Journal of Formative Design in Learning (2019) 3:16–26
https://doi.org/10.1007/s41686-019-00026-1

Teachers Reasoning About Students' Understanding: Teachers Learning Formative Instruction by Design

R. Benjamin Shapiro[1] · Peter Samuelson Wardrip[2]

Published online: 11 March 2019

Abstract
Formative assessment and instruction have turned out to be a remarkably difficult practice to implement in schools. Fundamental to this challenge is the fact that formative assessment is inherently a local, concrete instructional practice, as is the work or transforming assessment data into pedagogically responsive action. This paper explores teachers' thinking in their uses of a new data analysis tool to enact evidence-based instructional practices. Furthermore, this paper describes the possible relationships between teachers' existing beliefs, expertise, and routines and their construction of new practices. We show how current theories of assessment do not account for important aspects of formative instruction in practice and discuss the implications for teacher learning.

Keywords Design · Educational technology · Teacher thinking · Formative assessment

Formative instruction refers to the practice of using evidence of student thinking and learning to responsively shape ongoing instruction in order to support content mastery by all students (Black and Wiliam 1998b). We focus on formative instruction rather than formative assessment alone to emphasize the crucial role of instruction in responding to student information; it is insufficient merely to understand students' needs and we must develop pedagogies that use formative assessment to improve teaching and learning. Reviews and meta-studies have consistently demonstrated that formative instruction is one of the most powerful reforms known with strong effects across content areas and grade levels (Black and Wiliam 1998a). Yet, despite this recognition of the *potential* impact of formative instruction, there is, as Black and Wiliam (1998b) describe it, a "poverty of practice." We know little about how to enact it at scale (Black et al. 2003; Black and Wiliam 1998a, b; Ratnam-Lim and Tan 2015).

✉ Peter Samuelson Wardrip
wardrip@wisc.edu

R. Benjamin Shapiro
Ben.shapiro@colorado.edu

[1] Roser ATLAS Center, 1125 18th St. 320 UCB, Boulder, CO 80309-0320, USA

[2] UW-Madison, 210 Teacher Education Building, 225 North Mills Street, Madison, WI 53706, USA

Studies of teachers' efforts to implement formative instruction have revealed that it can be intensely difficult work, being stymied by cognitive challenges, such as teachers' lack of expertise about how to design classroom practices that elicit information about student thinking (Halverson and Shapiro 2012) or about data analysis (Black and Wiliam 1998b). In addition, school conditions play a prominent influence in the implementation of formative instruction, such as political conflict (Johnston et al. 1995), lack of data-competent leadership (Wayman et al. 2006), teacher isolation (Roehrig et al. 2008), and lack of resources, including adequate access to data about students (Wayman et al. 2004) and time to use it (Wayman et al. 2012). Further, many data systems in schools are geared around standardized, summative data, not data situated within individual teachers' routines or reflective of different students' passions (Halverson and Shapiro 2012).

This study presents a case of teachers' thinking about their students' understanding and what-to-do-next-in-instruction to further their students' thinking. The teachers were aided by a co-designed web-based tool that allowed them to look at students' annotations of text and then aggregate/filter the students' work across the whole class. Thus, we examine how teachers' use of this tool, combined with teachers' existing knowledge of students, teaching, and content made students' needs, and then possible instructional responses, "ready-to-hand" (Heidegger 2008; Wheeler 2013). This study speaks to recent calls from the field for research to document teachers' use of student data

in practice (Coburn and Turner 2012a, b, c; Halverson and Shapiro 2012; Herman et al. 2012; Shapiro and Wardrip 2015; Wardrip and Shapiro 2016). In addition, this study also surfaces the knowledge teachers need to use student data in practice (Mandinach 2012; Mandinach and Gummer 2013).

How We Can Learn What We Need to Know

The Role of Theory

Developing a knowledge base about formative instruction practice is not a theoretical exercise. Rather, it affords to apply current theory to understand practical problems, and using the study of practice to refine theory, both of domains and of assessment.

Domain Theories Domain models can tell us what is important in a particular teaching situation, i.e., what should be attended to. For example, Toulmin's (1958) theory of argument suggests that, when supporting students' argumentation, teachers should attend to their students' statements of claims, warrants, rebuttals, etc. The theory also suggests ways in which teachers could interpret such utterances to understand whether students are master arguers. This might include the characteristics of a good warrant and of a coherent argument as a combination of a warrant and a claim. However, despite the apparent clarity of the theory, assaying the ways in which students' classroom arguments relate to different ways of thinking is painstaking work that requires the development of additional theoretical machinery (Berland and Reiser 2009). And, even in that well-trodden domain, we do not yet know the reasons (including existing patterns of discourse) why different patterns of argumentation (e.g., arguing to win, arguing to learn from one another) emerge in different classrooms (Berland 2008). This information can be crucial to teachers' reasoning about student thinking (van Es and Sherin 2002). This example suggests that it is possible for domain theories to describe the characteristics of high-quality products without offering insight into the student thinking involved in creating those products, in which case, those theories are more felicitous to summative assessment than formative. Researching the pragmatics of formative instruction in domains should be able to, in the manner of Berland and Reiser (2009), illustrate ways in which current theories do not account for all that matters in teaching situations and, thereby, stimulate the development of stronger theory. Developing disciplinary theories of formative assessment would go a long way toward improving the state of the field (Coffey et al. 2011).

Assessment Theories Assessment theories can offer insight into what is necessary for teachers to know and do in order to assess student understanding, in addition to domain models. For example, the National Research Council (NRC) assessment triangle (Pellegrino et al. 2001) describes assessment as the conception and application of cognitive, observational, and interpretive models. This is seen in Fig. 1.

In the triangle, a cognitive model describes the set of things of which mastery comprises. It is essentially a domain theory of the sort described above. An observational model describes the tasks that are used to elicit data about student thinking and the ways in which those data are parsed in order to call out salient details. An interpretive model refers to the analytical strategies that are used to make claims about student thinking, in the language of the cognitive model, using the data produced by the application of the observational model.

