THE UNIVERSITY OF CHICAGO

TALKING SCIENCE TO SCHOOLS: ORGANIZING THE RESEARCH-PRACTICE NEXUS IN EARLY 21ST CENTURY AMERICAN EDUCATION

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPARATIVE HUMAN DEVELOPMENT

BY

SHENGHE YE

CHICAGO, ILLINOIS

AUGUST 2021

For Paula, who is not my friend

Table of Contents

List of Figures

List of Abbreviations

In alphabetical order.

BC2	Better Curriculum for Better Classrooms (Institute study)
COG	Classroom Observation Group (Empire University)
DOE	US Department of Education
ESRA	Education Sciences Reform Act of 2002 (established IES)
GSE	Golden School of Education
IES	Institute for Education Sciences
Institute, the	Golden Mathematics Institute
NCES	National Center for Education Statistics (housed in IES)
NCLB	No Child Left Behind Act of 2001 (passed Jan 2002)
NRC	National Research Council
NSF	National Science Foundation
OERI	Office of Educational Research and Improvement (replaced by IES)
RCT	Randomized Controlled Trial
SBR	Scientifically Based Research

Acknowledgements

A great deal of effort by a good many people was essential to the production of this finished dissertation.

The support of my interlocutors on the "PD team" has been invaluable. I am grateful to their willingness to allow a graduate student intern to begin studying the work that they do. Their intellectual insight, openness, curiosity, and generosity served as both a doorway to and a model for my own work. They showed a great deal of trust in me, even as I stumbled my way through my fieldwork, graciously allowing me to learn and grow from my mistakes.

I have also had the tremendous benefit of guidance from and generative conversations with E. Summerson Carr, Lindsey Richland, Eugene Raikhel, Constantine Nakassis, Lisa Rosen, Elizabeth Mertz, John Lucy, Susan Gal, Michael Silverstein, and Richard Taub. Many of the aforementioned have patiently seen me through a number of starts and stops in the trajectory of this project. This project has also grown through my involvement with several University institutions, chief among them the Semiotics and U.S. Locations workshops, and the Michicagoan conference. Being presented at these venues, my work was improved by lively discussion by all participants, and the commentary provided by Jessica R. Greenberg, Kate Graber, Susan Phillips, Rob Gelles, Karlyn Gorski, Colin Halverson, Briel Kobak, and Parysa Mostajir. Thanks as well to Jan English-Lueck for her comments during our 2017 AAA panel.

I am grateful to have been a part of, and to continue to be a part of, a wonderfully collegial department, the Department of Comparative Human Development, ably led by Jennifer Cole. The financial support offered by Susan and Lawrence Gianinno, in particular, has been a tremendous boon in surviving graduate student life, alleviating economic stress in times when other sources of stress have been abundant. I wrote this dissertation alongside fellow CHDers

Aron Marie and Sarah Cashdollar; and I am confident that without our weekly writing group, I would never have reached this point. I was able to work through and discuss many of the more metaphysical notions that were stirred up in the writing of this dissertation with what I have been calling the Pragmatism reading group, though I do not believe any official name was ever settled for this informal group of Peirce enthusiasts.

I expect these acknowledgements will suffer from recency bias, as this dissertation has been a long time in the making. I have also been sparing for fear of leaving people out of a more comprehensive list. I hope to be able to express my full appreciation for those who have been included and those who may have been omitted in our future meetings.

Throughout this long process I have been extremely lucky to have the support of friends and family who have embraced me for all of my faults—Rob, Carly, Sonia, Tristan, Haymarket, and beyond—and of course, Paula, without whom I may have never learned that I could be worthwhile.

Abstract

This dissertation is concerned with the "research-practice gap," the notion that the development of scientific knowledge in education has been relatively ineffective in addressing the practical problems of schooling. Research and Practice have been differentiated on the basis of actors, practices, expertise, organizational contexts, interests, and more; such are the apparent factors driving the Research-Practice Gap. Yet, many education researchers, policymakers, and practitioners are hard at work in attempts to bridge the two.

In imagining the bridging of the Research-Practice Gap as a precursor to the closure of the Achievement Gap, Research is figured as the external source of a pipeline of salutatory scientific interventions, which, provided appropriate translational work, can flow effectively into Practice. If proven effective, these interventions can be "scaled up" as a significant support of racial equality in the US.

An investigation of the Research-Practice relation brings up questions that are crucial not only to the operation of US education, but also the organization of social life in general, and the organization of a democratic society in particular. Can people meaningfully and effectively work together across difference? Can a nation professedly built on diversity be meaningfully United? This work is an investigation of communication and cooperation in the project of making American education scientific, itself a project of making America.

Central to this dual scientific and political project are trust and objectivity. Trust and objectivity are complementary modes of organizing social relations: Where a lack of trust threatens to sever lines of cooperation, objectivity enters to tie them back together. Historians of science have remarked that objectivity is distinctive in its "exclusion of judgment," as a "struggle against subjectivity" (Porter 1995, p.ix). Notably, the governance of American education has

largely turned to objectivity as a means of making decisions without a decider. As Porter puts it, "in science, as in political and administrative affairs, objectivity names a set of strategies for dealing with distance and distrust" (ibid.).

As a set of methods for negating individual subjectivity, objectivity becomes a means of rendering unity out of diversity, of identifying signals amidst noise. In this vein, objectivity is asked to not only produce education research as a diverse but unified *scientific* community, but also to produce US education as a diverse but *democratic* system of schooling.

I put forward three key distrusts which are managed in-and-through the 21st century American project of governing education through science. The first distrust is a bipartisan distrust of politics in education. This distrust of politics drives legislative appeals to science as a source of objectivity, as exemplified by the No Child Left Behind Act of 2001 (NCLB) in its recourse to and codification of "scientifically based research." It also sets the stage for the latter two distrusts, which are the focus of this dissertation.

In finding it necessary to define Scientifically Based Research, NCLB points to the second distrust: governmental and academic distrust of educational research. This line of distrust takes the Research-Practice Gap to be the fault of Research—insufficiently objective; incoherent; easily captured by the political, marketable, and fashionable. It is met by another line of distrust which holds the Research-Practice Gap to be the fault of Practice: researchers' distrust of practitioners' ability to faithfully implement the solutions offered by Research.

This work is concerned with describing education researchers' management of these two latter distrusts by way of two strategies of objectivity: reliability and fidelity. Through demonstrations of reliability, researchers constitute themselves as a scientific community, as Research, authorized to study and govern the work of Practice. In the name of fidelity,

researchers turn the work of Practice into "implementation," permitting the capture of variation against a Research-given standard, again capturing Practice within the jurisdiction of Research.

In an ethnographic account of the real-time activities of producing fidelity and reliability in a randomized controlled evaluation of an educational intervention, I use a linguistic anthropological analysis to describe the role of such interactions in maintaining the organization of the Research-Practice nexus, extending to the maintenance of its troublesome Gap. In concluding, I expand on the implications of my analysis in suggesting that a scientifically democratic system of education must be a democratically scientific system of education.

Introduction: Gaps, Objectivity, and Trust

The Achievement Gap

American education reform in the twenty-first century cannot be understood without reference to what is popularly described as the "achievement gap." The Achievement Gap refers to evaluations of wealthy White American students as more academically successful than their minority counterparts living in poverty. Usually invoked to point to the racialized nature of American inequality, the Achievement Gap has served as a persistent index of the failures of the American democratic project (Ladson-Billings 2006). As such, the closure of the Achievement Gap has become a powerful rationale for the funding and development of a wide range of educational interventions. The Achievement Gap is appealing precisely because it draws educational attainment, race, and economic class into alignment; suggesting that differences in education are a primary driver of racial and economic inequality in the US, and therefore racial and economic equality can be fostered by way of direct improvements to educational quality and equality.

Math4All,[1] the intervention of note in this project, was one such project, seeking to promote social equality by targeting the academic performance of children in low socioeconomic status (SES) minority communities for improvement. In doing so, Math4All was expected to improve these children's school readiness, allowing them to get the most of out their formal schooling experience, and ultimately set the stage for their upward social mobility. In this regard, Math4All followed the logic of "investing in people," or "building human capital," which

[1] I use pseudonyms throughout for all proper names, of people, organizations, and interventions. For more, see the section "Studying Up" on page 14.

informs a great deal of educational interventions seeking to address the Achievement Gap.[2] Under the logic of human capital, unevenly distributed education and training contributes to the uneven development of "skills"—cognitive and "non-cognitive"[3]—and it is this unequal distribution of skills that drives the uneven distribution of earnings in the US.

Math4All's theory of change was projected as follows: use of Math4All in low-SES communities will tend to improve teachers' instructional practices, which will tend to improve students' mathematical skills, towards being on par with those demonstrated by students in high-SES White communities. Greater mathematical skills will tend to improve school readiness, tending toward greater overall educational attainment, better employment opportunities, and ultimately an overall societal tendency toward racial equality. In fact, the "below target" and "on target" skill benchmarks within the Math4All intervention were set by taking the average scores of, respectively, low-SES minority (non-Asian)[4] students and high-SES White students from a cross-sectional survey of American youth. During my time with the project, the Math4All PIs repeatedly emphasized this operationalization of Achievement Gap as a major selling point of their intervention design, which emphasized the directness of Math4All's commitment to closing the Achievement Gap.

---

[2] The "human capital" concept was popularized by Becker [1964] 1993, with an educational focus from the beginning. Heckman (2006) is one influential articulation of its livelihood in the present-day. Both Becker and Heckman are Nobel prize–winning economists associated with the Chicago School of Economics. Their work has been hugely influential in American education research and policy in the 20th and 21st centuries.

[3] See Heckman (2006).

[4] I did not inquire after the rationale of excluding low-SES Asian-American children from the sample. Suffice to say that the Math4All research team, based on their analysis of the dataset, did not believe Asian-Americans living in poverty to be suffering from the same deficit of mathematical skills as other low-SES minority children.

The Math4All view of the Achievement Gap is, however, not universally shared. The Achievement Gap has been recognized by a variety of actors, from a range of interested points of view, as signifying various problems calling for similarly various solutions. In particular, critics of human capital framings of the Achievement Gap as a "skills gap" have argued that such framings (1) fail to recognize the stratifying function of education (cf. Bourdieu and Passeron [1977] 1990; Meyer 1977), (2) cannot explain empirical evidence against a direct relationship between educational and racial parity (cf. Darity and Mullens 2020), and (3) treat middle- to upper-class White practices as the neutral standard against which racialized practices are erased or marked as deficient (e.g., Avineri et al. 2015). Appendix A expands on each of these lines of critique.

The Math4All intervention is certainly susceptible to the above critiques, and in studying and analyzing Math4All, I am not endorsing the Achievement Gap framing which motivates its intervention, nor evaluating the extent to which Math4All is successful in its goal of closing the Achievement Gap. Rather, I take Math4All's direct operationalization of the Achievement Gap, and its human capital approach to closing the Gap, to make it an excellent representative of a highly valued and highly fundable mode of intervention in the present moment.

The Research-Practice Gap

How should we organize ourselves in working toward the closure of the Achievement Gap? The No Child Left Behind Act of 2001 (NCLB), federal legislation explicitly directed at the closure of the Achievement Gap, suggests strategies of "accountability, flexibility, and choice" (NCLB 2002, STAT. 1425). One of the central means by which NCLB attempts to effect this accountability is through a reliance on what it calls "scientifically based research" (ibid., STAT. 1964). It is this effort to address the Achievement Gap through the application of

objective science to educational problems with which this project is primarily concerned. More pointedly, this project is concerned with the widely held notion that the development of scientific knowledge in education has been relatively ineffective in addressing educational problems like the Achievement Gap, what is known as "the research-practice gap."

The leaders of Math4All—its primary investigators (PIs)—were academic researchers not unaware of the criticisms of their mode of intervention. Yet they persisted, because they truly believed in the value of their research with respect to the problem of the Research-Practice Gap. Because their research was based in objective practices of scientific knowledge production, Math4All was not only particularly likely to be effective in Practice contexts due to its basis in objective findings from cognitive psychology, but its effectiveness (or lack thereof) could be objectively measured through randomized controlled experimentation.

In line with historians of science who have identified objectivity on the basis of its suppression of subjectivity (Daston and Galison 2007), and in line with institutionalized articulations of scientific standards for education research, the Math4All PIs turned to objective research designs to produce knowledge stripped of personal and political biases, knowledge that was generalizable, de-contextualizable, lending itself to portability and scalability. Math4All, built upon objective facts of human cognition and rigorous statistical analysis of empirical data, would then not only be capable of traveling across the Research-Practice Gap, but, once on the other side, could be "scaled up" for nationwide use, uplifting low-SES minority students across the US. Math4All was a project that used the best scientific knowledge available to tackle the most important issues of our time.

This dissertation examines objectivity not as a principle of research design, but as a mode of social organization. What are the social consequences of the on-the-ground activities

necessary to enact objective designs? That is, in my work, I discuss the practices required to objectively evaluate Math4All's crossing of the Research-Practice Gap—its efficacy—and argue for their role in maintaining said Gap. I argue that the research designs required to objectively demonstrate the efficacy and effectiveness of interventions like Math4All, necessitate practices that constrain the usefulness of said interventions for teachers[5] by preventing them from being participants in the project of making education scientific.

In particular, I look at the concept of "fidelity of implementation" which underlies efforts to produce objective evidence of intervention effectiveness. I argue that fidelity requires teachers to become mechanistic producers of data for researchers' knowledge projects, potentially alienating teachers from scientific practice, and unquestionably limiting avenues for scientific inquiry in education.

Moving beyond critique, I continue by contrasting the situation of fidelity with that of reliability. By describing how researchers work with each other in calibrating the reliable use of each other's work, I offer reliability as an example of an alternative mode of organizing Research-Practice relations which does not blockade scientific inquiry. I argue that trust is a key concept in understanding what differentiates relations of reliability and fidelity. While both are concerned with the standardization of method, reliability is premised on relationships of trust, and fidelity on relationships of distrust. In concluding, I will suggest that placing trust in teachers

[5] I focus exclusively on teachers and not others who might fall under the umbrella of Practice in this work. This is due entirely to the substance of my fieldwork, which did not include any school administrators, district officials, and so on. I fully acknowledge the importance of these other actors to the Research-Practice relation and to the operation of education in general. I am simply not well-equipped to speak on their situations given the constraints of my fieldwork.

is necessary to move us not towards a "bridging" of the Research-Practice Gap, but a more salutatory "closure," such as been prescribed for the Achievement Gap.

In America, We Distrust

Distrust has long characterized the governance of educational and other public affairs in the US. Bipartisan distrust of federal involvement in education has long been a feature of the political landscape (Meens & Howe 2015); while distrust among and amid researchers, educators, administrators, and families has troubled on-the-ground reform activities (Bryk and Schneider 2002; Louis 2007; Carless 2009). All involved mediators are understood to be untrustworthy: politicians fail to faithfully represent their constituents; scientists fail to faithfully represent nature and/or the classroom; teachers and administrators fail to stay faithful to 'the science.' In this American imagination, human mediators have a tendency to stray from their proper representational function, drawn to infidelity, as it were, by their human fallibility or craven inhumanity, that is, their "bias," their "interests." The entrance of personal and partisan preference into governance (…the entrance of "politics" into politics?) is popularly recognized as highly undemocratic, violating American notions that only the only fair form of representation is faithful representation.

In particular, *mechanical objectivity* has been the favored mode of American political life, that is, objectivity via rule-following. Porter (1995) and other commentators[6] have observed a lack of trusting relationships and a high level of political exposure in American government in

[6] Porter (1995) cites Hugh Heclo's *A Government of Strangers*, and James Q. Wilson, who writes: "The United States relies on rules to control the exercise of official judgment to a greater extent than any other industrialized democracy" (as quoted on p.194).

particular, resulting in a characteristically American allegiance to rule-following as an index of impartiality:

> Congress [imposes] rules on every agency, dictating how to award contracts or hire and fire employees, as well as how to carry out its central mission. It sometimes even imposes such standards on itself. Cost-benefit analysis, for example, is a monument to the halfhearted desire of Congress to bind itself in red tape. As currently practiced, it is a distinctive achievement of American political culture. (p.195)

Outside of the legislative branch, the American judicial system also exercises of preference for explicit rules, and thus, prefers the testimony of those who deal with form (e.g. professional experts like academic researchers) over those who deal with content (e.g., teachers and other practitioners) (ibid.).

While Daston and Galison (2007) set the height of mechanical objectivity in the late 1800s, Porter (2008) observes a profusion of mechanical objectivity since the 1920s, continuing into the present day. He argues that Daston and Galison's focus on esoterically scientific developments leads them to miss the rich life which objectivity enjoys at the interface of science and public affairs, where procedures stand in for judgment in political decision-making. In the US, distrust of both politicians and researchers in education has created favorable conditions for mechanical strategies like fidelity and reliability to manage relationships of distrust in educational enterprises.

In the case of American educational governance, an additional obstacle to impartial decision-making is the situation of the purported experts. As Chapter One describes further, education researchers are themselves not free from accusations of allowing personal and partisan preference to enter into their work; drawing criticism from within and without of being insufficiently scientific. While educationists have been in a constant struggle against this line of distrust, their efforts to legitimize their work as scientific was catalyzed by the 2002 passage of

NCLB, which explicitly required education service providers to select programs based in Scientifically Based Research (SBR). This confluence of legislators' and researchers' political and professional interests in positing science as a privileged site of objectivity has conditioned the institutionalization of strategies of mechanical objectivity as not only the means of impersonal, bias-free decision-making, but also the means of setting boundaries on what counts as science, and who can be trusted to be an agent of science.

Though objectivity and science (and their respective histories) are often confounded,[7] I will not take them to be the same thing. Rather, I will treat objectivity as an ideology[8] of science (though we will see that it mediates much more than scientific practice). As a scientific ideology, objectivity describes a shared (i.e., social) set of beliefs about what science is, what it can and should do, who can and should do it, and so on. In this sense, objectivity is then also a *semiotic* ideology, a shared set of beliefs about what kinds of things can and cannot serve as *signs* of science.

In the American political landscape of distrust, demonstrations of objectivity—the production of signs of science—become the means for producing trust amidst distrust. However,

---

[7] For an extended discussion on the differences between objectivity and science, the history of objectivity and the history of science, and the history of objectivity and the history of epistemology, see Daston and Galison (2007), Chapter One, "Epistemologies of the Eye"

[8] By ideology, I mean a structure of belief which both enables and emerges from some situated collective interest, e.g., a scientific community. I do not mean the "false consciousness" of Marxist critique. Rather, I take all social action to be mediated by ideology, for better or worse. For instance, when Coburn and Talbert (2006) find that the beliefs that shape practitioner use of research are non-random, but *structured by institutional and organizational contexts*, they are describing the work of ideology. My use of the term "ideology" is also a claim of belonging to a community of linguistic anthropologists, situating myself as working within, and building on, a particular intellectual tradition. Schieffelin, Woolard, and Kroskrity (1998) documents formative discussions in the development of the ideology concept within this tradition, stemming from the foundational work of Michael Silverstein (1976, 1979, 1981, 1985) on the ideological mediation of social organization and discursive interaction.

in this case, the trust produced by objectivity is not an interpersonal trust between human actors, but a trust in collective processes. Education researchers are not individually objective, but are able to demonstrate their objectivity as a *scientific community*—as Research—via productions of reliability. Furthermore, trust in objectivity is always produced alongside a distrust of subjectivity. The need for implementation to be reined in by fidelity is as much an index of Research's objective trustworthiness as it is Practice's subjective untrustworthiness. Indeed, my analysis will not take objectivity to be a self-evident state of affairs but a professional and political claim made by education researchers *against* the claims of other groups, like teachers, over the jurisdiction of American education. I will describe how this claim of Research's objectivity has organized the Research-Practice nexus; and how the maintenance of that claim relies on the interactional achievement of, among other things, relationships of fidelity and reliability.

Math4All and the Research-Practice Nexus

Math4All was an educational intervention developed within the School of Education at Golden University, a private American research university, based in faculty research in psychometrics and cognitive psychology. It comprised a series of formative assessments[9] intended to improve math instruction and student outcomes. Teachers would use the Math4All materials to assess their individual students, input student results into the Math4All website, and

[9] In education circles, "formative" assessment stands in contrast to "summative" assessment. Summative assessments are designed to test students' learning at the end of some instructional period. The results of summative assessments might be kept as part of a student's academic record. In the current post-NCLB era, summative measures might further be used as metrics of "accountability" regarding any level of educational performance: teacher, school, district, state, national. Formative assessments are designed to test students' learning *in the midst of* some instructional period. Notably, the results of formative assessments are intended for instructor use only in informing future instructional strategies.

wait for the website's statistical backend to return information about the students' ability levels in various mathematical skills.

The Math4All research team presented Math4All's development as a formative assessment as an index of their alignment with teachers' interests. Math4All would deliver information about student abilities for the purposes of improving instruction, not for the purposes of institutional evaluation. Math4All was to be a helping hand in the classroom, delivering knowledge which teachers wanted and needed, but had neither the time nor expertise to develop on their own. Indeed, the Math4All PIs occasionally mentioned their latent concern that Math4All might eventually be taken up as a summative assessment and/or that it would be used to evaluate teacher performance, both highly undesirable outcomes. For the PIs, Math4All was about giving teachers critical knowledge about their students' needs and abilities. Any work that was being done in perfecting Math4All as an assessment, was ultimately in the service of providing teachers the best information as practically possible.

Math4All represented the translation of academic insights in psychometrics and cognitive psychology into an assessment instrument for teachers' classroom use. Though Math4All was a tool for teachers, however, it did not position teachers as agents of knowledge production. Rather, the Math4All website facilitated the exchange of teacher-produced data about student performance for researcher-produced knowledge about student abilities. Crucially, due to the assumptions of Math4All's psychometric statistical model, the meaningful translation of student data into valid knowledge about student abilities was fully dependent on the teachers using the Math4All assessment with fidelity.[10]

[10] Assessment administration is only one part of Math4All implementation, and it is the part I discuss in this dissertation. The other components must be left to later expansions of this work;

Math4All was not only imagined to be a vehicle for delivering knowledge about student cognitive abilities, but also a vehicle for the delivery of knowledge about young children's mathematical development in general. In-and-by growing familiar with the assessment, its tasks, and the website's visualizations and explanations, teachers were expected to grow familiar with the cognitive psychological findings about the range and developmental trajectory of various strands of mathematical thinking.

During my fieldwork, Math4All was undergoing an efficacy study, testing whether it would produce its desired outcomes under conditions conducive to high levels of fidelity of implementation. Several dozen classroom teachers had been recruited as participants in a randomized controlled trial, and the PIs, who had little to no experience working with teachers, hired a "professional development (PD) team," to train these teachers in faithful administration.

Half the PD team comprised individuals hired for their experience as former teachers and as professional development leaders at the Golden Mathematics Institute (henceforth "the Institute"), an educational research and development center affiliated with the Golden School of Education (GSE). Indeed, the logic of creating a PD team was to bring on individuals experienced not only in working *as* teachers, but also in working *with* teachers. The other half of the PD team came from the ranks of Math4All's research staff, including the project manager and several graduate research assistants.

When I first came to Math4All project, it was through an existing working relationship with several members of the Institute staff who came to be part of the Math4All PD team,

---

they include the interpretation the Math4All website's visualizations of student knowledge and the leveraging of such knowledge in the design of differentiated classroom instruction. These aspects of implementation were addressed in the second and third meetings of the teacher workshops, respectively.

Barbara, Rose and Anna. I had been working with these three on their own study of teacher use of curriculum materials—the Better Curriculums for Better Classrooms (BC2) project—until they were no longer able to secure research funding (I worked as a volunteer, but as Institute staff, they needed financial support in order to allot their worktime toward the project). It was through the characteristic generosity of these three, I was able to join them on the Math4All PD team in order to study how beliefs about what language is and can do shaped efforts to translate research for practice contexts. In exchange, I would continue to help them keep the BC2 project moving while unfunded, and assist in the more mundane elements of their Math4All work: creating teacher resources, making up graphics for sample lessons, charting alignments between Math4All and various common curricula, and creating a website to host such resources. Perhaps unsurprisingly, I regret to say that I quickly lost interest in the non-urgent work around the unfunded BC2 project, and became much more invested in, as it were, the flurry of activity around Math4All.

In following the PD team, I became acquainted with the three members from the "psych side," that is, from the GSE Department of Educational Psychology: Leah, research professional and Math4All project manager, and Ashley and Taylor, two graduate research assistants. Knowing these three led me to countless opportunities to become better acquainted with the Math4All project, as the psych side always seemed in need of extra support. I visited schools to conduct pre-testing, helped assemble the binders and materials that constituted the physical Math4All assessment, and eventually became part of the classroom observation team. I also secured permission to sit in on Math4All research team meetings, which gave me a better high-level understanding of the project and its goals.

The PD team itself had quite a lot on its plate. In addition to running a series of teacher workshops for three separate groups of teacher-participants and developing resources for teachers; the design of the efficacy study required them to follow-up with individual teachers over the course of two or more individual coaching sessions. As a result, I had the opportunity to observe and participate in workshop planning meetings, lesson writing meetings, manual writing meetings, teacher workshops, and individual coaching sessions.

The Research-Practice Nexus

I came to the Research-Practice nexus concept given a consideration of the enormous amount of activity around the Math4All study during my fieldwork. Initially, this flurry of activity led me to think of re-framing the Research-Practice Gap as the Research-Practice "crowd," to emphasize how many people worked within the figuratively empty "gap." However, I eventually realized that the crowd concept still emphasized a middle space between Research and Practice, while the activities enabling the Math4All study itself could not be contained by the space "between" Research and Practice.

