of the order of operations was reduced to a binary variable (whether they "got it" or not). At the start of this episode, Rachel looked at a list of assessment items by the percentage of students who answered correctly. She pointed to one of the lowest-scoring items on the list and noted that the item assessed Standard 3E, which addressed students' fluency with the order of operations. After identifying a problem area, Devon launched into a possible instructional strategy: teaching a checklist to apply the order of operations (Turn 6). Rachel then offered a similar strategy involving sticky notes (Turn 19). Reducing mathematical understanding to a binary and focusing on procedural reteaching strategies ends up simplifying complex phenomena of teaching and learning.

What is omitted. In other studies of data use under similar circumstances, there are a few instances of educators leveraging other resources – such as student work samples or recollections of classroom events – to make more nuanced interpretations (e.g. Horn et al., 2015). At Riverview, in contrast, the strict adherence to available data limited teachers' analysis of students' mathematical understanding. Devon and Rachel did not, for instance, seek out other questions about the order of operations, examine the distribution of students' responses, or even read the item in question. They also did not consider students' understanding, sensemaking, or funds of knowledge – resources that could be used to reframe the data in terms of the diverse assets and ideas that students bring to the classroom. Instead, students were only brought into the conversation insofar as they could follow a procedure "line-by-line." Since the teachers did not coordinate any other pieces of evidence, such as the format of the item, student work, or students' performance in class, there was little space for richer or more nuanced depictions of students' thought processes.

Remediation over instructional improvement

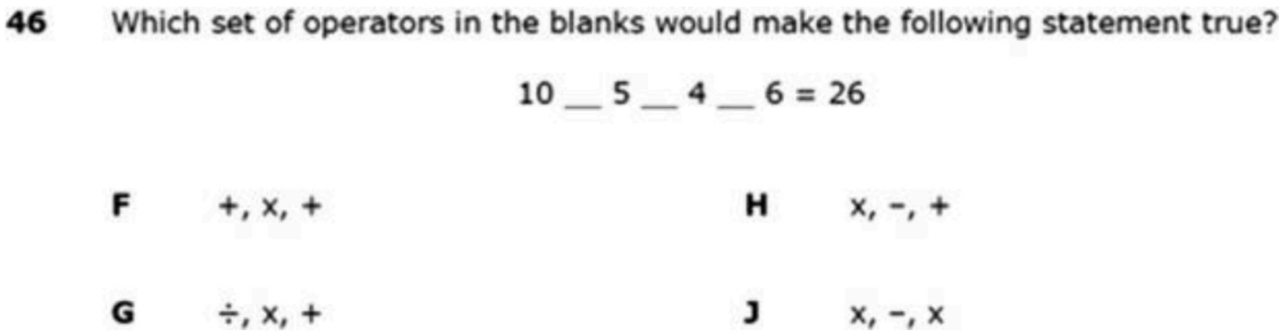
Accountability policies' emphasis on test scores as the primary metric of student learning creates a perverse incentive for teachers and schools. Ostensibly, schools could raise test scores through instructional improvement (Horn et al., 2015) – that is, by providing students richer and more rigorous learning opportunities and by supporting teachers to develop better instructional practices. This might involve using data (test results and student work) to consider students' sensemaking and to reflect on instruction. Such an approach allows teachers to design instruction and interventions that build on students' funds of knowledge and provide opportunities for all students to be successful. Improving instruction in these ways is likely to lead toward long-term gains in student learning. Yet instructional improvement efforts require time, resources, and supports for teacher learning (Darling- Hammond, 2007; Jackson et al., 2014). Accountability policies set ambitious goals for improvement that do not allow time for the false starts and uneven progress that are typical of workplace improvement efforts.

Instructional management approaches (popularly referred to as “teaching to the test” or “gaming the system”) are most pervasive in classes and schools with low-performing students (Diamond and Cooper, 2007). Since students who perform poorly on standardized assessments are disproportionately those in historically marginalized subpopulations, this works to reinforce and exacerbate educational inequities. In a desperate effort to raise test scores and avoid sanctions, teachers spend a great deal of time on test preparation and remediation, rather than teaching disciplinary content. Thus students from groups that have been historically marginalized in the US education system – Black, Latinx, Indigenous, low-income, and special education students – are systematically denied opportunities for rigorous mathematics instruction. This is Campbell’s Law in action: The more a quantitative social measure is used for decision-making, the more susceptible it is to corruption.

Remediation over instructional improvement in practice. In many EDRs, the Riverview teachers planned to reteach content based on specific items that students missed. This approach to data use prepares students to answer individual test items correctly but often does not support their deeper mathematical learning – or even prepare them to answer other non-isomorphic questions that address the same standard. In the EDR described above, the Riverview teachers identified the order of operations as a difficult standard for their students and described ways to reteach the content without examining the assessment item. Toward the end of the same episode (Table III.2), Crystal asked Rachel to look at the item in question (Figure III.3).

Table III.2. Episode of Data Reasoning I: Examining an item that assessed the order of operations.

37.	Crystal:	Where's – which – where's the problem at?
38.	Rachel:	I haven't even looked it up. Um, what? 3E it's 30, no 46
39.	Crystal:	46? Okay, I'll show you
40.	Rachel:	They have to figure out what part goes where. Oh, that one's a different kind of problem, I bet we haven't done many of those
41.	Crystal:	I've never done one of those
42.	Rachel:	Okay, so we need to make a note by that
43.	Crystal:	Yeah, what's 46? Is that 46, you said?
44.	Rachel:	Yeah. They have to put, putting symbols
45.	Crystal:	What we kind of do, indirectly, remember when they had to find the error?
46.	Rachel:	Yeah, yeah yeah. ((*Writing*)) "Putting symbols in to find answer"
47.	Crystal:	((*Organizing Devon's papers*)) Oooh, Devon is so messy. We have to teach him some organization
48.	Rachel:	Number 46. That's part of the problem, because I don't think they've done much of that
49.	Crystal:	Yeah
50.	Rachel:	There is somewhere, I saw a resource that has this. Where was that? It might be in [a test-prep book]. I can't remember where I saw that, but I know there's something that has a lot of that in it



46 Which set of operators in the blanks would make the following statement true?

10 __ 5 __ 4 __ 6 = 26

F +, x, + **H** x, −, +

G ÷, x, + **J** x, −, x

Figure III.3. An item from a sixth- grade district benchmark assessment assessing students' knowledge of the order of operations.

What is examined. In this exchange, Crystal and Rachel planned to remedy students' difficulty on an unusual order of operations item by providing additional practice on the same type of item. They noted that the item is atypical (Turns 40-41). A more typical item might include a numerical expression to simplify or a word problem requiring a multi-step calculation. Rachel identified a problem in that students have not seen many questions like Item 46 (Turn 48). Ultimately, she suggested finding similar items in a common test-prep book (Turn 50), ostensibly to use in future lessons. Discussion of the order of operations ended there; it was unclear from the Data Day discussion which of the competing lesson strategies (the checklist/sticky note strategy or problems similar to Item 46) the teachers intended to use with students. But both plans were organized around providing additional practice for students, and neither addressed students' mathematical understanding of the underlying principles of the order of operations.

What is omitted. Preparing students to answer specific test items might support their ability to answer similar items in the future but will not necessarily aid their success on other items on the same topic. For instance, strategies to determine the correct answer to Item 46 might not support students' ability to solve a word problem involving the order of operations. Furthermore, repeated practice on one type of item is unlikely to support students' deeper understanding of mathematical content. These approaches to data use are typical of instructional management logics (Horn et al., 2015). They limit both teachers' opportunities to learn to improve instruction and students' opportunities to learn mathematics more deeply.

Enacted faith in instrument validity

For decades, the educational reform community has expressed sincere skepticism over the validity of standardized assessments as measurements of student knowledge, connecting testing to behaviorist perspectives on learning that value product over process (Fenstermacher and Richardson, 2005; Schoenfeld, 1987; Stroup and Wilensky, 2000). Further research suggests that the most powerful predictor of a student's test score is their previous test score; students' performance is remarkably consistent across teachers, across years, and across content areas (Stroup, 2009). This indicates that rather than assessing students' knowledge of a particular subject in a particular year and using it as a proxy for teacher's instructional quality, it may be more accurate to consider standardized tests to be an indicator of students' test-taking ability.

There is further evidence that students from historically marginalized groups underperform on tests because of the tests themselves: That is, the assessments used to measure academic achievement are biased against Black and Latinx students, low-income students, and emergent bilinguals (Pham, 2009). Standardized tests are artifacts of a system centered on White, middle-class language and culture. Mathematics assessment items are often presented as word problems where familiarity with the context is taken as shared. Students' performance is, in part, a reflection of their facility with the items' linguistic and contextual demands, not solely the mathematical demands (Helms, 1992). To those familiar with school mathematics and White, middle-class culture, these items may seem innocuous; yet the assumed neutrality of the contexts and language used in mathematics assessments works to reinforce structural racism through colorblind ideology (Battey and Leyva, 2016). Using biased assessments as the markers of success or underachievement of marginalized populations continues the tradition the institutional injustices and disenfranchisement that these communities have historically faced and presently endure more broadly in society.

Enacted faith in instrument validity in practice. One characteristic of test-based accountability systems is that teachers have little control over the content of assessment items selected at the state and district levels. At Riverview, the teachers administered benchmark assessments created by district-level instructional leaders. These assessments were designed to mimic state tests, but were not psychometrically validated. Even so, the Riverview teachers acted on the assessment results at face value, without balking at their reliability.[5] This approach was not unique to Riverview; indeed, most workgroups and administrators in the larger study took assessments as valid measures of student learning. And yet, this is a reasonable approach: no matter what concerns educators may have about assessment items, they are still held accountable to students' scores; they have little recourse under the test-based accountability system.

During the following EDR, the sixth-grade teachers analyzed an assessment item that the majority of students missed (Item 31; Figure III.4). The item was intended to assess students' ability to generate equivalent forms of rational numbers – in this case, converting a percent (12.5%) to a fraction (1/8). However, this item presents an unusual context for the mathematical content: Discounts are usually given in whole percentages (e.g. 15 percent off, rather than 12.5 percent), and they are rarely represented as fractions. The syntactical structure of the item, with a dependent clause and multiple prepositional phrases, also places fairly high language demands on students (Zevenbergen, 2000) (Table III.3).

Table III.3. Episode of Data Reasoning II: Discussing students' understanding of rational number conversion.

1.	Crystal:	((*Reading*)) "Jeremy bought a skateboard on sale for $28, which was 12.5% off the original price. What was the discount as a fraction of the total price?" Well all they need to know is the 12 and a half percent

[5] There was at least one Riverview teacher (not in the focal sixth-grade group) who noted the disparity between the performance of two students who have similar grades in his class. As he held up their score sheets, he observed, "She always has a book in her hand. He hates to read." This suggested that he saw a conflation of math achievement and reading level, yet he proceeded to act as if the scores only reflected the former.

2. Devon: The majority of students chose B

3. Crystal: Why:

4. Devon: 12 over 5

5. Crystal: That's CRAZY

6. Rachel: All they did was take the numbers and – that makes me mad ((*laughs*))

7. Devon: Well, [unintel]

8. Crystal: But I totally agree with what Ms. Cardwell is saying that, even when they do stuff like that, that still suggests that –

9. Rachel: They don't know it

10. Crystal: – they are not confident, and that they don't know it, so they're going with the next best thing in their mind

11. Rachel: Yeah, see – the percent has a decimal in it, and they haven't learned how to move that decimal or figure out what 12 and a half percent really means

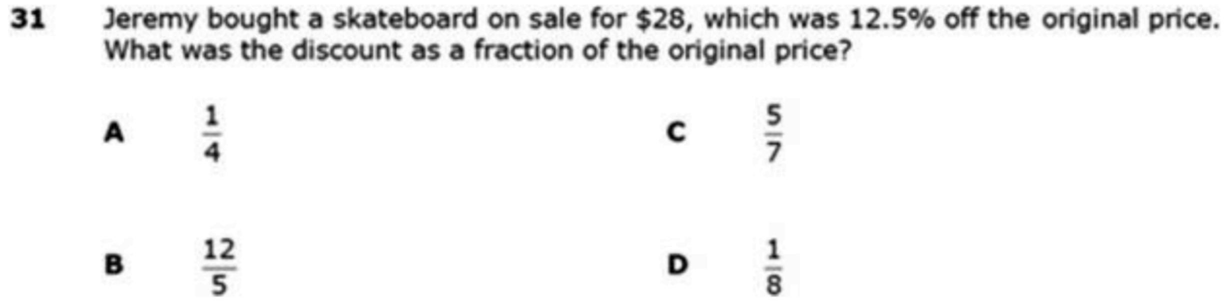
12. Devon: But we're always doing conversions between decimals and percents – we did yesterday and the day before that and the day before that and the day before that And we're always [unintel]

13. Crystal: We did a project and they still didn't get it

14. Rachel: We can find [...]

15. Devon: Told you

16. Rachel: I can see why they missed that. It's because they had to ((*5 sec pause*)) Okay, we need to do some problems with percent that has a decimal in it.



31 **Jeremy bought a skateboard on sale for $28, which was 12.5% off the original price. What was the discount as a fraction of the original price?**

A $\frac{1}{4}$ **C** $\frac{5}{7}$

B $\frac{12}{5}$ **D** $\frac{1}{8}$

Figure III.4. An item from a sixth-grade district benchmark assessment assessing students' knowledge of rational number conversion.

What is examined. As the teachers discussed this item, they accepted the contextual and linguistic features of the task. Crystal, for instance, read the problem and immediately stated "Well all they need to know is the 12 and a half percent" (Turn 1), suggesting that she found the wording of the item to be straightforward and unproblematic. Devon noted that most students chose 12/5, which was farthest from the correct answer numerically (Turns 2 and 4). Rachel implied that students moved the numbers in 12.5 percent to arrive at the answer 12/5 (Turn 11); later in this EDR, she asserted that students were "goofed up" by the decimal in the percent because they "haven't seen that enough," despite doing repeated practice and a project on rational number conversion. The teachers concluded their initial discussion of this problem with Rachel's suggestion to "do some problems with a percent that has a decimal in it" (Turn 16).

What is omitted. Within this exchange, the Riverview teachers spent little time considering student thinking around the problem. Though students' equating of 12.5 percent and 12/5 raises a valid concern about their understanding of the relationship between decimals and percents, there were other potential sources of confusion on Item 31, including students' interpretation of the complex syntax of the word problem. Perhaps, for instance, students abandoned hope of interpreting the problem because of its complex structure or unusual context and looked for something that was the same in the question and answer. Without considering the various ways that students might interpret an assessment item, teachers are limited in their ability to adjust instruction to meet their students' learning needs.

Implications: Bringing in a critical lens

Recall that a goal of this analysis is to bring a critical lens to teachers' data-use practices, reflecting on the meanings created in light of broader social-historical processes. Given the purported equity goals of accountability policies – which largely drive teachers' use of benchmark assessment data – there ought to be space to consider equity in teachers' data-use conversations. Yet this critical examination of accountability policy's influences on the Riverview teachers' data-use practices reveals that their approach leaves little room for considerations of students' agency, experiences, or funds of knowledge, which are important building blocks for more equitable instruction.

As noted above, accountability policies reduce students' identities and knowledge to quantitative variables. Since the Riverview teachers analyzed item-level data (aggregated across all students in the grade), they did not have data on student characteristics immediately available. As such, they largely ignored students' identities or funds of knowledge throughout the Data Day. This approach inadvertently reinforced such assessments as culturally neutral, rather than acknowledging them as artifacts of a deeply inequitable system. The lived experiences, counternarratives, and funds of knowledge that students from non-dominant backgrounds could bring to bear in their academic work are not recognized on assessments and are thus undervalued by accountability policies (Solórzano and Yosso, 2001). In a diverse classroom or school environment, such variation in perspectives could be leveraged to provide for more rigorous and more equitable learning opportunities. By emphasizing test scores as measures of student learning and achievement gaps as evidence of groups' inadequacies, test-based accountability policies leave little room to build upon the sociocultural resources that students bring to the classroom.