Similarly, evidence-centered design (ECD) is an approach to making evidentiary arguments from assessments. ECD itself is an assessment design framework that supports collaborative design with rigor (Mislevy et al. 2003; Halverson and Shapiro 2012; Herman et al. 2012). Specifically, ECD fosters rigor by emphasizing the coherence and fit between evidence collection and the interpretation of results for the learning goals. This coherence is fundamental in order to make broader claims about students because it can integrate multiple types of evidence. Importantly, achieving this coherence comes through social negotiation between stakeholders.

Thus, the triangle model and ECD expand assessors' attention beyond content knowledge to pedagogical knowledge and pedagogical content knowledge. However, as a theory that is about knowing whether students know the set of things contained in the cognitive model, it still neglects a number of important details, such as what teachers need to know to reason about *why* they are seeing what they are seeing, including what local contextual information may be relevant to interpreting students' performances. This might include past instruction or the backstories students bring to the classroom (Wardrip et al. 2011; Wardrip and Herman 2018). And while the triangle points too much of what may be important, it does not provide a way to conceptualize students' affection, motivation, and creativity as objects of teacher scrutiny. Moreover,

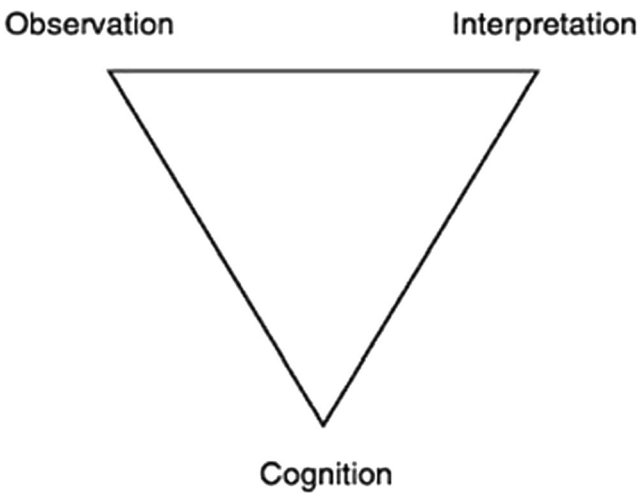


Fig. 1 NRC assessment triangle (Pellegrino et al. 2001)

it is a theory of assessment and not a theory of assessment-in-practice. Therefore, it does not model what teachers need to know in order to act on their interpretations of student thinking.

Current theoretical models about assessment pay little attention to how teachers integrate assessment into the other aspects of their work, including how they balance the problem of assessing each student's thinking with other pragmatic concerns, such as classroom management (Tomanek et al. 2008). Further research into the pragmatics of enacting formative instruction, grounded in current theory, is necessary in order to develop more robust theory. Thus, an inquiry into those pragmatics should support not only the development of theories of formative instruction practice but the refinement of existing theories of assessment, including remediation of ways in which existing theories neglect crucial aspects of practical problems. As Rothman (2004) notes, the use of interventions to improve theory "not only makes the justification for the intervention clear, but also improves the likelihood that investigators will recognize when their and their colleagues' efforts have focused consistently on a single or limited aspect of a given theory," such as how theories of assessment can speak to ways in which information could describe student thinking, but do not adequately conceptualize the practical mechanisms—and challenges thereof—through which such information could be gathered or applied.

Case Study: Teachers' Different Constructions of Formative Instruction Practice with Prototype Data Analysis Tools

Our work took place within a larger study of integrating literacy support strategies into content area instruction. This study spanned 15 months and included intensive professional development and in-class support in science, social studies, mathematics, and English/language arts with a team of sixth-grade teachers. As part of this study, we created a set of online tools for participating teachers to use to analyze their own students' work. Our goal was to enable teachers to understand their students' thinking and enact more responsive instruction.

Prior to the development of the tools, and as part of the work to support literacy in the content areas, the teachers had made annotation a routine component of reading in their classes. Annotation is a cognitive strategy that can improve students' understanding of texts (Liu 1996), especially when done in a manner that draws readers' attention to the most important structure and content within a text. As an activity that produces external representations about individual use of text, annotation seems like an ideal window into student thinking *for teachers*. However, little is known about what teachers should attend to, and how they should reason about it when looking at students' annotations in order to understand student thinking and then to shape follow-up teaching.

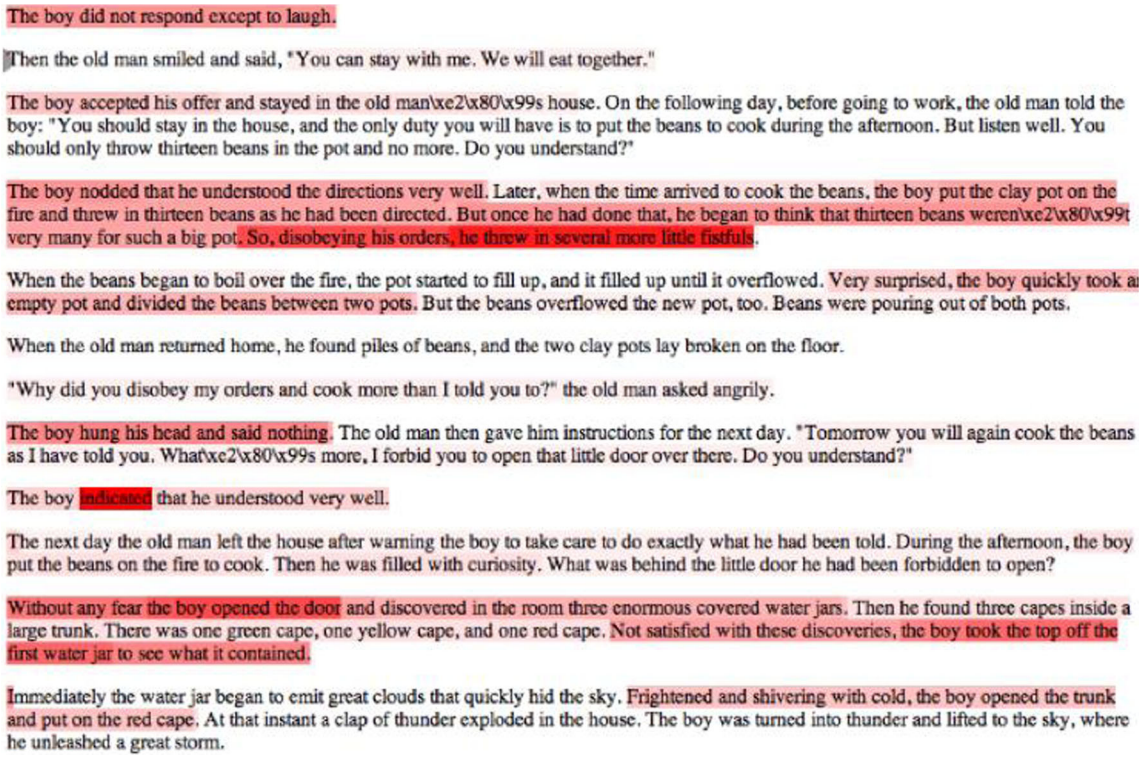
Understanding students' thinking through their annotations is a difficult problem, largely due to the (relative) enormity of the data sets with which teachers using annotation must contend, and the different perspective through which these data can be analyzed (e.g., as we show below, it is possible to notice students' effort, interest, time spent, understanding of rhetorical structure, and/or content understanding through their annotations, singly, and in aggregate). Any given annotated page of text might have a dozen (a number that is typical in our team's past work) or so details highlighted and annotated. For a short reading of three pages, that is about 36 annotations per student. Across the hundred students that a typical teacher may have, there might easily be over 3600 data points generated by a single assignment. Given this level of detail, how can teachers productively use students' annotations in order to understand their thinking? What can they understand? What is hard to see? Under which circumstances (including, but not limited to, content domains and already identified student needs) should teachers prioritize noticing some details over others?