Moving away from notions of 'betweenness' or liminality, the nexus concept emphasizes the empirical continuity of educational activity, in which researchers work with other researchers, researchers work with intermediaries, intermediaries work with teachers, etc. To describe intermediary/translational activities as taking place 'in the middle' fails to capture their continuity with the activities at the 'top' and 'bottom.' Activities in the Research-Practice nexus are linked not only with respect to their consequentiality (e.g., what happens in the course of iterative development can produce new imperatives for basic research), but also with respect to the actors and artifacts that participate in them. The Math4All PIs held meetings with the PD

team, who held workshops with the teachers, who were visited in their classrooms by the observation team, who were trained in observation by another research group, and so on.

A major affordance of the nexus concept's attention to the interactions which animate educational projects, is its ability to capture activities taking place 'at the top' on the Research end, allowing for the study of reliability in Research collaborations as a comparison case for fidelity in Research-Practice collaborations. I explore reliability in the context of the Math4All study in their collaboration with the Classroom Observation Group (COG), an external research group which trained Math4All researchers in classroom observation for the collection of outcome data. As part of the Math4All observation team, I was an active participant in both the training and conduct of classroom observations.

Studying Up

This research presents a case of "studying up," as described by anthropologist Laura Nader (1972). Studying up presents a departure from traditional conceptions of anthropological enterprise by focusing on the middle and upper echelons of societies which anthropologists themselves often call home. This presented some advantages for me, in that I hardly stood out in my field site, being a twentysomething Asian-American female graduate student (though the Asian women whom I encountered during my fieldwork were more often involved in the statistical side of educational research, and less likely to be involved in teacher-facing work). I had already spent the majority of my life in White educational contexts, and so had little trouble navigating or making sense of cultural customs and conventions.

The situation also presents some distinct disadvantages. As recounted by Diana Forsythe (1999), anthropologists who study up 'at home' work with subjects who are very capable of reading the work written about them, and very much in the position of influencing the career of

those who write critically about their practices. While it is difficult to not be concerned about the consequences of engaging in immanent critique before my "career" can be taken out of scare quotes, I am convinced of the Institute staff's sincere desire to reflect upon and improve their practice, and I am convinced of the Math4All PIs' commitment to free inquiry.

Studying up also means that some of my subjects have a heightened risk of identification. The exclusive and elite nature of academia, and its incentivization of self-promotion (Burkhardt and Schoenfeld 2003), means that academics have less anonymity than others might. As such, the Math4All PIs are perhaps most at risk of identification in this work. Ironically, they are the individuals whom I spent the least time with in my fieldwork, as they did not actively participate in the training interactions that most interested me. Accordingly, I discuss them very little in this work, and instead I offer throughout, in lieu of the PIs' voices, various prominent actors' articulations of the case for the objective regime of evidence production in education research. The Math4All case presents no meaningful departures from the logic and perspectives that such sources offer.

I would urge any and all readers to curb their desires to identify Math4All and the researchers associated with it. A lack of concern with personal identities would not only potentially foster an environment more open to instances of studying up, but also comport with the argument that I am trying to make in this work, which is that the Math4All case is not interesting in its particularities, but in its representativeness of highly valued form of research, which plays a central role in organizing the landscape of American education. Indeed, the generalizability of my analysis rests on Math4All being a remarkably standard project, and the interactions that I recount being exceedingly normal, in the most literal sense of the word.

Education is a field in which self-reflection is vaunted. This is especially true for teachers, who are always encouraged to be reflective and reflexive practitioners, to always be learning and improving their practice. I take this ethos to be the most laudable characteristic of educators, and I believe that researchers, including myself, should embody the same ethos. This is not an ethos of self-flagellation or blame, but of true accountability and improvement.

What do I mean by true accountability? In *Human Nature and Conduct* (1922), one of Dewey's central themes is a consideration of the distinction between means and ends. He addresses, for instance, the notion that 'the end justifies the means' is not merely ruthless, but ruthless due precisely to the fact that it selects only one set of consequences to deem 'the ends.' Indeed, the need to 'justify' the means suggests that there are several other, possibly 'intermediate,' consequences which were not considered as ends, and thus jettisoned from the evaluation of the morality of some course of action (pp.227–30). Dewey's recommendation then, in promoting our ability to more intelligently coordinate our actions, is to turn unflinchingly toward the examination of the full range of the consequences of our activities. It is in this spirit that I undertake my analysis of education research.

Dewey's call to broaden our view of consequentialities corresponds with his fundamental prerogative to expand the significance of the present (pp.265–9). That is to say, the greater our understanding of the multitude of consequentialities that any activity participates in, the more meaningful that activity itself becomes. For instance, if we have studied the consequentiality of children's experiences for their cognitive, socio-emotional, moral, linguistic, and physiological development, an actual child at play before us becomes much more meaningful, and potentially more demanding, than if our understanding was limited to just one dimension of development. Similarly, if education researchers are able to take a fuller view of the consequentialities of

reform activities beyond their effects on student outcomes, we can correspondingly expand the meaningfulness of those activities. With this more capacious insight, we may be able to more intelligently organize our collective efforts at improving the American system of education.

Fidelity and Reliability

Fidelity and reliability are both strategies of mechanical objectivity. Mechanical objectivity is defined by an adherence to rules of action, rules which act as "a check on subjectivity," and which, in their uniform application, are the basis for recognizing mechanical objectivity as "rigorous" (Porter 1995, p.5). Fidelity names uniformity between the procedures intended by researchers' in the design of their interventions and the procedures performed by teachers in their actual use of that intervention; while reliability names procedural uniformity among researchers in their practices of data collection and analysis.

Fidelity and reliability are similar in that they both describe a standardization of practice under the SBR regime of objectivity. In establishing reliability in classroom observations, researchers standardize their coding practices to enable the production of valid and commensurable data. In establishing fidelity of implementation, teachers are called on to standardize their administration practices toward the production of valid and commensurable data.

Yet despite the similarities between fidelity and reliability, the education literature rarely, if ever, describes teachers' engagement with interventions in terms of reliability; only fidelity. Reliability appears to be a thing that researchers do, not teachers. I aim to explain the sociological reality indexed by this uneven discursive distribution, with reference to the dynamics of trust which fidelity and reliability participate in given their roles in the political life of education research.

I take the terms "fidelity" and "reliability" directly from the educational literature, and I will be describing them as they have been institutionally and interactionally discussed and enacted within the context of contemporary US education. Fidelity and reliability are not simply two strategies of objectivity writ large, but a historically specific moment in the life of objectivity, as it has been institutionalized during the 2000s within the arena of education research.

Reliability and Scientifically Based Research

Figure 1: Defining "Scientifically Based Research." No Child Left Behind Act of 2001, Pub. L. 110-107, Section 803.37. Emphasis added.

A. research that involves the application of rigorous, systematic, and objective procedures to obtain **reliable and valid knowledge** relevant to education activities and programs; and
B. includes research that:
    i. employs systematic, empirical methods that draw on observation or experiment;
    ii. involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;
    iii. relies on measurements or observational methods that provide **reliable and valid data** across evaluators and observers, across multiple measurements and observations, and across studies by the same or different investigators;
    iv. is evaluated using experimental or quasi-experimental designs in which individuals, entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition controls;
    v. ensures that experimental studies are presented in sufficient detail and clarity to allow for replication or, at a minimum, offer the opportunity to build systematically on their findings; and
    vi. has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review.

The word "reliability" appears twice within the NCLB definition of Scientifically Based Research. In Figure 1, reliability appears as a direct modifier of both knowledge and data, but not as a direct qualifier of scientific practice or practitioners. However, the requirements of reliability for scientific practice surround its explicit invocations. Point iii., for instance, establishes that the reliability of data is associated with the quality of practices *across* evaluators and observers, *across* multiple measurements and observations, and *across* studies by the same or different investigators. That is to say, as a strategy of objectivity, reliability requires a social organization to enables a self-negating multiplication of subjectivity. The objectivity of data and knowledge *across* evaluators, *across* observations, *across* measurements, *across* studies is assured by the very biases that are assumed to operate at each point of evaluation, observation, measurement, and investigation. At each of these points, a group of actors (evaluators, observers, etc.) is expected to achieve procedural uniformity—reliability—ensuring that the data they produce will be independent of the biases of any single actor. If each of these internally reliable groups produces the same findings, then a very strong claim for objective knowledge can emerge, where objectivity is indexed by the quality of constancy across variation, giving objective knowledge a universal or general character. To say that reliability is a strategy of objectivity then, is to say that reliability is a means of organizing social actors and activities in a fashion which is understood to produce knowledge unmoored to any individual actor. In understanding how reliability is not just an epistemic principle, but a mode of social organization, one can begin to appreciate how objectivity as a semio-scientific ideology mediates much more than scientific practice.

A Collective Burden

Though Daston and Galison (2007) describe late 19th century mechanical objectivity as inhering in an ethic of individual self-discipline—"a will strong enough to bridle itself" (p.187)—the mechanical objectivity of SBR does not call for the mitigation subjectivity through *individual* practices of self-refusal. Though individual scientists are still expected to check their biases, they are assumed fallible by virtue of their humanity, with objectivity only emerging through the routine aggregation of those faults. Reliability, validity, replication, random assignment, control groups, peer review—these all name means of collective defusal over individual refusal at the levels of research design and evaluation.

Education researchers themselves recognize and articulate the centrality of collectivity to their scientific and professional enterprise. In a presentation prepared for a Department of Education conference on SBR following the passage of NCLB, Raudenbush (2002) asks himself a ventriloquized audience question, "Isn't it a little polyannish to expect researchers to police themselves in this way? After all, researchers are human beings with biases." He responds by holding up the role of "the scientific community."

> The burden of the objectivity does not fall entirely on the shoulders of the individual researcher. […] The methods of a study should be available to public scrutiny and data should be available for re-analysis. Findings should be subjected to rigorous peer review. And key conclusions emerge typically from convergent results over multiple studies conducted by multiple investigators whose personal viewpoints typically differ. A healthy scientific community is essential in examining the results from such streams of research. (p.6)

Not only is objectivity recognized as a collective burden for education researchers, but it is further recognized that objectivity has the potential to be the basis for the constitution of education researchers as a meaningful and effective collective, as a scientific community. In the wake of NCLB, the American Educational Research Association's flagship journal, *Educational*

Researcher, published a spotlight article describing the necessity of "consensus building" and "a strong, self-regulating culture" in response to NCLB's "unprecedented" call for the use of "scientific evidence as the key driver of educational policy and practice" (Feuer, Shavelson, and Towne 2002, p.9; p.4). In order to do so, "the multiple perspectives in education" must be unified under "a clear, commonly held understanding of how scientific claims are warranted" (p.9).

In this call, Feuer, Shavelson, and Towne articulate education researchers' self-conscious understanding of the centrality of consensus for the status of their profession (cf. Porter 2003, p.252). An inability to articulate a consensus would serve as an index of their inability to take up the position of leadership that NCLB put on offer. While the formal responses to the spotlight article evince education researchers' ambivalence toward the SBR paradigm, it has since become the *de facto* "clear, commonly held understanding" which defines the scientific community of education, at least from the point of view of the government patrons of education research.

While the AERA itself issued more or less methods-agnostic guidelines for "reporting on empirical social scientific research" (AERA 2006), major funding agencies like the US Department of Education's Institute for Education Sciences (IES) and the National Science Foundation (NSF) issued guidelines for educational research and development that made explicit the pride of place of experimental design.[11] Thus, despite a diversity of views among education researchers, SBR-style objectivity has become the institutional standard for what counts as scientific practice in education.

[11] IES and NSF (2013). See Chapter One section, "A 'Pipeline' of Evidence"

From Soft to Hard Science

The institutionalization of this mechanical model of objectivity is notable in how it leverages education research's alleged weakness as the "softest of the soft" sciences (Labaree 2010b, p.17). Teaching being a historically feminized profession, the academic study of education has been hard-pressed, as it were, to escape an association with the soft, feminine subjectivity of 'women's work' in order to achieve legitimization as scientific.[12] Through mechanical techniques of multiplication, SBR converts a threatening excess of subjective softness into a resource for objective hardness.

Random assignment (Figure 1, iv) exemplifies this statistical notion that, given a random sample, uncorrelated variations will defeat themselves upon aggregation.[13] Random assignment is also the defining feature of "experimental design," the acme of SBR. The randomized controlled trial (RCT) is experimental design in motion, and the oft-cited "gold standard" in education research. Tellingly, the random design described by SBR is not new to educational research, and in fact, its invention has been attributed to education researchers. Randomization is usually credited to R.A. Fisher's work in agricultural research, and contemporary education researchers tend to credit medical research's use of RCTs as their own model for action.[14] However, Dehue (1997) argues that random assignment was actually developed by educational psychologists "to support procedural objectivity in educational administration" (p.653).

---

[12] See Lagemann (2000) on the historical development of the separation between a male-dominated education science faculty and female-dominated teaching workforce. I also address this further in Chapter One. See Keller (1987) for more on the enduring opposition of science as objective, hard, and masculine against non-science as subjective, soft, and feminine.

[13] See the Chapter One section, "RCTs and Causal Inference." Note that, for this project, I am less interested in the logic of statistical reasoning so much as its institutionalization and its consequentiality for actual events of interaction.

[14] See Chapter One section, "The Classroom and the Clinic"

As Porter puts it, “Because applied and educational fields were such important sites for the introduction of statistical methods to psychology, it would be more nearly correct to say that statistics made its way up the hierarchy of prestige rather than down, from the applied and practical side of psychology to its more ‘basic’ and experimental forms” (2003). That is to say, SBR is not merely a symptom of some statistical hegemony making its inevitable appearance in educational research; rather, SBR is a powerful articulation of a statistical mode of knowledge production which was first developed within education research itself, for the very same purpose of enabling impersonal decision-making in educational governance. It is, in a sense, a homecoming.

As a form of objectivity continuous with a history of mechanical recourse in educational science, SBR takes method as king, and the RCT as king among methods. Note that going forward, I will use Scientifically Based Research or SBR to describe not only this 21st century American moment in the history of objectivity, but also the forms of social activity which mediate and are mediated by its institutionalization. Though SBR appeared 111 times in NCLB (Neuman in DOE 2002) and was frequently discussed in the immediate wake of NCLB, it is now rarely used in the educational literature. However, I will continue to use “SBR” as a convenient way of indexing the political/social/historical specificity of the productions of objectivity which this project describes. By productions of objectivity, I mean those activities which are licensed with reference to objectivity, and which simultaneously reproduce objectivity as an index of legitimacy in their enactment.

Fidelity and The Evidence Pipeline

Readers may have noticed that “fidelity of implementation”—what I refer to simply as “fidelity”—is not mentioned in the NCLB definition of SBR. This may seem even stranger given

NCLB's discussion of SBR was directed not at education researchers, but practitioners. NCLB defined SBR in order to direct education service providers to turn to SBR as the basis for their program selection and provision. While one might imagine educational service providers' use of SBR-based programs would be a prime site for a discussion of fidelity, fidelity is not mentioned in NCLB. I suggest two major reasons for this omission. First, even mentioning fidelity could potentially trouble the promise of a scientifically backed system of education. After all, the very necessity of measuring fidelity is the expectation of infidelity. And second, bringing up the issue of fidelity would demand a means of accounting for it.

Both education researchers and fidelity would eventually be directly addressed by federal legislation with the passing of the 2002 Education Science Reform Act, which established the Institute for Education Sciences (IES) within the US Department of Education (DOE).[15] The newly established IES continued the work of extending federal interest in science as a mode of apolitical politicking by articulating a vision of "evidence-based" decision-making which called for the use of Scientifically Based Research in evaluating the effectiveness of SBR-based interventions. Described as "a pipeline of evidence," the IES, in collaboration with the NSF, put forth a model of educational research and development which centered the objective evaluation of intervention effectiveness, i.e., "evidence," as the ultimate function of education research (IES and NSF 2013).[16] That is, in the Evidence Pipeline, the final steps in the development of an intervention are concerned principally with producing evidence of effectiveness.

---

[15] The IES replaced the Office of Educational Research and Improvement (OERI), itself having replaced the National Institute of Education (NIE) in 1980 as the federal agency governing education research. The "turbulent history" of NIE and OERI is one of the chief complaints in Kaestle (1993, p.29) in accounting for the "awful reputation" of education research.

[16] The Evidence Pipeline is addressed in more depth in Chapter One.

The Evidence Pipeline model fully aligns with NCLB's mission of strict accountability in education, arguably its most fundamental contribution to the American educational landscape (Schneider and Keesler 2007). A recursive application of SBR, the Evidence Pipeline continued an adherence to the principle of mechanical objectivity, heralding the entrance of fidelity. Fidelity insists that teachers follow the rules set forth by researchers in their design of interventions. In doing so, it renders all instances of intervention "use" evaluable as instances of "implementation." While "use" may be subject to any of myriad regimes of evaluation, "implementation" posits researchers' intentions as *the* evaluative standard for teachers' practice. Asserting the necessity of fidelity in aligning teacher practice with researcher intention is what holds together any intervention as a distinct, de/re-contextualizable (scalable!) intervention having properties as a result of its own design, being 'the same' intervention across its various contexts of use. That is, in a mechanistic fashion, *the* intervention is taken to be a set procedure whose uniform enactment across diverse contexts enables the statistical attribution of cause and effect. Even in cases of "adaptation"—fidelity's sparring partner—the effectiveness of interventions is typically still evaluated with respect to researchers' intended effects and purposes,[17] and so, for our purposes here, such a view of adaptation is merely a variation on the theme of fidelity.

In order to fully understand the role of both reliability and fidelity in the Evidence Pipeline, one must understand that while the pipeline produces evidence, no evidence actually flows through it. Rather, the Evidence Pipeline comprises two related pipelines, a science-to-

[17] Blakely et al. 1987, p.263; Century and Cassata 2016, p.200

school pipeline and a school-to-science pipeline. The former describes a model of knowledge distribution, while latter describes a model of data production.

Translation: From Science to School

The front-end of the Evidence Pipeline is what I call the "science-to-school" pipeline. Its model of knowledge distribution is the basis of this dissertation's title, *Talking Science to Schools*, itself a play on the 2007 National Research Council publication, *Taking Science to School*. My work is interested in the conceit implied by both turns of phrase: that science is exogenous to schools, and must be delivered to them. This figuration of a science-to-school pipeline, in which the origin of science is located outside of the school, has been described variously by education researchers as, to give but a few examples, "the knowledge hypothesis" (Kennedy 2005), "the linear model" (Coburn and Stein 2010), and "improvement by design" (Cohen et al. 2013). Each describes a common vision of education reform driven by externally developed scientific knowledge, applied to the context of the school. As psychologists discover how children learn to count, and as economists and sociologists discover what factors drive early drop-out rates, the resultant scientific knowledge can be translated into programs and policies of improved schooling.[18]

The science-to-school pipeline is concerned with the work of turning academic knowledge into instrumental knowledge. Academic knowledge can be the outcome of "basic research" or "applied research," that is, research which is conducted without a practical application in mind, or research which arises in order to address some purported problem. Academic knowledge typically circulates in written form in journal articles, books, funding

[18] Psychology, economics, and sociology are perhaps the most prominent disciplines in mainstream educational research.

proposals, white papers, and so on; and it is generally deemed to be of little use for practice,[19] having difficulty crossing the Research-Practice Gap.

In the pipeline model, academic knowledge can only come to flow (i.e., affect schooling) if it is translated into instrumental knowledge. Instrumental knowledge is academic knowledge developed for Practice in the form of "useful" or "usable" interventions.[20] Math4All is such a form of instrumental knowledge, in the development of which, for example, findings from cognitive psychology determined what Math4All would assess (mathematical development) and how it would be reckoned (what skills exist and what are the developmental benchmarks for each).

Note that I am not saying academic knowledge does not have instrumental uses in the broad sense.[21] Rather, I am differentiating the two categories in an artificial (ideological!) fashion as part of my unfolding argument about the organization of the Research-Practice nexus. I attempt to distinguish academic and instrumental knowledge based on the projected contexts of use which shape their (coming into) existence. Per my definition, academic knowledge is made by Research for use in Research contexts, that is, to be published, presented, cited, instrumentalized, or any such professional academic activity. By contrast, instrumental

[19] In my experience, virtually all academic discussions of the Research-Practice Gap dismiss out of hand the notion that teachers should be expected to read journal articles. See Models 1 and 2 in Burkhardt and Schoenfeld (2003, p.3-4). Researchers have done empirical work to find what types of research educators seek out or find useful in their actual operations. For instance, Kennedy (2005) discusses teachers' resource seeking behaviors, and Penuel et al. (2018) finds evidence that district leaders prefer books written by practice-engaged scholars over peer-reviewed impact studies in guiding their decision-making.

[20] Burkhardt and Schoenfeld (2003) describe this as a shift from a "science" to an "engineering" approach to education research (p.4).

[21] See Tseng (2012) for an insightful discussion of how education researchers have thought about and described various "uses" of research.

knowledge is made by Research specifically for use in Practice contexts, to be implemented by teachers, administrators, etc.

Like other models of translation, the science-to-school pipeline is built around the notion that the academic knowledge which informs the production of instrumental knowledge, is *still contained* within that instrumental knowledge; that the move from academic to instrumental is fundamentally a formal trans*form*ation, retaining the original *content* of the academic knowledge (cf. Gal 2015). Translation is possible under the SBR regime of objectivity because objective knowledge is understood by its adherents to be unattached to any subjective point of view. Objective knowledge exhibits a universality, portability, decontextualizability, which allows it to "travel" along pipelines. This translation however, can only be consummated as the delivery of academic knowledge to Practice, if and only if, its instrumental form is faithfully implemented.

All that being said, this work is not principally concerned with the activities taking place along the science-to-school pipeline—the production of psychological findings, the development of psychometric models, the development and testing of tasks. Rather, my project, and the Math4All efficacy study which I followed, begins at the transition from the science-to-school pipeline to its other half, the school-to-science pipeline, and continues down the length of the latter.

Recapture: From School to Science

The school-to-science pipeline is concerned with the work of transforming intervention use, *qua* implementation, into evidence. This is the situation in which both the cases of reliability and fidelity that I describe in this dissertation come into play. During my fieldwork, teachers' use

of Math4All became a source of data[22] for researchers' further production of instrumental and academic knowledge, in what I describe as processes of "recapture." Reliability in classroom observations allowed for teachers' Math4All use to be recaptured as evidence, a form of academic knowledge. Fidelity of implementation allowed for teachers' Math4All use to be recaptured as knowledge about student abilities, a form of instrumental knowledge.

The *instrumental recapture* of teachers' Math4All use describes the basic intended operation of Math4All as an intervention. In their faithful administration of Math4All, teachers would produce data about student performance, input that data into the Math4All website, and then—at least during this relatively early stage of development—someone on the Research side would run the latest statistical model, rendering the received student data into instrumental knowledge about students' math skills (henceforth "student knowledge"). As a form of instrumental knowledge, student knowledge was intended by researchers to be practically useful for teachers,[23] and it manifested as a part of Math4All itself, appearing on the Math4All website in the form of various graphic visualizations—icons and boxes marking out different points and portions of fields of gradient color.

The *academic recapture* of teachers' Math4All use describes the basic intended operation of the Math4All efficacy study in the production of evidence, itself a type of academic

[22] Data is an $n - 1$ order of knowledge which is not yet reached the status of academic or instrumental knowledge as I define them. That is, data is anything which is treated as having the potential to become knowledge. If the production of data is mediated by (textual) practices of order n which break up the world into meaningful pieces, then the production of knowledge is mediated by (textual) practices of order $n + 1$ which find meaningful patterns in that data (cf. Silverstein 2003). I describe data as being of order $n - 1$ because data is often conceived of in respect of its mediation of future knowledge, as the 'before' stage of a 'before-and-after' transformation.

[23] In this case, knowledge about individual student abilities was expected to be useful in helping teachers plan differentiated, targeted instruction.

knowledge.[24] Use of Math4All was the treatment condition of its RCT evaluation: teachers' use of Math4All was hypothesized to drive changes in instructional practices and student outcomes which would not be observed in classrooms led by teachers not using Math4All. In order to produce evidence of efficacy, researchers again depended upon teachers' work to generate data; in this case, data was not generated in the course of teachers' work in using Math4All, but in their primary role of leading classroom instruction, *under the condition of* having used Math4All or not. Unlike the case of instrumental recapture, teachers did not do the work of data collection in academic recapture, they only participated in generating the possibility of data collection. Instead, researchers collected both student and teacher data, through reliable practices in pre- and post-testing of students and classroom observation.

I use the term "recapture" to emphasize that, at the point of use, instrumental knowledge becomes a potential site of distributary knowledge flows.[25] For instance, as Chapter Three describes, teachers may develop novel uses for Math4All as a technology for creating their own knowledge, in ways not circumscribed or prescribed by the psychometric design of Math4All nor the RCT design of its efficacy study. However, the logic of fidelity recaptures use as implementation in order to keep data flowing from schools to science, rather than allowing it to wend away from the academy, and feed into, for example, schooling-internal professional learning communities. In directing teacher-sourced data back into Research, processes of

[24] Evidence often manifests as "peer-reviewed impact studies," and some educational researchers have called into question the role of evidence in decision-making processes, in line with the aforementioned skepticism over the mobility of academic knowledge. See earlier footnote on Penuel et al (2018); also Coburn et al (2009).