The perverse incentives created by an overemphasis on test scores disproportionately affect low-performing schools, which frequently have higher proportions of students from historically marginalized groups (Diamond and Cooper, 2007). This contributes to the paradox of accountability as a strategy to redress educational inequality. Teachers of students who underperform on standardized achievement tests are incentivized to "reteach" instead of being incentivized to teach for deeper understanding. Indeed, this played out in many EDRs across the Riverview Data Day. The overall effect of these perverse incentives is that, instead of meeting the promise of improving instruction for historically underserved students in under- resourced schools, accountability policies perpetuate pedagogies of poverty (Oakes, 1990) by emphasizing low-level skills and memorization. This systematically limits learning opportunities for students who are Black, Latinx, Indigenous, emergent bilinguals, or from low-income families, thereby reinforcing existing inequities.

The Riverview teachers placed faith in the test validity, in part, because they were not positioned to critique the appropriateness of assessment items In considering the systemic processes underlying their conversation about Item 31 (Figure III.4), it is important to note that a critique about syntax would have limited traction in the teachers' workplace. They had little say in the creation of district benchmark assessments, no mechanism for "talking back" to poorly constructed items – let alone opportunities for input on state assessments. Item 31 was meant to assess students' capacity to represent percents as fractions, yet that skill may have been obfuscated by a complex syntactical structure in the item. One might imagine a student who is able to represent a percent as a fraction and who could show success on a rational number project, yet has trouble parsing a complex word problem with an atypical context. But to school and district leaders, the salient metric of the teachers' effectiveness in teaching this standard is a single item on a benchmark assessment, whether or not it is well designed.

Discussion

This inquiry into the Riverview teachers' data-use practices, which are consistent with the structures of accountability policy and its underlying techno-rational logic, suggests that student data from state- and district-mandated standardized tests do not necessarily support meaningful considerations of diversity and equity in the classroom. The teachers focused on content – namely standards and related test items – in their discussion of data, almost to the exclusion of student thinking. Their patterns of conversation presented students as monolithically interacting with test items in static, one-dimensional ways. Indeed, the representations of data that the teachers had available encouraged this sort of analysis. Students' mathematical knowledge and ways of thinking were reduced to responses on multiple-choice items This limited opportunities for teachers to discuss who students are, how they think mathematically, and what forms of knowledge and experiences they may bring to bear in mathematical contexts. Consequently, students' cultural identities, experiences in schools, and funds of knowledge had no weight in the data analysis or impact on plans for future instruction.

This analysis reflected on each excerpt through the lens of critical theory, highlighting that data-use conversations limit teachers' possibilities to develop culturally responsive teaching practices or otherwise subvert inequitable policies. Of course, there may be other avenues in which the Riverview teachers pursued more equitable instructional practices and worked for social justice at their school. But given the purported equity goals of accountability policy, it is deeply troubling that the policy's techno-rational logic disallows considerations of students' experiences and funds of knowledge, provides perverse instructional incentives that encourage remediation, and creates structures that position teachers to have blind faith in the validity of assessment instruments.

Even with the unusual luxury of an entire day devoted to data use, there were virtually no opportunities for issues of equity to be brought into activities like the Data Day, since the stakes associated with standardized tests are high. Scores are used for a variety of important decisions, including grade-level promotion, teacher tenure, and funding for schools and districts. This creates intense, unfair pressure for teachers and administrators to raise scores in a system that does not support success for students of color, emergent bilinguals, and students from low-income families. For instance, decisions are often made to place students in more or less advanced curricular pathways (higher or lower tracks) in mathematics based on test scores (Datnow and Hubbard, 2015). Since students of color, emergent bilingual students, and students from low-income families are most likely to perform poorly on these assessments, they are disproportionately represented in these lower-tracked classes (Horn, 2007; National Science Board, 2014; Oakes, 1985). In lower-tracked classes, students receive more procedural and didactic teaching and are consistently denied access to ambitious instruction. Teachers then tend to perceive the students in lower-tracked classrooms as less capable, and their assumptions about student deficits appear in their curriculum and teaching practice (Horn, 2007).

For data use to serve the goals of educational equity, teacher leaders, administrators, and coaches need to steer teachers' conversations about data away from mere reteaching towards considerations of student thinking, students' experiences and resources, and their cultural funds of knowledge so that differences can be leveraged and acknowledged. Mechanisms need to be set up for teachers to confront systems that (re)produce inequities: They might "talk back" to inappropriate items to increase validity and reliability of instruments or resist efforts to use high-stakes assessments for tracking. In order for this to happen, scholars and practitioners need to assume a critical stance toward the testing apparatus and accountability policies. The broader educational community must question the measuring stick used for evaluating teaching and learning, considering the likelihood that "achievement gaps" reflect test design rather than students' differing abilities. Likewise, teachers and administrators should be supported in asking critical questions of assessment structures and encouraged to use a variety of data points to inform instructional practices. Reducing the high stakes associated with testing is the next step, as the pressure associated with assessments creates perverse incentives for improving test scores rather than instruction. Moreover, instructional leadership, curricular resources, and future research designs should be oriented toward instructional improvement, starting by bringing the focus of data use back to students, their experiences, their thinking, and their identities.

References

Aguirre, J.M. and Zavala, M. (2013), "Making culturally responsive mathematics teaching explicit: a lesson analysis tool", Pedagogies: An International Journal, Vol. 8 No. 2, pp. 163-190.

Ah-Teck, J.C. and Starr, K.E. (2014), "Total quality management in Mauritian education and principals' decision-making for school improvement: 'driven' or 'informed' by data?", Journal of Educational Administration, Vol. 52 No. 6, pp. 833-849.

Battey, D. and Leyva, L.A. (2016), "A framework for understanding whiteness in mathematics education", Journal of Urban Mathematics Education, Vol. 9 No. 2, pp. 49-80.

Bertrand, M. and Marsh, J.A. (2015), "Teachers' sensemaking of data and implications for equity", American Educational Research Journal, Vol. 52 No. 5, pp. 861-893.

Darling-Hammond, L. (2007), "Race, inequality and educational accountability: the irony of 'no child left behind' ", Race Ethnicity and Education, Vol. 10 No. 3, pp. 245-260.

Datnow, A. and Hubbard, L. (2015), "Teachers' use of assessment data to inform instruction: lessons from the past and prospects for the future", Teachers College Record, Vol. 117 No. 4, pp. 1-26.

Delpit, L. (1988), "The silenced dialogue: power and pedagogy in educating other people's children", Harvard Educational Review, Vol. 58 No. 3, pp. 280-299.

Derry, S.J., Pea, R.D., Barron, B., Engle, R.A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J.L., Sherin, M.G. and Sherin, B.L. (2010), "Conducting video research in the learning sciences: guidance on selection, analysis, technology, and ethics", The Journal of the Learning Sciences, Vol. 19 No. 1, pp. 3-53.

Diamond, J.B. and Cooper, K. (2007), "The uses of testing data in urban elementary schools: some lessons from Chicago", Yearbook of the National Society for the Study of Education, Vol. 106 No. 1, pp. 241-263.

Emerson, R.M., Fretz, R.I. and Shaw, L.L. (2011), Writing Ethnographic Fieldnotes, University of Chicago Press, Chicago, IL.

Fenstermacher, G.D. and Richardson, V. (2005), "On making determinations of quality in teaching", Teachers College Record, Vol. 107 No. 1, pp. 186-213.

Gay, G. (2010), Culturally Responsive Teaching: Theory, Research, and Practice, Teachers College Press, New York, NY.

Goldstein, B.E. and Hall, R. (2007), "Modeling without end: conflict across organizational and disciplinary boundaries in habitat conservation planning", in Lesh, R., Hamilton, E. and Kaput, J.J. (Eds), Foundations for the Future in Mathematics Education, Lawrence Erlbaum Publishing Company, Mahwah, NJ, pp. 57-76.

Gutiérrez, R. (2008), "A 'gap-gazing' fetish in mathematics education? Problematizing research on the achievement gap", Journal for Research in Mathematics Education, Vol. 38 No. 4, pp. 357-364.

Gutiérrez, R. (2009), "Framing equity: helping students 'play the game' and 'change the game'", Teaching for Excellence and Equity in Mathematics, Vol. 4 No. 1, pp. 1-3.

Gutiérrez, R. (2013), "The sociopolitical turn in mathematics education", Journal for Research in Mathematics Education, Vol. 44 No. 1, pp. 37-68.

Gutiérrez, R. (2016), "Strategies for creative insubordination in mathematics teaching", Teaching for Excellence and Equity in Mathematics, Vol. 7 No. 1, pp. 52-60.

Hall, R. and Horn, I.S. (2012), "Talk and conceptual change at work: adequate representation and epistemic stance in a comparative analysis of statistical consulting and teacher workgroups", Mind, Culture, and Activity, Vol. 19 No. 3, pp. 240-258.

Helms, J.E. (1992), "Why is there no study of cultural equivalence in standardized cognitive ability testing?", American Psychologist, Vol. 49 No. 9, pp. 1083-1101.

Horn, I.S. (2007), "Fast kids, slow kids, lazy kids: framing the mismatch problem in mathematics teachers' conversations", The Journal of the Learning Sciences, Vol. 16 No. 1, pp. 37-79.

Horn, I.S. (2016), "Accountability as a design for teacher learning' sensemaking about mathematics and equity in the NCLB era", Urban Education.

Horn, I.S. and Little, J.W. (2010), "Attending to problems of practice: routines and resources for professional learning in teachers' workplace interactions", American Educational Research Journal, Vol. 47 No. 1, pp. 181-217.

Horn, I.S., Kane, B.D. and Wilson, J. (2015), "Making sense of student performance data data use logics and mathematics teachers' learning opportunities", American Educational Research Journal, Vol. 52 No. 2, pp. 208-242.

Jackson, K., Cobb, P. and Rigby, J.G. (2014), "Instructional improvement and instructional management: district leaders' orientations towards improving mathematics teaching and learning", paper presented at the University Council for Educational Administration, Washington, DC, November 22.

Jennings, J. (2012), "The effects of accountability system design on teachers' use of test score data", Teachers College Record, Vol. 114 No. 11, pp. 1-23.

Jordan, B. and Henderson, A. (1995), "Interaction analysis: foundations and practice", The Journal of the Learning Sciences, Vol. 4 No. 1, pp. 39-103.

Ladson-Billings, G. (2006), "From the achievement gap to the education debt: understanding achievement in US schools", Educational Researcher, Vol. 35 No. 7, pp. 3-12.

Ladson-Billings, G. (2009), The Dreamkeepers: Successful Teachers of African American Children, John Wiley and Sons, San Francisco, CA.

Lave, J. and Wenger, E. (1991), Situated Learning: Legitimate Peripheral Participation, Cambridge University Press, Cambridge.

Lynch, D., Smith, R., Provost, S. and Madden, J. (2016), "Improving teaching capacity to increase student achievement: the key role of data interpretation by school leaders", Journal of Educational Administration, Vol. 54 No. 5, pp. 575-592.

Marsh, J.A. (2012), "Interventions promoting educators' use of data: research insights and gaps", Teachers College Record, Vol. 114 No. 11, pp. 1-48.

Milner, H.R. (2013), "Analyzing poverty, learning, and teaching through a critical race theory lens", Review of Research in Education, Vol. 37 No. 1, pp. 1-53.

Moll, L.C. and González, N. (2004), "Engaging life: a funds of knowledge approach to multicultural education", in Banks, J.A. and McGee Banks, C.A. (Eds), Handbook of Research on Multicultural Education, Jossey-Bass, New York, NY, pp. 699-715.

Monzó, L.D. (2016), "'They don't know anything!': Latinx immigrant students appropriating the oppressor's voice", Anthropology and Education Quarterly, Vol. 47 No. 2, pp. 148-166.

Munter, C. (2014), "Developing visions of high-quality mathematics instruction", Journal for Research in Mathematics Education, Vol. 45 No. 5, pp. 584-635.

National Science Board (2014), “Science and engineering indicators 2014”, National Science Foundation (NSB 14-01), Arlington, VA.

No Child Left Behind (NCLB) Act of 2001 (2002), Pub. L. No. 107-110, § 115, Stat. 1425. Oakes, J. (1985), Keeping Track: How Schools Structure Inequality, Yale University Press, New Haven, CT.

Oakes, J. (1990), “Opportunities, achievement, and choice: women and minority students in science and mathematics”, Review of Research in Education, Vol. 16 No. 1, pp. 153-222.

Ochs, E., Gonzales, P. and Jacoby, S. (1996), “ ‘When I come down I’m in the domain state’: grammar and graphic representation in the interpretive activity of physicists”, in Ochs, E., Schegloff, E. and Thompson, S. (Eds), Interaction and Grammar, Cambridge University Press, Cambridge, pp. 328-369.

Pham, V.H. (2009), “Computer modeling of the instructionally insensitive nature of the Texas assessment of knowledge and skills (TAKS) exam”, unpublished dissertation, University of Texas at Austin, Austin, TX.

Philip, T. and Garcia, A. (2013), “The importance of still teaching the iGeneration: new technologies and the centrality of pedagogy”, Harvard Educational Review, Vol. 83 No. 2, pp. 300-319.

Philip, T.M. and Garcia, A. (2015), “Schooling mobile phones assumptions about proximal benefits, the challenges of shifting meanings, and the politics of teaching”, Educational Policy, Vol. 29 No. 4, pp. 676-707.

Philip, T.M., Olivares-Pasillas, M.C. and Rocha, J. (2016), “Becoming racially literate about data and data-literate about race: data visualizations in the classroom as a site of racial-ideological micro-contestations”, Cognition and Instruction, Vol. 34 No. 4, pp. 361-388.

Philip, T.M., Martinez, D.C., Lopez, E. and Garcia, A. (2016), “Toward a teacher solidarity lens: former teachers of color (re)envisioning educational research”, Race Ethnicity and Education, Vol. 19 No. 1, pp. 182-199.

Rollock, N. and Gillborn, D. (2011), “Critical race theory (CRT), British educational research association”, available at: www.bera.ac.uk/researchers-resources/publications/critical-race- theory-crt (accessed March 22, 2015).

Sacks, H. (1967/1992), “Omnirelevant devices; settinged activities; ‘indicator terms’ (February 16, 1967)”, in Jefferson, G. (Ed.), Lectures on Conversation: Volumes I and II, Blackwell, Oxford, pp. 515-522.

Sacks, H., Schegloff, E.A. and Jefferson, G. (1974), “A simplest systematics for the organization of turn-taking for conversation”, Language, Vol. 50 No. 4, pp. 696-735.

Schegloff, E.A. (1992), “On talk and its institutional occasions”, in Drew, P. and Heritage, J.

(Eds), Talk at Work: Interaction in Institutional Settings, Cambridge University Press, Cambridge, pp. 101-134.

Schoenfeld, A.H. (1987), Cognitive Science and Mathematics Education, Lawrence Erlbaum Associates, Hillsdale, NJ.

Solomon, R.P., Portelli, J.P., Daniel, B.J. and Campbell, A. (2005), “The discourse of denial: how white teacher candidates construct race, racism and ‘white privilege’ ”, Race Ethnicity and Education, Vol. 8 No. 2, pp. 147-169.

Solórzano, D.G. and Yosso, T.J. (2001), “From racial stereotyping and deficit discourse toward a critical race theory in teacher education”, Multicultural Education, Vol. 9 No. 1, pp. 2-8.

Stevens, R. (2010), “Learning as a members’ phenomenon: toward an ethnographically adequate science of learning”, Yearbook of the National Society for the Study of Education, Vol. 109 No. 1, pp. 82-97.

Stroup, W. (2009), “What bernie madoff can teach us about accountability in education”, Education Weekly, Vol. 28 No. 25, pp. 22-23.