Electronic data analysis tools might support teachers in analyzing their classroom data by making it easier to manage the volume of information. For example, digitizing the work could cut down on time-consuming paper shuffling. It may also aid in finding conceptually related student work (e.g., annotations of the same passage within a text), or knowing about broad patterns in student thinking (e.g., the places in the text about which students have the most questions).

While we created several tools to assist teachers with using students' annotations in order to support their instruction, teachers elected to use one of them, the heatmap (Hill et al. 1992), almost exclusively. The teacher tools re-represented data produced by a student annotation tool, which were automatically uploaded, in real time, to a central database for access by the teacher tools.

The heatmap tool was intended to aid teachers' data analysis by making it easy for them to see which passages of a text were most frequently annotated by their students and to easily access students' annotations of any given place in a text (Fig. 2).

The heatmap augments a text with information about annotation frequency by shading words in the text with a background color, ranging from white, which meant that no student annotated to dark red, which meant that all of the students annotated. Thus, passages that have been annotated frequently are "hot." In addition, teachers can click anywhere on the text to see any/all students' annotations of the place clicked. We hoped that these affordances would not only help teachers to easier access their student work but *allow us to understand how teachers could enact data analysis practices* if some of the impediments to doing so, imposed by paper, were removed (Fig. 2).


The boy did not respond except to laugh.

Then the old man smiled and said, "You can stay with me. We will eat together."

The boy accepted his offer and stayed in the old man\xe2\x80\x99s house. On the following day, before going to work, the old man told the boy: "You should stay in the house, and the only duty you will have is to put the beans to cook during the afternoon. But listen well. You should only throw thirteen beans in the pot and no more. Do you understand?"

The boy nodded that he understood the directions very well. Later, when the time arrived to cook the beans, the boy put the clay pot on the fire and threw in thirteen beans as he had been directed. But once he had done that, he began to think that thirteen beans weren\xe2\x80\x99t very many for such a big pot. So, disobeying his orders, he threw in several more little fistfuls.

When the beans began to boil over the fire, the pot started to fill up, and it filled up until it overflowed. Very surprised, the boy quickly took an empty pot and divided the beans between two pots. But the beans overflowed the new pot, too. Beans were pouring out of both pots.

When the old man returned home, he found piles of beans, and the two clay pots lay broken on the floor.

"Why did you disobey my orders and cook more than I told you to?" the old man asked angrily.

The boy hung his head and said nothing. The old man then gave him instructions for the next day. "Tomorrow you will again cook the beans as I have told you. What\xe2\x80\x99s more, I forbid you to open that little door over there. Do you understand?"

The boy indicated that he understood very well.

The next day the old man left the house after warning the boy to take care to do exactly what he had been told. During the afternoon, the boy put the beans on the fire to cook. Then he was filled with curiosity. What was behind the little door he had been forbidden to open?

Without any fear the boy opened the door and discovered in the room three enormous covered water jars. Then he found three capes inside a large trunk. There was one green cape, one yellow cape, and one red cape. Not satisfied with these discoveries, the boy took the top off the first water jar to see what it contained.

Immediately the water jar began to emit great clouds that quickly hid the sky. Frightened and shivering with cold, the boy opened the trunk and put on the red cape. At that instant a clap of thunder exploded in the house. The boy was turned into thunder and lifted to the sky, where he unleashed a great storm.

Fig. 2 Heatmap

Study Design

We studied teachers' uses of the heatmap in order to better understand the problem of understanding students' thinking through annotations and how the specific cases of each teacher's particular enactment were related to that teacher's pre-existing routines, beliefs, and expertise. We analyzed each teacher's work qualitatively, using the learning to notice framework (van Es and Sherin 2002) to identify important aspects of teachers' uses of information. This analysis included how teachers called out details, how they interpreted those details, and how they then reasoned about those details using contextual knowledge.

When attempting to understand *why* teachers might enact the practices they did, we took a cultural-historical approach (Cole and Engeström 2007), juxtaposing the details of past practices and articulated beliefs with new practices, in order to understand how the former may be related to the latter. That is, we assumed that teachers' new tool-using practices would not be enacted de novo but would, instead, build upon what they already believed and did, including the schemes by which they saw significations of meaning in student-produced classroom materials (Bryk et al. 2006). Therefore, in the following account, we precede the description of how teachers used the tools with a brief account of pertinent details about their pre-existing practices and beliefs. Then, we relate our observations of the new practices to those details.