[25] A distributary is the opposite of tributary. A distributary is formed by the splitting of a river, creating a separate path from a formerly unitary stream; while a tributary is initially disparate and flows into another stream, joining into a unitary stream.

recapture renew Research's place as the proper—and exclusive—site of educational knowledge production.

Studying Fidelity and Reliability as Relationships of (Dis)Trust

Fidelity is notable as a mode of social coordination that does not require trust. It is fundamentally premised on a distrust of teachers' ability to realize the carefully calibrated designs of researchers' interventions. Indeed, recourse to fidelity arises in a historical context of competition between researchers and teachers for professional jurisdiction over education, expanded on further in Chapter One. In this history, researchers' objectivity is repeatedly set against practitioners' subjectivity; researchers' trustworthiness legible against teachers' untrustworthiness. The fidelity relationship is an operationalization of this trustworthy/untrustworthy, objective/subjective distinction toward the licensure of researchers' authority over teachers' work. That is, my analysis of fidelity will describe how a reliance on "rules, calculation, and fact-finding" is not simply characteristic of bureaucracy, but, as Porter (1995) puts it, "a defense against meddlesome outsiders and a strategy for controlling far-flung or untrustworthy subordinates" (p.194).

On the other hand, relationships of reliability do require trust, which is notable within the distrusting climate of education research in which researchers operate. I do not claim that reliability turns distrust into trust. Rather, I hope to show how demonstrations of reliability leverage existing trusting relations among researchers themselves, toward the outward-facing production of Research as an objective, and therefore trustworthy, scientific community. As noted earlier, that education researchers establish a reputation of scientific objectivity, indexed by their own collective coherence (reliability!), is key to authorizing their participation in shaping educational practice under the NCLB/SBR regime. In my work then, reliability will

serve as a site for the investigation of education researchers' professional self-constitution as a trustworthy scientific community, and, ultimately, as a potential alternative model for organizing the Research-Practice nexus as a whole.

Theorizing Trust

What do I mean by trust? My interest in trust is in line with a vast array of sociological literature which takes trust to be an essential component of the necessarily cooperative practices of human social life. I am indebted to Sztompka (1999) and Shapin (1994) for providing extensive reviews of what they each consider to be major contributions to the scholarly discussion of trust. Indeed, the two are useful in providing characterizations of two prominent, differing stances on the utility of trust as a social scientific analytic. I will describe both before offering up my own theorization of trust.

Rational Choice Theory

Rational choice theorizations of trust turn principally on predictability as a condition of calculability. In his expansive review of sociological scholarship on trust, Sztompka chooses to follow James Coleman's rational choice theorization of trust, and defines trust as "a bet about the future contingent actions of others" (1999, p.25; cf. Coleman 1990). This gambling frame makes trust the result of rational decision-making on the basis of situated cost-benefit calculations. Within this choice model, the interests of the trusting party, as a rational individual, are central in establishing the basis for formulating costs and benefits.

This view of trust treats the conditions of trust as its problematic, with researchers in this school tending toward the analysis of what conditions and factors make individuals more or less

likely to trust.[26] Coleman (1990), for instance, suggests that the closure of a social network—that is, the interconnectedness of its constituent individuals and the frequency of their interactions—is equivalent to its "trustworthiness." Increased closure produces an increasing number of sites of trust-enforcement, positively influencing trust calculations. For instance, under this model, a teacher is more likely to trust a student with a take-home test, if he knows that the student is likely to be found out and admonished by their caregivers if they do not adhere to the test protocol. Both the number of caregivers and the frequency of caregiver interactions strengthen the likelihood of breaches of trust being accounted for, incentivizing the student to act in a trustworthy manner.

Moral Community Theory

Shapin's moral community theorization of trust, on the other hand, gives the conditions of trust in its definition. Trust is defined as the moral relationship between members of a community. The very existence of communities implicates trust, because the bonds that hold the community together are what we call relationships of trust. In this case, one need not decide whether to trust; instead one must determine mutual belonging in some community. Trust is then the result of social practices of belonging and identity, rather than the individual calculus of gains and losses. To be clear, I do not use identity in the sense of individual personality or character, but as indicative of one's group affiliation(s), e.g., you are identified as a teacher because you are recognized as belonging to the group of people known as teachers.

The relevant question then becomes, how do people establish that they are part of the same community as one another, that they are the same type of people? For Shapin,

[26] See Chapter 4 of Sztompka (1999)

communicative practices are central to the processes of identification as sites for the enactment of social types. In my research, I will similarly focus on communicative activity as the site of claims of mutual belonging and trustworthiness in the case of reliability, and claims of difference and untrustworthiness in the case of fidelity.

Distrustful Cooperation

As previously noted, the fidelity case is particularly interesting as an instance of cooperation across distrust. Theorizing distrustful cooperation presents a departure from Shapin's ethnomethodological orientation which tends to read trust in all practically cooperative activity.[27] As he puts it: "cooperation presupposes a moral bond [=trust]" (1994, p.23). Rather than taking cooperation itself as an index of trust, my work describes how groups deal with the necessity of cooperative activity with untrustworthy others. I argue that education researchers have, in-and-by productions of objectivity, been able to override the professional claims of teachers by incorporating them into a division of labor which renders teachers as untrustworthy subjects in need of governance by a more trustworthy group.

This relationship is much like the one Shapin describes between Robert Boyle and his technicians, in which Boyle's authority is literally manifest in the single authorial voice he used to report the results of the collective efforts of himself and his technical staff (1995, pp.372–407). While this relationship was based in the established norms of master-servant relationships, such arrangements are not looked upon kindly in the contemporary US context, necessitating fidelity as an objective (read: apolitical) strategy of social organization. In analyzing fidelity's logic and its practical achievement, I will argue that the researcher-teacher relationship is much

[27] Shapin 1994, Chapter 1; cf. Garfinkel 1963

closer to the scientist-technician relationship than the producer-user relationship which is often described. Where Shapin describes the conditions which widely legitimized 17th century gentlemen as the natural spokespeople of science, my work describes how, absent an explicit aristocracy, within a putative democracy, contestation over scientific authority sees Practice subsumed under the jurisdiction of Research within a regime of objectivity.

My analysis presents a situation in which distrustful cooperation is a condition and consequence of education researchers' claims of authority over teachers' work. Further, I argue that the activities necessary to facilitate that distrustful cooperation drive the persistence of a Research-Practice Gap.

Defining Trust

How should one identify trust, if not via cooperation?

Trust arises in the condition of being blind to another party's action, whether that blindness results from the unknowability of future action; or the mere unobservability or unverifiability of past and present actions.[28] Trust is a belief on the part of one party about what another party has done, is doing, or will do, given a fundamental uncertainty with respect to others' actions. That is, trust arises in response to problems of alterity and interdependence. We are uncertain as to others' actions, but being affected by them, we must, in order to go about our own lives, have some kind of assurance about the nature of those actions.[29] Trust highlights both a lack of control and an inability to fully know another. It is not just about what another party has

[28] Sztompka (1999) posits the future orientation of trust, but it is not clear to me why trust should not apply to any unobserved situation, regardless of its temporality. Unable to verify or control, one must trust.

[29] Sztompka (1999) elaborates on these points with special reference to the work of Nicholas Luhmann, Adam Seligman, Anthony Giddens, and Jack Barbalet. See, in particular, pp. 21–26.

done, is doing, will do—but what they have done, are doing, would do, *left to their own devices*. That is, where objectivity disavows discretion, trust marks its entrance.

This emphasis on effective autonomy marks a departure from rational choice theory which allows for all sorts of "trust-enforcement" mechanisms, described in terms of "accountability," to enter into the calculation of trust decisions. One may, under rational choice theory, trust the babysitter, given the installation of a nanny cam. By contrast, under my model, the nanny cam is a clear index of distrust.

Further, I do not take predictability as satisfactory for a definition of trust, again departing from the rational choice theory which treats trust as a bet about others' future actions. Instead, I hold that trust must have a moral dimension in distinguishing between better or worse courses of action. In introducing this moral dimension, we can come to a stronger definition of trust as a belief on the part of one party that a second party, left to their own devices, would take *right actions*. In taking this tack in defining trust, we are committing ourselves to the nonsensical nature of statements like, "I trust you to betray me."

Finally, my model of trust recognizes the asymmetry of the trust relation in the determination of morality. That is to say, when A trusts B, A trusts that, if left alone, B will act in accordance with A's own values. This is a departure from the ethnomethodological treatment of trust which implicates always already existing norms in the maintenance of social coherence, the existence of which can be determined, for example, by breaching experiments (Garfinkel 1963). Instead, I emphasize the asymmetry of the trust relation in the determination of norms, and the very contestability of norms themselves, in order to understand the role of trust in political processes of social differentiation. Crucially, I suggest that processes of social

differentiation may take place through the selection of norms which render others untrustworthy, such that they may become governable.

With this working conceptualization of trust, we can pragmatically approach the problem of identifying trust by empirically investigating the degree of freedom of any party in relation to another with respect to some activity. To what extent does A leave B to their own devices? Trust is asserted in activities of allowing or limiting autonomy, as the allowance of autonomy is taken to index that whatever actions B might take, A believes that they need not intervene, as B's actions will be morally right from A's point of view. In describing the relationships structuring the Research-Practice nexus, this evaluation of practical autonomy will be the basis for describing reliability as a relationship of trust, in which cooperation ends with a parting of ways, and fidelity as a relationship of distrust, in which cooperation forever tethers one party to the interests of another.

In order to account for the very asymmetries that entail the inability of one party to "allow" or "disallow" activities of another, evidence of trust may also be approached by analyzing how parties behave in situations where they have the opportunity to identify or dis-identify themselves with the other. The fundamental assumption here is that a willingness to tie one's identity and reputation to another bespeaks a belief that their actions will be morally good or redeemable, from one's own point of view. Here, too, the fidelity concept bespeaks a lack of trust in that it fundamentally problematizes the identification of teachers' efforts with researchers' designs. Meanwhile, in the reliability case, professional citational practices require the COG and Math4All groups to be willing to be identified with one another in the case of publication of any observational findings.

Regimenting Textual Practice

My description of the real-time achievement of fidelity and reliability, and their role in facilitating the flow of knowledge and data, will center around the analysis of textual practices. I posit that the teachers' and researchers' textual practices are key vectors of force which, in their alignment, enable currents of knowledge and data flow so directed as to give the impression of being contained by a "pipeline."

Textual practices concern ways of consequentially differentiating the stuff of experience into meaningful segments, that is, texts.[30] Textual practices can be thought of analogically to literacy practices, like reading and writing, but they are unrestricted to encounters with linguistic signs. However, given the centrality of language use as a medium for social activity, textual practices do often end up literally(!) taking place in-and-through practices of reading and writing. For example, Chapter Three describes how the COG team helps the Math4All team become reliable by training them in a specialized textual practice: differentiating the behaviors they witness in classrooms as "events of instruction." This practice is enabled by a network of related textual practices which were discussed at length during our training, including how to meaningfully and effectively[31] write an "event log" based on observed activities, and how to read an event log in order to code it correctly. The ability to consistently and uniformly enact these textual practices underlies the ability of the Math4All team to achieve reliability, and thus their ability to turn classroom phenomena into statistically valid quantitative data.

---

[30] "Entextualization" is the linguistic anthropological term used to describe this process of turning experience into text. See Silverstein and Urban (1996).

[31] I will continue to use the terms "meaning" and "effect" in conjunction throughout this work, but I want to emphasize their interchangeability. From a pragmatic point of view, effects are the basis for meaning, so the meaning of anything is equivalent to its effectiveness. In this use, effectiveness refers only to consequentiality, unbound by intention.

The centrality of textual practices to knowledge travel can be observed in the easy movement of numbers, conditioned by the widespread uniformity of textual practices around numbers, that is, the existence of a "widely shared" and "highly disciplined" (meta-)discourse of mathematics:

> Since the rules for collecting and manipulating numbers are widely shared, they can easily be transported across oceans and continents and used to coordinate activities or settle disputes. Perhaps most crucially, reliance on numbers and quantitative manipulation minimizes the need for intimate knowledge and personal trust. Quantification is well suited for communication that goes beyond the boundaries of locality and community. A highly disciplined discourse helps to produce knowledge independent of the particular people who make it. (Porter 1995, p.ix)

Perhaps curiously, this work about SBR will feature very little discussion of numbers, despite critics' contentions that SBR is effectively synonymous with quantitative methodology (e.g., St. Pierre 2006). Rather, I will discuss the conditions which enable the production and travel of quantitative knowledge, conditions exceeding the formal logic of mathematics or statistics. I argue that the production of statistically meaningful student data and outcome measures in the Math4All study are the result of the *discursive regimentation of textual practices*, that is, talk about how to see the same things in the same way.

Education researchers have described the structured contingency of textual practices in research use and curriculum enactment with analytics like "sensemaking" or "modes of engagement" (Coburn and Talbert 2006; Remillard 2011). Such "use research" can be contrasted with "fidelity research" which tends to be concerned with intended models, measuring fidelity to those models, and associating fidelity with outcome measures (Century and Cassata 2006). Use research tends to be more interested in describing the organizational and interactional processes by which some external intervention comes to have actual effects, that is, to develop a meaning-

in-context.[32] As such, they have developed not only taxonomies of different types of use (Tseng 2012), but different lexical means of emphasizing the *process* by which uptake occurs: sensemaking, engagement, transaction, appropriation, creep and accretion.[33] Each of these theorizations rejects rational actor models of decision-making as a discrete event driven by the dispassionate analysis of evidence, and emphasizes the ways in which policymakers' and practitioners' biographical trajectories, institutional contexts, and organizational roles structure their textual practices.

As a result of their investigations, some use researchers have called for greater attention to the regimentation of textual practice in promoting more effective use of research, that is, a focus on "teaching" rather than "telling" (Honig 2012, Honig et al. 2014), or "talking to" rather than "talking through" teachers (Remillard 1999, p.328). This project is about such practices of teaching and talking to. Furthermore, by taking on the unique comparison case of researchers teaching other researchers, I describe contrasting divisions of textual labor across fidelity and reliability.

Actors and Artifacts

Textual practices, like reading and writing, involve creative encounters with artifacts. Like literacy practices—a subset of textual practices dealing with language-in-use—textual practices are not strongly determined by the artifacts which participate in them. This

[32] Examples of use research in Practice contexts: Honig and Coburn 2008; Coburn et al. 2009, Honig et al 2014, Kennedy 2005. In Policy contexts: Dumont 2019; Finnigan and Daly 2014; Weiss 1980, 1995; Yanovitzy and Weber 2020.

[33] Coburn and Talbert (2006) take sensemaking from Karl Weick; Remillard (2011) draws on Elizabeth Ellsworth in theorizing modes of engagement, and Louise Rosenblatt in describing teacher transactions with curriculum materials; Honig et al. (2014) describes different levels of appropriation; and Weiss (1980) describes "knowledge creep and decision accretion."

indeterminacy of uptake has been recognized by both science studies researchers concerned with the role of technology in social life (e.g., Oudshoorn and Pinch 2003) and education researchers focused on the role of artifacts like curriculum materials in educational improvement (e.g., Remillard 2011).

In my work I will use the term "artifacts" to describe entities that are taken up in the production of texts (meaningful apportionings of experience) by way of textual practices. For example, teachers' engagement with printed curriculum materials (text-artifacts) involves various textual practices which result in the enacted curriculum (text). As Ball and Cohen (1996), among others, note, these textual practices are mediated by teacher beliefs (p.7). That is, generally speaking, textual practices are a form of ideological work, being mediated by structured beliefs arising from collective interests, in this case, teachers' interests. Indeed, Ball and Cohen give examples of how teacher beliefs about textbook use have been structured by their collective interest in professional autonomy, and how the figuration of autonomy has been constructed in opposition to textbooks, themselves figured as instruments of control, given historical contests between teachers and those who have attempted to control teachers' work (p.6).

I will also use "artifact" to refer to the material precipitates of textual practices themselves. For example, textual practices of curriculum enactment may involve the production of further artifacts like lesson plans. Less obviously material productions of curriculum enactment—such as teacher speech, bodily movement, and so—may also be taken up as artifacts in other actors' (e.g., administrators, researchers) textual practices of classroom observation. That is to say, what counts as a text or text-artifact cannot be determined outside of the analyst's textual practice, itself a form of ideological work which privileges a particular point of view

which abductively[34] selects some slice of reality as meaningful. (Statisticians' texts, for example, are called "constructs.") All texts are materially mediated, and all artifacts may be the product of preceding or proceeding textual practices, or the object of potential future textual practice.

I will use the term "actors" to describe things that take up artifacts in the production of texts. So, a textbook may be an artifact, but an actor is required to take up the textbook in the production of a text, e.g., a lesson plan or instructional event. As Chapter Five points out, speech is also an artifact, if a rather ephemeral one. Textual practices around speech include its evaluation as instantiating a type of speech (you always seem to have *a counterpoint*), indexing a type of speaker (were you *a debater* in high school?), or enacting a type of interaction (I wish you would stop treating everything like *a debate*). Each of these textual practices, again, entail the differentiation of empirically continuous sonic vibrations in the air with reference not only to some grammatical code (e.g., English), but also cultural models [=stereotypes] of speech and speakers (how "debaters" act).

Just as with text-artifacts and texts, the difference between an artifact and an actor is their positioning within a particular, consequential event of textual practice. In such a relationship, the actor, and not the artifact, is engaging in a purposive, agentive activity by way of the artifact, in a way that the artifact is not. This does not mean that the material properties of the artifact do not contribute to the quality or contingency of the practice (cf. Keane 2003), only that the event would not have occurred without the actor, and would not have occurred in the same way given an actor with a different purpose.

[34] By this, I mean that the "selection" of some slice of reality is really a working hypothesis that such a slice exists, rather than the self-evident identification of some existing thing (cf. Peirce [1898] 1992).

A review of the education literature on the design of educational materials and the STS literature on the design of other technologies evince an aspiration toward, or description of, efforts to "blackbox the user" (Pinch 1993) – that is, to use artifactual design to contain the range of actors' possible purposes and practices. While initial scholarship on the social construction of technology framed users as a crucial component in closing off the "interpretive flexibility" of a technology soon after its introduction (Bijker et al. 2005), Actor Network theorists attempted to recoup artifacts' purported agency in proposed activities like "configuring the user" (Woolgar 1991), "scripting" (Akrich 1992), or "prescription" (Latour as Johnson 1988). In such theorizations, humans can "delegate" actions to non-human actors by designing them to lend themselves to particular kinds of access and use (ibid.).

In following the achievement of fidelity and reliability, I have found that the "materiality and infrastructural properties"[35] of artifacts like the COG observation app and the Math4All assessment binder and score sheets contribute very little to said achievements. Rather, fidelity and reliability have proven to be products primarily of the discursive interactions in-and-through which actors are entrained in particular value-laden ways of interacting with particular artifacts.

Communities of Textual Practice

Building on linguistic anthropological scholarship, my analysis will treat the regimentation of textual practice in the production of fidelity and reliability as a site for understanding the real-time interactional organization of the Research-Practice nexus. In such an analysis, I posit a dialectical relationship between textual practice and social organization, such

[35] Star 2010, p.613

that an actor's social location conditions their textual practices,[36] while textual practices participate in the positioning of actors in relationship to one another as particular social types within some social/moral order.[37]

A fundamental premise of this approach is that textual practices are fundamentally collective practices, practices which condition or are conditioned by membership in some community, i.e., are ideologically mediated. For example, Elizabeth Mertz (2007) has described the ways in which learning to "think like a lawyer" in law school takes place through learning to read, write, and speak like a lawyer. That is to say, Mertz shows that what legal professionals describe as "thinking like a lawyer" is made up of an array of textual practices which are themselves mediated by a linguistic ideology which takes legal texts to be neutral sources of authority, though they highlight particular textual features over others. In doing so, the enactment of these textual practices does the ideological work of erasing "content, morality, and social context," and privileging "form, authority, and legal-linguistic context" (p.4).

A second, related, premise of this approach is that the discursive regimentation of textual practices is the means by which a social group constitutes itself as a group. As Goodwin (1994) puts it:

> Discursive practices are used by members of a profession to shape events in the domains subject to their professional scrutiny. The shaping process creates the objects of knowledge that become the insignia of a profession's craft: the theories, artifacts, and bodies of expertise that distinguish it from other professions. (p.606)

---

[36] In education research, see for instance, Cynthia Coburn's discussions of sensemaking practices as being conditioned by actors' institutional and organizational contexts, e.g., Coburn and Talbert 2006. In cultural analysis, Stuart Hall (1973) presents a classic model of how social location conditions meaning-making activities.

[37] In science and technology studies (STS), work on "co-production" which recognizes knowledge practices as practices of social organization is perhaps the most related body of work. See Jasanoff (2010) which briefly surveys and evaluates the STS literature with reference to the idiom of co-production.

Recall that textual practices act to differentiate the continuity of experience into meaningful and discrete texts, what I read Goodwin describing as "objects of knowledge." Goodwin's point here, which my approach agrees with, is that the distinctiveness of these objects of knowledge, and the textual practices which produce them, is the basis for the distinctiveness of the social group (e.g., a profession) that concerns itself with said objects of knowledge. *We* are concerned with this (the science of education), *they* are concerned with something else (a political agenda). So, not only can individuals be identified as belonging in some group given their textual practices, but those textual practices are instrumental in creating the objects of knowledge—the texts—by which a group becomes distinguishable as a group, in their production and treatment of group-defining matters in group-defining ways. Again, it is then appropriate to describe textual practices as forms of ideological work, linking up collective interests and collective organization.

A final premise of my analysis is that, as foregrounded in the Goodwin quote, an empirical investigation of textual practices finds its natural site in the analysis of discursive practices, and to put a finer point on it, discursive *interactions*. Even as linguistic anthropological analysis has expanded to non-linguistic modes of semiosis, discourse remains a key site for both the doing and analyzing of social life, analysis being but one mode of doing (cf. Keane 2003, 2018). As my empirical data corroborate, textual practices tend to take place through practices of language use like talking and writing. They are also regimented through the very same practices, e.g., talking about how to write, writing about how to talk. In this view, the primary function of communicative practices is not reference (describing the world or representing one's internal state) but *social organization*: defining, differentiating, and aligning social groups (their objects of knowledge, their textual practices, their members and non-members) with respect to one another.

For example, as described in Chapter Three, using the COG observation system involves the ability to perform several specialized textual practices like *reading*, *writing*, and *coding* "event logs" (written descriptions of classroom phenomena). At the same time, training involved a number of pedagogical discursive practices, many of which paralleled the practices which they hoped to support: the COG team *wrote* "practice scenarios" which the Math4All team *read* and *coded*; and most importantly, the two teams spent a great deal of time *talking about* how to code, how to write, and how to read practice scenarios. As our textual practices on the MAth4All team became more and more like the textual practices of the COG team, we were able to reach agreements about coding certain scenarios which had previously presented seemingly impassable obstacles to reliability (Chapter Three). Further, the Math4All team's ability to "get reliable" with the COG team (to align our textual practices) indexed our mutual belonging to Research, a properly *scientific* community which takes classroom activities as one of its objects of knowledge.

This discussion may bring to mind Lave and Wenger's "community of practice" theory of learning (1991). Indeed, the reliability case of the COG and Math4All research groups conforms quite well with the notion of learning as participation in a community of practice (ibid.). In this formulation, learning is a process by which an individual is able to meaningfully experience the world only by virtue of be(com)ing a member of some community, and engaging in the meaning-making practices which define it as a community. In this case, the Math4All teams' engagement in the training was a form a "legitimate peripheral participation" which, should they meet the standard of reliability, would lead to their "full participation" as classroom observers, able to independently collect observational data (ibid.).

What I have to offer devotees of "situated learning" in my comparative analysis of two cases of legitimate peripheral participation, or training, is an analysis of *trajectories of participation.* Lave and Wenger (1991) themselves call for such an analysis in their initial discussion of the useful ambiguities of the peripherality concept. They describe "legitimate peripherality" as "a complex notion, implicated in social structures involving relations of power," a potentially "empowering" or "disempowering position," "at the articulation of related communities" (p.36). They go on to write:

> [P]eripherality, when it is enabled, suggests an opening, a way of gaining access to sources for understanding through growing involvement. The ambiguity inherent in peripheral participation must then be connected to issues of legitimacy, of the social organization of and control over resources, if it is to gain its full analytical potential. (p.37)

Lave and Wenger are here arguing that the ambiguity of peripheral participation is precisely what makes it an excellent site for the analysis of issues of legitimacy, social organization, and resource control. That is, one should be able to describe a landscape of power relations by asking the open empirical question: "where does peripheral participation lead?"

What my work will show is that full participation under reliability looks very different from full participation under fidelity. While the Math4All team's full participation in COG-style classroom observation (as marked by formal calculations of reliability) saw our increasing independence from the group who had trained us, and an ability to produce our own knowledge according to our own agenda; the full participation of teachers in Math4All at the hypothetical point of complete fidelity would always remain tied to researchers' projects of knowledge production. Fidelity complements reliability's claim of objectivity by holding up the subjectivity of teachers as licensing researchers' authority over their work. Where reliability authorizes knowledge, fidelity authorizes power on the basis of that knowledge.

Outline

This work comprises two parts. Part I concerns fidelity, and Part II reliability, each part comprising two chapters. Preceding both parts is Chapter One, which sets the scene for all that follows. Entitled "Contests of Objectivity," Chapter One tells two entangled stories about objectivity as a claim of scientific and political legitimacy. The first story concerns the efforts of research universities and their faculties to claim educational authority, and the second story concerns the efforts of federal legislators to govern education in the course of nation-building projects. SBR and the Evidence Pipeline are presented as emerging at the convergence of these two stories.