Stroup, W.M. and Wilensky, U. (2000), “Assessing learning as emergent phenomena: moving constructivist statistics beyond the bell curve”, in Kelly, A.E. and Lesh, R.A. (Eds), Handbook of Methods for Research in Learning and Teaching Science and Mathematics, Routledge, New York, NY, pp. 877-911.

Tate, W.F. (1997), “Critical race theory and education: history, theory, and implications”, Review of Research in Education, Vol. 22 No. 1, pp. 195-247.

Villegas, A.M. and Irvine, J.J. (2010), “Diversifying the teaching force: an examination of major arguments”, The Urban Review, Vol. 42 No. 3, pp. 175-192.

Yin, R.K. (2009), Case Study Research: Design and Methods, Sage Publications, Thousand Oaks, CA.

Zevenbergen, R. (2000), “ ‘Cracking the code’ of mathematics classrooms: school success as a function of linguistic, social and cultural background”, in Boaler, J. (Ed.), Multiple Perspectives on Mathematics Teaching and Learning, Greenwood Publishing Group, Westport, CT, pp. 201-223.

CHAPTER IV

THE EPISTEMIC FOUNDATIONS OF TEACHERS' DATA USE: WHAT DO THE DATA SAY?

In the era of test-based accountability policies like the No Child Left Behind (NCLB) Act and the Every Student Succeeds Act (ESSA), U.S. educators are held accountable to increasing students' scores on standardized assessments, particularly in the frequently-tested content areas of mathematics and English-language arts. Teachers are encouraged to use data to "drive" instruction, but the details of this process vary across contexts (Datnow & Hubbard, 2015): What counts as data? Which data are emphasized, and for which students? How do educators draw conclusions from data? The answers to these questions have serious implications for teachers' practice and students' learning opportunities.

The details of data use for instructional improvement are frequently left unspecified. Policymakers, administrators, and educational leaders often describe data-driven decision-making (DDDM) as rational and objective. There are many DDDM models available (e.g., Mandinach, 2012) that outline similar processes: Educators collect, organize, analyze, and synthesize data, and then make and implement a decision. While such models suggest a logical path to organizational improvement, they oversimplify data use by assuming that educators have shared understandings about what data represent, what can be learned from data, and how to respond to data. Ethnographic studies of scientists have shown that analyzing data to make evidence-based decisions is far more complicated than the DDDM rhetoric suggests (Pickering, 2010; Goldstein & Hall, 2007). The ways that educators analyze, learn from, and respond to data are similarly complex — and likely even more so, considering the inherently social and interactive nature of teaching and learning.

Data use is an inherently interpretive activity, as educators identify patterns within the data and draw conclusions about what data mean. They use data to learn about their instructional efficacy and students' progress, and adjust future instruction in response. But the education profession lacks shared norms and practices for generating such knowledge — particularly for generating knowledge that would support instructional improvement (Glazer & Peurach, 2015). Without a coherent epistemic community to guide reform efforts, bureaucratic controls such as test-based accountability politics and DDDM have little hope of improving teachers' practice or student learning outcomes. Indeed, "data-driven" reforms have had mostly distortive effects on teaching and learning, as educators working under accountability pressure attempt to raise test scores by teaching to the test, narrowing the curriculum, and emphasizing the success of some students at the expense of others (Booher-Jennings, 2005; Jennings & Bearak, 2014; Lee, Louis, & Anderson, 2012). Such distortions exacerbate systemic racism and classism at the district and school levels, as accountability pressure is strongest for students from historically marginalized communities (e.g., Horn, 2016; Khalifa, Jennings, Briscoe, Oleszweski, & Abdi, 2013).

Yet the solution is not to disband with data use entirely. The underlying notion of data-driven decision-making has merit: Teachers should (and often do) use knowledge of students' progress to inform their instructional choices. Good teaching — what many scholars have come to call "ambitious" teaching — requires that teachers adjust instruction based on what students know and are able to do (Lampert, Boerst, & Graziani, 2011). Furthermore, almost all successful school improvement efforts cite data use as a central strategy for improving instruction (e.g., Schaffer, Reynolds, & Stringfield, 2012; Villavicencio & Grayman, 2012). So it seems that good instruction relies on student learning data, but using student learning data does not always result in good instruction. What accounts for this disconnect? Why does data use improve teaching and learning in some settings, but distort teaching and learning in others?

I argue that this is an issue of epistemics: what teachers can know about student learning, how they come to know it, and how they use it to inform instruction. In this analysis, I compare two cases of teacher workgroups analyzing data under the guidance of expert instructional coaches. Though both coaches describe an ambitious vision for instruction, they approach data use from different epistemic stances. This shapes the workgroups' data use practices, their plans for future instruction, and ultimately students' learning opportunities. This analysis sheds light on what is often a tacit assumption that educators hold about data use. By investigating the underlying epistemics of educators' data use practices, this analysis can help researchers and teacher educators better understand how teachers can use data to inform their instruction.

Using data to inform instruction

The push for data-driven decision-making is a relatively new phenomenon, but the underlying principle — that is, using evidence of student learning to inform instruction — has long been a part of teachers' practice. Thoughtful teachers have always considered students' prior knowledge and skills as they planned for instruction (Black & Wiliam, 1998), as is captured in the adage, "Start with where the students are." Teachers collect information about what students know and are able to do from a variety of sources, including conversations with students, observations of students during class, written work, and various formal and informal assessments (Young & Kim, 2010). Each of these pieces of evidence could be considered data. But over the last two decades, the term *data* has narrowed to refer primarily to quantitative data from multiple-choice assessments (Marsh, Pane, & Hamilton, 2006). Even though teachers can (and do) still consider other sources of evidence of student learning, quantitative assessment data is typically what they refer to when they describe their use of data or being "data-driven."

This shift in the meaning of *data* has come hand-in-hand with the U.S. education system's move toward standards-based education and test-based accountability. This move is most visible through policies like the No Child Left Behind Act of 2001 (NCLB), the Race to the Top initiative of 2009 (RttT), and the Every Student Succeeds Act of 2016 (ESSA). Under these programs, each state is required to administer high-stakes end-of-year assessments to all students; students are tested in mathematics in Grades 3-8 and at least once in high school. State tests typically include primarily (though not exclusively) multiple-choice assessment items (Yuan & Le, 2012). Data from state tests are typically not available until after the end of the school year, so they are not useful for informing day-to-day instruction. As a result, many schools and districts administer interim "benchmark" assessments at regular points throughout the school year (Datnow & Hubbard, 2015). Benchmark assessments are meant to mimic the state test in both format and content; they, too, typically comprise multiple-choice items Thus, the data that are most consequential for gauging student success — their performance on high-stakes assessments and tests meant to mimic high-stakes assessments — are primarily quantitative.

Within the context of test-based accountability, the U.S. Department of Education has promoted DDDM as a key element of educational reform (Duncan, 2009; Hamilton, Halverson, Jackson, Mandinach, & Supovitz, 2009; Means, Padilla, & Gallagher, 2010; National Forum on Education Statistics, 2011). The logic behind DDDM efforts is straightforward: After assessing student learning, educators can use data to identify trouble spots (e.g., "failing" students or schools, achievement differences between groups of students, or content that students are struggling with). Then teachers can adjust instructional choices to address (or “remediate”) these gaps. By regularly engaging in this process, students will make gradual progress toward end-of-year goals. To ensure that educators will be invested in this process and that students have access to high-quality schools, states threaten sanctions if improvement efforts are unsuccessful.

In some cases, data use works as intended. For instance, Diamond and Cooper (2007) found that among schools in a district with a high-stakes testing system, some teachers (largely those in schools that were not in danger of failing to meet accountability goals) used data to reflect on instruction and change their teaching practice to support deeper student learning. Even in high-pressure contexts, data use has been a part of many successful school turnaround efforts (e.g., Schaffer et al., 2012; Villavicencio & Grayman, 2012). Educators in these settings used data to identify students who could benefit from additional supports, like additional instructional time or small-group tutoring. They also used data to identify the knowledge and skills that students found difficult, and used that to address specific academic skills within the support structures. Based on positive examples like these, data use is touted as the most effective way to support instructional improvement and student learning (Duncan, 2009; Hamilton, Halverson, Jackson, Mandinach, & Supovitz, 2009; Means, Padilla, & Gallagher, 2010; National Forum on Education Statistics, 2011).

But the road from data collection to student learning is neither clear nor direct. Despite the optimistic promise of using data to promote achievement and equity, DDDM efforts can distort teaching in ways that dry-dock teachers and limit students’ learning opportunities. In the name of data-driven decision-making, educators engage in educational triage and reinforce deficit orientations toward “failing” students (Jennings, 2012; Horn, 2016). They narrow the curriculum, emphasizing procedural skills and test-prep strategies (Jennings & Bearak, 2014).

These and other distortions — what I call Distortive Data Use Practices in Education (DDUPEs; Paper 1) — are exacerbated in schools facing the greatest pressure from test-based accountability policies (Diamond & Cooper, 2007).

Reconciling these different data use stories is not a matter of determining *whether* data use is promising or dangerous (it is both). Rather, it is a question of understanding *how* educators use data and *under what conditions* data use supports instructional improvement and student learning. Until recently, there have been few studies of educators' data use practices (Coburn & Turner, 2011; Little, 2011). The emerging literature in this area shows that educators' sensemaking is consequential for the types of decisions they make (e.g., Bertrand & Marsh, 2015; Horn, Kane, & Wilson, 2015; Park, Daly, & Guerra, 2012). I build on this research base by examining educators' epistemic stances on data: what they can know from data, how they come to know it, and why it is of value. I argue that the effectiveness of educators' data use stems primarily from their epistemic stances on data.

Data Use as an Epistemic Quandary

Data use is an inherently epistemic endeavor, as educators examine data in order to draw conclusions about the world. Some DDDM models and guides (e.g., Boudett et al., 2005; Mandinach, 2012) present data use as a technical process, wherein educators collect, organize, analyze, and synthesize data in straightforward and rational ways that reveal clear-cut next steps for improvement. But the ways that educators actually engage in that process — their *data use practices* (Coburn & Turner, 2011) — are laden with choices: What data should we collect? How should we organize and analyze them? What conclusions can we draw from the data? What are our goals, and what changes will support progress toward those goals?

Educators' answers to these questions are based, in part, on their *epistemic stances* around data — that is, their perspectives on what can be known from data, how to know it, and why it is of value (Horn et al., 2015; Horn & Kane, 2015). Many advocates of DDDM elide these questions in their calls for teachers and schools to be more "data-driven." They often animate data as the subject of DDDM, as did former Secretary of Education Arne Duncan when he touted the "the power of data to drive our decisions" because data "tells us where we are, where we need to go, and who is most at risk" (Duncan, 2009, para. 5). But data cannot drive or say anything; it is people who interpret and draw conclusions from data.

Because data are so often presented as the subject of data-driven decision-making, rather than as the object of data use, educators' epistemic stances on data often remain tacit. Framing data use as a matter of listening to what data "say" allows educators with different epistemic stances to "hear" very different things. A C+ on a test, for instance, could be interpreted as a bad score (depending, perhaps, on the student's prior performance). A bad score could be attributed to a student's lack of understanding, a teacher's ineffective teaching, a test-writer's poor item design, or any number of other issues. The epistemic stances from which educators approach data use shape the conclusions that they find probable (or even possible). But rhetoric that positions data use as an objective process — driven by data rather than human interpretation — invites educators to overlook these differences. As a result, educators use terms like "data-driven decision-making" to refer to a diverse set of data use practices that stem from various — and sometimes contradictory — epistemic stances.

It is, perhaps, unsurprising that there is a lack of consensus around data use in education, as the profession lacks a coherent epistemic community (Glazer & Peruach, 2015). That is, educators are not organized around shared tools, theories, or methods of communicating new expectations or generating new knowledge. This impedes all educational reform efforts, but it is especially salient for DDDM. Unlike topics like classroom management or pedagogical methods, data use is rarely addressed in teacher preparation programs (Mandinach & Gummer, 2013). Professional development efforts to support in-service teachers' data use typically emphasize practical and logistical concerns like accessing data management software, instead of addressing the epistemic demands of data use (Jimerson & Wayman, 2015). And federal, state, and local policies often articulate an expectation for DDDM without describing any particular data use practices (e.g., Means et al., 2010). As a result, there is little consensus within the profession about how to use data to generate knowledge about student learning.

Data Use Conversations as Spaces for Teacher Learning

Teachers' collaborative workgroups are one setting with the potential to support the development of an epistemic community, at least on a small scale. In an effort to support teachers' professional development, many districts and schools have organized the school day to provide teachers time to meet with colleagues who teach the same grade level or content area. Workgroup meetings are a common site for data use, as teachers analyze and discuss data from classroom assignments, exit tickets, district benchmark assessments, and other sources. When data use is a regular element of their practice, workgroups can develop shared ways of interpreting and generating knowledge from data.

Drawing on the work of Horn and colleagues investigating professional learning opportunities in teacher workgroups, I focus on mathematics teachers' discussions in workgroup meetings, often under the facilitation of a principal or instructional coach. I take a situative view of teacher learning, assuming that learning happens in interaction (Engeström & Sannino, 2010). To study teachers' professional learning opportunities, I examine how interactions (a) marshal conceptual resources for teachers and (b) mobilize teachers for future work (Hall & Horn, 2012; Horn et al., 2015). Most workgroup meetings mobilize teachers for future work, as they share instructional strategies or align the pace of instruction, often without developing concepts through their discussion. But the richest learning opportunities occur when teachers collectively develop concepts about pedagogical issues, and then connect the concepts to their future instruction (Horn, Garner, Kane, & Brasel, 2017). Horn and colleagues (2015) identified four key elements of workgroup conversations that shape the nature of learning opportunities: 1) activity structures, which are the ways tasks are carried out; 2) frames, which are the ways issues or problems are defined; 3) representations of practice, which are ways of sharing aspects of instruction; and 4) epistemic stances.

In the present inquiry, I build on this literature by analyzing the ways that teacher workgroups make sense of student assessment data to plan for future instruction, paying particular attention to the role of epistemic stances in shaping the workgroups' data use practices.

Methods

Research Context

This analysis stems from a larger design-based research study of instructional improvement in middle-school mathematics. For eight years beginning in 2007, the Middle-School Mathematics and the Institutional Setting of Teaching (MIST) project investigated large-scale support of mathematics teachers' development of ambitious and equitable instruction (Cobb, Jackson, Henrick, & Smith, 2018). Our research team partnered with large urban districts that were committed to instructional improvement in middle-school mathematics, as evidenced by their investment in supports such as professional development and high-quality curricula. To understand the mechanisms of instructional improvement efforts from district offices to schools and classrooms, we collected a variety of qualitative and quantitative data, including observations of classroom instruction, measures of educators' knowledge and beliefs about mathematics instruction, and interviews with participants at all levels of the district.

Two of our partner districts invested in collaborative teacher workgroups as a key strategy for supporting instructional improvement. As a result, a new line of inquiry emerged: to understand the ways that workgroup conversations support teachers' learning opportunities. In the last four years of the study, we purposively selected focal teacher workgroups based on the presence of supports or expertise that could function as catalysts for teacher learning, such as an experienced instructional coach or a promising protocol for workgroup meetings (Horn et al., 2017). Though we intended to create longitudinal cases across multiple years, high rates of teacher turnover and other sources of institutional churn required us to engage in participant selection each year. We selected approximately eight groups in each of four years. During each school year, we recorded between four and six meetings from each group and conducted site visits. Meetings were typically an hour long; approximately 110 hours of recorded meetings form the primary corpus of data.