Background of Setting and Data

While the larger study of this work included all three members of a sixth-grade teacher team, the math teacher declined to fully integrate annotation into her classroom's reading practices. Thus, only two of those three teachers were candidate users of the learning analytics tools described above. For each of these two focal participants, we describe below (1) what the practitioner's practice was like at the outset; (2) what their beliefs were about teaching, learning, and assessment; (3) how they adapted annotation (i.e., what did they set themselves up to notice?); (4) how they analyzed students' annotations (i.e., what did they notice and how did they interpret/reason about it?); and (5) how that analysis informed instruction. While topics 3 and 4 are fundamentally about cognitive, observational, and interpretive models as specified in the NRC assessment triangle, it is critical to note that teachers' decisions about these are not independent of the choices that they make about instruction. This unique micro-historical context of their classrooms (including past teaching practices and areas of comfort) inextricably shaped their implementation of formative instruction: The decisions that teachers made about how to teach and how students should annotate determine what data will be available for later analysis.

An example of this pre-determination of available data was how both teachers canonized annotation into one procedure that all students in their classroom should follow. For instance, one of them required that every student highlight and

summarize in a word everything that the main character in a fictional story says and does. These procedures, which were intended by the teachers to be easy enough for all students to follow, shaped the kind of data that were available for analysis. But these procedures, while perhaps suitable for students struggling to decode the literal events in a story and capable of providing information about the same, are unlikely to provide information about higher order reasoning about the text, such as inferences about motivation or cause and effect. As such, teachers' decisions about how to enact instruction not only enabled certain kinds of student learning and analysis thereof, but may have also constrained the kinds of learning, and analysis, possible.

Our descriptions of teachers' work below are drawn from observational and interview data collected over 15 months, with an especially intense period of data collection occurring during the final 2 months of the school year when teachers used the online learning analytics tools. During that period, on a weekly basis, the two teachers completed questionnaires about student mastery goals; researchers audio recorded any "pre-teaching" of texts, then audio recorded teachers' "think-aloud" analyses of student data while completing a follow-up questionnaire about student understanding and follow-up instructional plans, any follow-up instruction was audio recorded, and, when possible, teachers were debriefed after that instruction about their impressions about its success.

The examples presented here are selected to exemplify important tensions in making formative instruction routine. To the extent that they do not portray a number of other contrasts and similarities between the teachers, they are not necessarily a representative sample of all of the teachers' practices, though the claims made here are not contradicted by other data.

As depicted in Fig. 3, the system of practice reported on is the product of combining a set of *Base Instruction Practices* (i.e., the set of instructional routines teachers already had), drawing upon *Base Beliefs & Expertise*, and *Base Texts*, with an outside reform message, *Annotation*, part of a larger push for attention to literacy in the content areas. Teachers then reified *Annotation* by creating a set of *Classroom Annotation Procedures* (in the examples we present, teachers used the same *Base Texts* as they had in past years). As constructivists, we hypothesize that the canonical ways of annotating that teachers constructed and prescribed would not be random but would, instead, draw upon their existing routines, expectations, expertise, and beliefs and that the ways that they did so might parallel aspects of their existing teaching practice. We wanted to understand how teachers drew upon these same beliefs and expertise when analyzing student work, reasoning about its implications for a subsequent instruction, and enacting it. Accordingly, we present our results below by moving left to right through Fig. 3, quickly summarizing salient features of teachers' existing beliefs and routines, then describing how they enacted new routines.

Results

Participants

Both teachers described here were part of the same sixth-grade teacher team at an ethnically and economically diverse Midwestern American K-8 school. Jane teaches science and social studies. Greg teaches English/Language Arts. Both teachers have about 5 years of teaching experience, most of it in the same school.

Base Beliefs and Practices

In interviews and classroom observations conducted over the course of the school year, we learned that Jane and Greg had

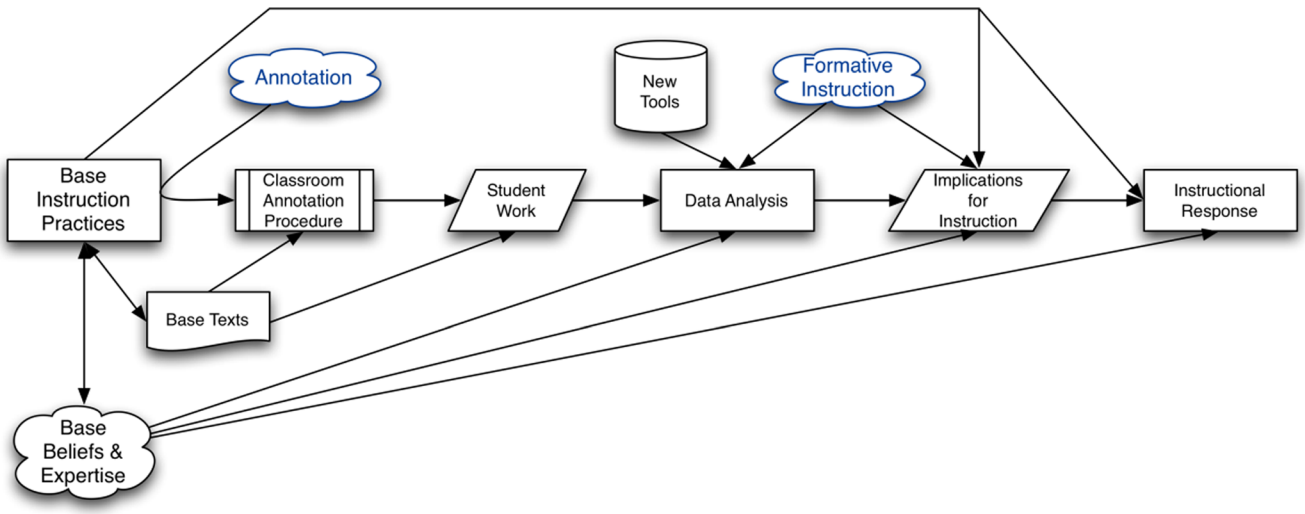


Fig. 3 Genetic model of formative instruction

sharply differing beliefs about student learning and teaching. Jane talked about students' academic success, including her own, as largely the result of learning to learn, of developing valuable study skills, which enable them to succeed academically. In contrast, Greg seemed to have a fixed capacity view of students. At one point, he explained that some of his students are "pencils" and that others are "computers" and inherently more capable than the "pencils." He used this belief to rationalize, among other things, a sliding scale grading system where some students were graded easier than others. This contrast seems consistent with the incremental and entity theories, respectively, described in the motivation literature (Dweck and Leggett 1988; Blackwell et al. 2007).