Part I, "Fidelity," describes both PD team planning efforts around teacher training workshops and what occurs during an actual Math4All workshop. In Chapter Two, "Teaching Teachers," the PD team discuss what textual practices will be most effective in cultivating administrative fidelity, focusing on two methods of 'reading' videos of Math4All in action. In Chapter Three, "Experimental Vision, Faithful Administration," I describe how these textual practices were facilitated in the course of actual teacher workshops toward the cultivation and then foreclosure of teachers' experimental impulses, indexing their subordinate position as instruments, and not agents, of science. I argue that the real-time production of objectivity, via the production of administrative fidelity on the part of teacher-participants, worked to maintain the Research-Practice Gap by highlighting the distrusting division of labor governing researchers' and teachers' respective role responsibilities.

Part II, "Reliability," examines the Math4All observation team's efforts to "get reliable" with the COG team and among themselves. It presents a comparison case to Part I, detailing how relationships of reliability organize intergroup cooperation in contrast to relationships of fidelity.

Chapter Four, “Training Trust,” documents how researchers persist in working towards reliability despite the emergence of a seemingly irresolvable conflict about how to code a written description of a classroom event. I argue that trust between the two groups is vindicated and perhaps strengthened, by the resolution of said conflict. Chapter Five, “Adaptive Strategies,” follows what happens when the trainees leave training. What challenges does adaptation introduce? Allowed to adapt the COG observation tool towards our own ends, we, on the Math4All observation team, work to get reliable on our modified system. I suggest that the reliability model can help education researchers imagine a means of closing, rather than bridging, the Research-Practice Gap.

The conclusion describes the present organization of the Research-Practice nexus around the production of objective evidence as impeding the development of scientific system of education by foreclosing the scientific potential of teachers—an enormous, well-positioned, component of the educational workforce. I point to several existing proposals for alternative systems, and suggest that this work itself offers a model for how they might be evaluated.

Chapter One: Objectivity, Democracy, and the Nation

Why aren't teachers considered experts in teaching? This was the question that originally motivated my interest in education researchers. How did researchers come to lay claim to expertise in something so divorced from their day-to-day activities?

As Carr (2010) has noted, expertise is not simply something that one has, but an ideologically and institutionally mediated claim that must be enacted in interaction. Key concepts in education researchers' claims to educational expertise are "science" and "objectivity." Since its beginnings in the late 19th century, the academic enterprise of education research has continually established its authority over educational matters by asserting its scientific, objective character as distinct from, and superior to, the non-scientific, subjective work of teaching (Lagemann 2000).

As Gieryn (1983) notes, what qualifies as science is not simply deducible from established norms (cf. Merton 1973); rather, the boundaries of science are managed by scientists and related others in line with their professional interests. Furthermore, this "boundary-work" of differentiating science from non-science is not just about selecting norms which enable one to claim "science" ('*we* checked the boxes, *we* do *science*'); but norms which project a less favorable view of competitors as "non-science" ('*you* didn't check the boxes, *you* do *pseudo-science*').

In Gieryn's analysis, scientists engage in boundary-work when they are engaged in professional contests: when they must defend their autonomy, attack jurisdictional counterclaimants, and/or capture authority and resources (pp.791–2). This is perfectly in line with Abbott's (1998) contention that professions are defined by their engagement in

jurisdictional contests of control over some arena of activity, not merely as a result of meeting "professionalization" benchmarks—a state-sponsored system of licensure, a professional association, a code of ethics, and so on. The character of professions is constituted in-and-through their methods of attack and defense, that is, the work of differentiating themselves from their opponent as a superior claimant to authority over some jurisdiction.

Thinking about professionalization in the American educational context may first bring to mind work concerned with teachers' professionalization (or lack thereof). However, I am primarily concerned with describing education researchers' professional claims and their consequences. Of foremost concern is education researchers' claim of educational authority, that is, that they are the experts on educational matters, and they should be the ones consulted on questions of educational governance. As jurisdictional claims are fundamentally a matter of interprofessional competition, researchers' claims—should they be taken up—are consequential not only for researchers, but for the other actors who might otherwise have grounds for a competing claim, such as teachers. Such is one of the principle conceits of this study, that researchers' professional projects are played out in-and-through the governance of teachers' work. In studying the activities taking place within the Research-Practice nexus, we bear witness to the means by which Research distinguishes itself from Practice as the authoritative site of knowledge production. In studying the activities by which the Research community constitutes and distinguishes itself, we learn something valuable about the constitution of the Research-Practice relation.

The key insight driving the arguments of both Gieryn and Abbott is that activities of *social differentiation* are the means by which jurisdictional disputes are negotiated and provisionally settled into more or less durable social structures and power relations, that is,

institutionalized. Furthermore, claims of authority on the basis of objectivity, for example, always occur within actual historical events of contestation, and thus the characteristics of objectivity are always subject to transformation given the actual events in which they are invoked. Thus, in considering the "ideological work"[1] of differentiating actors and activities as objective or subjective within education, we can understand claims of objectivity to have consequences for both social organization (group boundaries drawn along lines of objectivity; vectors of control arising across differences in objectivity) and ideology (what is objectivity) itself. Within such an analysis, the objectivity claimed by education researchers in the 1920s can be taken to be continuous with, but not identical to, the objectivity claimed by education researchers in the early 2000s. However, my project does not seek to detail historical transformations in ideology. Rather, I am interested in describing the present-day claims of objectivity in education research, and their consequences with respect to organization of the Research-Practice relation and its avowed Gap.

No Child Left Behind

Interventions like Math4All sit comfortably atop the 21st century institutionalization of American education research since the No Child Left Behind Act of 2001 and the Education Sciences Reform Act of 2002 which established the Institute of Educational Sciences (IES) in the US Department of Education. Under the jurisdiction of these legislative reforms, US educational services must be based on what they call Scientifically Based Research (SBR), exemplified by the randomized controlled trial (RCT).

[1] I take this phrase from Susan Gal and Judith Irvine (2019). My focus on differentiation is largely a result of the overall influence of their scholarship, especially the Irvine and Gal (2000) article, "Language Ideology and Linguistic Differentiation."

The No Child Left Behind Act of 2001 (NCLB) was passed on January 2002,[2] under the second Bush administration. Overwhelmingly co-sponsored by Republicans in the House,[3] the bill passed with bipartisan support.[4] In its imperative to close the Achievement Gap, NCLB shared in the spirit of its predecessor, the 1965 Elementary and Secondary Education Act (ESEA), of which NCLB was one in a line of major amendments and reauthorizations. As part of Lyndon Johnson's War on Poverty, the 1965 act allocated a billion dollars in federal funding specifically to schools and programs serving disadvantaged students, and has been credited with "redefining the federal role in education" in its commitment to equity in education (Nelson 2016, p.359). Other amendments of the 1965 ESEA act include the Bilingual Education Act of 1967 and the Equal Educational Opportunities Act of 1974, both passed during the Civil Rights Movement, recognizing the need to change learning environments to produce similar outcomes for students with different needs, and prohibiting discriminatory practices in education, respectively.

Prior to the 1965 ESEA, educational policy had been primarily the province of state and local governance. As such, attempts to legislate educational policy at the federal level has always been accompanied by jurisdictional conflicts. As Meens and Howe (2015) point out, the federal/local conflict is notable in that vocal advocates of local control exist on both sides of the political spectrum, leaving federal educational policies with few constant allies. Both conservatives and liberals have taken up the mantle of 'local control' towards their own political

[2] Because the NCLB passed on Jan 8, 2002, it is sometimes erroneously cited as the No Child Left Behind Act of 2002 (including on its current Wikipedia page).
[3] See <https://www.congress.gov/bill/107th-congress/house-bill/1/cosponsors>
[4] Results of the roll call can be found on the webpage of the Clerk of the House of Representatives <http://clerk.house.gov/evs/2001/roll145.xml>

ends, whether in resisting desegregation efforts or toward the empowerment of marginalized communities (p.13).

NCLB would only serve to heighten the federal/state conflict in education with its introduction of nationwide standards-based accountability, building on the incorporation of standards-based reform in the 1994 reauthorization of ESEA under the Clinton administration. The aspects of NCLB which have been popularly recognized as constituting a "quantum leap" in federal involvement in education are the sanctions which it attached to new measures of "adequate yearly progress" (AYP) for individual schools (Groen 2012). Based on students' standardized test scores, any school not meeting the AYP benchmark would be subject to restructuring, conversion to a charter school, or even closure. Commentators popular and academic have blamed the centrality of "high-stakes testing" to the NCLB regime for the narrowing of curriculum and instruction, as well as changes in staffing policies, and administrator and teacher evaluation (Au 2007, AERA 2015, Groen 2012).

The Teacher and the Scientist

Indeed, teacher evaluation and performance were especially contentious points in the debate over NCLB and high-stakes testing. Notably, this controversy over the role of teachers was in keeping with the history of US federal intervention into education during the Cold War, which is illustrative in its similarities with the present case.

After World War II, disciplinary scientists, starting with physicists and expanding outward, became heavily involved in K–12 curriculum development. For the physicists of the Physical Sciences Study Committee (PSSC), involvement with the K–12 science curriculum was

a matter of public relations.[5] The image of the scientist was in major disrepair after the devastation wrought by scientific advances like the atom bomb during WWII, and physicists especially were worried that the high level of funding which they had enjoyed during the war might disappear at any moment. If the public could be taught what science was really like, then they would be more likely to support continued funding for basic research.

At the same time, the US government was embroiled in the Cold War, and the threat of Soviet technological supremacy was made starkly concrete with the launch of Sputnik in 1957. In response, the US put the problem of outcompeting the USSR to the American education system. In 1958, Congress increased funding to the NSF, and Eisenhower passed the National Defense Education Act (NDEA), provisioning federal funds for the improvement of science, math, and foreign language education. Education became a site of confluence for the reputational interests of both physicists and the US government.

What this meant in practice was that scientists' curriculum development projects were funded by the NSF, while the NDEA provided funds for schools and districts to purchase those curricula. Crucially, this marked a point for natural scientists to mark their superiority by distinguishing themselves from education specialists and professionals. Because the professional education community was blamed for the "soft" and "intellectually puerile" life adjustment curriculum; almost every project funded by the NSF, itself founded to preserve funding for "basic research," was led by a non-educationist academic from a major university (Kliebard 2004, p.226).

[5] This discussion is based on Rudolph (2002), which offers a thorough account of the PSSC, its methods and motivations.

The end of NSF-funded curriculum development arrived with a K–12 anthropology curriculum developed by Jerome Bruner, *Man: A Course of Study*. McCarthy-style controversy around its content was leveraged by politicians to legislatively bar the NSF from funding curriculum development entirely (Lagemann 2000, p.174–6). But lack of funding and political controversy was not the only problem for curriculum as a mode of intervention. The pre-packaged curricula developed by PSSC and others were at the center of concerns about "teacher-proofing" and the "deskilling" of teachers. Scientists had designed their curriculum such that, so went both the pitch and the critique, anyone could implement them, with or without any scientific or pedagogical experience. With the right curriculum, any literate person could step into the classroom and become a teacher.

With reference to their development of "teacher-proof" curriculum, scientists were criticized for the devaluation or erasure of teachers' professional status, abilities, and judgment. So sociologist Michael Apple describes the increased academic efforts in curriculum development following WWII as part of "the history of the state, in concert with capital and a largely male academic body of consultants and developers, intervening at the level of practice into the work of a largely female workforce" (1986, p.37). Apple further claims that the failure of these curricular reform efforts to actually produce their intended effects in teaching practice was not for naught (see also Cuban 1984 for an account of how US teaching practices have stayed fairly constant over time). Rather, they helped to legitimate external control and technical forms of state intervention regarding teaching.

Though actual teachers appear sparingly within my project, the history of jurisdictional contestations between teachers and researchers (of whatever discipline)—that is, the history of teachers being treated as the instruments of science—is crucial to informing the present-day

structure of the Research-Practice nexus.[6] Cold War-era educational projects intervened in teachers' work as part of convergent projects of nation-building and reputational repair. The educational projects of the NCLB-era are no different. At the turn of the 21st century, the nation-building project turns on the image of democracy (the Achievement Gap),[7] and the reputation at stake is not that of physicists, but of education researchers themselves.

This dissertation will corroborate the continuation of a power relationship between researchers and teachers, even as researchers now strive to develop not teacher-proof, but "useful" and "educative" materials for teachers. Crucially, my pragmatic semiotic analysis of how materials become useful and educative, will show that this process takes place through the same events of interaction which maintain the authority of researchers over teachers' work as scientifically ordained. Further, I contend that the same types of communicative practices which maintain the institutional superiority of Research to Practice also work to continually re-authorize and re-legitimate Research as a distinctly objective and trustworthy scientific community.

The Politics of Education Science

With the passing of NCLB, education researchers were busy discussing not only "adequate yearly progress," but also another novel and consequential NCLB codification: "scientifically based research." Later that year, *Educational Researcher*, the flagship journal of the AERA, would dedicate their November 2002 issue entirely to the topic of scientific research in education. In it, they featured a spotlight article from Richard Shavelson and Lisa Towne,

---

[6] Again, see Abbott (1998) on the "system of professions."
[7] It is also concerned with American supremacy globally, but I have chosen not to focus on this aspect here. My work on the *Next Generation Science Standards* further discusses the international context of standards-based reform.

chairs of the National Research Council (NRC) Committee on Scientific Principles for Education Research; and Michael Feuer, the director of the NRC Center for Education. Shavelson and Towne also served as editors of a 200+ page NRC report, *Scientific Research in Education* (2002), published two days prior to the passing of NCLB. But why spill all this ink over the state of research in education? Is it not redundant to clarify that academic research must be scientific? And why did the US government feel the need to weigh in on the matter?

Since its inception in the late 19th century, education research in the US has had difficulties in legitimation (Lagemann 2000). It has been described, by insiders and outsiders alike, as a field of incoherent, fractured research programs of extremely variable quality, vulnerable both to fad and fancy, unable to produce knowledge that is scientific, non-obvious, and/or practically useful.

A recent example of the field's introspection on its image is educational historian Carl Kaestle's oral history of the "awful reputation of education research" (1993). In this, Kaestle's interlocutors, agency officials and researchers alike, explain the "weakness" of education research ("a lack of chutzpah") as a direct result of unstable Congressional and agency support (p.29). This instability is in turn attributed to political misunderstandings and conflicts regarding the value and impact of education research. For instance, a commonly described obstacle to the legitimacy and necessity of education research is encapsulated idiomatically as "everybody's been to fourth grade" (p.27). That is to say, since everyone—policymakers, the public, other researchers—went to school for a good portion of their lives, they all believe that they have an intimate and warranted understanding of the problems of education and their proper solution. Congress members were not convinced that putting money towards education would produce a worthwhile return on investment. Why should they pay researchers to spend several years

working, before returning to tell them what they already knew? Had education researchers ever produced anything which worked to improve the state of American education?

In all this, Kaestle remains a staunch defender of the field, and never wavers from the conviction that education research was suffering primarily from a problem of *reputation*, not of method or quality of research. In line with this framing, his article recommends solutions directed at reputation, namely in creating better relationships between researchers, practitioners, and policymakers.

Four years later, in 1997, AERA government liaison Gerald Sroufe wrote his own article entitled, following Kaestle, "Improving the 'Awful Reputation' of Education Research" (1997). He begins by recounting the President's Committee of Arts and Sciences' "dissing" of education research as plagued by "mediocrity," and the "disrespectful" assertion that education would benefit from an infusion of researchers trained in science, math, and engineering (i.e., NOT education). In the words of the committee chair: "Most education R&D is applied and anecdotal; it permits gleaning to support one's perspective" (p.26).

Following this recollection of indignities, however, Sroufe does not follow Kaestle in making the case for a mismatch between reputation and actual quality of research. Rather, Sroufe proceeds to join with his committee of detractors in asserting that education researchers would have to produce a higher level of research in order to garner the respect they desired. The problem is not reputational, but methodological, and as such, Sroufe recommends that education research "make more extensive use of techniques that are commonplace in other behavioral sciences" (p.28). It was this latter diagnosis that was ultimately taken up in NCLB, five years after the publication of Sroufe's article, in the form of Scientifically Based Research.

The above lines of internal and external criticism of education research's scientific *bona fides* explain in part the arguable redundancy of Scientifically Based Research, but such criticism does not address the inclusion of such a provision within federal education legislation. While the technocratic zeitgeist of contemporary life may make it appear commonsensical that doing better science is the obvious answer to problems of educational and perhaps racial and economic inequality, the role of science in educational improvement has not historically been straightforward. In order to understand why Scientifically Based Research was included in the same legislation which advocated for the realization of equality amidst diversity, one must understand the "wars" taking place around education reform in the 80s and 90s. The standards-based reform that NCLB advocated for (school accountability via a nationally mandated program of standardized testing), was a type of outcomes-based education (OBE) reform. OBE, as instantiated in standards-based reform, hoped to cut the knot of federal/state conflict in the US—outcome measures could be set at the national level in the form of standards, while states, districts, schools, and teachers retained the flexibility and autonomy to get students to those outcomes in whatever way they chose.

However, OBE, despite appearing to be a solution to the problem of contested educational governance, only gave rise to heated conflicts over the content of schooling, especially raising the ire of Christian conservatives (Schrag 2001). At the same time, the standardization movement was inciting intra-disciplinary "wars" in all areas of schooling. While the "Math Wars" are perhaps among the most memorable, similar contests occurred in language

arts, social studies, and science.[8] Education, it appeared, had become too highly politicized, with competing interest groups fighting for the enshrinement of their methods into standards, textbooks, and curriculum. Given these concerns about political bias, the Bush administration's move to present Scientifically Based Research as objective—that is, unbiased—furnished a putatively apolitical means of governing education. In this historical context, upholding the objectivity of Scientifically Based Research would serve not only the political goals of the US government, but the professional goals of education researchers.

It is worth noting that the hard turn to objectivity within education research was not merely imposed from outside via NCLB, but was already alluring to many who conduct and fund education research (Eisenhart and Towne 2003). As Porter (1995) and Abbott (1988) point out, movement away from expert judgment and toward more aspirationally mechanical or impersonal means of evaluation is typical of "weaker" professions, whose perceived expertise is not sufficient to allow for their autonomous self-governance.

The Teacher and the Scientist Revisited: The School and the Laboratory

Under NCLB, all educational services and programs were to be based upon Scientifically Based Research (see Figure 1). However, this formulation of what would count as "science" was not uniformly celebrated within world of education research. The aforementioned *Educational Researcher* spotlight article, "Scientific Culture and Education Research," begins:

> To rejoice or to recoil: That is the question faced by educational researchers today. Unprecedented federal legislation exalts scientific evidence as the key driver of education policy and practice, but—here's the rub—it also inches

[8] Selected readings: on the Math Wars (California focus), Rosen (2000); the Phonics-Whole Language debate in language arts, Hempenstall (1997); the Social Science Wars, Evans (2004); and constructivism in science education, Phillips (2000).

> dangerously toward a prescription of methods and a rigid definition of research quality. (Feuer, Towne, and Shavelson 2002, p.4)

In education research circles, the primary cause for recoil over Scientifically Based Research was its apparent creation of a hierarchy of research methods in which the randomized controlled trial (RCT) sat at the top. Elizabeth St. Pierre (2004) puts forward the following 2001 request for proposals by the OERI—the former research arm of the US Department of Education—as evidence of such a methodological hierarchy:

> The proposal must indicate method and why the approach taken optimally addresses the research question. Any approach must incorporate a valid process that allows for generalization beyond the study participants. Proposals must indicate which of the following approaches is to be used:
>
> 1. Experiment (control group); randomized assignment—both required
> 2. Quasi-experiment (comparison group, stratified random assignment, groups comparable at pretest, statistical adjustment for comparability)
> 3. Correlational study (simple, multiple/logistic regression, structural equation modeling, hierarchical linear modeling)
> 4. Other quantitative (e.g., simulation)
> 5. Descriptive study using qualitative techniques (e.g., ethnographic methods; focus groups; classroom observations; case studies; single subject designs)
>
> The design of studies should be clear: independent and dependent, or predictor and criterion, variables should be distinguished. Proposed research is expected to employ the most sophisticated level of design questions that is appropriate to the research question. For research questions that cannot be answered using a randomized assignment experimental design, the proposal should spell out the reasons why such a design is not applicable and why it would not represent a superior approach (compared to the selected design). (US Department of Education, as quoted[9] in St. Pierre 2004, pp.131–2)

This type of solicitation conveniently diagrams the hierarchy of methods, arguably in the order of their institutionally recognized epistemic authority. Indeed, the supporting text makes explicit that if option one, an RCT, is not held to be the best method for answering the research question,

[9] I have removed St. Pierre's added emphasis and corrected what appeared to be a copy-paste error in St. Pierre's reproduction of the original text, which is no longer hosted on the Department of Education's updated website.

explanation is required. Otherwise, the superiority of RCT is assumed and does not require justification. To put it another way, the superiority of RCTs is indexed by the fact that they do not require justification.

Crucially, this methodological hierarchy is tightly bound with the vision of research that dominates American education. While both "contestation and change" have occurred with respect to both the federal definition of Scientifically Based Research, and when and where such a definition should hold sway (Eisenhart and Towne 2003); there has been less contestation as to the notion that education research should strive to become more scientific. In the history of American education research, the question has not been *if* education should become more scientific, but *how* it should do so. What should a science of education look like?

Educational historian Ellen Condliffe Lagemann (1989) describes the history of American education research as the loss of John Dewey to Edward Thorndike.[10] While Dewey remains perhaps the more popularly known of the two, as he certainly was in his lifetime, educational historians have argued that Thorndike's vision is the one which has guided the actual institution of American education (see also Tomlinson 1997).

Thorndike was a behavioral psychologist enamored with measurement and statistical analysis, who came to education explicitly as a fertile field for the application of psychology.[11] He is often quoted as saying, "Whatever exists, exists in some amount."[12] Thorndike understood human behavior in terms of probabilistic relationships of stimulus and response, once describing

[10] Labaree (2010a) recounts Dewey's loss to yet another opponent, social efficiency proponent David Snedden.
[11] Lagemann 2000, pg.58
[12] Jonçich (1968), the canonical biography of Thorndike, discusses this line extensively in the footnote on p.283.

teaching as "the art of giving and withholding stimuli with the result of producing or preventing certain responses."[13] With a strong belief in heredity and eugenics, Thorndike saw schooling as the means of preparing individuals for the roles they were best suited to play in life, as determined by their natural aptitudes. That is to say, he believed in the scientific management of schooling for purposes of social efficiency. From his point of view, one should study stimulus-response bonds in the laboratory, and then apply that knowledge to improve schooling practices in the discernment and cultivation of individual aptitudes. The school was no place to do scientific work, being inhospitable to the precision Thorndike saw as necessary for science:

> To have in education the real benefits of quantitative science, we must spend arduous years in devising, testing, and standardizing units of measurement, in searching for convenient arbitrary zero-points, and eventually, for… the errors of measurement.[14]

For Dewey, on the other hand, the school was the natural site of scientific experiment in education. It was the environment in which teaching and learning took place, and thus the context in which they must be studied. The school was its own social world, and the interactions that took place within it could not be reduced to stimulus and response pairs. Laboratory schools were to be places of experiment, experiment towards the realization of a more unified, democratic society.[15] Within such laboratory schools, teachers were perfectly positioned to be key players in both research and practice; indeed, Dewey advocated for educational research and practice to be contained by every person working in schools, i.e., "the adoption of intellectual initiative, discussion, and decision throughout the entire school corps."[16]

---

[13] Quoted in Tomlinson (1997), p.371
[14] Quoted in Lagemann (2000), p.59
[15] See Dewey [1916] 1997
[16] Quoted in Lagemann (2000), p.50

Dewey and Thorndike were at odds with respect to proper nature, aims, and practitioners of educational science. Critically, these notions of science were entangled with their visions of the relationship between science and schools. While Dewey saw the school itself as the site of educational science; Thorndike was instrumental in establishing the externality of educational science to schooling.

Further, in the history of education research in the US, the scientific legitimacy of education has always been tied up with the success of male academics in differentiating themselves from the predominantly female teaching workforce (Lagemann 2000). In taking up the Thorndikean view of educational science, academia was further entrenched as the objective masculine domain of "education" science, in large part being recognizable as having these qualities inasmuch as education researchers could make their science differentiable from the subjective feminine work of "teaching." (One may begin to recognize the irony of education researchers attempting to scientize a domain whose valuation as unscientific has been and continues to be a major condition of their very own scientific status.)

Indeed, the Cold War era and the post-NCLB era have not been the only times at which teachers and academic researchers have found their professional fates at odds with one another. Their struggle dates back to the very beginnings of educational research in the US, at the end of the nineteenth century, which saw both the feminization of the teaching force and the emergence of the American research university and its fight for control over the US education system (ibid.). This meant that the political ambitions of male university administrators were staked in both the distinctive authority of their male education faculties and control over a female teaching force. It is no coincidence that labor organizing and unionization among teachers began amidst

this centralizing movement to subordinate the work of teaching to the expertise of an elite group of "successful men" (Tyack 1974, p.126; cf. Murphy 1990).

While a lively and diverse collection of research methods like practitioner inquiry and design experiments have produced and continue to produce valuable work which positions teachers as researchers and experimenters, I am concerned here with describing and evidencing an institutionally sanctioned view of Research and Practice as distinct arenas in the educational division of labor, siting the origin of scientific knowledge outside of schools, and giving teachers themselves little license to make institutionally valuable knowledge claims.