Data use as a strategy for improvement. Particularly in the last four years of the larger study, data use emerged a key strategy for instructional improvement in both districts. District leaders established goals for increasing student achievement on end-of-year state assessments. They outlined various strategies for achieving these goals, including implementing data-driven teacher workgroup meetings. Principals in both districts were expected to organize teachers' time so that they could meet regularly with colleagues teaching the same grade level and content. In many schools, instructional coaches were tasked with managing assessment data and ensuring that teachers had access to data from benchmark assessments. Teachers, in turn, were expected to analyze student data in workgroup meetings. Occasionally, an instructional coach or administrator would facilitate these meetings, but unfacilitated workgroups were also common.

In line with district leaders' expectations, data use was a common activity in workgroup meetings. Approximately 25% of sampled meetings were devoted exclusively to data use, but all participating teachers reported that they analyzed data in workgroup meetings throughout the school year. When asked about the sources of data that they analyzed, most workgroup participants referred to data from district benchmark assessments, locally developed common assessments, and commercially available multiple-choice assessments. Many teachers also reported that they analyzed student work in workgroup meetings, but they did not always name this as a data use activity. This suggests that many participants in the larger study used a narrow definition of the term *data*, to refer to quantitative assessment data rather than qualitative data like student work.

Despite consistent expectations for teachers to use data in workgroup settings, district leaders did not specify *how* teachers should use data to inform instruction. As researchers have found in other studies, educators in our participating workgroups used data in various ways. Some participants used data to identify ways to maximize student achievement scores without necessarily changing instruction; this is what Horn and colleagues refer to as an *instructional management logic* (Horn et al., 2015). Others used an *instructional improvement logic*, as they used data to reflect on instruction and make changes to support student learning (Horn et al., 2015). Though these logics are not mutually exclusive, they were often in tension, as educators negotiated expectations to increase student achievement in the short-term while improving instructional practice in the long-term.

Focal workgroups. For the present analysis, I selected two workgroups that generally used different logics in their data analysis. I selected groups from the same district for a comparative case study (Yin, 2013): the 7th-grade math teachers at Cypress[1] Middle School from the 2014-15 school year, and the 6th-grade math teachers at Magnolia Middle School from 2012-13 (Table IV.1). To identify these cases, I looked across our sample of data use meetings to identify workgroups that shared similar characteristics, despite their different approaches to data use. I sought out groups that made a concerted effort to use data to inform instruction, rather than groups for whom data use was a compliance activity. I also looked for groups with comparable leadership and expertise, to avoid setting up a comparison that would necessarily put one approach in a more flattering light than the other.

The Cypress and Magnolia workgroups represent cases in which teachers, with the support of expert instructional coaches, intentionally and regularly analyzed assessment data in order to inform future instruction. They did so within the context of test-based accountability pressures; both schools had been identified by the state as needing to improve. The teachers in each group had been teaching for similar numbers of years, and their instructional quality was comparable (Boston, 2012). The facilitators in each group also described similarly sophisticated visions of high-quality mathematics instruction (Munter, 2014) and had experience teaching and coaching in middle-school mathematics. Yet despite these similarities, the facilitators at Cypress and Magnolia approached data use from very different epistemic stances. As a result, they organized very different data analysis cycles, which shaped their workgroups' data use practices, designs for future instruction, and professional learning opportunities.

[1] All school and participant names are pseudonyms.

Table IV.1. Participant summary. All names are pseudonyms.

School	Participant	Role	Experience
Cypress	Diane Butler	Coach	4 years coaching 7 years teaching
	June Farrow	Teacher (7th)	13 years teaching
	Greta Malone	Teacher (7th)	11 years teaching
	Marissa Winters	Teacher (7th)	5 years teaching
Magnolia	Lindsay Millard	Coach	4 years coaching 7 years teaching
	Gerard Donovan	Assistant Principal	2 years administrator 5 years teaching
	Deanna Callahan	Teacher (6th)	9 years teaching
	Tasha Engle	Teacher (6th)	8 years teaching
	Shonda Banks	Teacher (6th)	6 years teaching

Cypress Middle School. Cypress is a middle school in a large urban district in the southern U.S. Like many schools participating in our study, Cypress struggled to meet the accountability goals established under NCLB; they never made Adequate Yearly Progress (AYP) and were targeted for restructuring. But in the year prior to data collection (2013-14), the state's accountability policies changed and Cypress met their accountability goal (which shifted from AYP to AMO, or Annual Measurable Objective) for the first time. Even with this success, accountability pressures were still high in the year of data collection (2014-15): The state Department of Education labeled Cypress as a focus school and a school in need of improvement. This blue wave of pressure trickled down through the school's leadership: The principal and assistant principal described increasing student achievement as their primary goal for the year.

The instructional coach, Diane Butler, echoed this sentiment in describing her overall goal for the 2014-15 school year: “to make sure Cypress meets AMO.” To do this, she organized and copied assessment data to share students’ progress with teachers and administrators. She also designed a data analysis protocol for workgroups to use after district benchmark assessments. With the 7th-grade math team, Coach Diane organized Targeted Re-teaching Days (TRDs) as a way to re-teach content after benchmark assessments. The administrators and teachers described TRDs as an effective strategy for using data to promote student achievement, and they discussed plans for implementing TRDs in other grade levels and content areas.

Coach Diane had been a middle-school math teacher for seven years (including three years at Cypress) before becoming an instructional coach; this was her fourth year as a coach. Based on data collected as part of the larger study, Coach Diane demonstrated fairly high mathematical and pedagogical expertise. In interviews, she described sophisticated visions of high-quality mathematics instruction (Munter, 2014). She articulated goals for students to work on rich tasks with multiple solution paths, and for them to engage in rich mathematical discussions that involved argumentation and justification. She described the importance of teachers’ questioning strategies as they facilitated groupwork, rather than “standing at the front of the room teaching the entire time.” Among coaches participating in the larger study, Coach Diane demonstrated greater-than-typical experience and expertise for supporting mathematics teacher learning.

Coach Diane often facilitated meetings for the three 7th-grade math teachers: June, Greta, and Marissa. The workgroup had a long history of working together: June was in her 13th year teaching, Greta was in her 11th year, and Marissa was in her 5th year. All three had been teaching at Cypress for nearly their entire careers and were fairly accomplished teachers. In interviews, they each described somewhat sophisticated visions of high-quality mathematics instruction. Like Coach Diane, they noted the importance of facilitating group discussions and pressing students to justify their thinking. And although they described mathematical tasks that emphasized procedural skills, they did use rich mathematical tasks in observations outside of the TRDs. Overall, the Cypress teachers were relatively accomplished teachers, with above-average supports for instruction.

Magnolia Middle School. Magnolia is a middle school in the same large urban district as Cypress. Like Cypress, Magnolia never made AYP under NCLB and was under intense pressure to increase test scores. As part of NCLB-related sanctions, Magnolia was labeled as persistently low-achieving and was targeted for restructuring. Just prior to the year of data collection (2012-13), the district school board ordered the Magnolia principal to re-staff at least half the teachers in the school in an effort to improve student achievement. The principal and assistant principal were acutely aware of the accountability pressure, and even feared for their job security. But they also emphasized the importance of supporting teacher learning in order to raise test scores. The principal, for instance, outlined an expectation for workgroups to "deepen understanding of what their content is and what is being assessed" and to analyze student work together. Assistant Principal Gerard Donovan said that his primary goal was to help teachers — especially new teachers — feel supported in their work.

In response to Magnolia's consistently low student achievement, the state Department of Education hired Lindsay Millard as an instructional coach at Magnolia. Like Mr. Donovan and the principal, Coach Lindsay emphasized the importance of supporting teacher learning. She described her role as "helping Magnolia develop and maintain systems that improve student achievement...systems that improve instruction and therefore improve student learning." With

Mr. Donovan, Coach Lindsay facilitated math workgroup meetings organized around a three-week cycle of analyzing content standards and student work. Coach Lindsay also developed professional development sessions for Magnolia teachers and worked one-on-one with them to work on various instructional improvement goals.

Coach Lindsay was in her first year at Magnolia, but her fourth year as an instructional coach in the district. She also had seven years of experience as a middle-school math teacher in the district. In interviews, Coach Lindsay described very sophisticated visions of high-quality mathematics instruction. She articulated goals for students to work collaboratively on rich mathematical tasks, so that students "are actually doing the mathematics...there's inquiry, they're asking questions of each other, they're trying things out, they're drawing pictures." She described teachers as facilitators who support students' engagement in tasks by asking questions, rather than giving answers. Mr. Donovan also had a mathematics background; he taught middle-school math for five years before becoming an administrator. And he described a fairly sophisticated vision of high-quality mathematics instruction, though not as detailed or rich as Lindsay's vision. The Magnolia workgroup facilitators, especially Coach Lindsay, had unusually deep experience and expertise for supporting mathematics teacher learning.

Coach Lindsay and Mr. Donovan worked with Magnolia's 6th-grade math teachers: Deanna, Tasha (who specialized in teaching special education inclusion classes), and Shonda. Another school-based instructional coach, Tabitha, occasionally joined the workgroup meetings. Even though many teachers at Magnolia were new that year, the 6th-grade teachers were all veterans: Deanna was in her 9th year teaching at Magnolia, Tasha was in her 8th, and Shonda was in her 6th. In interviews, Shonda and Deanna described fairly sophisticated visions of high-quality mathematics instruction, noting the importance of students' discussions with each other and their justification of their work.[2] In observations, they also selected rich tasks for students to work on and supported small-group and whole-class discussions. Overall, the Magnolia teachers were fairly accomplished, with deep expertise and experience.

Summary. Among the workgroups in our study, Cypress and Magnolia stood out as settings that had the potential to support teacher learning and instructional improvement through data use, in spite of accountability pressures. Both workgroups incorporated data analysis as part of their regular practice through coach-designed data use cycles. They made concerted efforts to use evidence of student learning to inform instruction. And both groups were facilitated by expert instructional coaches who outlined ambitious goals for mathematics instruction. Yet, the workgroups' data use cycles — and the resulting data use practices, teacher learning opportunities, and instructional designs — were very different. I conceptualize these workgroups as archetypal examples of two approaches to data use. Their many shared characteristics (particularly competent and knowledgeable leadership) bring into sharp relief the importance of epistemic stances in using data for instructional improvement.

Data collection

The primary data for this analysis are four videotaped meetings from Cypress (2014-15) and four videotaped meetings from Magnolia (2012-13), including transcripts and artifacts. To situate the workgroups' conversations in their school and district contexts, I draw on interviews with workgroup participants, school administrators, and district leaders in which they described their goals and expectations for mathematics instruction and data use. Additional interviews and emails with Coach Diane and Coach Lindsay provide further detail on their approaches toward

[2] Tasha was not a full participant in the larger study and thus did not participate in interviews or classroom observations, but there is no evidence within the meetings that her expertise differed significantly from Deanna's and Shonda's.

data use. I also draw upon site visits to two rounds of the Cypress Targeted Re-teaching Days (including field notes and artifacts) as evidence of how that group's data use practices influenced their instruction. I also draw on supplemental data collected as part of the larger study, including videotaped classroom observations and participants' visions of high-quality mathematics instruction (Munter, 2014).

Analytic approach: Understanding epistemic perspectives

To investigate teachers' learning opportunities in workgroup settings, I build on the work of Horn and colleagues, who use Episodes of Pedagogical Reasoning (EPRS; Horn, 2007) as the primary unit of analysis. EPRs are topically-bounded segments of talk in which participants discuss or raise a question of teaching practice. For analyses of teachers' learning opportunities in workgroup meetings that center data use, I adapt this unit of analysis by parsing meetings into Episodes of Data Reasoning (EDRs; Garner, Thorne, & Horn, 2017), which are also topically-bounded units of talk. But in EDRs, participants discuss or raise a question about data, which does not necessarily implicate any considerations of pedagogy. EDRs are often bounded by discussions about particular questions or standards. For instance, an EDR might begin when one teacher asks, "How did they do on #1?" and end when the group moves on to discuss another assessment item. In many instances across our large corpus of workgroup conversations, EDRs included participants' reasoning about data without any discussion of teaching or learning (e.g., "#1 was really bad. It was the worst question. What about #2?"). This finding necessitated the new unit of analysis (Garner et al., 2017).

I analyzed video data of EDRs in the tradition of sociocultural historical studies of joint interaction in workplace settings (e.g., Goldstein and Hall, 2007; Hall and Horn, 2012; Horn et al., 2015). Drawing on the methodologies of sociolinguistics (Hymes, 1974) and interaction analysis (Jordan & Henderson, 1995), I used video data to examine educators' talk, gesture, and tool use as they collaboratively analyzed student learning data and plan for future instruction. I paid attention to what participants considered relevant for sensemaking, treating learning as a members' phenomenon that is situated in social contexts (Lave & Wenger, 1991; Sacks, 1967/1992; Stevens, 2010).

I analyzed EDRs using a constant comparative method (Glaser & Strauss, 1965), attending to what participants considered relevant and consequential for sensemaking (Sacks 1967/1992; Stevens, 2010). Preliminary analyses of multiple workgroups' data use practices suggested that participants approached data from different perspectives (Garner & Horn, 2016). Even though they used similar terms to describe their processes (e.g., "using assessment data to drive instruction"), they engaged in very different practices. To conceptualize this phenomenon, I applied Horn and colleagues' framework of conversational resources that shape teachers learning opportunities: activity structures, frames, representations of practice, and epistemic stances (Horn et al., 2015). This level of coding indicated that differences in workgroups' data use practices were related to differences in their epistemic stances on data. There seemed to be two prevailing stances on data: Some groups used data as a direct measurement of what students know and are able to do, while other groups used data as an indirect indicator of student learning.

To strengthen my interpretations, I situated participants' conversations within the larger body of data collected as part of the MIST study. In particular, I juxtaposed educators' talk-in-interaction with the ways that they described their collaborative meetings and data use practices in one-on-one interviews. This shed light on what participants found most salient or relevant in their work. I also examined the larger school and district contexts: Whenever possible, I consulted interviews with participants' supervisors and colleagues to understand the local pressures, constraints, and contexts shaping their work. I also took a longitudinal perspective by consulting multiple years of supplemental data, especially interviews, to situate my analysis in the participants', schools', and districts', recent histories.

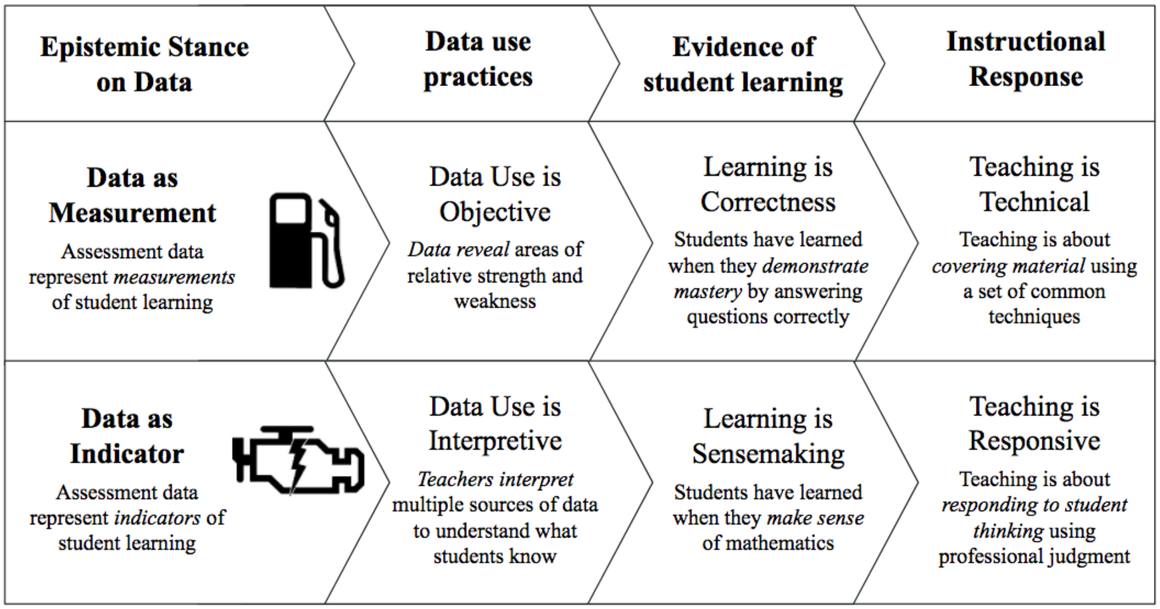
I also considered the collegial relationships among workgroup participants. The focal workgroups were each facilitated by designated leaders — instructional coaches and an assistant principal. Even though I conceptualize the workgroups' conversations as joint accomplishments, I recognize that the workgroup facilitators played an important role in shaping the learning opportunities in workgroup conversations. Instructional leaders determine activity structures (e.g., data use protocols, cycles of analysis), establish goals that frame data use (e.g., identify students for an intervention, plan next week's instruction), and provide teachers with specific representations of practice (e.g., tables of numerical data, copies of the assessment, student work). Their epistemic stances are embedded in these choices and, therefore, in the workgroup's activity. The teachers in these workgroups seemed to share similar epistemic stances to their facilitators, and so I situate my analysis at the group level. But I recognize a possible limitation in that within-group differences in epistemic stance may be hidden. Such an analysis would likely require additional data (e.g., follow-up interviews with teachers) that are not available within this study.