The teachers also articulated different beliefs about whether and how they should differentiate instruction for their students. Jane described how lesson planning should account for the perspective and needs of the range of her students. For example, she explained how:

> I can't really be planning a lesson without thinking, is this really going to stimulate or inspire, you know, this high [achieving] student? Is there enough in it for her? And then, how is it accessible for this particular low [achieving] student? … So, I think, you know, when I'm planning a lesson too, I'm thinking about those little random things that happen. Like, 'Oh, how's so and so going to feel about this?' Like, how can I [engage each student]?

In contrast, Greg described how there are a priori right and wrong ways of teaching and learning content and that the right ways should not be deviated from, even if they do not work for some students (because of their current level of skill or their innate capacity):

> Going into it, one of the reasons why I do this the way I do it with the lecture – and it's very structured – is whether you're at my low end or my high end, I want them to hear the same words. You might not get 'author and you' right now because you're not there. But I don't want to dumb it down. I try to believe as a teacher that I set the bar high for all kids. *But some kids are never going to get there*. Some words are over their head. We have an example of a few in our class that don't get it. Not that I'm not going to try to reteach it. Not that I'm not going to sit down and reteach, preteach, whatever. But I don't want to start coming up with smaller, baby explanations because if and when eventually they get it, *they're never going to get to the same place*.

Observations of the teachers' instruction were similarly contrastive. Jane's class was highly interactive and question centric, with an emphasis on students' ideas about content. She taught by asking her students questions and by addressing questions that they had. In fact, she described using her students' questions to motivate her own reading about science and regularly checked out new books from the library in order to research answers to students' questions. In contrast, Greg's class was lecture-driven, with relatively little student talk and therefore fewer opportunities for Greg to understand his students' thinking.

These gross differences in teaching style seemed to align with their stated beliefs. Jane talked in the abstract about the need to make instruction responsive to her students' interests and needs. Her teaching included frequent opportunities to see how her students thought and to respond to those observations. Greg talked about right and wrong ways to teach things, knowable by him a priori, and diminished the importance of adapting instruction to students' needs, then enacted instruction with relatively few places to see and respond to students' thinking. He expressly disclaimed the possibility of all students mastering content regardless of how it was taught. Below, we describe how teachers' canonizations of annotation, and their analyses of students' annotations, seemed to parallel these contrasts.

Texts Used and Canonizations of Annotation

One strong influence on the ways in which the teachers adapt annotation for their students, and are able to see students' thinking through their annotations, is the texts used in the classroom, for they are central tools to both students' and teachers' activity. During the time period studied here, teachers used texts drawn from their usual repertoires, all of which were in district-mandated textbooks or on district-suggested reading lists. Jane used her science textbook (Space Science, by McDougal-Littell), while Greg used a number of fictional short stories. While Jane's texts were short, comprised of litanies of facts grouped in short paragraphs, Greg's texts were long and contained supernatural ambiguities, such as whether or not a character returned from the dead.

Jane's adaptation of annotation was to instruct her students to highlight all section headings, write a question that the heading inspires next to each, and then to annotate any text within the following section that helps to answer the question. Greg instructed his students to highlight everything the main character says or does and to write a summary adjective (e.g., "rude") beside each statement or action. Greg's approach necessarily affords considerably less discretion to students than Jane's approach does. Whereas Jane's students might select different text based upon whatever questions they asked of headings, Greg's students should all annotate the same literal events of the story.

Thus, both teachers seemed to adapt annotation in ways that were congruous with their existing routines. Questions had been a central feature of Jane's pedagogy and remained

so in the new activity of annotation. Greg, who had insisted on one shared way of learning in his class, created a system of annotating where there was one correct way to annotate. At the same time, the teachers' annotation styles seemed mismatched to the texts that they used. Jane's annotation style, which was about questions and explanations, was paired with a text that regarded science mastery as knowing a set of facts. Greg's annotation style, which was about identifying facts in a narrative, was paired with stories about the mystical and unexplained. In both cases, the teachers' base practices seemed to drumpf what the texts alone would suggest. This supports our constructivist hypothesis about the self-similarity of teachers' practice during reconstruction with new tools and is also consistent with the literature on school reform (Coburn 2001).

Analyses of Student Work

Both teachers used the annotation analytics heatmap tool to analyze their students' work. They used the tools to enact multi-day instructional arcs wherein the teachers introduced texts to their classes, students read and annotated those texts at home, and then the teachers analyzed those annotations using the tools in order to plan post-reading instruction. Jane and Greg completed questionnaires at multiple points in those arcs, including at the very beginning about learning goals and after analyzing their students' work where they wrote about what they believed the students did and did not understand and the instructional implications of those beliefs.

Both teachers completed research questionnaires about what students should be able to know from and do with the texts read (i.e., learning goals). However, Jane's level of description was more concrete than Greg's. For example, Jane identified, among others, the following goals for the "Stars have life cycles" text: "Understand how stars form and change," "Understand differences in the life cycles of higher and lower mass stars." In contrast, Greg described more abstract skills, such as "Inference," "Understand cause and effect," and "Motivation." Yet, the goals that Greg identified were inconsistent with the ways that he asked students to attend to the text (i.e., picking out the literal events of everything the character says and does) and none of his in-class activities elicited these practices. In fact, as we illustrate, his in-class discussions with students were unsupportive of many students' reasoning. More puzzlingly, while there were stark differences between teachers' goals and the ways that they taught annotation, we cannot fully account for the possible impact of these styles of annotation on teacher thinking. As we illustrate, both teachers attended to only a subset of the goals they described, suggesting a need for further investigation into what analyses different annotation approaches afford.

Jane Jane assigned her students a section from her textbook entitled "Stars Have Life Cycles", which began with a heading of the same name. Her students annotated that heading with questions like "What happens in the cycle?" "What are these life cycles like; what happens in them?" "How long are they?" "What is a star's life cycle like?" Students did this annotation as homework, and early the following morning, prior to the start of school, a member of our research team sat with Jane and asked her to explore the data, explaining what she noticed and what she might do about it in that day's class.