RCTs and Causal Inference

For now, let us return to the OERI list to see what kind of knowledge claims are institutionally valuable when the production of scientific knowledge about education is taken to be a schooling-external endeavor.

1. Experiment (control group); randomized assignment—both required
2. Quasi-experiment (comparison group, stratified random assignment, groups comparable at pretest, statistical adjustment for comparability)
3. Correlational study (simple, multiple/logistic regression, structural equation modeling, hierarchical linear modeling)
4. Other quantitative (e.g., simulation)
5. Descriptive study using qualitative techniques (e.g., ethnographic methods; focus groups; classroom observations; case studies; single subject designs)

At the top of the list is the randomized controlled trial (RCT), in which individuals are assigned at random to experimenter-designed treatment and control conditions. Second are quasi-experiments, often called "natural experiments." Quasi-experiments take advantage of (policy) events not purposively engineered by experimenters (e.g., a state-wide minimum wage increase).

Given careful selection and treatment of data connected to these events, researchers can create meaningful comparison groups without the benefit of random assignment.[17]

Experiments and quasi-experiments are at the top of the methodological hierarchy because they allow researchers to make a singularly valuable type of knowledge claim: "causal inference." Experiments, and quasi-experiments to a lesser extent, turn on the premise of randomized comparison, which allows researchers to attribute cross-group differences to particular cause and effect relationships. Unlike correlational studies, which are next on the list, randomized experiments do not rely on researchers to carefully identify and attempt to control for all possible confounding variables. Randomization does that for them.

In February 2002, at a US Department of Education conference convened to discuss Scientifically Based Research,[18] this point about randomized trials and the importance of causal inference is made multiple times. (The audience for this conference was not composed of other researchers, but educators interested in the implications of Scientifically Based Research.) Valerie Reyna, Senior Research Advisor at the OERI, in a presentation on the "logic" of "scientifically based evidence" describes randomization in this way:

> Clinical trials [=RCTs] in fact are the only way to really be sure about what works in medicine. The logic of it—and the other speakers are going to go into far more depth than I really have the time to do, the logic of it is basically the following: You have a group of people that you want to make a conclusion about.

---

[17] Rosenbaum (1999) offers an excellent account of quasi-experimental research.

[18] The conference theme was "The Use of Scientifically Based Research in Education." In attendance were representatives of the US Department of Education, representatives from the National Research Council's Center for Education, and a handful of researchers from academic and independent research organizations. From what is discernable in the transcript, the independent organizations include Westat and the American Institutes for Research (AIR, erroneously referred to as "AEIR" in the proceedings). A transcript of the conference proceedings can be found online at <https://www.govinfo.gov/content/pkg/ERIC-ED466791/pdf/ERIC-ED466791.pdf>.

> You want to say this intervention—whatever it is, if it's a new reading technique, or whatever—works for this group or not.
>
> So, what you do is you take members of that population and you flip a coin essentially as to whether they are going to be in the group that actually gets the intervention or gets some kind of comparison, like what you would have done had you not done this new thing. Standard treatment, that's a common control.
>
> The idea is that if you do this enough times and you get big enough groups, you've got two groups, the fact that you're flipping a coin ensures that these two groups, if you have enough people in them, are going to be comparable in every way except the intervention you're interested in.
>
> Why is that? Because there was nothing that put one person in one group as opposed to the other. It was all by chance alone that you ended up in the reading intervention group as opposed to the control group. And, so, all the ways in which people do in fact differ, and people do differ, should be represented in both groups. They should be comparable in every way, except the one thing that you made different in their lives, therefore, we can isolate the effect of the outcome and trace it to that intervention uniquely.
>
> This is the only design that allows you to do that, to make a causal inference. Everything else is subject to a whole bunch of other possible interpretations.
>
> Now if you have too small a sample, obviously the logic doesn't follow. Because you can have all the smart people in one group, the not so smart people in the other if you only have a few. If you do this enough times, you get a big enough group, they will be representative. That has been proven mathematically by things like – well, we won't get into that! (DOE 2002, pp.7–8)

Stephen Raudenbush, an invited speaker from the University of Michigan, in his presentation titled "Identifying Scientifically-Based Research in Education," similarly makes the point that randomized control trials are only ideal in the case of "causal question[s]," which are not the only questions that education research is concerned with (ibid., p.35). In his prepared remarks, which contain both more and less than remarks he delivered during the actual conference itself,[19] his conclusion includes the following point (emphasis in original):

> *Making valid causal inferences about the impacts of our interventions* is, in my view, the key challenge facing us now. Lots of good work using surveys and

[19] Raudenbush (2002) is a separate document from the transcription of conference proceedings (DOE 2002). I located it through a Google Scholar search. It appears to be a typed copy of the comments to be delivered. It is not a transcription of the actual comments delivered during the conference itself, though it includes much of what was said during the actual conference. It includes the quoted excerpt in the "Conclusion" section which was not delivered live.

> qualitative inquiry can help us identify unsolved problems – that is, targets of intervention, and also promising new ideas about practice. At the end of the day, however, we must judge our research enterprise by its track record in sorting out claims about the impact of educational interventions on student learning. (Raudenbush 2002, p.11)

These explicit rationales for the unique epistemological status of causal inference corroborate a reading of the OERI methods list as ranked with respect to the ability to make causal inferences. At the end the list are methods that are less clearly able to make causal claims in the statistical idiom. The last two items are miscellaneous quantitative methods and all qualitative methods. The former are not preemptively evaluable as they are indeed miscellaneous and thus variable in quality. The latter, both in their broad-brush treatment and dead-last placement, are positioned as least helpful in the project of producing causal inferences.

Debates about quantitative and qualitative methods in education research have long filled the pages of the AERA journal, and it was highly expectable that concerns and anger over the devaluation of qualitative methods would spike with the codification of Scientifically Based Research. Even Schneider and Keesler (2009), which celebrates both NCLB and quantitative research, remarks that SBR signaled a potential "social movement" within education research away from qualitative methods: "The U.S. federal legislation advocating a specific scientific approach to educational research suggests a transfer of power from other types of research paradigms that were more popular, such as, for example, intensive qualitative case studies" (p.213).

As such, figureheads in the conversation around scientific research in education were quick to provide reassurances that they—fellow education researchers—believed that any method could be scientific, if it held to the norms of scientific inquiry (e.g., Feuer, Towne, and Shavelson 2002). Evaluating what counts as science was not a question of which method, but of

the suitability of a particular method to a particular line inquiry. Different methods suit difference purposes.

Similarly, here I am trying to elaborate on not just what methods have been most institutionally valued by the agencies which govern and fund research in US education, but how those valuations have been made as a part of, and as a result of, agentive construals of the role of educational research in educational improvement. How are our beliefs in research's role in educational improvement tied up with the valorization of RCTs and causal inference as the gold standards of Scientifically Based Research? That is, what is the relationship between our beliefs about what research can and should do, and our beliefs about what research should look like?

The Classroom and The Clinic

In imagining the role of research in the project educational improvement, contemporary education research has drawn its inspiration, not from physics, the discipline most often valorized as the most scientific or objective, but from medicine.[20] Following Porter's (1995) contention that tendencies towards quantification have less to do with validity than with the social position of researchers, education researchers' hard turn to mechanical objectivity was not based on the success of physics in modeling the natural world, but the success of medical researchers in claiming authority over medical practice—a *professional* success that education researchers wished to replicate. (This is not an indictment of the validity of education research, only an observation of the socio-historical factors at play in birthing Scientifically Based Research and its kin.)

[20] As the Introduction describes, this appears to be a case of education researchers re-discovering or re-appropriating their field's own historical innovation. Medical research being more prestigious than education research, this circularity points to the low prestige of education in its continued inability to claim randomization for itself, despite arguably originating it.

In both the Reyna and Raudenbush presentations, medical research is explicitly cited as the model for Scientifically Based Research in education. Reyna, for example, describes clinical trials in discussing the value of randomization over "anecdote" or "personal experience" (DOE 2002, pp.5–6). Raudenbush opens his presentation discussing his attendance at an American Academy of the Arts and Sciences conference on improving the quality of education research. The conference was convened not by education researchers, he recounts, but by two medical researchers, Howard Hyatt and Frederick Mosteller. The twoo, as he goes on to say, were critical in advocating for the use of scientific research in medicine, over and above "the clinical judgment of the seasoned practitioner" (DOE 2002, p.31, l.26).

In the case of education, SBR pits the scientific objectivity of the education researcher against the professional judgment of the seasoned classroom teacher. This opposition is only indirectly alluded to in the written version of the presentation which includes a comment from a "well-known" (but unnamed) colleague who Raudenbush (2002) recounts as accusing him of "totalitarian thinking that unethically denies parents and teachers their rights" (p.2). Indeed, in positing a parallel between "the debate in medicine then, and the debate in education now" (ibid.), Raudenbush is explicitly most concerned with the ethical or unethical nature of randomized experimentation as a mode of knowledge production. Is it unethical to push aside practitioner experience? He puts it this way in his written comments:

> Hyatt and Mosteller noted that, in many cases, the profession really doesn't know what the best treatment is for a given disease. In that situation, it is unethical for us NOT to use the best available scientific methods, including experiments, to find out what works best. Once we know how best to deal with a given disease, many will benefit, revealing the true ethical character of the decision to conduct experiments. (ibid., capitalization in original)

Given a lack of practitioner knowledge, researcher inaction is unethical. This line of thought follows the Thorndikean tradition of keeping Research separate from Practice,

researchers separate from practitioners. The scientific problem is one of measurement, of developing objective evidence of "what works best." The Deweyan response to this conundrum of unknowledgeable practitioners would be to set groups of practitioners to work as communities of research-practitioners, to systematically and collectively build knowledge within the context of, and given the experience of, active practice. However, the model of intervention that education research has appropriated from medicine, and which is well-aligned with the present capacities of the educational division of labor, puts researchers to work furnishing an evidence base to guide decision-making.

In playing this evidence-producing role, causal inference is valuable as a form of knowledge production because of its ability to claim objectivity. As Reyna explained in the earlier lengthy quote, randomization precludes bias. By contrast, personal experience is full of bias. Earlier in her presentation, Reyna described doctors bleeding patients as a form of medical treatment. She asks, why didn't personal experience dissuade doctors from continuing the practice? She says that bleeding actually contributed to George Washington's death, so why did doctors not notice that bleeding did not work?

> There's been research done about when you ask people to report about things they have directly observed and directly witnessed and the biases that can creep into that type of reporting. These are normal human biases that are generally adaptive, but they have predictable pitfalls. So, if you rely on your memory for past events, we know that that memory will be biased, and so on. Drawing simply on your personal experience alone is not a solid foundation for generalization. Drawing simply on your personal experience alone is not a solid foundation for generalization. (DOE 2002, p.6)

In this line of thought, bias is derived from individual, personal experience. Such personal bias precludes generalizability, but can be mitigated via randomization. Therefore, the benefit of RCTs is not only the elimination (or reduction) of bias, but the production of

generalizable knowledge. By contrast, qualitative methods are not only too subject to bias, but their biased nature prevents their generalizability.

Generalizability is important because knowledge produced in Research contexts must be able to 'travel' to Practice contexts. The generalizability of science is the promise of a unified nation, no longer threatened by gaps. The generalizability of insights from cognitive psychology at the scale of humanity augurs the closure of the Achievement Gap. The decontextualizability of objective knowledge presages the closure of the Research-Practice Gap. The objectivity of SBR indexes aspirations for portable knowledge; it attempts to circumvent the problem of trust by appealing to the universality of truth. In the next section, I discuss the life course of such objective knowledge, as currently institutionalized.

A "Pipeline" of Evidence

Following the January 2002 passing of NCLB, the Bush administration passed yet another piece of federal education legislation, this time strictly targeted at education research. The Education Sciences Reform Act (ESRA) was passed in November 2002, the same month which saw the publication of *Educational Researcher*'s special issue on Scientific Research in Education. The act jettisoned the OERI, the previous DOE agency overseeing educational research, and replaced it with the newly formed Institute of Education Sciences (IES).

As the research arm of the DOE, the IES has overseen the institutionalization (as it were) of the contours of Scientifically Based Research through the present day. In collaboration with the NSF, it has issued a set of "Common Guidelines for Education Research and Development" which outlines a taxonomy of research types which are also legible as phases of research as SBR makes its way from Research to Practice (IES and NSF Joint Committee 2013). These Common Guidelines lay out the logic which informs IES and NSF funding decisions, so that the two

agencies most responsible for the disbursement of federal funds for education research have a shared basis for decision-making. In producing the Common Guidelines, the Joint Committee also consulted the AERA "Standards for Reporting on Empirical Social Science Research in AERA Publications" (AERA 2006). The Common Guidelines, then, are the articulation, in both senses of the word, of state and professional interests in a particular vision of education research. In the first sense of articulation, the *Common* Guidelines manifest the alignment and imbrication of state and professional interests in objectivity. In the second sense, the Common *Guidelines* make explicit the conceptual apparatus of this joint vision, not least of all in the research typology[21] that follows:

1. Foundational Research
2. Early-Stage or Exploratory Research
3. Design and Development
4. Efficacy Study
5. Efficiency Study
6. Scale-Up Study

Importantly, these six types are also sequential phases within the IES/NSF "pipeline" of evidence (p.8). While the Joint Committee acknowledges that actual R&D may be much messier and less linear than the pipeline metaphor suggests, their use of the metaphor suggests that they still believe it to be a useful one. Certainly, it will be useful for us in coming to grips with the model of intervention which underwrites the remunerated activities of education researchers. While the Common Guidelines acknowledge the complexity of the research process,[22] and that

[21] Again, this research typology is based on the medical model of evidence-based practice. This appropriation is acknowledged within the document itself in the "References Consulted by the Joint Committee" section, which lists several sources concerning "comparative effectiveness," "knowing what works," and "the reliable source of evidence" in health care contexts (pp.25–6).
[22] "Knowledge Generation and the Complex Connections among Research Types" on page 10 elaborates on the following caveats: non-linear feedback loops occur, new technologies may make some phases skippable, individual studies may span multiple types.

these six research types do not represent the whole of "useful investigations in education," they do not offer any support for those falling outside of their R&D model.

The six types can be divided into two groups of three, as is frequently done within the Common Guidelines document itself. The first three types indicate earlier phases of R&D which culminate in the development of a "solution," such as an "intervention" or "strategy" (p.9). The latter three phases describe the by-degrees expansion and recursive evaluation of said solutions toward the production of evidence. By "recursive," I mean that the same SBR methods used in the development of the intervention are again used to evaluate the intervention's impacts.

The first two research types are often grouped together, both being research types which provide suggestions for possible interventions. Foundational Research is functionally synonymous with what is generally known as "basic research," research having no immediate application in mind, but done 'merely' for the sake of producing new knowledge, what the Guidelines call "core knowledge" (p.9). Grouped with Foundational Research is Early-Stage or Exploratory Research which "examines relationships among important constructs in education" and is usually "correlational rather than causal" (ibid.). That is to say, Early-Stage Research begins to describe relationships between variables such as 'A is positively correlated with B,' potentially setting the stage for interventions in the form of 'improving A will improve B.' This stage also covers another important type of question: does A respond to intervention? Is it "malleable" (p.12)?

The third research type, Design and Development, begins to move into the production of instrumental knowledge, concerning itself with the practicalities of intervention, of goal-oriented solutions. Building on existing academic knowledge from Foundational and Exploratory Research, interventions should be designed with a "well-defined theory of action" and "a well-

defined end user"; implementation measures should be developed; the practical feasibility of the intervention should be explored (p.12). This phase encompasses both the testing of individual intervention components and the pilot testing of fully developed interventions (p.9). This phase is also expected to be iterative. Testing may indicate the necessity of returning to and revising foundational theories; or it may "indicate that the intervention or strategy is sufficiently promising to warrant more advanced testing" (ibid.).

The latter set of three phases are all concerned with this "more advanced testing" of an already piloted, fully developed intervention. Each phase progressively marks out larger and larger sets of testing sites, with less and less "developer involvement" (ibid.). In Efficacy Research, the intervention is tested in "ideal" conditions (ibid.). In the case of Math4All, which was in the Efficacy Research phase during my fieldwork, ideal conditions involved a higher level of implementation support, and a smaller number of students for each teacher to assess, than would be the case 'normally.' If an intervention is successful in showing efficacy, it moves onto Effectiveness Research. This phase is concerned with how an intervention performs "in the target context" (ibid.). The Common Guidelines emphasize that this means less developer involvement, at least no more than would be provisioned under "typical" circumstances" (ibid.). That is, researchers should not bias the results by offering more support than participants would expect to find outside of an evaluation study.

The final test for any intervention is "scaling up," being implemented across "a wide range of populations, contexts, and circumstances, without substantial developer involvement in implementation or evaluation" (ibid.). Developers are again reminded in the Common Guidelines that, just as with Effectiveness Research, they should stay out of it and allow the implementation to go forward as it would if the researchers' intervention was not being evaluated. Indeed, all

these comments mark a recognition that researchers have a lot to gain by the success of their interventions, an entrance of interest that would get in the way of the dispassionate objectivity of Scientifically Based Research.

Crucially, the latter three phases of the Evidence Pipeline should all be conducted as RCTs. The Pipeline is structured around the superiority of RCTs and causal inference in making claims about intervention impacts. All other forms of evidence are best leveraged towards designing interventions which can then be evaluated by RCT.

The Evidence Pipeline is not just about using research to design interventions, but the generation of evidence about the effectiveness of interventions. In this sense, the Evidence Pipeline is not a straight line, but a U-bend: academic knowledge flows into classrooms by way of its transformation into instrumental knowledge, an intervention. Intervention implementation in turn generates data which flows back to researchers for transformation into evidence. (Then this new evidence can then be fed back into the pipeline.) The IES and NSF—funding agencies—like causal inference because it helps guide, in a putatively apolitical manner, investment decisions. That is to say, RCTs and causal inference are highly complementary with a mode of intervention (Evidence Pipeline) which is primarily concerned with efficient resource allocation and accountability, in the sense given in rational choice theory.

The latter three phases of the Evidence Pipeline are each expansions of each other, with each phase being a "scaled-up" version of the preceding phase. The intervention under evaluation moves from a small, highly supportive context; to a middling, typically supportive context; to a large, typically supportive context. The Pipeline then, in economic terms, is not just concerned with return on investment, but market size. The ideal market size is all 50 states, because the problem of "scaling up," of effectiveness across diverse contexts, is precisely the

American problem of unity across diversity indexed by both the Achievement Gap and Research-Practice Gap. This political need to scale interventions nationally again leads us back to the RCT as mode of *objective* knowledge production. Scalability is the presumed affordance of the generalizability of the objective knowledge which RCTs provide.

Though I have been using more economic terms, I am not implying that profit or financial gain is the primary driver of educational research.[23] Rather, first, I am pointing out that the selection of projects (to fund) follows an economic logic of calculating investment risk and returns which makes the RCT the best mode of knowledge production; and second, that the ultimate goal of scaling-up is not just a matter of efficiency, but it is fundamentally tied up with nation-building projects which scale education as a *national* problem, e.g., closing the Achievement Gap (cf. Carr & Lempert 2016).

[23] It was occasionally joked that the successful scaling up of Math4All would be "when we all become multimillionaires." From what I could tell, multimillionaire status was not something any of the intermediaries who I worked with thought was actually going to happen, in the near or distant future. However, it was an attendant consequence of an imagined future in which Math4All was widely used and adopted by school systems the nation over. While this imagined future perhaps informed PIs' decisions (developing a statistically ironclad research design, finding ways to "scale up" and add in additional sites in future years, pursuing possible leads with publishers, etc.), it did not play much of a role in structuring the actions of individuals I worked with, who, as this dissertation will describe, had much more immediate concerns. In general, I personally never got any sense that commercial profit was a primary motivating factor in the development of Math4All, though this motive has been suggested to me by more cynical individuals outside of Math4All. Rather, Math4All appeared to me to be driven by a charitable desire to produce a specific kind of research, with an understanding that this kind of research could play a useful (or central) role in reforming the education system and decreasing racial and economic inequality. That is, the idea that SBR was a key player in closing the Achievement Gap, was sincerely held by the PIs. However, the PIs were certainly not naïve as to the implications of a commercially successful intervention, and often acknowledged the financial pull of other educational products when they came up in conversation.

Math4All

At this point, we have developed enough of a historical and institutional framework to understand Math4All as instantiating a highly valorized form of knowledge production in a particular moment in US history. Math4All is not representative of the majority of education research, which, despite the prestige of the RCT, does not appear to dominate the field in numbers. However, the RCT remains the most fundable and ideologically valuable form, because it is the standard against which all other research is evaluated. As we saw with the OERI list, all other research must effectively justify why it is not an RCT. Math4All is representative, not of a population of studies, but in its approach of an idealized form of research. Math4All is not just a case study, but a "best case" study: What happens when you try to do everything right?

During my fieldwork, Math4All was moving into the stage of efficacy research. The efficacy study, as described in the Common Guidelines, had an experimental design (RCT), recruiting a pool of 100 or so classroom teachers and randomly assigning them to treatment (Math4All use) and control groups (no Math4All use). In order to prevent researcher bias favoring the treatment group, the efficacy study was designed as a double-blind experiment, such that neither researchers nor teachers would know which teachers were in the treatment group and which were in the control group.

First described to me as an "existence proof" meant to show the mere existence any intended effect, the efficacy study sought to produce "ideal" conditions by securing as much manpower as possible to support implementation. This manpower took the form of a "professional development team," hired to execute a rigorous training and coaching schedule for the treatment group. This level of implementation support was above and beyond what would be available to teachers under "normal" conditions, that is, when and if Math4All was adopted "at

scale," i.e., once it was packaged, published, and distributed for widespread use across the US. If Math4All was successful in in demonstrating effects in its efficacy study, it would move on, as described by the Common Guidelines, to an effectiveness study. Indeed, by the time I left the Math4All project, the research team had secured several additional study sites, and the PD team had been put to work designing an online teacher support system which was feasibly scalable under normal implementation conditions, that is, without hiring additional implementation support staff.

The remainder of this dissertation will argue that the above research design which makes Math4All so politically and professionally desirable and imminently fundable, when put in action, ends up highlighting the difficulties which SBR emerged to evade, in particular the subordination of teachers' work and interests to political and professional agendas. In order to show how this happens, my fieldwork will examine the activity with which the latter phases of the Evidence Pipeline are so concerned: implementation. What can we discover if we do not narrow the lens of accountability to the correspondence of intention and effect, but widen it out to the examination of a broader scope of social consequentiality attending to the production of objective evidence? My work will follow how objective strategies of reliability and fidelity were enacted as part of the Math4All study, in the production of classroom observation data by researchers and the production of student data by teachers. I will argue that the practices involved in producing reliability and fidelity did not just condition the production of objective evidence, but actively worked to maintain the organization of the Research-Practice nexus.

Conclusion

Scientific efforts to govern knowledge are also sociopolitical efforts to govern people. This "co-production" of epistemic and political structures can be seen in the joint historical

development of US education research and US education policy (Jasanoff 2010). The promotion of RCT and causal inference as the gold standards in education research emerged at the articulation of researchers' professional projects and politicians' nation-building projects, producing a hierarchy of knowledge claims which re-inscribes the distinction between an objective realm of Research and the subjective realm of Practice. In the rest of the dissertation, I will describe the role that reliability and fidelity play in maintaining this organization of the Research-Practice nexus within the "best case" context of the Math4All efficacy study.

Part I: Fidelity

While academic researchers led the design of Math4All's development and evaluation, they had little to no experience in teacher training. They therefore hired a "professional development (PD) team" for teacher-facing support activities. The goal of creating a PD team was in large part, though not exclusively, to ensure fidelity of implementation as a condition for the production of objective evidence. "Fidelity of implementation" is a term of art, used in translational and implementation research, to describe the degree of match between how an innovation is actually carried out and how it was designed to be carried out (Century and Cassata 2016, O'Donnell 2008). But what does the work of supporting fidelity entail for the relationship of Research to Practice?

Half the PD team comprised individuals hired from the Golden Mathematics Institute, an educational research and development center affiliated with the Golden School of Education (GSE). The Institute staff were valued for their experience in leading teacher professional development and their former work as classroom teachers. Barbara and Rose, both White women, had both taught in K–3 classrooms before seeking out less structured, if more precarious, [1] work which would allow them more flexibility in raising their own children. They were now both school support specialists at the Institute, though Rose, the younger of the two, was more like a teacher-in-residence than Barbara, who had been with the Institute much longer, and occupied a senior leadership position. Rounding out the Institute staff was Anna, a White psychology PhD, who typically worked on applied research projects at the Institute. Anna hoped to gain some skills in teacher-facing work, and so was engaged in a self-styled apprenticeship

[1] Institute operations were funded through "soft money," and so staff positions were only as secured as the grants that funded them.

under Barbara and Rose with respect to the training aspects of the PD team's work. My involvement with the Institute, and then Math4All, was facilitated by the characteristic generosity of this trio.

For the PD team, Barbara, Rose, and Anna were joined by three additional White women from what they described as the "psych side," i.e., being employed by the Math4All PIs as part of their labs at GSE. Our colleagues from the psych side comprised Leah, research professional and Math4All project manager, and Ashley and Taylor, two graduate research assistants. Ashley, Taylor, and I all joined Math4All during the same summer, with Leah specifically bringing Taylor onto the Math4All PD team based on her experience as a classroom teacher.