Findings: Epistemic stances shaped cycles of data use

The Cypress and Magnolia workgroups engaged data use practices that shared many structural similarities. In line with the district's expectations, the teachers administered benchmark assessments at regular intervals. The workgroups' facilitators incorporated data analysis into their regular meeting routines, with Coach Diane and Coach Lindsay developing data use cycles based on the district benchmark assessment schedule. The workgroups planned future instruction to address gaps in students' knowledge and support students' success on the end-of-year state tests. The routines and data use practices that emerged in both groups' conversations met the district's vague expectation for data-driven decision-making.

Despite these similarities, the Cypress and Magnolia workgroups approached data use from different epistemic stances (Figure IV.1). The Cypress group, led by Coach Diane, treated data as a *measurement* of student learning. The Magnolia group, led by Coach Lindsay and Mr. Donovan, treated data as an *indicator* of student learning. These fundamentally different perspectives on what data represent reverberated through other aspects of the educators' work, including their data use practices, evidence of student learning, and plans for future instruction. Even though the epistemic stance taken up by Coach Diane and the Cypress teachers is reasonable, given the pressures of test-based accountability policies, it is unlikely to support teacher learning or the development of more ambitious instructional practices. On the other hand, the epistemic stance taken by Coach Lindsay, Mr. Donovan, and the Magnolia teachers has much greater potential to support teacher learning and instructional improvement. But using data as an indicator of student learning is more time-consuming and requires deep content knowledge and pedagogical expertise; this makes it more difficult for teacher workgroups to treat data in this way.

In the following sections, I first describe the Cypress and Magnolia workgroups' epistemic stances on data. Then I illustrate the ways that their epistemic stances were reflected in their data use practices, evidence of student learning, and instructional responses. I center these illustrations on meetings in which the workgroups both discussed proportional reasoning in order to avoid a comparison that is skewed due to differences in content.



Epistemic Stance on Data	Data use practices	Evidence of student learning	Instructional Response
Data as Measurement Assessment data represent *measurements* of student learning	Data Use is Objective *Data reveal* areas of relative strength and weakness	Learning is Correctness Students have learned when they *demonstrate mastery* by answering questions correctly	Teaching is Technical Teaching is about *covering material* using a set of common techniques
Data as Indicator Assessment data represent *indicators* of student learning	Data Use is Interpretive *Teachers interpret* multiple sources of data to understand what students know	Learning is Sensemaking Students have learned when they *make sense* of mathematics	Teaching is Responsive Teaching is about *responding to student thinking* using professional judgment

Figure IV.1. Two epistemic stances on data, which shape educators' data use practices, evidence of student learning, and instructional responses.

Epistemic stances on data

The primary distinction between the Cypress and Magnolia workgroups' epistemic stances on data was their perspective on the ontological status of data — that is, what data are and what data represent. The Cypress workgroup treated data as a measurement of student learning, while the Magnolia workgroup treated data as an indicator of student learning. I relate these approaches to how drivers respond to different warning lights in a car.

Data as measurement. Treating data as measurements of student learning is analogous to how a driver typically treats the fuel tank light in a car. The fuel light is activated in response to a single data point: the level of gasoline in the tank. It reveals a clear and objective problem: there is not enough gas, and the car will soon run out. This problem has a straightforward technical solution: add more gasoline to the tank. Though the driver makes some choices in determining a solution path, as they select a gas station and grade of gasoline, these choices are trivial. Each time the fuel light is activated, the driver's response is essentially the same: put more gas in the car.

Educators, like those at Cypress, who use data as measurements of student learning act as though assessment items reliably capture students' knowledge. They often use quantitative data showing the percent of correct answers: higher scores demonstrate mastery and lower scores signal a lack of knowledge. Scores that fall below a particular threshold (e.g., questions that more than 50% of students missed or students with a score below 66%) are treated as warning lights, signaling an objective problem: students do not have enough knowledge. A common solution is to re-teach the content, but conversations about how to re-teach are often limited to superficial discussions, like whether to re-teach during a warm-up or after school, or whether it should be for all students or just a few. The instructional response is almost always the same: refill students' mental gas tanks.

Within the context of high-stakes accountability policies, this is a reasonable approach to data use, even though it is unlikely to support instructional improvement. Students scores on end-of-year tests are taken as measurements of their mathematical knowledge; students are often categorized into groups such as "proficient"[3] or "novice" based solely on the number of questions they answered correctly. The details of students' mathematical thinking are not as relevant as their ability to find correct answers. Re-teaching content often includes explaining mathematical procedures, teaching test-preparation strategies, or completing additional practice problems. Though these efforts may support short-term success, they are unlikely to support deeper student learning or instructional improvement.

[3] The Cypress Parent-Teacher Association even distributed T-shirts with the slogan "I am Cypress. I am Proficient" for students who passed the state test.

Data as indicator. In contrast, treating data as an indicator of student learning is analogous to how a driver typically treats the check engine light in a car. There is no singular data point that triggers the check engine light, and so the light does not point to any specific problem. Instead, the driver or their mechanic needs to collect multiple sources of information and triangulate them to determine the underlying issue. Interpreting the data requires mechanical expertise and knowledge of cars to determine whether there a failing catalytic converter, a faulty spark plug, or any number of other issues. In some cases, a check engine light is activated because of a relative non-issue, like a loose wire or a missing gas cap. The driver's solution is developed in response to the data analysis; different diagnoses require different solutions.

Educators, like those at Magnolia, who use data as indicators of student learning act as though assessment items give only a partial view of students' understanding. Quantitative data only reveal part of the story, as students could answer a question incorrectly for various reasons. Educators turn to other data — such as students' work, responses to other items, or past performance — to determine how students are thinking about the content. Perhaps students had a conceptual misunderstanding, had difficulty parsing word problems, or made a procedural error. In some cases, the problem might not be one of student understanding: A poorly-worded question or running out time could cause students to answer incorrectly. Interpreting data in this way requires pedagogical content knowledge as well as knowledge of students, but it allows teachers to respond to student thinking and support deeper learning.

This is a more complicated approach to data use, but it is more likely to support instructional improvement. Collecting and making sense of multiple data sources takes time, which is teachers' most precious resource. Educators must be able to consider multiple approaches to an assessment item and make sense of students' unique (and sometimes surprising) responses. Once they make sense of student thinking, educators must use their pedagogical judgment to determine how to respond instructionally. If students exhibit a particular misunderstanding, teachers must design a way to press students' thinking on that issue. This often requires re-thinking instruction and changing teachers' practice to support students' deeper conceptual understanding. In this way, using data as an indicator of student learning has the potential to support student learning and instructional improvement.

Data use practices

The workgroups' epistemic stances on data were reflected in their data use practices, or the ways that they collected and analyzed data in workgroup meetings. Both groups developed data use cycles that became a routine part of their collaborative meetings. The Cypress workgroup used data as an objective measurement of student learning in the design and enactment of Targeted Re-teaching Days. As a result, they relied primarily on quantitative data and made only superficial attempts to discern student thinking. In contrast, the Magnolia workgroup used data as an indicator of student learning that need to be interpreted in the context of multiple sources of information. Their three-week cycles of data use prioritized qualitative data, which allowed the workgroup to engage in deeper analyses of student thinking. In the following sections, I illustrate how the workgroups' underlying epistemic stances on data played out in the their data use practices.

Cypress: Targeted reteaching days. The Cypress workgroup's epistemic stance on data as a measurement of student learning was evident in their design and enactment of Targeted Reteaching Days (TRDs; see Figure IV.2). In the 2014-15 school year, district benchmark assessments contained three questions on each of three Focus Topics; Focus Topics were groups of two or three related standards. The Cypress group agreed on a threshold — two out of three questions on each Focus Topic — to distinguish between students who “got it” and those who “missed it.” For the planning meeting, Coach Diane compiled lists of students who “got” and “missed” each topic, as well as the percent of students answering each question correctly. The workgroup used these data to identify trouble spots — students in need of intervention and the topics that required re-teaching — to address during the TRDs. Of the three topics on the benchmark, they selected the two Focus Topics with the most students in the “missed it” category. Then they organized one TRD for each of those two Focus Topics.

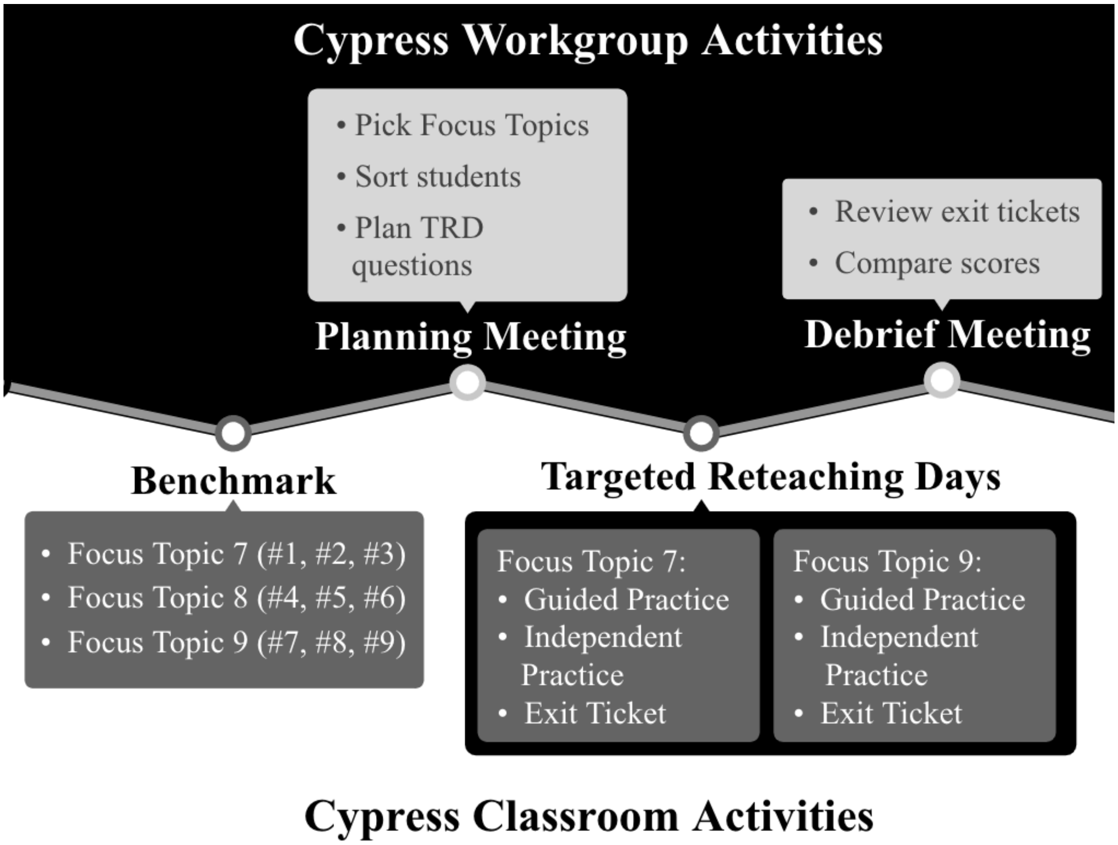


Figure IV.2. Timeline of the Cypress workgroup and classroom activities surrounding Targeted Re-teaching Days.

During TRDs, Coach Diane “mixed up” all of the students who took math during each class period (e.g., all 7th grade students with math in Period 1) and assigned them to a classroom based on whether they “got” or “missed” that day’s topic. Typically, one teacher (e.g., Marissa) taught an “enrichment class” for students who got it while the other two (e.g., June and Greta) taught an “intervention class” for those who missed it. Occasionally, other staff were recruited to help teach an additional group of students to avoid overly full classes. During intervention lessons, teachers retaught the Focus Topic as it was covered in the benchmark assessment. The enrichment class typically addressed an especially difficult benchmark question or another related topic (e.g., calculating percent increase and decrease was the enrichment topic for a TRD on proportional reasoning). At the end of the period, students in both the intervention and enrichment classes completed an exit ticket with three questions; the Cypress teachers gauged the effectiveness of the TRD by the number of students who answered at least two of the three questions correctly.

The TRD cycle reflected the Cypress workgroups’ epistemic stance on data as measurement the ways that they took assessment results as an objective reflection of what student knew and were able to do. In the meeting on February 10, 2015, they analyzed data from a benchmark assessment covering Focus Topics 7, 8, and 9. Their data consisted of the percent of students who answered each item correctly and the number of students who “got” and “missed” each Focus Topic. They also had copies of the assessment. Though additional data, such as the distribution of students’ responses, were available online through the district’s data management system, the workgroup did not seek out that information. Coach Diane recorded notes on the group’s meeting in a data analysis template that she created.

The Cypress workgroup typically treated all assessment items on a Focus Topic as equal in measuring students’ knowledge of the topic. Similarly, they treated all incorrect responses as demonstrating the same lack of knowledge. To identify Focus Topics for TRDs, they used Coach Diane's counts of students who "missed" and "got" teach topic:

Table IV.2. Cypress: Seven and nine are the worst.

30.	Greta:	[Focus Topics] 7 and 9 are the worst, right? Like 8's probably okay?
31.	Coach Diane:	Seven? There were 290 kids that missed that whole topic.
32.	Greta:	That missed all three?
33.	Coach Diane:	Either three or two out of the three.
34.	Greta:	Okay. Okay.
35.	Coach Diane:	290. The Focus Topic 8 is what they performed the best on.
36.	Greta:	Yeah.
37.	Coach Diane:	There was like a hundred kids that missed that.
38.	Greta:	Okay.

This brief exchange is typical of the ways that the Cypress workgroup identified topics and organized students for the TRDs. Greta turned the group's attention to the "worst" topics, which Coach Diane interprets as the topics that the most students "missed" (Turns 30-31). These identifications, which persisted throughout their meetings, were based solely on the number of correct answers. The Cypress group made no distinction among which questions were missed or what answer choices students selected.

This approach to data use reflected an epistemic stance on data as measurement. The Cypress workgroup took (the correctness of) students' responses to the three items on each Focus Topic as an objective measurement of their knowledge about the topic. They treated students who had fewer than two correct answers (that is, who "missed" the topic) as having a warning light that signaled insufficient knowledge of the Focus Topic. This allowed the workgroup to identify the topics with the most warning lights — in this excerpt, Topics 7 and 9 — and repair students' lack of knowledge by reteaching. The response for the "worst" topics on each benchmark was essentially the same every time: schedule a TRD and re-teach the content.