We quickly discovered that the paragraphs throughout the texts were almost uniformly pink on the heatmap, reflecting the fact that there was no uniformity within the students doing the assignment about what text to annotate. Jane described this phenomenon as "what you would expect," reflecting her expectation that students would focus on different sections of text, depending upon their interests and questions.

On both of the arcs from Jane's class that we studied, as Jane analyzed students' annotations, she attended primarily to evidence of students' *engagement* with the text, rather than whether students correctly understood what they were reading. For example, "We Identify Stars By Their Characteristics", another section from her textbook, contained the following passage:

> It is hard to get a sense of how large stars are from viewing them in the sky. Even the Sun, which is much closer than any other star, is far larger than its appearance suggests. The diameter of the Sun is about 100 times greater than that of Earth. A jet plane flying 800 kilometers per hour (500 mi/h) would travel around Earth's equator in about two days. If you could travel around the Sun's equator at the same speed, the trip would take more than seven months" (Space Science 2004, p. 462).

Jane called out one of her students' annotations of a section of the passage bolded above, which read "What about gas to keep the plane going?" as follows:

> Jane: He was thinking a lot, like he is really engaged. That is good. I mean it is – you know what I mean.
> Researcher: Yeah, it is both, right? It is – because when I looked at it, I thought, okay, well, he is thinking about what he is seeing and trying to imagine it, right? The question does not have anything to do with the overall point of the reading, though.
> Jane: I mean, he is thinking super hard. Like, look at this. *Jane points at a passage of text that says* "Some stars are much larger than the Sun", *about which the same student has annotated* "Dose (sic.) that mean that they are hotter"
> Jane: "Does that mean they are hotter?" I mean, who cares if it means that? He is really thinking about that.

Jane repeatedly took this stance, that actual factual understandings matter less than "really thinking," in one-on-one conversations with me and in presentations of her classroom data to her colleagues. There is a reason to believe that her claim here is partly pragmatic because Jane struggled to ascertain from looking at students' annotations precisely what they did or did not understand. For example, when attempting to complete a research questionnaire about the night's homework that explicitly requested information about her students' most and least understood content, Jane asked me "Can I just say 'everything else' because in a sense I guess I do not know if I can tell that?" explaining further that "It is harder to tell what they do not understand than what they do understand… I am not ready to write that [an understanding goal she identified in advance] is something they least understood."

In short, while Jane could easily point to evidence of students' effort to read her classroom texts, she struggled to make claims about students' conceptualization of the disciplinary ideas that she identified as important for the unit. In a follow-up conversation, Jane also could not identify how what she saw using the data analysis tools should affect her plans for teaching.

Greg Greg's data analysis process was markedly different from Jane's. Whereas Jane had regarded pinkness on the heatmap as "what you would expect," Greg recognized it as problematic because he expected his students to annotate all of the same things (which would result in dark red shading). When he analyzed students' work, he tended to first examine the distribution of annotations through a text (looking, for example, at whether students' attention petered out by the end of a text, indicating that they "got lazy") and whether students annotated the "correct" details, then examine what students actually wrote in their annotations.

A striking example of Greg's mode of analysis occurred at a critical passage in the story "Lob's Girl," an English short story about a dog named Lob (main character) that adopts Sandy, the daughter of the Pengelly family. Midway through the story, Lob and Sandy are struck by a truck while walking down the street. Sandy is hospitalized. Lob is killed and the Pengelly brothers wrap his body in a chain and throw him from their boat, burying him at sea. The supernatural twist to the story is that Lob seems to reappear at the hospital to comfort Sandy.

Before he assigned the text to his students, Greg wrote on a questionnaire that "students need to be able to understand and comprehend that they may have to take a reader's 'leap of faith' to believe the story". As I show below, this leap of faith proved to be a sticking point for Greg's analysis of student thinking, even while his annotation canonization seemed to only highlight clear-cut literal elements of the story.

Lob's reappearance in the story reads as follows: "By that afternoon it became noticeable that a dog seemed to have taken up position outside the hospital, with the fixed intention of getting in." When studying students' annotations, Greg noticed that that sentence was pink, which he regarded as problematic. He explained:

> Every student – it's a non-negotiable. There's no inferencing there. The dog is outside whining." … "Are these students annotating the right things? My quick gut feeling is no. It was simple. What does the dog do? It sits in front of the hospital. And five of our ten kids did it.

Unlike Jane, who struggled to come up with actionable conclusions about the data she saw, Greg easily identified the non-unanimous annotation of the passage as indicating that students either did not understand his instructions or did not understand the story. Because of this, he planned to specifically address this perceived confusion—a misunderstanding of what he believed to be the literal events of the story—in his next class.

From Analysis to the Classroom

Both teachers drew on their data analyses in their subsequent teaching. However, they did so in sharply different ways.

Jane Jane, who had trouble identifying what students did or did not understand from the data, did not identify ways in which what she saw in students' work should cause her to change her lesson plans, but nonetheless found surprising value in her examination of the data. She used her recollection of the students' work to reason about students' in-class participation. For example, she compared the understandings, ascertained through in-class dialog, of students who did do the homework with those who did not do the homework and had to read the text in class ("You guys are understanding and remembering what you read, even those of you who did it last night").

She also used her knowledge of students' annotations to reason about the appropriateness of her own instructional choices. When an in-class discussion with students revealed that many students did not understand a text, she used her knowledge, gleaned through annotation analysis, that many of those same students had read the text to determine that assigning the text as independent reading was a mistake, because it had been too hard. She explained her conclusion:

> I think it obviously wasn't something that they should just read independently. There are things in this science course all year long that I make decisions about what can they read independently and what do I just need to teach in a different way, and this wasn't one of those

things to read like that and to just—I really don't think it was. I don't know.

That episode revealed a potential impact of data analysis, and the new tools, on instruction that had been unanticipated by the model depicted in Fig. 3. Whereas the model anticipated that data analysis could inform instruction by enabling teachers to identify a set of focal constructs for teaching, it did not account for the way in which a teacher could use aggregate information, not contemplated by the assessment triangle, about in-class participation to reason about assessment data.