Indeed, the logic of creating a PD team was not only to bring on individuals with experience working *with* teachers, but also with experience working *as* teachers. Former teachers like Rose, Barbara, and Taylor were expected to bring with them experiential knowledge about the situation of teachers. Math4All meetings which included former teachers would often see even the PIs—typically treated as the most knowledgeable and authoritative people in the room—deferring and turning to these individuals to answer questions regarding teachers' preferences, desires, knowledge levels, working lives, and so on.[2]

The Teacher Support System

As it happened, the Math4All research team gave the PD team very little instruction on the conduct of teacher training, a discretionary berth which was as much a deferral to their

[2] This is not to say that Rose, Barbara, or Taylor were always able to assert their expertise as teachers, as the domain of any issue was always contestable. For instance, concerns over teacher compensation based in experience working with and as teachers might be re-framed as concerns of research design and proper incentive structures. Therefore, any enactment of "teacher" expertise could be potentially contested as falling under the purview of the PIs rather than the PD team (cf. Carr 2010).

expertise as it was a relinquishment of responsibility over implementation. In fact, the PIs did not institute any formal measures of fidelity of implementation though such measures were considered best practices at the time (O'Donnell 2008, p.35). I do not know the reasons for this decision, though it was highly uncharacteristic for the PIs who were otherwise quite preoccupied with the specificities of research design and their implications for the objectivity of the efficacy study.

The PIs primary concern was that the PD team carry out the schedule of supports detailed and budgeted for in their successful grant application. This schedule promised three workshops (one each on administration, data interpretation, and instructional design) and at least two individual coaching sessions for each of the several dozen teachers in the study treatment group. Integrated into the experimental design of the project, this schedule could not be overruled by the preferences of the PD team, even when it exceeded the terms of their employment.

For instance, at least one member of the PD team voiced a desire to assign the treatment condition (Math4All use) to teachers located closer to the Institute. This would mean less time spent driving to and from individual coaching sessions (control teachers did not require coaching). This suggestion was unequivocally shot down by the PIs, because random assignment was a crucial factor in the preserving the objectivity of the RCT. As a result, the PD team frequently felt overwhelmed by the time investment entailed in training and coaching teachers, but, unable to overrule the research design or to find more money in the study budget, simply did their best with what time they had, often working overtime without additional compensation.[3]

[3] I am given to understand from conversations with colleagues and other researchers that this is a commonplace situation among mid- and lower-level staff on large research projects. It is beyond the scope of this work to offer an analysis of how prevailing or acceptable these conditions are, though my personal opinion is that they are bad and, if pervasive, even moreso. However, here I

The PD team's work was part of a years-long effort to pilot and test Math4All with actual teachers, including liaising with teacher-collaborators the year prior. I do not cover much of the work which fell under their purview in this dissertation, not for lack of interest but lack of time. Instead, this chapter presents a few thin slices of one part of the PD team's work—planning and running introductory workshops—to illustrate a situation not atypical of the PD team's intermediary position, in which they must work through the living tensions between the PIs' objective research designs and their commitment to supporting and building teacher capacities.

The System

The Math4All system was quite extensive by design. It was built to provide a vast array of detailed information on individual students' math abilities. Not only did this entail the use of a series of assessments, but familiarity with the wide range of tasks contained by each assessment. Teachers also needed to be familiar with the website into which they would input their data, and from which they would receive detailed graphical representations of their students' abilities.

Each Math4All assessment was to be administered one-on-one to one student at a time. Each comprised a different grouping of *tasks*, with tasks appearing in one, some, or all assessments at differing degrees of difficulty. Each task required the use of a different script, different bodily movements, and often different sets of materials. While tasks typically utilized a central *stimulus booklet* featuring various pictures and graphics on each page, some tasks required no materials, or came with extra materials in the form of *manipulatives* like sets of cards or counters.

---

offer this example principally to illustrate the primacy and authority of the research design in preserving objectivity above all other considerations.

Each task had an attendant *score sheet*, which featured *administration instructions and prohibitions*, including what the teacher should say, what they should wait for the child to do, how they should respond during "training items" where kids were being taught the task, how they should respond during "test items" where kids were actually being assessed on the task, common administrative mistakes, what words they should avoid saying, how they should use their hands, and how they shouldn't use their hands. Various features of the score sheet were styled differently to distinguish the actions, ideas, or items to which they referred, e.g., **things to be said out loud were in bold**, and *training items were italicized*. At the very end of the sheet was space for the teacher to record student responses to each item (a prompt for student response) in the task, and a "Notes" column, for information the teacher might want to record for their own future reference.

Fortunately, each assessment came with all of its task score sheets stapled together in a pre-assembled packet, which was also referred to as "the score sheet."[4] It was rare for any single score sheet to live alone, except for training purposes, and so, in the context of actual administration, "score sheet" most often referred to the whole packet, or to whatever page said packet was open to at the moment.

[4] There was discussion of referring to this amalgamated "score sheet" as an "administration and scoring packet" or some similar phrase, given that score sheets consisted mostly of administrative instructions and tips, far outweighing the relatively limited section where teachers would "score" or write down student responses. However, the more descriptive name gained little traction, as only Barbara and Rose attempted to use it briefly during introductory workshops. Meanwhile, the document was a "score sheet" in many more places, not only in the habitual speech patterns of the PD team, but also the computer filing system where the digital copies of said score sheets were named, modified, and stored. In my fieldnotes, I continuously referred to these artifacts as score sheets, forgetting over time that there had ever been a debate over their name. I will continue to refer to entire packets and individual sheets both as "score sheets."

Chapter Two: Teaching Teachers

Administrative Infidelities

What problems did the PD team anticipate in teachers' administration of the Math4All assessment?

The Math4All assessment hoped to establish a miniature laboratory environment in each one-on-one session of assessment, recommending that teachers find administration environments as sheltered from the everyday chaos of the classroom as possible.[1] A controlled environment was necessary for the production of objective knowledge about the cognitive development of individual students. Math4All effectively treated cognition as a phenomenon of individual minds. Its careful design and elaborate administrative protocols were meant, to the extent possible, to isolate individual mental activity by controlling the contextual inputs which might mediate its behavioral expression. As such, the PD team, tasked with controlling administrative inputs, was most concerned with teacher actions which (1) deviated from the task protocol as set forward in the score sheet/manual, and/or which (2) might cause students to modify their behavior in ways that would obfuscate the elicitation of their 'actual' (individual) ability.

The PD team sometimes used the term "fatal mistakes" amongst themselves to refer to mistakes which would irreparably tarnish data collection, requiring student responses be thrown out and the task to be re-administered. However, it was rare that an actual event occurred which PD team members labeled as a fatal mistake. Rather, fatal mistakes were more like boogeymen they sought to avoid. Most of our time was spent discussing non-fatal mistakes like sweeping one's hand across possible answers in the stimulus booklet in an overly suggestive fashion. If

[1] This was more or less impossible for some teachers, but others were able to use spaces like storage rooms, offices, conference rooms, and hallways for assessments.

one's hand appeared to linger over some portion of the page at the end of the sweeping motion, it was speculated, children might take the lingering as an indication of the location of the correct answer.

By contrast, a fatal mistake would be to consistently offer praise for correct answers, and no praise for incorrect answers. This would provide information to the student about their performance which might influence their future responses. Another fatal mistake might be using language that "suggests a counting strategy" during a task not designed to measure counting ability. For instance, by using number words ("one," "two," "three") in a non-counting task, children might be prompted to count the items presented to them, rather than responding using some other strategy that they might have otherwise been predisposed toward (e.g., subitizing, matching). Children were free to count if they would like, but the teacher-administrator should not bias them towards that strategy. Most of these infidelities were imagined not as matters of teacher incompetence, but as outgrowths of teachers' well-intentioned habitual disposition toward facilitating student learning.

Another concerning vector of infidelity was teachers' ability to respond appropriately to children's (mis)behavior during the relatively long period of administration (up to 30 minutes). As members of the PD team often commented, it was one thing to administer the assessment to an adult, and another thing entirely to administer it to a child, who by virtue of "being a kid," or even, "being a *real* kid," was prone to a whole range of unpredictable behaviors. The problem that faced the PD team was not in managing children's behavior. Indeed, it did not cross anyone on the PD team's mind that children should be made to behave any differently than they did.[2]

[2] On the contrary, children's strange behaviors were often a source of delight for members of the PD team, a few of whom would jump at the chance to retell humorous stories of their encounters

Rather, the PD team sought to find ways to manage *teachers' responses* to child behavior. The unpredictability of child behavior was only a problem inasmuch as it exploded the range of possible scenarios which the PD team should prepare teachers to respond to.

For example, during a shape recognition task, teachers were to hold up a series of cards with shapes on them to show children, who would then respond with what kind of shape it was. It was essential to the task design that the cards be oriented a certain way, so that the orientation of the shapes would be uniformly presented across administrations. During a PD team meeting, collectively reviewing the teacher's manual description of this task, Leah wondered what they should write about what to do if a child were to take, or attempt to take, the card from the teacher's hand. The card should not leave the teacher's hand, everyone agreed, but if the child successfully took the card, the teacher should take it back, *re-orient it correctly*, and then re-prompt the student for an answer. Such were the contingencies that concerned the PD team in their task of ensuring faithful administration. The problem was, for every task, of which there were over a dozen, a different set of task-specific concerns emerged.

The range of possible concerns was then too great to permit distinct address of any given concern, especially if children lived up to their reputation as chaos agents. Contingencies exceeded both count and capture, and so could never be fully addressed across any number of training sessions, let alone a single three-hour workshop. Instead, the PD team endeavored to address the range of possible mistakes and scenarios by helping teachers understand "conceptually" why tasks were designed the way there were. Given a "deeper" understanding of

---

with specific children during their own practice administrations, the fascinating and funny behaviors they observed, and their compulsory reactions of equanimity in the moment.

task design, teacher would then be able to knowledgeably tailor their redirections of child behavior, beyond what the PD team might be able to specifically prepare them for.

Educative Practice

Cultivating greater teacher sympathies toward Math4All was also considered as having another possible benefit. Teachers who were able to understand Math4All's design rationale, and the reasoning behind the multitude of sanctions around administration, might be dissuaded from perceiving the Math4All system and/or the Math4All PD team as "bossy." In discussions of "bossiness," the PD team anticipated resistance or otherwise negative reactions to Math4All's strict administrative guidelines and the PD team's active promulgation and enforcement of those guidelines. In my understanding, the PD team believed that negative reactions toward perceived bossiness might (1) inhibit the PD team's ability to meaningfully communicate with teachers and/or (2) inhibit teachers' investment in or enthusiasm for Math4All. Either of these possibilities could potentially lead to less fidelity in use, and, perhaps more importantly for many members of the PD team, no one—teachers, students, or researchers—would see any benefit from all their hard work.

In the education literature, a prominent proposal in response to the problem of managing curriculum enactment—that is, managing textual practices around curricular materials—suggests doing so through the design of "educative" materials which speak directly "to" teachers rather than "through" them [to students] (Remillard 1999). Recognizing popular beliefs among teachers that curriculum materials are agents of control ("bossy"), and that good teachers do not use textbooks, for instance, these proposals recommend that textbooks and teachers "partner" with one another (Remillard 2016; Ball and Cohen 1996).

Rather than attempt to constrain users' textual practices, these approaches to curriculum design attempt to cultivate more "educated" engagement through elaborated forms of direct address to teacher-readers, interpellating teachers as co-participants in curriculum enactment. In this view, curricular materials should include teacher's manuals that comment on all those factors which are imagined to influence enactment: teacher beliefs, student responses, and so on, even looping teachers in on the decision-making process of the curriculum developers in their design of the materials (Ball and Cohen 1996). Engagements with such "educative materials" are envisioned as sites of teacher learning.

Researchers like those cited above tend to recommended that these interventions manifest in the design of curriculum materials themselves, that is, through changes in the language or generic conventions of text-artifacts like teacher's guides, in attempts to imbue such artifacts with greater agency by making them more elaborate and self-referential.[3] This recourse to design over training is not due to a lack of belief in the helpfulness of training, but a resignation with respect to the "realities of schools" and staying within the confines of teachers' current working conditions (Davis and Krajcik 2005, p.4). Yet this design approach appears to merely reproduce the problem of enactment to a higher order, referring the creative agency of human actors toward their engagements with the higher-order text-artifacts. Will the teachers' guide have its own guide on how to read it?

The Math4All case offers the opportunity to study the conditions mediating the possibility of teachers "partnering" with an intervention, or the possibility of an intervention

[3] Linguistic anthropologists might describe this as "textuality" — my work on the *Next Generation Science Standards*, not included in this dissertation, comments on this phenomenon more deeply.

functioning educatively in offering teachers learning opportunities, in a situation where a non-negligible amount of resources can be devoted to training.[4] In line with the above proposals, the PD team imagined that treating teachers as co-participants in the Math4All project would ultimately improve their instructional practices (the same goal as the above curricular interventions). Their working hypothesis posited that treating teachers as co-participants would increase buy-in and enthusiasm, promoting faithful use of the assessment, and thus enhancing whatever benefits they would receive through their use of Math4All. This chapter will address the PD team's deliberation over the mechanism by which teachers should be educated into their role as co-participants in the course of the introductory workshops. The next chapter will describe the use of those strategies during the unfolding of the workshops themselves.

The Videos

Assessment videos were identified as a key resource in the PD team's workshop planning. Barbara—the most senior of the Institute staff, and an experienced PD leader—established early on in the planning that teachers should be actively engaged during workshops, that they should not simply be lectured at for the duration. It was in brainstorming said engaging activities that Leah, senior psych-side representative, brought up the possibility of using assessment videos.

The PD team had access to an archive of videos of Math4All tasks being administered to children in a lab setting at the GSE. The test subjects (young children) had been recruited by Math4All for purposes of testing new assessment items, providing the PD team opportunity to

[4] In the case of the Math4All efficacy study, there was yet no entertaining the idea that the human interactional element in training could be dispensed with. This would become less and less the case as attempts to scale-up the intervention continued, during which training was to be offloaded onto web-based video recordings.

practice assessment administration, and producing videos of assessment administration for "educational purposes," i.e., teacher training. Because all the most up-to-date videos were of PD team practice administrations, the videos to be used for workshop purposes all featured a PD team member as the administrator.

The PD team imagined that video could provide teachers a valuable window into what Math4All administration looked like in action, between an actual child and actual adult. This consensus on the pedagogical value of videos however, did not straightforwardly correspond to a consensus on how videos should be used during workshops. Rather, discussion of what textual practices should surround the 'reading' of videos disclosed two competing visions of teacher education.

Two Readings

There were two distinctive ways the PD team imagined engaging with videos during workshops. The first was a *cognitive reading* which proposed treating video depictions of children's behavior as evidence of cognition. In viewing the videos in this way, teachers would inhabit the role of the Math4All researchers, the role of the Math4All assessment itself: they would attempt to discern the developmental stages of the children being assessed, becoming curious about children's cognitive development. The second was an *administrative* reading, which proposed treating the videos as documenting events of administration (rather than events of cognition). In viewing the videos in this way, teachers would inhabit the role of a Math4All administrator, attempting to discern how they would be expected to act, becoming curious about the assessment itself.

Each textual practice, or "reading," presented a different mode of producing a "denotational text" (what the video was "about") from an encounter with a video. And, in-and-by

producing different denotational texts, each textual practice also entailed a different "interactional text," positioning the viewer in different social location (e.g., researcher/assessment, administrator) in relation to the events depicted in the video.[5]

Both of modes of reading emerged out of a desire to position teachers as co-participants in the Math4All project, to demonstrate to teachers the usefulness of Math4All, to generate teacher buy-in and enthusiasm; but, as forms of ideological work, the competing practices presented differences emerging out of the PD team members' different social alignments, mediating differently interested claims about what teacher engagement looks like, the straightforwardness of administrative practice, and the relationship of learning about administration to learning about cognition.

In what follows, I will describe the interactional unfolding of a planning meeting in-and-through which these claims take shape.

Video Talk

Leah, Ashley, and Taylor would carpool from the GSE to the Institute about once a week for PD meetings. PD meetings were the only ones that took place at the Institute, which occupied one floor of a modern office building housing several other unrelated, but also Golden-affiliated enterprises. Everything else related to Math4All—lab meetings, observation meetings, funding meetings—took place at the GSE, one of the older buildings on the Golden University campus. PD meetings took place in spacious, well-lit conference rooms, appointed with various rarely

[5] On denotational and interactional texts, see Silverstein 1997.

used technological affordances for virtual meetings; quite a contrast to the cramped, dim spaces the observation team often found ourselves meeting in at the GSE.[6]

Leah usually provided the agenda for our PD meetings, and the Institute staff were very happy for her to take on the responsibility of leadership. They were there to meet the needs of the Math4All project, not to determine those needs. At this particular meeting, we begin by brainstorming activities for the first teacher workshop, and arrive at the topic of assessment videos.

Leah starts us off by suggesting a cognition-centered exercise,[7] "I mean one thing you could do is show a video of kid doing [the assessment], and have [the teachers] be in little groups and talk about like what they think it means—what they think this means about what the child understands."

Rose responds to Leah's cognitive approach ("what they think this means about what the child understands") with an administrative approach: "Or even like practice scoring videos—like watch a video, practice scoring, based on what you see this kid doing."

Barbara joins in, enthusiastic about the idea of using videos as a means for engagement: "I like the idea of […] for some of them, have that be their first thing, then have like, 'What do you notice about this task?' 'What questions do you have?'"

As Barbara and Rose begin to expand on this idea, Leah interrupts to ask for clarification: "Barbara, what idea did you like? The, um, having them score it?" In asking which idea Barbara liked, Leah begins to position cognition- and administration-centered practices as distinct, and

[6] There were also larger, more distinguished conference rooms in the GSE, but they were usually reserved for meetings with the PIs in attendance.
[7] See Appendix B for an unbroken transcript of the conversation of note.

possibly conflicting, exercises. Barbara responds without evincing any distinction between the two:

> I think having them, before we tell them much of anything, or maybe there's like an overview, some way we use the manual text, I don't know, that before we talk too much at them, we actually have them score something using a video, and then you know, talk to each other, and then use that to [move into asking] "What questions did you have about this task?", "What did you notice about it?," "What-" you know what I mean? I think we need to engage them.

For Barbara, the priority is engagement, and there is no need to distinguish between any form of engaging with the video. Note however, that for Barbara the subject of discussion is administrative, it is the task itself ("what question did you have about this task? what did you notice about it?"), and not "what the child understands" as Leah put it, that remains at the center of her prompts.

Following Barbara's administratively inclined suggestion, Anna tacks back to a cognitively focused activity:

> Or even if you want to take a step back and ask them, what kind of mathematical thinking do you see here? […] especially for something like the [redacted] task, show them that and be like, "What sorts of things is the kid saying?", "What is he doing?", […], "What kinds of thinking do you see happening here?" So that they're sort of thinking about the kid and what they're getting out of watching the kid do this, than this is the right way to score this.

Anna takes the same position as Leah here, not only in construing the videos as depicting behavioral evidence of cognition (what kind of thinking can be "seen" in noticing the kid's speech and actions), but in distinguishing attending to cognition as more fundamental ("take a step back") and more worthwhile an activity than attending to administration ("*than* this is the right way to score this"). For her, a major payoff of the cognitive approach is that it draws teachers' attention to "what they're getting out of watching the kid do this," that is to say, it makes apparent the usefulness of Math4All as a tool for teachers.

Rose follows her good friend Anna by, like Barbara, implicitly disavowing any conflict between a cognitive and administrative reading of the videos:

> Well then, that's giving a great- I like that [Anna's activity] in terms of motivating like turning back to the manual, like, let's read, let's see- then you're kind of primed for like, "ooo I wonder what *is* going-" like, "this is what I think is going on"

Here Rose brings another assessment material into the conversation (beyond the implied score sheets): the teacher's manual. The PD team had spent much of the summer working on a teacher's manual which described each Math4All cognitive skill and the tasks associated with it. (Note that Barbara's earlier response to Leah also included mention of "making use of the manual text.") Notably, in her contributions, Rose continually evinces a tendency to link workshop activities back to Math4All itself as an assemblage of materials that one must directly engage with in becoming familiar with the assessment.

Indeed, in the midst of further discussion about whether suitable videos could be found to support the proposed activities, Rose brings up scoring again, underlining the *materiality* of scoring, in two senses of the word:

> So teachers actually have the score sheet [waving paper] for that particular task, they're practicing scoring, cos that's going to be something, that they're, that they need to do, and then, after it's done, then you have this conversation about whatever

She emphasizes that score sheets are a *tangible component* of the assessment by waving a piece of paper standing in for the scoresheet in the air, and does this while explaining the *practical necessity* of scoring for teachers. Furthermore, in line with Barbara, she discusses scoring alongside video discussion, failing to recognize a conflict between the two activities ("after it's done, then you have this conversation about whatever"). Without saying as much, she

also contests her friend Anna's ranking of activities, positioning scoring as primary ("after it's done"), and other discussion as secondary ("then […] whatever").

Leah responds by reintroducing the cognitive/administrative distinction:

> So, right and I, and I honestly, I think having them score it is fine, I'd almost rather them not be sitting there focusing on the score sheet and just have them watch it, and then um, … because there's a lot of rich things happening and then, and the scoring I think, *is* pretty straightforward, um, I mean we could have them score it afterwards.

Leah again discursively reverses Rose's ordering to back to Anna's positioning of scoring as secondary to looking for mathematical thinking. What's more, specific qualities have now been invoked in justifying this regime of value: where the video happenings [child behavior] are "rich," scoring is "straightforward." Further, Leah positions scoring and watching the videos as competing for limited attentional resources, such that one must be done at the expense of the other ("sitting there focusing on the score sheet" vs "just" watching).[8]

In response, Barbara draws another distinction, "passive" and "active" engagement with the videos:

> I think it's less about the scoring and more about their active engagement with the video, rather than passive engagement with the video, I think that's why, so it's not so much that the scoring is hard, it's that we have to have ways to help them not be passive learners.

While Leah discusses the video's happenings as demanding a high level of attention, Barbara turns this around by figuring the act of watching as a passive form of engagement. By contrast, scoring, being literally hands-on, is figured as an active form of engagement.

---

[8] Toward Leah's point, she did recall at another point that the collaborating teachers from the previous year were unenthusiastic and bored when it came to practicing administration. In response, Barbara made further inquiries as to why this was the case, but I have not yet fully attended to that exchange (and/or the conversation wandered off elsewhere), and so I cannot offer more insight about it here.

Two Models

The Cognitive Case

Advocates for a cognitive reading of the videos—Leah and Anna—figured engagement with children's mathematical thinking as the key factor in getting teachers to appreciate Math4All's vision. Math4All was a system for delivering knowledge about mathematical thinking, so, in order to realize the value of this system, teachers should come to appreciate the richness of children's mathematical thinking. Teachers should learn to see children as Math4All saw them, as growing bundles of cognitive abilities. Teachers should come to directly experience such vision as valuable, and then, they could apply themselves to the administrative tasks involved.

As we have seen in the preceding conversation, cognitive advocates frame looking for evidence of mathematical thinking as both more rewarding and richer than practicing scoring. Leah re-iterates this regime of value in answering a question from Barbara about what she wanted to prioritize during the workshops:

> In some way, we need to talk about yeah, thinking more deeply about children's thinking in that area [math], rather than necessarily doing the task correctly.

In returning to my fieldnotes and recordings, I found it quite remarkable that Leah would say that teachers' deep thinking was more important than correct task administration. Of all the PD team members, Leah was the one most concerned with faithful administration, and certainly, as the Math4All project manager, she was the most professionally invested in the legitimacy of the efficacy study. She often returned to the idea of creating a document of overall administration guidelines, an initiative that inspired little enthusiasm among the rest of the team. She even

ultimately secured a spot at the end of introductory workshop to go over administrative best practices, despite repeated attempts by others to simply leave the idea behind.

However, Leah's focus on mathematical thinking does not present a contradiction given a working hypothesis that teachers who literally take on Math4All's vision are more likely to become a part of that vision. Under this model, guiding teachers' attention toward student cognition is an essential driver of enthusiasm about, investment in, and fidelity to Math4All.

Anna puts it this way:

> One of the goals would be to get [teachers] enthusiastic about what they could learn about their students, and caring about this stuff [mathematical thinking] in seeing it, and honestly, I think building enthusiasm is more important than anything else, like "This is useful to me, I want to know this stuff about my students, I want to try this with my students, and this is different than stuff I've done in the past"

There was a distinctly egalitarian ethic evident in Anna and Leah's advocacy for teachers to see like developmental/cognitive psychologists. Both of Leah and Anna found it personally and professionally rewarding to engage in the textual practices that they wished to facilitate in teachers. In suggesting cognitive exercises, they hoped to share with teachers the experiences that motivated their own enthusiasm about Math4All.

The cognitive reading strategy is understood by its advocates to be a means of securing teacher enthusiasm and buy-in. Teacher co-participation in Math4All is hypothesized as being conditioned by the ability to share a single vision, to see the same things in the same way. Teachers would learn about Math4All—its tasks, its design—by engaging in the same textual practice as Math4All itself—meaningfully distinguishing elements of student behavior as evidence of cognitive development.

The Administrative Case

In making the case for administratively focused exercises, Barbara responds to Anna's discussion of fostering teacher enthusiasm with skepticism about the PD-teacher relation as the site of learning:

> We could go on for 6 hours or 4 days about these things [e.g., mathematical thinking], but the power of the assessment, is that you can learn about them [math skills] through and with your kids, not from us, standing up here with slides. I think that that's the hook for them. I think the real way you're going to learn about this stuff is by actually using the assessment.