The process of identifying topics and planning an intervention lesson included only brief considerations of student thinking. When she described TRDs in an interview, Coach Diane said that they "look at the test questions that the kids missed and try to figure out what it is that the kids don't understand." But in practice, the group addressed student thinking only in brief and superficial discussions. During meetings, considerations of why students missed questions was typically reserved for a section on Coach Diane's data analysis template that asked for "reason(s) questions were missed." This was presented as a generic checklist for the entire benchmark, rather than something tied to any particular questions. Near the end of the February 10 meeting, Coach Diane asked the teachers for reasons that students missed questions. They offered suggestions from the checklist, including "poor wording" and "students' misconceptions;" Coach Diane marked the appropriate lines on the checklist. But no one in the workgroup offered any evidence to support their claims, nor did they tie their claims to a specific answer. Rather, checking "student' misconceptions" seemed to mean that "students' misconceptions resulted in incorrect answers somewhere on the benchmark," which is so vague as to be nearly meaningless. Without identifying specific misconceptions to address (or even specific questions that revealed misconceptions), the teachers could not tie their analysis to any instructional choices.

Even with a checklist to identify reasons that students missed questions, the Cypress workgroup's overall epistemic stance on data was as a measurement of student learning. Their data use practices treated assessment results as straightforward and objective measures of student learning: The number of correct answers revealed students' level of knowledge, and the number of students who "missed" a topic revealed which topics needed to be re-taught. Though the workgroup listed some reasons that students missed questions, they did not use this superficial analysis in planning TRDs. Rather, they engaged in the same objective data use routine for each

TRD: Use a benchmark assessment measure student learning, examine the results to find areas of relative weakness (students who "missed" Focus Topics and topics that were "the worst"), schedule a Targeted Re-teaching Day, and re-teach the "worst" topics to the students who "missed" them.

Magnolia: Analyzing student work. The Magnolia workgroup's approach to data use contrasts sharply with the Cypress Targeted Re-teaching Days. In line with their epistemic stance on data as an indicator of student learning, Coach Lindsay and Mr. Donovan designed a three-week cycle of data analysis (Figure IV.3). In the 2012-13 school year, the district's benchmark assessments covered material that was recently taught (though items were not strictly organized into Focus Topics); some benchmarks also included review questions from content taught earlier in the year. Yet this was not the primary source of data that the workgroup analyzed. After each benchmark assessment, they identified "bubble kids,"[4] or students just on the cusp of mastery. But then they spent the majority of their data use efforts analyzing student work — particularly work from the bubble kids. The Magnolia data use cycle was three weeks long, which meant that they could typically complete two cycles in between benchmark assessments.

[4] Some schools identify bubble kids for problematic reasons — for instance, to target them for intense test-prep interventions and to ignore the needs of students who are expected to fail the state test (e.g., Booher-Jennings, 2005). Note that the Magnolia workgroup used bubble kids in a very different way.

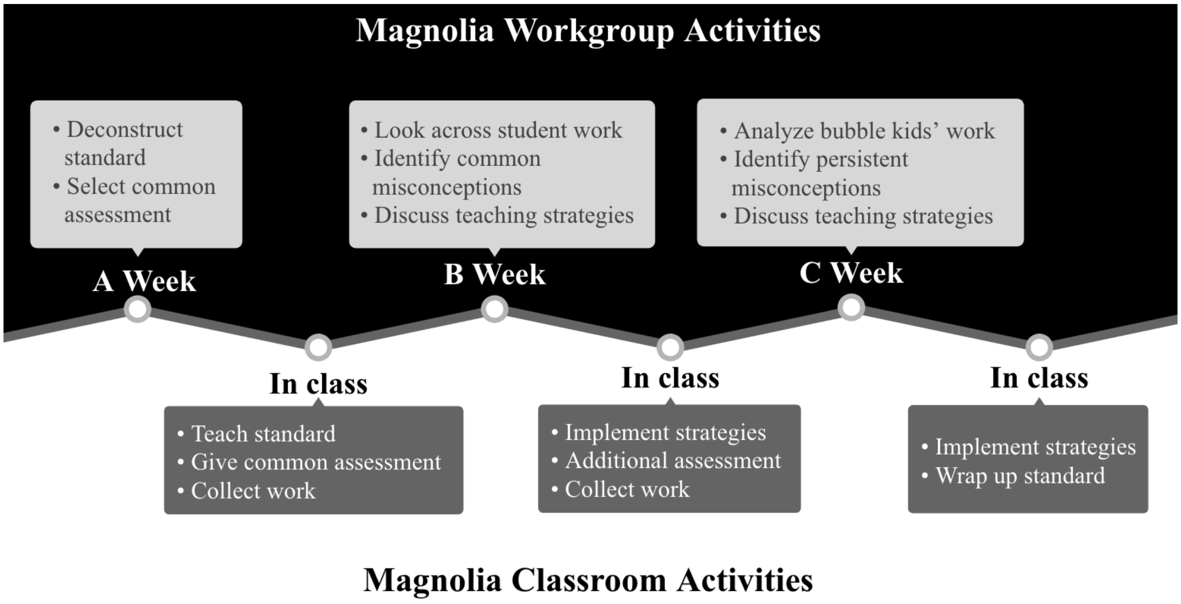

Magnolia Workgroup Activities
• Deconstruct standard
• Select common assessment
A Week
• Look across student work
• Identify common misconceptions
• Discuss teaching strategies
B Week
• Analyze bubble kids' work
• Identify persistent misconceptions
• Discuss teaching strategies
C Week
In class
• Teach standard
• Give common assessment
• Collect work
In class
• Implement strategies
• Additional assessment
• Collect work
In class
• Implement strategies
• Wrap up standard
Magnolia Classroom Activities

Figure IV.3. Timeline of the Magnolia workgroup and classroom activities during a three-week data use cycle.

The first week of the cycle was an A Week. In A Weeks, the Magnolia workgroup “deconstructed” a standard into objectives to determine what students need to know and how they could demonstrate understanding. In an interview, Coach Lindsay framed this as a way to support teachers’ learning. When she arrived at Magnolia, she noticed that teachers frequently deconstructed standards into smaller daily objectives, but that they (and their students) had trouble putting them back together:

> Coach Lindsay: During A Weeks, we’re deconstructing a standard. And so we’re really digging into the standards, which really wasn’t happening before… I saw a lot of deconstructing the standards to these tiny little chunks, which were the objectives, but then never putting it back together again and assessing the whole standard, or doing what I said earlier, where I teach all these little pieces, and then the kids can’t put it back together again when they’re given an assessment. So I tried to redirect things so that we’re focusing on the standard. We do some deconstruction, but then we have this section where we reconstruct and we look at a common formative assessment which assesses the standard — not any chunk of the standard, but the whole standard.

In A Weeks, Coach Lindsay and Mr. Donovan supported teachers in analyzing the standard to determine what the entire standard entailed and how students could show their understanding on a common formative assessment. For instance: What does it mean for a student to understand unit rate? What are the different representations and contexts that a student should be familiar with? How can a student demonstrate conceptual understanding of unit rate, in addition to procedural knowledge?

In B Weeks, the teachers brought in students' work from the formative assessment to identify common solution strategies and misconceptions. Mr. Donovan began the B Week meeting on February 12, 2013 with this framing:

> Mr. Donovan: It is a B Week and so we are gonna talk about the formative assessment that we all agreed upon from the A Week, okay? As you remember, we wanna look at what are the misconceptions. What are the questions we wanna ask, like for giving [students] feedback. What are the objectives? And what're the teaching tasks we're gonna do to address this? Also, Lindsay and I have prepared some sample student work because we didn't know if all of the different possibilities of solving this problem were gonna be shown. So, we just wanted to share some strategies.

During this meeting, the Magnolia teachers shared the various strategies that their students used on the formative assessment about unit rate, as well some common mistakes that they made. They discussed issues like keeping track of units (e.g., five *dollars* for each *hour*) and understanding the multiplicative relationship between the values. Coach Lindsay and Mr. Donovan helped the teachers made connections across their students' strategies and also shared additional representations (like tape diagrams and double number lines) that could support deeper conceptual understanding in the coming week.

In C Weeks, teachers again brought student work to the meeting, but they focused on analyzing work from their "bubble kids." At the beginning of the C Week meeting on February 19, 2013, Mr. Donovan described their logic for using bubble kids:

> Mr. Donovan: We have our bubble kids, and we're looking at the student work from them, and then we're gonna see whether or not they mastered [the standard] or not. And the important part with this is our bubble kids are kids closest to proficiency… If we're looking at their misconceptions, there's a good chance that those misconceptions

> permeate throughout the room… Because these are the kids who are the closest kids for getting it naturally… So, it's not that we're forgetting about the other kids, it's just, if you don't have high numbers of kids mastering the standards, with these kids, then, more than likely, your whole room's suffering, and we need to reteach and do some interventions.

The Magnolia group used bubble kids like canaries in a coal mine: If the bubble kids had a misunderstanding, it was likely that other students did, as well. In C Week meetings, the group analyzed work and then brainstormed ways to respond to student thinking and support students' conceptual understanding. In the following week, they began a new cycle with a new standard. After each benchmark assessment, the teachers identified new sets of "bubble kids."

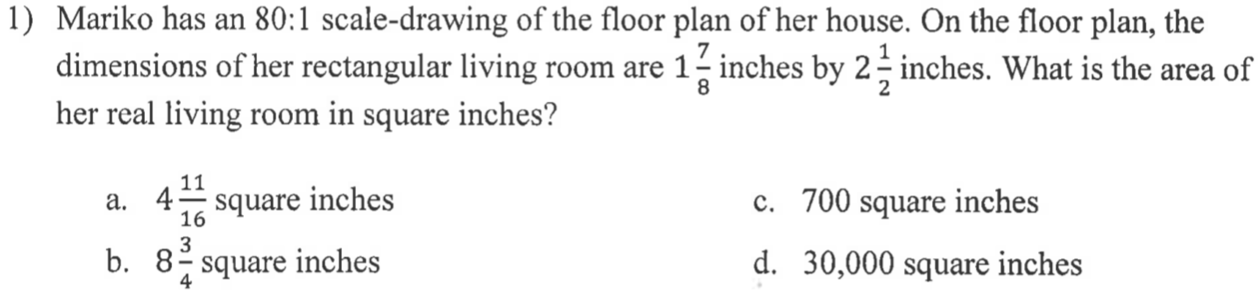
The data use cycle developed by Coach Lindsay and Mr. Donovan demonstrated an epistemic stance on data as an indicator of student learning. Rather than use students' responses as a binary metric of whether they "got" or "missed" the content, the Magnolia workgroup analyzed students' work to uncover different strategies and sensemaking processes. Though an incorrect answer often signaled a problem or a warning light (as incorrect answers did for the Cypress group), the Magnolia group looked deeper to figure out what the incorrect answer might mean. They used multiple sources of information — including different parts of each formative assessment and assessments over multiple weeks — to look "under the hood" and figure out how students were thinking mathematically. As they deconstructed standards and analyzed student work, they tied their conversations to future instruction, with the goal of gradually deepening students' conceptual understanding over the course of the three weeks.

Evidence of student learning

The workgroups' different epistemic stances on data also shaped what they took as evidence of student learning. The Cypress workgroup, using a data-as-measurement approach, equated correct answers with student learning: If students answered enough questions correctly, they knew (or “got”) the Focus Topic, regardless of the specific content of the questions. In line with their stance on data as an indicator, the Magnolia workgroup’ framed learning as a sensemaking process. As a result, they tried to understand students’ reasoning, opening up the possibility for a student to have an incorrect answer and still know the content. In the following sections, I illustrate this phenomenon with excerpts from the workgroups’ meetings.

Cypress: Learning as correctness. As they planned and implemented TRDs, the Cypress workgroup typically equated correct answers with student learning. This aligns with their epistemic stance on data as measurement: They considered a student to have learned a Focus Topic if (and only if) they answered at least two out of the three benchmark items correctly. In general, they did not differentiate among items within a Focus Topic, even if they assessed students understanding in different ways, or if they addressed different parts of the topic. Focus Topic 7, for instance, required that students compute unit rates with fractional values in a variety of contexts, including with ratios, lengths, and areas. The three benchmark items (including Number 1, Figure IV.4) covered area (within the context of a scale drawing), ratios (in the context of weight conversion), and comparing speeds. Each item could reveal something different about students’ understandings of unit rate, but the Cypress workgroup treated the questions as assessing the same thing.

They also treated each incorrect answer as equivalent. The Cypress workgroup used data that showed how many students answered correctly or incorrectly, but they did not use data on which answers students selected. Treating responses as a binary — correct or incorrect — prevented the workgroup from developing more nuanced understandings of student thinking. For instance, consider the following benchmark item (Figure IV.4), which the workgroup discussed at the beginning of their meeting on February 10, 2015. This was the most frequently-missed item on the assessment: Only 10% of students answered it correctly.



1) Mariko has an 80:1 scale-drawing of the floor plan of her house. On the floor plan, the dimensions of her rectangular living room are $1\frac{7}{8}$ inches by $2\frac{1}{2}$ inches. What is the area of her real living room in square inches?

a. $4\frac{11}{16}$ square inches

b. $8\frac{3}{4}$ square inches

c. 700 square inches

d. 30,000 square inches

Figure IV.4. Number 1 on a benchmark assessment used at Cypress. The correct answer is D (30,000 square inches).

According to the test blueprint, this item was marked as assessing unit rates, which was included in Focus Topic 7. However, the question requires additional knowledge beyond proportional reasoning, which makes interpreting incorrect responses more complicated. In order to find the correct answer, students must be able to: calculate area, multiply mixed numbers, and interpret the context of the question, in addition to understanding the unit rate expressed as a scale factor.[5] A student who understands scale factors (and can thus find the dimensions of the living room), but calculates the perimeter of the room instead of the area would select answer choice C (700 square inches). This error could be related to the students' geometric understanding. A student who knows how to calculate area (and thus multiplies 1 $^{7}/_{8}$ by 2 $^{1}/_{2}$) but does not use the scale factor to find the dimensions of the room would select answer A (4 $^{11}/_{16}$ square inches). This error could be related to the students' understanding of proportionality or their interpretation of the phrase "80:1 scale-drawing." A student who does not know how to multiply mixed numbers or who does not understand the context of the item might guess an answer, which would further

[5] Of course, students could recognize that only one answer choice is a reasonable size for a living room. But square inch is an uncommon unit for measuring the area of a room, which makes such estimation more complicated.

complicate an interpretation of their choice. Number 1 is a complicated item, requiring multiple steps and connecting multiple mathematical ideas. And so there are many reasons that students would answer this question incorrectly; treating all incorrect answers as stemming from the same misconception is an oversimplification.

Typically, the Cypress workgroup accepted answers as unproblematically measuring students' knowledge of the associated Focus Topic, with little discussion or fanfare. But during the planning meeting on February 12, 2015, they almost came to question the appropriateness of Number 1 in assessing students' knowledge of unit rate. At the beginning of the meeting, the group turned to Focus Topic 7 in general and Number 1 in particular. June said that she "needed to do more with unit rate" and Marissa acknowledged "Number 1 gets [students] all the time, though. It just throws them off." They started to move on to discuss the next question, but Coach Diane redirected the conversation back to Number 1:

Table IV.3. Cypress: A complicated question.

18.	Coach Diane:	So that particular question, Number 1,
19.	June:	Hmm hmm.
20.	Coach Diane:	was over included area.
21.	June:	So, it's scale drawing that they had to find the area of the real house by finding the two dimensions of the real house. So, they had to find scale to find it.
22.	Greta:	It was very complicated.
23.	Coach Diane:	I thought so, too.
24.	June:	I thought so, too.
25.	Greta:	I mean, it's not just completely testing their ability to do the scaled drawing, I mean, yeah, like you're saying, they have to know area.
26.	June:	So, I mean, it's just saying, let's look at that focus topic. It's saying Focus Topic 7. ((*Reading*)) *"Compute unit rates with fractional values, compute unit rates with ratios, lengths, and areas."*
27.	Marissa:	That, I mean, it hits it,
28.	June:	Hmm hmm. ((*Continued reading*)) *"In like or different units of*

measure.”