Greg Greg was able to identify specific understandings or interpretations of the text that shaped his plans for subsequent in-class instruction. In the example above, Greg noted as problematic the fact that many of his students did not annotate a particular passage of text. As planned, Greg raised the discrepancy between his expectations and students' work in class:

Greg: Tell me why you didn't annotate that part.
Student: Was this at the hospital when he was sitting outside?
Greg: At the hospital. It says that the dog, it was sitting outside the hospital.
Student: Maybe after you read it, you didn't know if that was really Lob.
Another Student: Yeah, it might not have been Lob.

In the above exchange, students proffered an explanation for why they had annotated the text in the way that they had. Their explanation was reasoned, not haphazard, and challenged Greg's interpretation of the data. Whereas Greg believed that the meaning of the text was unambiguous (i.e., that the dog in the story was Lob), some students were unsure (based upon textual evidence that the dog was buried at sea and their knowledge that dead things do not return to life). The story is intentionally unambiguous. The students' explanation revealed that they annotated as they did for reasons that Greg had not contemplated when he analyzed the data. That is, Greg's interpretations of students' work had not fully accounted for the cognitive possibilities of the task, including students not making the supernatural leap that the author offers.

Greg was unable to integrate this new information into his subsequent teaching. During a later exchange with students about the same passage of text, Greg scolded his students for not annotating the passage:

Greg: 'By that afternoon it became noticeable that a dog seemed to have taken a position outside the hospital.' Ok. Who's taken up the position? Lob.
Student: I didn't believe it was Lob.
Greg: So, at this point, this was the sentence I was talking about, why I didn't understand why everyone who did this assignment didn't have that annotated...
Student: But –
Greg: At this point, I thought it was Lob as most people probably do.
Student: I think it was Lob, but I didn't know.
Greg moved on to another topic.

Even in this follow-up discussion hours later, Greg seemed utterly unable to respond to the ways in which students' legitimate interpretations of a classroom text challenged his own.

By any definition, Greg was enacting formative instruction: he studied student work in order to identify misunderstandings, then planned and enacted instruction meant to remediate those misunderstandings. The problem was that the cognitive model employed by Greg was flawed. It did not account for multiple ways of understanding a story and an incorrect interpretation of student data. This leads to mistaken observational and interpretive models and subsequent instruction.

Discussion

Summary

Both teachers' adaptations of annotation seemed to mirror their base instructional practices and beliefs. Jane's classroom instruction made frequent use of questions to elicit information about student thinking and she canonized annotation in a manner that was question-centric. Jane discussed, in the abstract, the importance of engaging every student in instruction and, when she analyzed students' data, she attended primarily to evidence of students' engagement. At the same time, she could not conclusively determine from the data whether students correctly understood the texts' content and noted that whether they did was less important than whether they were engaged or not. In class, Jane used her recollection of the data to reason about students' classroom participation and the effectiveness of her own instructional decisions.

Greg's classroom instruction was primarily lecture-driven, with Greg delivering content in a manner he could know, a priori, to be the one best way. His adaptation of annotation was intentionally homogeneous, with all students expected to highlight the same details. His analysis of students' annotations was, accordingly, largely consumed with whether students annotated the right things. When Greg noted what seemed to him to be mistaken in students' work, he did not seem to realize ways in which they possibly indicated other *legitimate* interpretations of the text and did not respond appropriately even when his students explained their thinking to him.

A number of studies have shown how practitioners' existing routines, beliefs, and expertise, and relationships can influence the ways that they construct new practices using new technologies (Barley 1986; Orlikowski 1992). As in those studies, both teachers' uses of the tools to enact new routines were deeply connected to existing expertise and beliefs. For example, Greg's flawed enactment of formative instruction included using the tools to analyze student thinking, but he used those tools to come to erroneous conclusions and to take problematic actions due, arguably, to lack of expertise. Jane enacted a focus on student engagement, consistent with her beliefs about the importance of doing so, but could not use the tools to determine what students did and did not understand. In both cases, the teachers' uses of the tools were simultaneously promising and problematic. On the one hand, both teachers thought the tools were valuable, in part because both believed that they were able to better understand their students' thinking than they otherwise would have been able to in the time available to them. On the other hand, Jane's analyses were insufficient to enable her to tighten her lesson plans around specific student needs and Greg saw misunderstandings where none existed. In order for the field to understand how to support formative instruction, whether it be through learning analytics tools or instructional coaching, we need to build models of formative instruction that take into account the variety of ways teachers reasonably enact formative instructional practices, and support teachers' development of more student thinking centered, evidence-based teaching.

Issues Raised

The purpose of the work here was to develop some rudimentary knowledge about how teachers could use annotation data in order to understand their students' thinking, and how teachers' choices about how to understand student thinking through annotation can be shaped by teachers' existing routines and knowledge. While there are too few examples in this paper to make causal claims about what matters most, the data presented above nonetheless raise important issues for future enactors of formative instruction practices, designers of learning analytics tools to support them, and designers of teacher learning opportunities because the challenges the teachers encountered here are ones with which any future effort will have to contend.

There are many domains in which we do not yet have established models of how best to call forth evidence of understanding and what to attend to in such data. Even seemingly, well-conceptualized activities, such as scientific argumentation, can be surprisingly challenging to assess and predict the course of (in light of past practice) in the classroom. There are many intellectually demanding activities where the value of choices about how to perform the activity, and the quality of the resulting product, is highly subjective, even in such seemingly "hard" disciplines as computer programming and statistics (Turkle and Papert 1992).

This also raises the question of whether or not an educational activity needs to be able to be assessed in order to have value. Annotation is far more difficult to an activity to assess than a multiple choice test because there are many ways one can engage with a text, not all of which can be easily identified as correct or not. There are quite likely ways of engaging with texts though annotation that can be productive for learners that would altogether defy systematic assessment about correctness. What does that mean for teachers and technology designers? What do those implications say about the relevance of assessment theories like the NRC triangle or the evidence-centered design for formative instruction?