Influenced by research on educative materials,[9] Barbara posited that "the power of the assessment" was its ability to facilitate teacher learning "through and with" interactions with students. That is to say, engagement in actual assessment administration, involving actual student-teacher interactions, were the medium through which Math4All would produce teacher learning and, presumably, if it was found useful, enthusiasm and buy-in. That is, the administrative reading of videos would be consonant with an approach to teacher training which takes administration itself to be not merely evaluable as "doing a task correctly" or "the right way to score this," but as an engaging site of learning.

As Barbara says earlier, it is not that "scoring is hard" or not, it's that the PD team should help teachers be active learners, and unlike Leah, she does not see watching assessment videos as part of an active learning process. Their disparate evaluations of what counts as "engaging" is a result of different orientations on Barbara and Leah's part about what teachers should be learning about. Teachers may be actively engaged in learning about mathematical thinking through attentive viewing, but they are not actively engaged in learning about the assessment itself.

[9] e.g., Ball and Cohen 1996, Davis and Krajcik 2005, Remillard 2000, Remillard and Reinke 2012

Rose's position is also concerned with active engagement with the assessment, not as a facilitator of teacher-student interactions, but as a material technology. Rose was the PD team member most explicitly concerned with respecting teachers' time and expertise, and consistently advocated for allowing teachers to do "actual work" during the workshop itself. What Rose meant by "actual work" was *required* work. If teachers would be required to familiarize themselves with scoring tasks, then they should be given time to do so during the workshops themselves. If teachers would be required to plan lessons around Math4All data, then they should be given time to do so during the workshops themselves. Besides making good use of teachers' time during the workshops, having teachers do actual work during workshops also meant that the PD team would be available to assist teachers in that work.

Rose was distinctive among the PD team in her previously remarked upon tendency to repeatedly tie activities back to practical interactions with the tangible aspects of the assessment system: score sheets, the teacher's manual. The actual work most salient for Rose then, in the introductory workshop, was the practical necessity of physically interacting with the Math4All materials. In continuing the conversation about building enthusiasm, she has this to say:

> A big part of it [building enthusiasm] is like, you know, honoring their experience and building off that, but then I think it's like constantly happening, like they're building enthusiasm when they're seeing kids doing really cool things in the videos, they're building enthusiasm when they're gaining confidence in like, well how do I even open this thing, and put it on the table so it doesn't flop down.

Rose's final example of gaining confidence describes teachers' use of the stimulus binder, which must be set up in a special way to support itself a tabletop. For Rose, here again, spending time with something potentially perceivable as mundanely administrative is not inherently less valuable than time spent explicitly discussing children's mathematical thinking.

Both are part of the same project of learning about and with Math4All, building enthusiasm, and gaining confidence, which is "constantly happening."

Professional Visions

Straightforwardness

Are the Math4All materials straightforward or not?

Leah's claim of straightforwardness is made possible by her central position in the Math4All project. At the point of the recounted conversation, Leah had not only been with the project for many years, but had also witnessed, supervised, and actively participated in the development and transformation of the score sheets over those years of Math4All's development.

Indeed, Math4All users were distinguishable on the basis of their relationship with score sheets. Those tasked with the production, reproduction, and maintenance of score sheets and their associated practices could often recite administration scripts unprompted.[10] The adept is able to administer the assessment with little to no reference to the score sheet itself. The amateur keeps the score sheet close by for guidance, with a trained eye scanning for important features, e.g., **boldface text to be spoken aloud**, a special icon indicating bodily movements. The novice however, may not yet be able to recognize these features as features. The carefully designed indexicalities which the formal regimentation of the sheet attempts to assert ("bold means speech!"), will not necessarily be noticed as discrete and differentiable elements, as having any effective meaning. As such, Leah's support of cognition-centered readings of the videos rests

[10] The tasks all featured repetitive verbatim prompts for each item for the sake of standardization. Having been trained on and administered Math4All tasks for my work on the BC2 study in the previous year, and having spent so much time with the PD team, I too could recite many of the prompts by heart despite not once administering Math4All during my time with Math4All proper.

upon the situated evaluation of scoring as straightforward,[11] made from position of experience within a community of textual practice.[12] That is to say, like all evaluations, straightforwardness is an ideological construal.

By contrast, the Institute staff—Barbara and Rose in particular—previously worked to support teachers' use of various instructional materials and professional development resources. Their experiences led them to quite a different conclusion than Leah: that despite being written in English, color-coded, and so on, all the assessment materials—binders, score sheets, manual—would initially be encountered as foreign objects, which would come into meaningfulness only through facilitated active engagement. Their support of an administrative reading of the videos is informed by their point of view (also situated, also ideological) that nothing about Math4All could be assumed to be straightforward.

As it happened, ample evidence of the non-straightforward nature of the score sheet emerged in the course of visiting teachers for individual coaching sessions after the introductory workshop. As the PD coaches watched teachers stare blankly and silently at the score sheets, they would gently remind them that everything that they needed to say out loud was in bold, that the materials list was at the very top of the sheet, that the color of the sheet indicated which task

---

[11] To be fair, the ability to score is not commensurate with a full familiarity with the score sheet, which also included task procedure, script, necessary materials, and so on. However, the scoring section of the sheet was inextricably attached to the rest of the score sheet. As such, the uptake of the scoring portion of the sheet cannot be assumed to be detachable from the uptake of the score sheet as a whole (cf. Keane 2003).

[12] I use the term "community of textual practice" here analogously to "speech community," as differentiable from "language community." As described by Silverstein (1996), speech communities are delimited by the distribution of patterns of speech, while language communities are delimited by their reflexive identification with or against a grammatical code, regardless of their actual speaking practices. Analytically, it is useful to have a concept which allows for the description of communities who do not reflexively think of themselves as communities or as engaging in community-bounded textual practices.

they were on. At least one coach repeatedly reassured teachers, "Everything you need to know is on the score sheet. When you start to feel lost or unsure, just stop, take a second, find your place on the score sheet, and then go from there."

It was not that the PD team had not pointed out features of the score sheet to teachers, as many aspects of the score sheets were discussed in the course of workshops. Rather many of the teachers had clearly not yet practiced before their coaching sessions, and had not spent very much time working with the score sheets. This lack of practice was an expected course of events, and indeed part of the reason that Rose advocated for scoring practice to be a part of the workshops themselves. (This is not to say following along with the score sheet would have been sufficient to guarantee teacher performance, only that it would be in line with the kind of practice necessary to improve.)

What Teachers Lack

The cases for cognitive versus administrative readings were also informed by the PD team members' differing professional understandings of the lack the workshops were meant to address.

Anna and Leah's point of view as PhD-holding psychologists was very much in line with Math4All's intended purpose of remedying a lack of knowledge. If teachers knew better, then they could do better; however, through no fault of their own, they had a deficit of knowledge about both cognitive development in general, and their students' cognitive development in particular. If teachers could come to see what they were missing, as it were, then they would come to recognize its value—in the same way that Anna and Leah did—and willingly and enthusiastically cooperate in the production of such knowledge.

Barbara and Rose on the other hand, were less immediately concerned with a lack of knowledge on teachers' part, and more concerned with teachers' lack of time. For these two former teachers and seasoned PD leaders, the problem of the workshops was not how much information could be covered, or how deeply teachers engaged in discussions about children's mathematical thinking, but how to make sure the workshops were not a waste of time. The workshops should not only be actively engaging so they did not *feel* like a waste of time, but they should also be practical so that teachers would leave with experience in using Math4All, having set up a binder and used a scoresheet.

Notably, though the cognitive case tends to sideline or set aside administrative concerns as straightforward or relatively unimportant; both Barbara and Roses' administrative orientations do not exclude learning about mathematical thinking. Indeed, in our time together over a year or so of fieldwork, Barbara, and especially Rose, had continually shown themselves to be personally quite passionate and invested in thinking about children's mathematical thinking and cognitive development. From both their points of view, an administrative orientation is at most, necessary for, and at least, complementary to, learning about cognitive development. I offer again this quote from Barbara:

> We could go on for 6 hours or 4 days about these things [e.g., mathematical thinking], but the power of the assessment, is that you can learn about them [math skills] through and with your kids, not from us, standing up here with slides. I think that that's the hook for them. I think the real way you're going to learn about this stuff is by actually using the assessment.

Conclusion

This chapter has discussed two competing visions for what textual practices will best facilitate teachers' movement to full participation in the Math4All project, marked by enthusiasm, buy-in, and faithful administration. Advocacy for cognitive versus administrative

textual practices is linked to individuals' biographical participation within specific professional communities. The next chapter will show how the PD team supported both cognitive and administrative textual practices during an actual teacher workshop. It will elaborate on how the work of regimenting of textual practices does not simply index pre-given social locations (as was commented upon in this chapter), but actively positions participants in various ways with respect to Math4All over the course of the unfolding workshop event.

Chapter Three: Experimental Vision, Faithful Administration

There were three groups of treatment condition teachers to train, each belonging to a different school network. Between the teachers and the PD team, there was only one man, a teacher. The PD team was all White, excepting myself, and all held advanced degrees. This was not the case for the teachers. Beyond demographic categories, the PD team was well-aware that we were, as a group, very unlike the teachers we were working with. Some members of the PD team remarked on the "Blue Lives Matter" signs that we would pass driving to school sites; I was personally taken aback by a memorial for aborted children outside one parochial school; and Ashley and I once visited a school on a day on which staff came dressed in military fatigues. When the November 2016 election came and went, there was speculation amongst the (anti-Trump) PD team as to whether some of the study teachers voted for Trump.

While I do not have information about the general characteristics of research intermediaries as a group,[1] there were several characteristics of the Math4All study which conditioned our distinctiveness with respect to the teachers we were working with. First, Math4All was an early childhood intervention, and as such, our participants were part of a population both less White and less educated than their K–12 counterparts.[2] Second, the two networks with the most teachers participating in the study were both Head Start–funded programs located outside Golden City, in more rural areas facing economic decline.[3] The third

[1] I do not know of any studies about the demographic characteristics of research intermediaries, professional development professionals, and their ilk, and would welcome any research on this subject.
[2] Compare Paschall et al. (2020) and NCES (2021).
[3] Head Start is a program administered by the U.S. Department of Health and Human Services to provide early childhood education and related services to low-income children and families.

network oversaw teachers working closer to, or within, Golden City, and who were relatively less distinct as a group in comparison with the PD team.

When we came together for workshops, there would be about twenty or so teachers in the room, with five to six PD team members in attendance, including myself. During the breaks, teachers tended to keep to themselves, and the PD team to ourselves. A few PD team members—typically Institute staff—did circulate and mingle with teachers during break times. Children, both those in the classroom and of one's own, were a popular topic of conversation that facilitated the participation of all parties. Indeed, these workshops would serve as a kind of proving ground for a fundamental question at the center of implementation studies and concerns about the Research-Practice Gap: Could a shared interest in the education and well-being of children overcome the multiplex differences distinguishing researchers from teachers?

This chapter will show how the very practices necessary for the production of fidelity, as a relationship of objectivity, served to re-affirm the distinction between researchers and teachers not simply along demographic lines—liberal and conservative; urban and rural; rich and poor—but as agents and instruments of science. As events of communication, these workshops briefly positioned teachers as researchers before replacing them to their subordinate position, down-and-out from the institutional locus of science.

Interestingly, this subordinate position was taken up differently depending on how teachers related to the project of making education scientific by way of academic research. The group of teachers working closer to, or within, Golden City, was notably more enthusiastic about being a part of the Math4All study than the other two. A set of merely six teachers, they were sometimes referred to as the "Golden Child" group. The Golden Child group appeared to be the only one whose members' participation was fully voluntary. Recognizing the engagement of

several of the Golden Child teachers during workshops, the PD team was the least concerned about their overall ability to faithfully administer Math4All. The members of this group—who also demonstrated the most pre-existing knowledge of academic research—tended to act as if they recognized the authoritative position of the academy as legitimate, understanding themselves as agents within a shared project of educational improvement.

While the precise circumstances of the other two groups' recruitment were never totally clear to us on the PD team, their general lack of enthusiasm around the assessment led to speculation that they may have faced more administrative or financial pressure to join the study than the Golden Child group. The networks which employed them would certainly gain a boost in reputation by partnering with Golden University and advancing cutting-edge research. These two non-Golden groups comprised the vast majority of the study participants, with those in the treatment condition numbering forty to fifty teachers in total, compared to the Golden group's six. The workshop which will be discussed at length in this chapter features one of these two larger groups, being more representative of the overall nature of teacher involvement in the Math4All study.

The Plan

Eventually, over the course of meetings of the type described in the previous chapter, the PD team decided to structure the introductory workshops around "deep dives" into three specific math skills and their associated tasks. (As Barbara pointed out, there would not be time to cover every skill, nor was every skill deserving of the same amount of time, so deep dives into a few important skills would have to suffice.). Deferring to the goals set forth by Leah, taken to be a proxy for the research side (the raison d'être) of the study, these deep dives would begin with cognitive readings of assessment videos, moving into a short, somewhat interactive lecture on

the skill centered in the task video and discussion around it. Following these cognitively oriented exercises, the teachers would be given a chance to practice administering the task to each other, with one teacher playing the student, and then swapping roles. The teacher playing the child role would follow prompts on cue cards we had developed, which offered scenarios like "Keep saying 'I don't know'," or "Take the card out of the teacher's hand."

In following the unfolding of a single three-hour workshop, I will describe three different video activities and how textual practices around the videos change as the activities and the workshop go on. In the first instance, Leah leads the teachers in the cognitive reading of a task video. Together, they discuss what the child in the video knows, mathematically speaking. Does she "know two"? Does she "know three"? This activity serves to put teachers in the same position as Math4All itself, the role of knowledge producer, to develop what I call "experimental vision." In the second instance, occurring later in the workshop, Barbara leads a group of teachers through another video viewing exercise. Now seeing with experimental vision, the teachers begin a discussion of the design of the Math4All assessment and how it might be flawed, or altered. Barbara must curtail this experimental impulse by drawing a distinction between acting as a teacher, who might experiment to find out what children know, and acting as an assessment administrator, who must remain faithful to the Math4All protocol. The third instance, coming at the end of the workshop, is once again led by Leah, who begins with a cognitive reading of a task video, before moving into an administrative reading in the course of the same viewing session. In my review these three moments, I call out the tension between the PD team's vision of recruiting teachers as co-participants in the Math4All intervention, and their charge to ensure faithful administration.

The Setting

Workshop space would typically be provided by the network that oversaw the particular group of teachers to be trained, in their regional offices. As none of these regions were particularly close to the Institute or GSE, we often carpooled to workshops, driving up to 45 minutes or so one-way. The offices were typically older buildings, made of brick and concrete, rather than glass and metal. We were often put in large basement meeting rooms, some brightly lit by fluorescent lights, others quite dim and gloomy. One such basement was large enough to have folding, retractable walls that could be moved in and out to change the size of open areas. The size of the network often dictated the quality of its accommodations, with larger networks sporting more modern office-style modular furniture, and smaller networks repurposing cafeteria-style 8-person rounds with classroom-style blue plastic chairs. One PD team member expressed concern about the air quality in the basements, being sensitive to mold and other allergens. In general, we were never sure which room we would be put in prior to the workshops, and would arrive hoping that our appointment had not been forgotten or overlooked.

With few exceptions (for paid professional development days), workshops were held in the evening hours after school, or on weekends, to accommodate teacher and PD team work schedules. This contributed to an environment in which the prevailing assumption was that no one was particularly excited to be there, teachers and PD team alike. Perhaps to counteract this unfortunate mood, the Math4All team ordered food for all workshops. Evenings were pizza, salad, and soft drink affairs; breakfast saw bagels, pastries and coffee; and lunches sandwiches, chips, fruit, and more coffee.

Experimental Vision

In the very first meeting of one of the larger teacher groups, Leah begins the first "deep dive" section with a video. She begins by describing what will occur in the video:

> This is going to go a little bit fast so I'm going to tell you what the task is. Basically, this is a task, this isn't exactly how it happens in the assessment, but it's the same idea, where the uh, adult is asking for a certain number of bears, in this case, so we're trying to see if she [the child] can produce a certain number of bears, so she's going to ask her for two, and then she's going to ask her for three.

Before the video even begins, Leah has already begun to engage in the discursive figuration of the events depicted in the video as "a task" in which "an adult" "asks" for "a certain number of bears" and the goal of which is "to see if she [the child] can produce a certain number of bears." That is, in our textual practice around this video, we should take it as an event of solicited, numerically specified, bear production between an adult and another party, which is concerned with the correspondence between the quantity requested and the quantity produced.

Other possibilities for our attention might have been the sound quality of the video, the clothes that each party is wearing, whether or not they speak with accents, and so on; but these have not been discursively highlighted as meaningful features of the video, and they are not relevant to the workshop goals. At this stage of Leah's unfolding presentation, however, the emergence of an object of knowledge, or what we're supposed to "get" from this video, is not yet fully discernable.[4]

Description over, Leah begins playing the video.

---

[4] Both "highlight" and "object of knowledge" are taken from Goodwin (1994) who writes this describing what I would consider to be textual practice: "An event being seen, a relevant *object of knowledge* [mathematical thinking] emerges through the interplay between *a domain of scrutiny* [a video] and a set of *discursive practices* [to be illustrated] being deployed within a *specific activity* [assessment training]." (p.606, italics in original, square bracket text altered) Highlighting is one of the discursive practices which he enumerates in the article.

In the video, Barbara sits across from a child at a small table, and the child has maybe ten or so small plastic bears on her side of the table. In front of Barbara is a toy house, with a large open doorway. Barbara asks the child, "Could you put two bears in the house?" The child puts one bear, and then another, into the house. Barbara softly says, "Okay, good job" as she slides the house off the table, and the bears fall out from its open base into her hand. She then replaces the bears on the child's side of the table. She asks the child, "Could you put three bears in the house?" and the child puts one bear, and then another, and then another, and then another, and then another, and then another, and then another, and then another, and then another, until she runs out of bears.

At this point, Leah stops the video, which has been running for 23 seconds. Barbara's prompt for three bears ends at 10 seconds, leaving a full 13 seconds during which the child is engaged in placing bears one by one into the house until she runs out of bears. (For those who do not often find themselves timing interactions, it may be worth setting a timer to get a sense of how incredibly long 13 seconds feels in interactional time.)

Leah asks, "So what happened first of all? What happened when she was asked for two?" The teachers, as one, respond that the child gave two. "And for three?" The teachers answer all at once, and their varied responses overlap one another.

In her questions, Leah is continuing to discursively figure the video as a series of solicitations and responses, highlighting the response portion in particular by inquiring after it specifically. In doing so, she asserts that a proper claim of "what happened" in the video—proper textual practice around the video—should be concerned with what the child does when she is asked for something by Barbara.

Leah continues, asking, "Okay, what do you think's going on? What–" she cuts herself off as teachers begin to answer. In response to this rather unspecified question, the teachers demonstrate clear uptake of Leah's discursive focus on the child by similarly focusing their responses on the child's behavior. One teacher says that the child knows that three is more than two, and so she knew to put more bears. A second teacher remarks that the child was not receiving feedback from Barbara, and so kept adding bears hoping to get some sign that Barbara was happy with her performance.

The teachers are preemptively doing some of Leah's work for her, by actively co-participating in the construction of an unfolding interaction which takes depicted child behaviors as signs of cognition. As we know, the child's mathematical thinking is precisely what Leah hopes to present as the proper object of knowledge in this exercise of professional vision. Helpfully, the reading of mental states from behavior is already a common textual practice in American culture outside of research psychology more broadly,[5] and so the teachers take quickly to collaborating in its emergence in this interplay of video and video commentary.

By following the real-time temporal unfolding of this viewing session, we will be able to witness the effective meaning of the video events change over time, as previous viewings and discussion of the same artifact become the context within which future viewings take place. For example, Leah decides to replay the video based on the second teacher's contribution, discursively highlighting *the child's attention to Barbara* as a central concern. Leah takes up the teachers' suggestion, "So she looked up at her [Barbara]? Okay, let's just watch that again."

[5] Andrejevic 2010 discusses this tendency to read the body as an index of inner states (often to bypass the untrustworthy character of language!) in American popular culture.

This second viewing of the video is no longer the same as the first. For all involved, their textual practices around the video have been transformed by all that has been said up to this point. The first viewing and the attendant discussion are now context for the second viewing. During this second viewing, as the child puts two bears in the house, Leah says out loud "that's two," marking the first of two now-established events of solicitation and response. The video continues with Barbara's prompt for three. The child begins adding more and more bears to the house, and though the teachers initially watched this portion in silence, the second time through, they begin talking amongst themselves as the video continues, presumably about the discursively highlighted issue: Is the child looking up at Barbara?

When the clip ends, Leah too returns to the highlighted concern, affirming the teacher's hypothesis and building on it: "So she's sort of looking up. Why do you think she's looking up?" Teachers suggest that she is waiting for positive feedback, and Leah echoes back, "So she's waiting for, okay" as another teacher hypothesizes as to what would have happened if Barbara had said something to the child.

Leah does not take up this contribution, but instead rather adeptly redirects discursive focus to the object of knowledge which this entire exercise has been directed at producing: "So what does she not know, or understand?" The teachers respond, "Three." Leah continues: "So she doesn't know what three is, do you think she knows what two is?" There is general agreement, but one teacher seems to not be so sure. Leah: "You don't think she knows what two is? Why do you say that?" The teacher equivocates, but wonders if Barbara said something to make the child stop at two bears. Leah responds, "Okay, you think she said—Let's just watch it one more time." As she turns to restart the video, she adds, "Sorry, I like watching videos again."

Upon this third viewing of the video—again, a new event of "seeing" shaped by the unfolding of the interaction up to this point—when the child puts two bears in the house in response to Barbara's prompt, multiple teachers say out loud, "she stopped, she stopped." Leah too, echoes, "she stopped." There is a murmur as teachers talk amongst themselves for a second, before Leah asks, returning to the child's mathematical knowledge as the emergent object of knowledge: "Does anyone think that *she didn't understand two* and kind of did it by chance?" (emphasis added).

At least one teacher confirms out loud that the child understands two, but another voice comes up with a new observation: Barbara smiled. A few teachers begin discussing this suggestion amongst themselves, before Leah cuts in, taking up the new highlight in order to frame it as yet another way into discovering the child's math knowledge: "Okay, alright, let's pretend she—let's pretend she knew three really well, she understands three. Would she be looking up at her [Barbara] and looking for the smile?" The teachers agree that this is unlikely. Leah continues, "No, she'd probably just put it right in and stop." She then proceeds to make explicit what the video should be taken to mean: "So we think she probably understands two, she doesn't really understand three." She continues: "And um, what do I mean when I say 'understand three'? What does that mean?", segueing into a presentation on the concept of cardinality.

What we see in the presentation and discussion of this video, then, is the unfolding, interactive, discursive entextualization of assessment videos as opportunities to discover and hypothesize as to the state of children's mathematical knowledge, based on a reading of their visible behaviors as signs of cognitive development. The screening of other task videos during the introductory workshops saw similar framing, especially in the form of prefatory prompts like,

"While you're watching the video, we want you to think about what kinds of mathematical thinking you might be seeing," or "I want you to just look at [child's name], and just see what you notice, okay? Keeping in mind, we're trying to understand her, uh, what she does and doesn't know about cardinality."

Anna and Leah, themselves psychologists by training, both very much enjoyed the opportunity to share with teachers the practices which brought them into psychology to begin with: observing and hypothesizing as to the meaning of children's behaviors with respect to their mental processes. Recall their appreciation of the "richness" of these videos, Leah's self-described proclivity for watching and re-watching the videos. In the above example, teachers themselves appeared to be quite engaged in such discussions and re-watchings, given their active, often unprompted, participation in these workshop segments. Not only were they responsive, but their suggestions were, at times, the impetus for replaying the video and the framing for subsequent discussion.

I have included so much of the back-and-forth between the Leah and the teachers not only to show the discursive regimentation of video uptake, but to show how, in the course of being entrained in the professional vision of researchers—what I call "experimental vision"—teachers exhibited a tendency to engage in one of the hallmark practices of researchers: hypothesis formation. In my analysis, I myself discursively figure teachers' contributions as "hypotheses" when they offer explanations as to why the child might have displayed the behavior she displayed, or how she might have behaved given some other set of conditions.

That is, experimental vision, or any form of professional vision, exceeds merely "seeing" or "reading." Said practices arise from certain social locations, and are thus imbricated with other practices characteristic of those locations. This imbrication is precisely what produced an

unfortunate difficulty for the Math4All PD team during the workshops. As teacher training progressed, the PD team began to be wary of how the impulse to hypothesize might interfere with their objective of ensuring faithful administration. Would the teachers break with administrative procedure in the name of experiment?

Curtailing Experiment

The below exchange comes from later in the same workshop excerpted earlier, with the same group of teachers. In the course of discussing a task item which asked children to point to a picture with the correct number of ladybugs, a few teachers ask Barbara how she had scored the third item, as there was some ambiguity as to where the student was pointing in the video. After answering their question, Barbara begins to tell a story from the Math4All workshop she had worked last week, with the Golden Child (GC) group.

Barbara: "Somebody in a training that we did last week said—and I just wanted to point out, in case you were thinking of it. *Hm, I think if those had been in a different order she would have gotten it right, if we had switched the three and the four.*" Barbara, speaking as the teacher, is referring to the order of the test items in the ladybug task, which ask students to identify first two, then one, then four, then three ladybugs as they progress [2,1,4,3]. The teacher in Barbara's example hypothesizes that if the test items asked for three before four, the student in the video would have gotten the last two items correct.