In this exchange, Coach Diane kept the teachers’ attention on Number 1, nothing that it “included area,” which was perhaps surprising since Focus Topic 7 covered unit rates (Turn 20). June elaborated on this, noting that students needed to use a scale factor to find the area (Turn 21). Greta pointed out that the question was “very complicated” (Turn 22) because it assessed students’ understanding of area as well as proportional reasoning (Turn 25). This potentially brought into question the appropriateness of Number 1 for assessing Focus Topic 7. But then June looked up the description of Focus Topic 7 and found that the topic included the use of fractional values to compute unit rates with areas (Turn 26). Based on this information, Marissa and June confirmed that the question fit within the Focus Topic (Turns 27-28). This settled the discussion: Since area was mentioned in the description of Focus Topic 7, the workgroup took Number 1 as a measurement of students’ knowledge of unit rate — and they treated it no differently than Number 2 or Number 3. After Turn 28, the group quickly turned to identifying Focus Topics to address in the TRDs (“Seven and nine are the worst,” above).

The group identified Focus Topic 7 as one of the topics to address in a TRD, though they did not discuss any of the items in further detail; they did not discuss Numbers 2 or 3 individually in the meeting at all. In her data analysis protocol, Coach Diane summarized their analysis of Number 1 as “Targeted Reteaching Day needs to address topic. Students did not understand the magnitude of what was being asked of them since they did not remember how to do area.” For Numbers 2 and 3 (which also addressed Focus Topic 7 and 25% of students answered correctly), Coach Diane recorded simply that the “Targeted Reteaching Day needs to address topic.” Recording the same phrase for each item in a Focus Topic was typical for the Cypress workgroup. For Focus Topic 8, for instance, Coach Diane recorded “Surprised that they did so well since this has not been covered as much" for all three items, even though only 40% of students answered Number 4 correctly (Numbers 5 and 6 had 86% and 59% correct, respectively). This provided further evidence that the group treated the items within a Focus Topic as equivalent in assessing students' knowledge of that topic.

Notably, this episode was the most in-depth discussion of any item on the benchmark assessment during our sampled meetings. At first, the group considered the possibility that Number 1 assessed something beyond Focus Topic 7. If so, it could have troubled their data-as-measurement approach. But once June read the description of Focus Topic 7, they decided that calculating area was embedded within the topic and did not probe any further. Coach Diane noted that area might have confused students on Number 1, albeit without consulting any other data (like the distribution of student's responses) to justify the claim.

In line with their data-as-measurement epistemic stance, the Cypress workgroup equated correct answers as evidence of student learning. With the exception of the brief consideration of Number 1, the Cypress group accepted the items at face value without further discussion. When identifying students and topics for re-teaching, they did not distinguish among the questions within any of the topics or the incorrect answers within a question. They also did not consider the possibility of any other confounding factors (e.g., reading comprehension or arithmetic with fractions). Instead, they equated correct answers with student learning (and incorrect answers as evidence of a lack of learning) on the relevant Focus Topic. Given that the state end-of-year test will be the most consequential metric of students' learning, this may be a reasonable approach (insofar as benchmark assessments can be used as a proxy for the state test, which is questionable). But equating correctness with learning in this way limited the depth of analysis that the Cypress workgroup engaged in and the instructional responses that they designed.

Without considering why students missed Number 1, the teachers could not plan for instruction in a way that would address student thinking.

Magnolia: Making sense of the content. Reflecting their epistemic stance of data as an indicator, the Magnolia workgroup frequently analyzed student work during their three-week data use cycle. They analyzed work with the goal of understanding student thinking. They tried to figure out how students approached questions in order to determine what students had mastered — what students understood mathematically, and what misconceptions they had. Learning, then, was about making sense of the mathematics, rather than arriving at a correct answer. This allowed the possibility of false negatives, where students understood a key concept, but still answered incorrectly (e.g., due to an arithmetic mistake). In our sample of meetings, the workgroup did not look for false positives (i.e., students who answered correctly but lacked conceptual understanding), but Mr. Donovan's description of bubble kids' misconceptions as "permeating the entire room" suggests that this may have been a possibility for the group.

As they analyzed at student work, the Magnolia teachers occasionally came across puzzling responses. In their C Week meeting on February 19, 2013 (see Figure IV.3, which shows the cycle), Mr. Donovan asked Deanna if one of her bubble kids, Tommy, had mastered unit rate. Tommy's response to one part of the task was confusing. He correctly found a unit rate of $5 per hour and completed a table relating time and money in the same context. But in a third part of the task, Tommy wrote "$3.05" (not $35) as the amount of money to be earned after 7 hours.

Table IV.4. Magnolia: What did the standard require?

157.	Mr. Donovan:	What about Tommy?
158.	Deanna:	I think, okay, Tommy got the table, like filling in the table, finding the unit rate, doing that, but then, when he had to apply it and figure out how much money for seven hours, it was supposed

to be $35, and he put $3.50 —

159. Mr. Donovan: Okay. So now, let's go back and look at the standard.

160. Deanna: — Or $3.05.

161. Mr. Donovan: Okay. So, when we look at the standard, what did the standard require us to do?

162. Deanna: He cannot apply, well, he found the unit rate, but I just think he's having issues with the, well, if he did seven times five, he should've got 35. So, I don't know why he's getting three point, where he's getting the decimal from.

163 Coach Lindsay: So, what was, what was it supposed to be? 35 cents?

164. Deanna: He may've — Seven times five, 35 dollars.

165. Coach Lindsay: 35 dollars.

166. Shonda: But sometimes they get confused on the calculator. You know, and he may've inadvertently put the zero, because you know sometimes you can set the mode or whatever?

167. Deanna: Yeah?

168. Shonda: Maybe he inadvertently did that.

169. Deanna: I mean, I think for the most part, he has the concept, but —

170. Coach Lindsay: Well, but it sounds like he has the procedure for how to find unit rate

171. Deanna: The procedure, yeah

172. Coach Lindsay: I'm not sure —

173. Deanna: He knows what that

174. Coach Lindsay: Did it — what did the question ask?

175. Deanna: He can’t — So, I, I'm gonna say no. It says how much money will he earn in seven hours and he put $3. And if he looks at the table, he should be able to figure out that's not right.

176. Mr. Donovan: And see, he, and if he knows what the unit rate is, it should make sense for him.

177. Deanna: Right — you're right. If he knew the unit rate's $5,

178. Coach Lindsay: So, I would say that's probably more of an issue of interpreting within the context —

179. Deanna: Right, and we've —

180. Coach Lindsay: Like what makes sense.

181. Deanna: Been trying to do that.

182. Coach Lindsay: Yeah.
183. Deanna: All right. So, he's a no.

Even though Tommy found an incorrect answer, Deanna did not immediately take that as evidence that he did not understand unit rate. She recognized that he correctly calculated the unit rate elsewhere, but did not “apply it” in context (Turn 158). Mr. Donovan referred to the previous A Week, asking Deanna to think back to what the standard required (Turns 159, 161). The workgroup considered the possibility of a misplaced decimal point (Turns 162-163) or a calculator error (Turns 166-168), which could be sources of confusion unrelated to Tommy’s understanding of unit rate. Deanna and Coach Lindsay briefly discuss whether the parts of the task that Tommy did correctly indicate procedural or conceptual understanding (Turns 169-171). Ultimately, Deanna decided that in order to demonstrate conceptual understanding of unit rate, students needed to be able to judge the reasonableness of an answer (Turn 175), and that Tommy did not meet that bar (Turn 183). Coach Lindsay called this “an issue of interpreting with the context,” and Deanna agreed, noting that she had been “trying to do that” with her students (Turns 178-181). Once Deanna determines that Tommy is “a no,” the group moves on to discuss other bubble kids.

Over the course of the meeting, Deanna and Shonda reported which of their bubble kids had and had not mastered unit rate. Though most bubble kids received a quick “yes,” the group analyzed many of the “no” students more closely. But these were not homogenous categories, like students who “got it” or “missed it” at Cypress. The Magnolia workgroup looked closely at the work of the “no” students and of those whose work was initially unclear, like Tommy. They analyzed the work to understand how students were making sense of the content, looking both at individuals and across the group. In this meeting, they determined that most of the “no” bubble kids demonstrated procedural knowledge on at least one part of the task, but they lacked a conceptual understanding of unit rate as a multiplicative relationship between two numbers. This misunderstanding prevented students like Tommy from "interpreting [values] within the context" of the task.

The Magnolia workgroup's close analysis of student work, particularly in B Week and C Week meetings, aligned with their epistemic stance on data as an indicator of student learning. Tommy's wrong answer was not immediately taken as evidence that he did not understand unit rate. Instead, Deanna and the others considered multiple sources of data to make sense of his answer. They even seemed to leave open the possibility that a student could answer incorrectly and still demonstrate understanding of unit rate — perhaps if Tommy made a calculator error or an arithmetic mistake that gave a more reasonable answer. Their interpretations of student work also allowed them to develop more nuanced understandings of student thinking. Even though Deanna concluded that Tommy was "a no," she did not conclude that Tommy doesn't understand unit rate. Rather, she and the rest of the workgroup determined what, specifically, Tommy and other students understood (e.g., the procedure for finding unit rate), and what they did not (e.g., what makes an answer reasonable and the multiplicative nature of unit rate). By treating data as an indicator and examining students' sensemaking for evidence of learning, the Magnolia workgroup was positioned to design instructional responses that specifically attended to student thinking and supported deeper learning.

Instructional Response

The workgroups' epistemic stances on data reverberated through their data use practices and evidence of student learning to shape their plans for future instruction. After identifying students and Focus Topics for re-teaching, the Cypress workgroup approached teaching as a technical issue. They planned TRDs in which they reviewed procedures that they expected students to use to answer questions like those found on the benchmark. The Magnolia workgroup, on the other hand, approached teaching as a matter of responding to student thinking. They used their interpretations of student thinking to plan instructional strategies that aimed to press on students' misconceptions and support deeper conceptual understanding. In many ways, the instructional responses that each group designed were coherent with their epistemic stances on data, their data use practices, and what they took as evidence of student learning; the Magnolia workgroup's approach had greater potential to support instructional improvement and student learning.

Cypress: Re-teaching as a technical problem. In line with their view on data as a measurement of student learning and mastery as getting correct answers, the Cypress teachers treated teaching as a technical issue. Teaching in a way that would support student success was a matter of covering material again or executing generic strategies, rather than using pedagogical judgment to respond to student thinking. In conversation, they typically described their work as a set of things *do*, thereby eliding the complexities of the *doing*. When Coach Diane asked questions like "What will we do for the students who got it," teachers gave suggestions like "Just have enrichment." Occasionally, someone would share a strategy or resource — June, for instance, suggested a website with activities that could be used for enrichment — but these suggestions included little detail beyond using the website to "do scale drawing."

During the February 10, 2015 meeting, the Cypress workgroup spent most of their time attempting to schedule TRDs around field trips and other school events. As a result, they did not have time to plan the TRD lessons or create the necessary materials. Marissa and June met later in the week (February 12, 2015) to create worksheets and exit tickets to use in the enrichment and intervention classes for Focus Topic 7 and Focus Topic 9. At the start of this meeting, they decided on a way to divide the labor:

Table IV.5. Cypress: You're doing 7.

4.	Marissa:	Alright, so you're doing 7 and I'm doing 9.
5.	June:	Well can I do 9?
6.	Marissa:	Oh yeah, sure.
7.	June:	Because I brought a bunch of stuff. But I also have some — 7 is unit rate, right? And enrichment was scale? Is that was we said?
8.	Marissa:	I think so.

In this meeting, "doing 7" and "doing 9" (Turns 4-5) meant writing problems for the intervention and enrichment classes to use during the TRD. The questions that the teachers wrote were typically isomorphic to the items on the benchmark assessment, involving similar contexts with different names and numbers. They reserved the most difficult questions (e.g., Number 1 for Focus Topic 7) for the enrichment class — what June referred to as "scale" (Turn 7). After this discussion, they turned to their individual computers to create worksheets. They spent most of the meeting working independently, though they occasionally consulted each other about formatting choices (e.g., where to include a space for students' names) or to discuss weekend plans.

Marissa and June wrote three sets of questions for the intervention and enrichment classes for their respective Focus Topics. During the TRDs, teachers reviewed one set of questions with students as a guided practice. They instructed students on how to execute procedures that would help them answer each question; they reminded students that they taught the same procedures in previous lessons. After the guided practice, students worked on another set of isomorphic questions on their own or in pairs. At the end of class, students received an exit ticket (with a third set of isomorphic questions). The Cypress teachers judged the effectiveness of the TRD by students' responses on the exit ticket. They used the same threshold as on the benchmark: Students who answered at least two of the three exit ticket questions correct "got" the Focus Topic. After being shown how to solve two similar sets of problems, students overwhelmingly "got it" on the exit ticket.

Throughout their planning and enactment of the TRDs, the Cypress workgroup treated teaching as a technical issue. They collapsed complicated practices, like planning and executing an entire lesson, into shorthand: "do 7" or "do enrichment." This left little room for a discussion of the different strategies that they planned to use, much less why they might use them. During the TRDs, the teachers sometimes taught slightly different procedures: Greta told students to "cross multiply" to solve a proportion, while Marissa called the same procedure the "fish method" and June called it "butterfly multiply." Although they were aware of the differences in each other's terms, they did not discuss the various approaches that they planned to use during the sampled meetings or during site visits. Within their workgroup, teaching was described as a set of things to "do" or techniques to execute, and so the reasoning behind their choices remained hidden.

Magnolia: Responding to student thinking. In line with their epistemic stance on data as indicator and their use of data to interpret students' sensemaking, the Magnolia workgroup planned instruction that was responsive to student thinking. After analyzing student work and identifying student misconceptions, the Magnolia workgroup designed ways to specifically address those misconceptions. Over the course of their three-week cycle, they sought to gradually deepen students' understanding of key topics. Particularly in B Weeks and C Weeks, when they examined student work, Coach Lindsay pressed teachers to tie their understandings of student thinking to instructional strategies.

Recall that in the February 19, 2015 meeting, the Magnolia workgroup analyzed work from Tommy and other bubble kids. They determined that students did not have a strong multiplicative understanding of unit rate. Though most students could follow procedures to get correct answers on most parts of the task, they did not have the conceptual understanding that unit rates describe a multiplicative relationship between two values (e.g., $5 for every hour). After discussing this, Coach Lindsay turned the group's attention to designing an instructional response:

Table IV.6. Magnolia: Getting them back on track.

269.	Coach Lindsay:	Okay. So, what, then, the question is what're we gonna ask them to get them back on track?
270.	Shonda:	I had, so, when we went over it, I said, "What you should've done was made your own table. When you showed your work, you should've made your own table and did your one, two, three, four, five, six, seven," and then, once they did that, when we went over this, they understood that, "Okay, it's going up by," I forget however many it was. Was it 10 or five?
271.	Deanna:	They're timsing, yeah.
272.	Coach Lindsay:	The, the
273.	Shonda:	And they're, their interpretation, "Well, it went up by however so many. And the three was out of order."
274.	Coach Lindsay:	Right.
275.	Shonda:	So then, once they saw that they make their own table, then, they were able to correct the graph.
276.	Coach Lindsay:	And I, I think that that's a nice strategy for them to see the multiplicative relationship, but I don't know if that really helps them interpret the context. Because I think that, I think that what that still, I think that what that's still focusing on is that you have to have the numbers in order to figure out what's happening, and really what we want them to focus on the unit, the unit rate, and the relationship between 10 and two.
277.	Deanna:	Right.
278.	Coach Lindsay:	So, I think the next step would be to draw their attention to, "What is the relationship between money and time?" And if I know money, but I don't know time, then how can I use that

relationship to figure out one or the other? Because in real life, you're not gonna have an ordinal table.