The data presented here suggest that even in domains where we do not know what "good" looks like; trying to enact formative instruction can still have value. Jane was able to know about her students' motivation and engagement and to use the annotation data to reflect upon her own instructional decisions. But, she did so by conceiving of data in ways totally outside the NRC triangle. Greg worked within the triangle but ended up generating spurious conclusions. As limited in scope as the data presented in this paper, they nonetheless challenge the appropriateness of tools like the NRC assessment triangle or the evidence-centered design assessment cycle (Williamson et al. 2004) as foundations for creating formative instruction in two respects: first, such theories usually assume that it is possible to know what mastery looks like but that may not be the case in many valuable activities (like annotation). Second, much of current assessment theory is grounded in psychometric determinations of a priori important constructs and neglects the full scope of information that can be interesting and useful to teachers in practice (such as engagement). Richer theories, incorporating the practical disciplinarily grounded possibilities for using annotation data productively, are needed.

Compliance with Ethical Standards

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

References

Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly*, 78–108.

Berland, L. G. K. (2008). Understanding the composite practice that forms when classrooms take up the practice of scientific argumentation. Unpublished Doctoral Dissertation.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26–55.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 13–14.

Black, P., & Wiliam, D. (1998b). Inside the Black box: raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. *Child Development, 78*(1), 246–263.

Bryk, A., Gomez, L., Joseph, D., Pinkard, N., Rosen, L., & Walker, L. (2006). *Activity theory framework for the information infrastructure system. Working paper*. Chicago: Information Infrastructure System Project at the Center for Urban School Improvement.

Coburn, C. E. (2001). Collective sensemaking about reading: how teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis, 23*(2), 145–170.

Coburn, C. E., & Turner, E. O. (2012a). Research on data use: a framework and analysis. *Measurement: Interdisciplinary Research and Perspectives, 9*(4), 173–206.

Coburn, C. E., & Turner, E. O. (2012b). The practice of data use: an introduction. *American Journal of Education, 118*(2), 99–111.

Coburn, C. E., & Turner, E. O. (2012c). Putting the "use" back in data use: an outsider's contribution to the measurement community's conversation about data use. *Measurement: Interdisciplinary Research and Perspectives, 9*(4), 227–234.

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching, 48*(10), 1109–1136.

Cole, M., & Engeström, Y. (2007). Cultural-historical approaches to designing for development. In Valsiner, J., & Rosa, A. (Eds.), *The Cambridge handbook of sociocultural psychology* (pp. 484–507). New York: Cambridge University Press.

Dweck, C. D., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256–273.

Halverson, R., & Shapiro, B. (2012). *Technologies for education and technologies for learners: How information technologies are (and should be) changing schools (No. 2012-6)*. Madison: Wisconsin Center for Educational Research (WCER) Working Paper.

Heidegger, M. (2008). *Being and time*. New York: HarperCollins.

Herman, P., Wardrip, P., Hall, A., & Chimino, A. (2012). Teachers harness the power of assessment. *The Learning Professional, 33*(4), 26.

Hill, W., Hollan, J., Wroblewski, D., & McCandless, T. (1992). Edit wear and read wear. *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 3-9.

Johnston, P., Guice, S., Baker, K., Malone, J., & Michelson, N. (1995). Assessment of teaching and learning in "literature-based" classrooms. *Teaching and Teacher Education, 11*(4), 359–371.

Liu, K. (1996). Annotation as an index to critical writing. *Urban Education, 41*, 192–207.

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.

Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science, 3*(3), 398–427.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: on the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Ratnam-Lim, C. T. L., & Tan, K. H. K. (2015). Large-scale implementation of formative assessment practices in an examination-oriented culture. *Assessment in Education: Principles, Policy & Practice, 22*(1), 61–78.

Roehrig, A. D., Duggar, S. W., Moats, L., Glover, M., & Mincey, B. (2008). When teachers work to use progress monitoring data to inform literacy instruction: identifying potential supports and challenges. *Remedial and Special Education, 29*(6), 364–382.

Rothman, A. (2004). "Is there nothing more practical than a good theory?" Why innovations and advances in health behavior change will arise if interventions are used to test and refine theory. *International Journal of Behavioral Nutrition and Physical Activity, 1*(11), 11.

Shapiro, R. B., & Wardrip, P. S. (2015). Keepin'it real: Understanding analytics in classroom practice. *Technology, Instruction, Cognition & Learning, 10*(2), 127–149.

Space Science. (2004). *Module E*. Evanston: McDougal Littell Science.

Tomanek, D., Talanquer, V., & Novodvorsky, I. (2008). What do science teachers consider when selecting formative assessment tasks? *Journal of Research in Science Teaching, 45*(10), 1113–1130.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Turkle, S., & Papert, S. (1992). Epistemological pluralism and the revaluation of the concrete. *Journal of Mathematical Behavior, 11*(1), 3–33.

van Es, E. A., & Sherin, M. G. (2002). Learning to notice: scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education, 10*(4), 571–596.

Wardrip, P. S., & Herman, P. (2018). 'We're keeping on top of the students': Making sense of test data with more informal data in a grade-level instructional team. *Teacher Development, 22*(1), 31–50.

Wardrip, P. S., & Shapiro, R. B. (2016). Digital media and data: using and designing technologies to support learning in practice. *Learning, Media and Technology, 41*(2), 187–192.

Wardrip, P. S., Herman, P., Gomez, L. M. & Greeno, J. G. (2011). *Knowing more about students' backstories: Rich data for instruction*. Round-table paper presentation for the annual meeting of the American Educational Research Association, New Orleans, LA.

Wayman, J., Stringfield, S., & Yakimowski, M. (2004). *Software enabling school improvement through analysis of student data*. Baltimore: Center for Research on the Education of Students Placed At Risk. CRESPAR/Johns Hopkins University.

Wayman, J. C., Midgley, S., & Stringfield, S. (2006). *Leadership for data-based decision making: Collaborative data teams*. In annual meeting of the American Educational Research Association, San Francisco, CA.

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.

Wheeler, M. (2013), "Martin Heidegger", *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL http://plato.stanford.edu/archives/spr2013/entries/heidegger/. Accessed 17 June 2018.

Williamson, D., Bauer, M., Steinberg, L., Mislevy, R., Behrens, J., & DeMark, S. (2004). Design rationale for a complex performance assessment. *International Journal of Testing, 4*(4), 303–332.