279. Shonda: Uh uh.

280. Coach Lindsay: In real life, you're gonna have these situations where, I don't even know what the context is, but in real life you're gonna have the situation where they're babysitting on Tuesday, and you babysit for a lot longer than you did on Wednesday. So, I think that it's good for them to see the connection between the order to, to make sense of it, but I don't, I don't, I caution in staying with that, and leaving it there, because what that perpetuates is the m-, the additive relationship that we're trying to get away from.

In this excerpt, Coach Lindsay elicited potential instructional strategies (Turn 269) and Shonda suggested an approach that helped many of her students: creating a table based on the unit rate (Turns 270-275). Shonda went on to describe this strategy and how it supported students' understanding. Creating a table helped her students identify errors in one part of the task (a graph; Turn 275), but also could have helped students like Tommy see if their answer was unreasonable or "out of order" (Turn 273). But Coach Lindsay found this insufficient to "get them back on track." She noted that Shonda's strategy reinforced an additive notion of unit rate, rather than a multiplicative one (Turn 276-280). Although the strategy was not bad — and likely helped students "see the connection... to make sense of it" — Coach Lindsay noted that it did not address the mathematical goal that the group identified (Turn 280).

As the Magnolia workgroup continued debating additional strategies, Coach Lindsay reinforced her commitment to supporting students' sensemaking in the following turn:

> Coach Lindsay: So, I'm wondering what can we do at this point, and what sort of — Because we've been — It's easy to go over it and do all the talking, but they [the students] need to do something with this. They need to make meaning out of it. So, what can we do, what sort of experiences can we give them, so that they can make meaning of using the strategy using, "Are you attacking?" [*a problem-solving strategy*]. And then, also make that connection between, what is unit rate, and how can unit rate help me complete a table when it's not in order?

The teachers were not sure what they could do to help support students' meaning-making in the ways that Coach Lindsay described. Deanna suggested using another task about unit rate, but noted that they had already done something similar. Then Coach Lindsay suggested giving students feedback on the task that the group was looking at that day. The teachers took up this idea, saying that they could give feedback since they had not yet returned any students' papers on that task. Deanna suggested creating another task about unit rate and having students work in small groups to analyze an exemplar response and compare it to their approach, with the teachers' written feedback. Coach Lindsay, Mr. Donovan, and Shonda helped clarify these plans by giving suggestions on what kinds of contexts to use and how to implement the lesson.

For the Magnolia workgroup, teaching was about more than simply executing techniques. Through their conversations, they framed teaching as a matter of using pedagogical judgment to respond to student thinking. As they shared ideas for instruction, they also backed up their suggestions with explanations of how they could support students' learning. Shonda, for instance, did not leave her suggestion at having kids "make a table;" she described what she meant by that, how students used it, and how it helped them assess the reasonableness of their own answers. Coach Lindsay's pushback on Shonda's idea, and their resulting discussion of how to use feedback, further demonstrated the importance that this group placed on connecting instructional ideas to supporting students' sensemaking. This allowed the group to plan instruction that stood to support deeper student learning.

The Magnolia workgroup's three-week data use cycle reflected a consistent use of data as an indicator of student learning: If data only indicate what students know and are able to do, then correct and incorrect answers are insufficient to determine what students have learned. Accordingly, the Magnolia workgroup used qualitative data (student work) to interpret students' sensemaking. Analyses of student thinking allowed the group to design instructional responses that stood to support students' conceptual understanding. Through these conversations, the Magnolia workgroup collectively developed pedagogical concepts that supported teachers' opportunities to learn, encouraging the development of more ambitious pedagogies. In these ways, the Magnolia workgroup's approach to data use had the potential to instructional improvement and student learning.

Discussion

The increased emphasis on standardized testing in mathematics has resulted in greater pressure to use data to drive instruction, particularly in urban schools facing sanctions from accountability policies. District leaders and principals encourage (or require) teachers to use benchmark assessment data to inform their instruction. Schools often address the data-use mandate through teachers' collaborative workgroup meetings. Yet there are rarely any expectations for data use beyond that. The details of what it means to use data to "drive" instruction are left unspecified. And so workgroups, like those at Cypress and Magnolia, develop their own data use practices and routines.

In many ways, the Cypress and Magnolia workgroups both made concerted efforts to incorporate the district's data use expectations into their collaborative work. They systematically collected data, organized it, and analyzed it during workgroup meetings. They used their analyses to inform instruction, and sought to support student learning through their instructional plans. And they made this a regular part of their practice by engaging in regular cycles of data use. Both groups invested a great deal of time and energy into using data throughout the school year, and they did so under the guidance of knowledgeable and experienced instructional coaches. But like two brothers, the workgroups' similarities only go so far: The workgroups' approaches to data use had very different implications for supporting instructional improvement and student learning.

The cycles of data use that emerged at Cypress and Magnolia were reflections of their epistemic stances on data. At Cypress, the epistemic stance on data as measurement led the workgroup to use benchmark assessment results as objective, straightforward measures of student learning. They treated students' correct answers as evidence of their learning and incorrect answers as evidence of a lack of mastery, without differentiating among individual items or responses. This allowed them to make clear-cut categorizations of students who "missed" or "got" each Focus Topic, as well as Focus Topics that were most in need of re-teaching. Re-teaching, then, became a matter of executing techniques to help students get correct answers. Though this approach supported students' short-term success, as measured by exit tickets during the Targeted Re-teaching Days, it was unlikely to support instructional improvement or students' long-term mathematical understanding.

At Magnolia, however, the workgroup took an epistemic stance on data as an indicator of student learning. Though they used benchmark assessment data to identify students just on the cusp of mastery ("bubble kids"), they sought out the bubble students' work and analyzed it closely in C Weeks. The workgroup looked beyond correct and incorrect answers to understand students' mathematical sensemaking as evidence of learning. They framed teaching as a matter of using pedagogical judgment to support student thinking. Analyzing students' work supported this endeavor by allowing them to design instructional responses that stood to support students' deeper conceptual understanding. In these ways, the Magnolia workgroup's data use cycles stood to support instructional improvement and students' mathematical understanding.

Based on the larger corpus of teacher workgroup meetings, I conjecture that using data as a measurement is a far more common epistemic stance than using data as an indicator. This is reasonable, given the pressures of accountability policies: States judge students, teachers, and schools based on data from end-of-year state tests. Students are categorized with labels like "novice" or "proficient" based solely on the number of questions they answer correctly. Schools are similarly marked with labels like "in need of improvement" based on students' results. Since there is no mechanism to "talk back" to test developers, teachers must accept these results at face value. And so it is, perhaps, unsurprising that educators — like those at Cypress — would treat benchmarks assessments similarly, they are meant to mimic the state test.

What is surprising about the epistemic stance that Coach Diane used, however, is that it seemed to be in stark opposition to what she described as high-quality mathematics instruction. In interviews across multiple years of the study, she consistently described the importance of students' developing conceptual understanding through engagement in rich tasks, and of teachers facilitating groupwork rather than explaining mathematical procedures. Yet when it came to using data to prepare students for success on standardized tests, she organized a data use cycle that consistently led to the development of low-level tasks and lessons where teachers reinforced procedures. This suggests that, even with a knowledgeable coach, data use efforts in the context of test-based accountability policies can distort teaching and learning in ways that inhibit instructional improvement.

In contrast, the epistemic stance used by Coach Lindsay and Mr. Donovan aligned with their descriptions of high-quality mathematics instruction and supported teachers' learning opportunities toward more ambitious forms of practice. However, consistently using data as an indicator of student learning takes a great investment of resources. Analyzing student work takes time; elsewhere I describe this as aiming for quality over quantity (Garner & Horn, 2018). Their three-week data use cycle allowed the workgroup to deeply examine students' mathematical thinking, but it took up time that could have otherwise been spent creating lesson plans or grading. In interviews, the teachers expressed some frustration with this, though they said they appreciated Coach Lindsay's support. Furthermore, this approach took an investment of Coach Lindsay's and Mr. Donovan's expertise. Many teacher workgroups do not have any expert facilitator; very few have two facilitators with a background in math education. And although Mr. Donovan had other duties as an assistant principal, Coach Lindsay's sole job was to support the math teachers at Magnolia. Though using data as an indicator of student learning does not necessarily require as much support as the Magnolia teachers had, it certainly requires above-average expertise in mathematics teaching and learning. Few workgroups in our study sustained such rich conversations across their meetings (Horn et al., 2017), and so expertise may be a limiting factor in using data in this way.

In addition to the requirements for mathematical and pedagogical expertise, the Magnolia workgroup's data use practices provide additional lessons for supporting an epistemic stance on data as an indicator of student learning. First, coaches and teachers need support to use data in ways that can support instructional improvement. Data use practices are more complex than the DDDM rhetoric would suggest, and using data as an indicator of student learning is likely quite rare. Even relatively expert coaches, like Coach Diane, would benefit from supports (e.g., professional development) that address educators' epistemic stances on data and their data use practices. Second, data use conversations should take on a wider definition of data. In line with developing an epistemic stance on data as an indicator, supports for educators learning should address a range of data that moves beyond quantifiable assessment data. Third, data use conversations should move past identifying problem areas to consider student thinking. With an expanded range of available data — particularly qualitative data like student work — educators can use data to understand how students are thinking about content and use those analyses to inform future instruction. Without addressing these concerns, however, efforts for DDDM are likely to continue to distort teaching and learning and drive the U.S. education system into an abyss of endless test preparation.

References

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: principles, policy & practice, 5(1), 7-74.

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

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal*, *113*(1), 76-104.

Boudett, K. P., Murnane, R. J., City, E., & Moody, L. (2005). Teaching educators how to use student assessment data to improve instruction. *Phi Delta Kappan*, *86*(9), 700-706.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 173-206

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., & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. *Journal of Education for Students Placed at Risk (JESPAR)*, *17*(4), 247-265.

Diamond, J. B., & Cooper, K. (2007). The uses of testing data in urban elementary schools: Some lessons from chicago. In *Yearbook of the national society for the study of education* (Vol. 106, pp. 241-263). Wiley Online Library.

Duncan, A. (2009, June). Robust Data Gives Us The Roadmap to Reform. Speech made at the Fourth Annual IES Research Conference, Washington, DC. Retrieved from https://www.ed.gov/news/speeches/robust-data-gives-us-roadmap-reform

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational research review*, *5*(1), 1-24.

Garner, B. & Horn, I.S. (2016, April). Learning from Assessment Data: Epistemic Foundations of Data Use. Paper presented at the annual research conference for the National Council of Teachers of Mathematics, San Francisco

Garner, B. & Horn, I.S. (2018) Using Standardized Test Data as a Starting Point for Inquiry: A Case for Thoughtful Compliance. In Barnes, N. & Fives, H.R. (eds.), *Teachers' Data Use: Cases of Promising Practice*. Routledge, New York City.

Garner, B., Thorne, J. K., & Horn, I. S. (2017). Teachers interpreting data for instructional decisions: where does equity come in?. *Journal of Educational Administration*, *55*(4), 407-426.

Glaser, B.G., & Strauss, A. (1995). The discovery of Grounded Theory. Strategies for Qualitative Research.

Glazer, J. L., & Peurach, D. J. (2015). Occupational control in education: The logic and leverage of epistemic communities. *Harvard Educational Review*, *85*(2), 172-202.

Goldstein, B. E., & Hall, R. (2007). Modeling without end: Conflict across organizational and disciplinary boundaries in habitat conservation planning. In Kaput, E. Hamilton, S. Zawojewski, & R. Lesh (Eds.), *Foundations for the future.*

Hall, R., & Seidel Horn, I. (2012). Talk and conceptual change at work: Adequate representation and epistemic stance in a comparative analysis of statistical consulting and teacher workgroups. *Mind, Culture, and Activity*, *19*(3), 240-258.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., Wayman, J. C., Pickens, C., Martin, E.S., & Steele, J. L. (2009). Using student achievement data to support instructional decision making.

Horn, I. S. (2007). Fast kids, slow kids, lazy kids: Framing the mismatch problem in mathematics teachers' conversations. *The Journal of the Learning Sciences*, *16*(1), 37-79.

Horn, I. S. (2016). Accountability as a design for teacher learning: Sensemaking about mathematics and equity in the NCLB era. *Urban Education*, 0042085916646625

Horn, I. S., Garner, B., Kane, B. D., & Brasel, J. (2017). A taxonomy of instructional learning opportunities in teachers' workgroup conversations. *Journal of Teacher Education*, *68*(1), 41-54.

Horn, I. S., & Kane, B. D. (2015). Opportunities for professional learning in mathematics teacher workgroup conversations: Relationships to instructional expertise. *Journal of the Learning Sciences*, *24*(3), 373-418.

Horn, I. S., Kane, B. D., & Wilson, J. (2015). Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. *American Educational Research Journal*, *52*(2), 208-242

Hymes, D. (1974). Foundations in sociolinguistics: An ethnographic approach. Philadelphia: University of Pennsylvania Press.

Jennings, J. L., & Bearak, J. M. (2014). “Teaching to the test” in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381-389.

Jimerson, J. B. & Wayman, J. C. (2015). Professional learning for using data: Examining teacher needs and supports. *Teachers College Record*, *117*(4).

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *The journal of the learning sciences*, *4*(1), 39-103.

Khalifa, M. A., Jennings, M. E., Briscoe, F., Oleszweski, A. M., & Abdi, N. (2013). Racism? Administrative and community perspectives in data-driven decision making: Systemic perspectives versus technical-rational perspectives. *Urban Education*, 0042085913475635.

Lampert, M., Boerst, T. A., & Graziani, F. (2011). Organizational resources in the service of schoolwide ambitious teaching practice. *Teachers College Record*, *113*(7), 1361-1400.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Lee, M., Louis, K.S., & Anderson, S. (2012). Local education authorities and student learning: The effects of policies and practices. *School Effectiveness and School Improvement*, *23*(2), 133-158.

Little, J. W. (2011). Understanding data use practice among teachers: The contribution of micro-process studies. *American Journal of Education*, *118*(2), 143-166

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71-85.

Mandinach, E. B., & Gummer, E. S. (2013). A systemic view of implementing data literacy in educator preparation. *Educational Researcher*, *42*(1), 30-37.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). Making sense of data-driven decision making in education.

Means, B., Padilla, C., & Gallagher, L. (2010). Use of education data at the local level: From accountability to instructional improvement. *US Department of Education.*

Munter, C. (2014). Developing visions of high-quality mathematics instruction. *Journal for Research in Mathematics Education*, *45*(5), 584-635.

National Forum on Education Statistics. (2011). Traveling Through Time: The Forum Guide to Longitudinal Data Systems. Book Four of Four: Advanced LDS Usage (NFES 2011-

802). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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.

Pickering, A. (2010). *The mangle of practice: Time, agency, and science*. University of Chicago Press.

Sacks, H. (1992). Omnirelevant devices; settinged activities; “indicator terms”(February 16, 1967). *Lectures on conversation: Volumes I and II*, 515-522.

Schaffer, E., Reynolds, D., & Stringfield, S. (2012). Sustaining turnaround at the school and district Levels: The high reliability schools project at Sandfields Secondary School. *Journal of Education for Students Placed at Risk (JESPAR)*, *17*(1-2), 108-127.

Stevens, R. (2010). Learning as a Members' Phenomenon: Toward an Ethnographically Adequate Science of Learning. *Yearbook of the National Society for the Study of Education*, *109*(1), 82-97.

Villavicencio, A., & Grayman, J. K. (2012). Learning from “turnaround” middle schools: Strategies for success. *New York: Research Alliance for New York City Schools.*

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation*, *19*(3), 321-332.

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, 18.

Yuan, K., & Le, V. (2012). Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests. Santa Monica, CA: RAND Corporation. Retrieved from the RAND website: http://www.rand.org/content/dam/rand/pubs/working_papers/2012/RAND_WR967.pdf