Barbara continues, "Maybe she [the child] would have [gotten it right], but could we change the order? What do you think our answer to that is?" Immediately the teachers responded with a chorus of "no"s. Barbara asks them, "Why not?" and multiple teachers begin to give an flurry of overlapping and largely indistinguishable responses.

Barbara responds, “Okay, so I couldn’t hear everybody, but I could hear you [a specific teacher] and I think it [the crowd’s response] mirrored that [specific teacher’s response]. Actually this [ordering] helps us better see how robust her understanding is. She really understands it when it’s in a different order. Because they always see them in the world in order.” Barbara is articulating the design rationale of the task: the reversed order of the test items, [4,3] rather than [3,4], as a purposeful strategy to test the “robustness” (i.e., decontextualizability) of the child’s understanding of those numbers.

Barbara continues, “So this is one of those cases where our teacher instinct, in teaching, might be, *If we show them in order that might help them build the understanding. It will scaffold it, it’ll be right there.* But for our assessment purposes, this order, even though it seems harder, um, is definitely intentional and pretty important to follow. It actually might do something different. So I think it was a really good question the teacher asked last week. I was super glad she asked it, and I wanted to raise it here.”

In this stretch, Barbara hits a theme which she raises several times over the course of the workshop, a role distinction which she sometimes refers to as wearing one’s “assessment administrator hat” versus one’s “teacher hat.” The women in the audience should use their experimental vision to recognize the intentional design of the task, and as such, should suppress their teacher impulses and act as administrators.

But again, in seeing like a researcher, the impulses of the researcher also arise. A teacher continues the conversation, asking if the training items (*not* the test items) for this task had been in order: [1,2,3,4,5]. Barbara responds that the purpose of the training items is to show how to do the task, to be “as supportive as humanly possible.” The teacher clarifies her point, formulating the hypothesis that perhaps the training items confused the student, because she was initially

exposed to the items in the order [3,4]. Barbara echoes back, "That's a very interesting thought, maybe she was following the order." The teacher restates her point, that perhaps the student was simply following the original order that she had been trained on.

Barbara continues, "That's a reasonable theory. I mean we did have two and one reversed here, so we sort of disrupted the order from the get-go. But I do think when kids learn this, it's really an ordered thing. So it's really great to *think* about the possible reasons, but again, it sort of measures that robustness of understanding."

This conversation is picked up by another teacher, who begins likening the situation with how the alphabet is taught, and so on. Barbara again affirms the teachers' thinking, and again ends by reiterating the purpose of the test item ordering: to "measure the robustness of understanding." Her refrain must always return to justification of the current system, even as the teachers extend their critical thinking beyond the limits of their station. In emphasizing that it's great to *think* about possible reasons, Barbara is in a sense indexing the inability of the teachers to *do* anything about their hypotheses. They cannot change the design of the task, and their critique of the training items is unlikely to be reported upwards. The Math4All research team is interested in teachers' feedback regarding the logistics of assessment administration, not experimental design.

In making the above point about administrative fidelity, Barbara cited an interaction from a previous workshop as a potential site of experimental infidelity. During that prior workshop, for the Golden Child group, the teacher who brought up the ordering of test items *did not* suggest that the test items might be swapped in the course of administering the assessment. Rather, she and her colleagues were also questioning the effectiveness of the training items in conveying the mechanics of the task to the child. Further, the teachers in the workshop featured in this chapter

were also immediate and unanimous in their agreement that the test items could not be swapped. By and large, the teachers all evinced a clear understanding of the limits of their administrative role within the Math4All project; however, having been guided to see like fellow researchers during the cognitive aspects of the workshops, they took seriously the commitments of that station too.

It was true, however, that earlier in this workshop (not the GC workshop), a teacher had asked if she was allowed to repeat a test item to make sure that a child understood. Leah, who was leading that session, noted that this would be a good strategy "in general," but when pressed again by the teacher whether it could be done during the assessment, she answered, "No, not in our assessment." In this light, we might understand Barbara's invocation and re-framing of the item-ordering question to be a 'just in case' tactic, a vigilant hedging against any possibility of what would certainly be a fatal mistake. Whether or not Barbara harbored any actual doubt of the teachers' understanding, she was employed to ensure the objectivity of the Math4All study via its faithful administration, and in such a position, she attempts to cover all potential sources of infidelity.

Faithful Administration

In general, the discussions in which teachers are able to engage in hypothesis formation were the liveliest moments in the workshops. This was in contrast to lengthy silences from teachers when the workshops tended towards a more regulatory mode of engagement, in which PD team members would expound on proper administration techniques. That is to say, if leading teachers in cognitive readings of task videos enlivened teachers' experimental curiosity, turning to administrative readings tended to emphasize the incompatibility of experiment and fidelity.

For the final deep dive of the day, Leah plays a task video for teachers, with instructions to focus on the child's behavior and thinking. Following a teacher's joke about the child being tired of the task, Leah and the teachers discuss what the child seems to know. At one point, they are so simpatico in their textual practice around the video, that the teachers and Leah seamlessly finish each other's sentences.

Then Leah plays the video for a second time, this time asking teachers to pay attention to Barbara's behavior, an administrative reading. She begins, "We're going to watch it again, and I want you to look at Barbara and see what she's doing."

She starts the video again, and when it stops, she asks, "So what, what is Barbara doing dur[ing]- in the-" A teacher says something about Barbara's gaze moving back and forth. "Uh huh," Leah backchannels, as the teacher trails off, and Leah starts to speak, "Yeah, she's sort of just remaining really neutral, just like, hanging out."

She voices Barbara, "I'm just going to wait til you're done, you know, you tell me, I'm going to like maybe pretend, I'm busy over here, until you tell me you're done, you know?"

At this point, Leah begins alternating between playing the video and going on at length about proper administrative disposition:

> Sometimes it helps not to, not to watch the child, or look at the items that they're doing, because you might accidentally like, sort of, you know, do— uh, smile or something when they get it to the right number, just sort of don't pay attention, be patient, they'll let you know when they're done.

Between the two viewings, teachers are encouraged first to do a cognitive reading of the video, before being led into an administrative reading. In the first, a lively back-and-forth occurs, in which Leah and the teachers are able to literally finish each other's sentences. Over the course of the second viewing however, Leah stops asking questions and begins to monologue about what good administration looks like. She actually goes on for quite a while after what I have

included here, eventually working her way back to connecting administrative best practices with producing knowledge about children's cognitive development.

But the problem with emphasizing how teachers can turn attention to student behaviors into knowledge about their cognitive development ("so she knows two") is that, in the Math4All system, teachers are not meant to translate events of assessment into student knowledge. That's Math4All's job. The teachers' job is to feed Math4All the data it needs to make student knowledge, through their faithful administration of the assessment.

Oppressive Administration?

My point is not that administration is an innately constraining or mechanical activity; or that centering administration must always be at the expense of exploration and experiment. Rather, administrative talk appeared oppressive during these workshops because discussions of administration tended to arise as correctives for the vision imparted by the cognitively centered activities. Events of assessment were first actively produced as sites of rich interaction by the cognitively oriented activities which began the workshop. Teachers were initially encouraged to become co-participants in the work of Math4All itself, the work of making knowledge about children's thinking. Only once teachers' experimental vision was recognized as a possible source of infidelity did the PD team resort to administrative discourse as a mode of warning teachers off of potentially inappropriate behavior. In this light, being relegated to administration becomes a demotion, an admonition.

In fact, all the workshops began with a number of administrative mundanities necessary to familiarize teachers with the scope and aims of the Math4All project. Yet, under Barbara's able facilitation, these segments still felt engaging, alive, and miraculously, even her use of question-and-answer routines around mundane logistical information in order to actively engage

teachers in learning did not feel hokey or patronizing. I was not alone in my admiration of Barbara's abilities—the entire PD team was rather awed by her facilitation prowess, and we all expressed some nervousness when we would have to run a workshop without her in attendance. It is unclear what an administration-forward workshop, in line with Barbara and Rose's proposals, would look like, or if it would have had a different effect on teachers' enthusiasm, investment, buy-in, fidelity, and so on.

What is more apparent, however, is that the quality of the workshops themselves and their effects on all of the above would not have any effect on teachers' working conditions, that is, the context in which they would be expected to implement the assessment. Even the most enthusiastic of the GC teachers had difficulty fitting the workshops into their schedules, and professed not having spent much time with assessment prior to their coaching sessions. Math4All's solution to the problem of not enough time, was to use item response theory to produce as streamlined an assessment as possible, to reduce the amount of time necessary to administer the assessment to each child. Ironically, this meant that teachers were required to administer the entirety of the assessment in order to get any feedback from the Math4All website, which required every data point from the streamlined assessment to produce any student knowledge at all. That is, researcher attempts to design a more feasible administration were directly tied to its rigidity.

Conclusion

Despite the inclusion of "professional development" in their name, and despite securing teachers compensation for professional development, one could argue that the PD team's charge was not professional development per se. As set out in the design, the Math4All teacher-facing events could not practically be expected to prioritize the building of teachers' professional

capacities. Rather, given the limited amount of time allotted, fidelity of implementation emerged as the central priority of teacher workshops and coaching. As this chapter shows, this focus usually ended up foreclosing, or even explicitly forbidding, other potential avenues for the professional development of teachers. The PD team members evinced different levels of awareness around this irony, but none wavered in their efforts to carry out the job that had been set in front of them: supporting fidelity of implementation.

Ultimately, this chapter points to a fundamental mismatch between how teachers were trained to become a part of the Math4All vision, and the role that they were expected to play as part of that vision. The "cognitive" strategy of teaching teachers to see like Math4All did not end up making teachers more sympathetic to the rationales underlying the strict procedure that they were expected to follow. Rather, in putting teachers in researchers' shoes, as it were, it provided teachers an opportunity to imagine how they would go about being knowledge producers, and to critically engage with the design of the experiment they were participating in. When teachers tried to follow through on the practical entailments of experimental vision, they were reminded of their actual role as instruments, and not agents, of experiment.[6]

Where some researchers have puzzled over the cooperation of people who see the 'same' thing in 'different' ways, this chapter's analysis suggests that seeing the same thing in the same way is no greater a guarantor of rosy cooperation (cf. Star and Griesemer 1989). Here, learning

[6] Though it was suggested that teachers could take notes on the assessment score sheets to inform their practices "in general" (i.e., "not in our assessment"), it was also acknowledged that they were unlikely to have much time to take notes while simultaneously minding the child's behavior and their own administrative practices. It was suggested that when they had their teacher hats on, and not their assessor hats, that they could "take risks" in instruction. However, instructional support mostly turned on facilitating the faithful use of assessment results, another site of regimentation which I do not discuss here.

to ‘see the same thing’ in ‘the same way’ does not put everyone on the same footing, but rather clarifies the differential role relation of teachers and researchers as organized by fidelity. Fidelity serves not only to recapture teachers’ work for the instrumental growth of Math4All in the form of student knowledge, but to recapture teachers’ work toward the objective production of evidence about the efficacy of an abstract entity, “the” Math4All assessment.

Part II. Reliability

It is March 2017, and the observation team I am a part of is in the midst of final preparations before being sent out on "reliability visits." Our trainer tells us that a reliability visit will look like this: Two team members will pair up and observe a classroom. Each will observe and record moments of mathematics instruction, each on their own tablet device. As observers, they are expected to avoid interacting with teachers and students alike, except to introduce themselves to the teachers, who should already be expecting them.

Of the pair, one of the observers will act as the lead. The other observer will follow the lead's movements around the classroom. The pair will move as one between activity centers where students fight over magnetic triangle toys, stack wooden blocks as high as they can, or sort bears by color. Even when the students move out to the playground for recess, the observers will go with them outside and listen in on exchanges between teacher and students about rocks and ice cream, hurriedly tapping their fingers against their tablet screens, taking notes.

The existence of a lead observer is necessitated by the expectation that both observers attempt to record and code 'the same thing,' that the pair are attending to the 'same' events. If there were no lead to set the mutual point of observation at any given time, one observer might make notes on block-stacking at one end of the classroom, while the other taps away about children playing at the sand station at the other. The congruity of the observers' attention is critical to the objective of the reliability visits: to ascertain the degree of agreement between the two observer's records, and hopefully, to confirm that the degree of agreement meets or surpasses the minimum threshold of professional scientific legitimacy known as "reliability." In this case, the threshold is ≥90% agreement on each of two consecutive reliability visits.

"Inter-observer reliability"[1] is used to describe the state of affairs in which independent observers assign the same codes as one another, to the same observed phenomena, almost all of the time. Some amount of disagreement is always expected, hence the 90% threshold of acceptability, rather than 100%. In fact, in the regime of mechanical objectivity which reliability participates, the accrual of measurement errors is not a bug but a feature. In the development of statistical thought during the 19th century, statistical methods were commonly referred to as "the combination of observations" (Stigler 1986, p.11). In astronomy, the true location of a fixed object like a star was determined to be the mean of the distribution of errors across observations. Errors were not only normal, in the lay sense, but they were "normal" in the statistical, Gaussian sense: in combination, they produced a normal curve, a bell curve. The center of that curve, the value around which all observations fell in a regular, symmetrical distribution, was the "true" location of the star.

In our classroom observations, the position and attention of the lead observer would index the fixed object with respect to which observations should be produced. The "star" of the classroom was the "event of instruction (EOI)" indexed by the lead observer's attention. Inter-observer reliability was the condition for the claim that given some agreed upon object, our numerous observations could be combined in producing enough noise to discern a true signal. However, the classroom scenario and the star example have a major difference. In our classroom observations, we were not in the business of observing any single star or event, but a series of events that we claimed to be fundamentally of the same *type*.

[1] There are analogous phrases for different, or more specific, data practices: interrater reliability, intercoder reliability.

As such, a necessary prerequisite for our achievement of reliability would be our intersubjective calibration with respect to our object, that is, discerning EOIs and their attendant properties. If the Math4All statisticians wished to draw conclusions across a whole set of independently conducted classroom observations, they must be sure that the observers were describing 'the same thing' within and across each observation. Correspondingly, the most heavily weighted metric of reliability within its algorithmic determination was the number of such EOIs recorded. Our ability to reliably discern the number of stars in our patch of sky was taken as an indicator of our ability to tell what was a star to begin with.

Producing intersubjective alignment is a lot of work, and it is work that occurs through events of discursively regimenting of textual practices, that is, events of talking about how to break apart the world into meaningful pieces.

"Getting Reliable"

We were conducting classroom observations in order to collect data how teachers taught math and how much math teachers incorporated into their overall classroom practice. The analysis of this data across treatment and control conditions would serve to detect any effects of teachers' use of Math4All on their instructional practice. That is, the data we collected would serve as evidence of Math4All's efficacy. My analysis here is not at all concerned with whether such evidence was generated or the quality of said evidence, but with the activities of "getting reliable" which grounded claims about the validity and reliability of the data to begin with.

"Getting reliable" is the phrase we used to discuss the achievement of inter-observer reliability while we were in the process of preparing for classroom observations. One could say, "How long will it take us to get reliable?" in reference to the reliability of the team as a whole, or "How long will it take for Lily to get reliable?" in reference to the reliability of a single

individual.[2] A person could also *be* reliable, so, in addition to discussing who had already "gotten reliable," we discussed people who "were reliable," or described ourselves as "reliable." The word "reliable" seemed to be constantly on our lips during the months of January and February, as we were immersed in the process of getting reliable.

When I would use this same language in describing my fieldwork to friends back home in Chicago, they would react differently based on their disciplinary training. Anthropologists used to only ethnographic fieldwork would balk at its grammatical strangeness. As one workshop attendee commented, "You *get* reliable…? I don't like that." Educationists, psychologists, and others with lab experience would laughingly commiserate about the unfortunate familiarity of the whole ordeal. Indeed, reliability is rarely accorded a formal role in authorizing the professional status of sociocultural anthropology, while it is an explicit professional standard in much of psychological and educational research (see AERA 2006, p.36).

This is because reliability is central to the scientific legitimation of so-called "soft" sciences like education and psychology, which are accused of being insufficiently objective, and, crucially, wish to refute such accusations. By contrast, the American Anthropological Association removed the word "science" from its mission in 2010, reportedly much to the chagrin of archaeologists and biological and linguistic anthropologists (Glenn 2010).[3]

---

[2] This example caused some curiosity among early readers as to whether I was perceived as particularly incompetent. It is merely an example of a question that might be asked of anyone, for a variety of reasons—positive, neutral, and negative. I have used myself in the example precisely to direct any potentially negative implications solely at myself. I will discuss how we made sense of potential or actual failures of reliability later on.

[3] Earlier in my training, I was surprised to find linguistic anthropologists listed among those who voiced their displeasure with this move. Now several years later, I find myself assuming that it was representatives from "my" group, the Silversteinian line of linguistic anthropology, who were against this move, seeing their work as part of a unifiable scientific project not at all divorced from the work of physicists, chemists, biologists, psychologists, and so on.

Meanwhile, many prominent education researchers are invested in establishing the scientific nature of their enterprise, sincerely and deeply invested in the belief that objectivity is the best way to drive real improvements to education, especially in lessening disparities across race and class.

In the post-NCLB period, reliability is a key social relation by which Research gives definition to itself, organizes itself, within the SBR regime of objectivity. I suggest that we can arrive at a better understanding of the Research-Practice nexus by developing an understanding of how Research itself comes into distinctness as a domain. Getting reliable is not the only practice through which Research distinguishes itself; it is only one type of activity among a whole interconnected range of activities through which researchers claim a discernable 'research community,' and in so doing, constitute themselves as members of said community.[4] However, reliability work is a particularly rich site for thinking about the formation of a scientific community because reliability describes the fundamental nature of a collective: a group of individuals who orient toward 'the same things' in 'the same ways.'

Departing from typical investigations of research use which look at the uptake of research by practitioners (e.g., Coburn et al. 2009, Honig et al. 2014, Penuel et al. 2018), this latter half of the dissertation discusses the uptake of research by other researchers, offering an alternative model of research use to that described under fidelity.

[4] This focus on the professional standard of reliability as a mediator of the orderliness of particular, interested, events of cooperation differentiates the linguistic anthropological approach from an ethnomethodological one, which, in its analytic eschewal of formal rules and regulations, makes it difficult to account for the role of reflexivity in mediating social continuity and change. See, for example, the explicit refusal to engage with reliability in Olszewski et al. 2007, an ethnomethodological study of coding.

Reliability as a Sign of Objectivity

Reliability is a product of methods standardization; it describes the reproducibility of measurement. Science studies scholars have written about (methods) standardization as an enabling condition for cooperation (Star & Griesemer 1989) or a ubiquitous part of the "infrastructure" of "everyday life" (Lampland & Star 2009). Here we will be dealing with the means by which standardization occurs in-and-through events of interaction. How does standardization happen in phenomenal time, through personal, agentive acts? While much has been written about the conditions and consequences of scientific standardization in various historical moments (e.g., Latour 1993, Porter 1995, Bowker & Star 1999), less has been written describing the real-time, contingent unfolding of actual events through which these states-of-affairs and their attendant consequentiality are achieved, and through which they may be continually contested.

In the context of education research, reliability is an important sign of objectivity, with objectivity itself taken to be the legitimating condition of any scientific claim. I suggest that in order to understand the separation of Research and Practice in education, it is instructive to consider the activities through which Research constitutes itself as coherent. Rather than presume the existence of distinct communities of practice which form relatively well- or poor-fitting contexts within which professional standards like reliability may operate (Lampland & Star 2009, pg. 7), my analysis takes communities and standards to be co- and re-produced in innumerable linked moments of communicative practice. In this view, every achievement of reliability participates in the constitution of Research; and "getting reliable" is a form of boundary work in which the scientific domain of Research is positively defined in the repeated and distributed enactment of objectivity (cf. Gieryn 1983).

The cases of reliability talk that I will analyze serve the purpose of demonstrating (1) the discursive work of regimenting textual practices and (2) how those discursive acts serve as acts of community formation, that is, of *communication*. In this case, the communicative achievement of reliability participates in the reflexive formation of a distinct domain of Research in American education. To speak of "the work of reliability," then, is to point to both the communicative work required for the achievement of reliability, and the work of community formation which the achievement of reliability mediates.

Chapter Four: Training and Trust

In January 2017, I had made my way onto the team of individuals hired to conduct classroom observations for the Math4All study. The collected observational data was to be used as a Math4All outcome measure, providing data on changes in teacher practices as a result of using Math4All.

The observation team comprised existing Math4All researchers, full-time and part-time, plus a few new RAs recruited from the Math4All PIs' other projects. I had already been working with the teacher support side of Math4All and learned about the classroom observations from Leah and Ashley, who worked with both the PD team and "the psych side" of Math4All.

Leah was the Math4All project manager, a petite White woman who alternately radiated urgency and exhaustion. This was perhaps unsurprising as she was situated in the dead center of the Math4All project, supervising psych side RAs, liaising with the website developers, and working with the PD team, all while reporting directly to the PIs, keeping them abreast of developments, soliciting their feedback, and carrying out their instructions. Besides the PIs, she had been on the Math4All team the longest, over five years at the point of my appearance. The rest of us—including Ashley—expected to move on to new and different projects in the coming years. A few of the observers were not even primarily RAs for Math4All, only temporarily joining Math4All in order to assist with classroom observation.

Ashley was Leah's right-hand man. She was a White woman a decade or so younger than Leah, completing her dissertation on teaching styles at GSE. She had been hired the same summer that I joined the PD team. Ashley was most concerned with the logistical aspects of the study. She rarely spoke during meetings—taking notes with pen and paper rather than on a laptop—but when she did, it was usually about the practical feasibility of a course of action

given when pre-testing needed to start, or the number of cars available, or how long it would take for the materials to get back from the printer.

Both Leah and Ashley were concerned about the tight timeline for observations—they needed to be completed before the end of school, while still leaving enough time for post-testing—and so I was able to take advantage of this time pressure in volunteering myself as an observer. Not only would an additional observer speed along the process, but I also came with the added benefit of not affecting the Math4All budget. I would work in exchange for the ability to document the experience for my own research.

The Chain of Reliability

The observation tool chosen for the task was developed by the Classroom Observation Group (COG) at Empire University, several states away from GSE. COG had been developing their tool over the course of many years, continuing to modify it over the course of their collaboration with various research groups interested in collecting classroom data. As such, they were well-versed in the difficulties and possibilities of both classroom observation and training strangers in classroom observation.

Before any training began, the Math4All observation team, including myself, received several emails from Leah. The first email came with several attachments: an instruction sheet for downloading the COG observation tool onto a tablet; a paper version of the tool; the COG observation system manual; an agenda for the upcoming training; and a document containing several "practice scenarios." We were instructed to review these documents before training began the next week, and for those of us with personal tablets, to take some time to play around with the observation tool.

In a typical training arrangement, Math4All would have paid for COG staff to be flown into Golden City, to conduct an in-person training with our observation team. Then a few key members of our team would become "reliability anchors" by going on a series of paired classroom observations with COG researchers and demonstrating reliability.[1] These anchors would then serve as the baseline against which the rest of the observation team would get reliable. At the end of this process, we would all be anchored to one another along a chain of reliability: the COG trainer anchoring the Math4All anchors, who would then anchor the rest of us.

However, at the point of classroom data collection, Math4All found itself short on both time and money, so the PIs of COG and Math4All came to an alternate arrangement. COG staff would conduct team training remotely, and Math4All would fly just Ashley out to Empire City to become our reliability anchor. Ashley was not only universally perceived to be highly competent and dependable, her personal and professional situations allowed her the flexibility to travel. Leah was likely the only other good candidate for the trip, but she had too many responsibilities to spend time away from Golden City. Instead, Ashley would train Leah to be our second anchor upon her return to Golden. Then, the two of them would split the responsibility of going out on reliability visits with the rest of the team. Thus, despite the limitations to the observation team's travel, the chain of reliability could still be forged.

The air travel Ashley undertook in order to forge a link between the COG and Math4All teams appeared to facilitate another kind of travel, the travel of the COG observation system itself to Golden City. In this chapter, I will show how the "travel" of the COG system was

[1] Again, reliability required a comparison of their records over two consecutive observations to each score over 90 percent on COG's reliability algorithm.

mediated not only by the reproduction of its material fractions—a mobile application, a series of Word documents—but in the reproduction of a community of users entrained in specific textual practices in their encounters with those materials. That is to say, the travel of the COG system—as a way of knowing about how teaching works and what teachers do—can be understood as a phenomenon of social reproduction. Each link in the chain of reliability represents an achievement in socialization, as enacted through the discursive regimentation of textual practices. That is, the COG system traveled along lines of reliability, and that reliability came about through events of talking about how to meaningfully engage with various artifacts in the COG way.

Anxieties and Assurances

As anyone who has been part of a lab or research group in the midst of qualitative coding can tell you, reliability talk often stokes conflict and anxiety. I would guess this is the reason for the dearth of scholarship on actual coding practice. Researchers simply prefer to keep its inevitable messiness in the family, as it were. In opening up the difficulties of the coding process, one might open oneself up to accusations of messy science or messy scientists. Here I assert the opposite, that the messiness of the coding process is a necessary part of producing good science. That messiness is both the foundation upon which and the material with which reliability is built.

As Sanders and Cuneo (2010) note in their insightful self-study, not only are disagreement and negative emotionality highly expectable occurrences in coding talk, but they can also aid in the achievement of reliability. The central example in this chapter will deal with an affectively fraught moment of disconnect that occurred during training between the COG and

Math4All groups, and build on Sanders and Cuneo's insight in showing the process by which disagreement and dis-ease become the proving ground for a trusting relationship of reliability.

As previously mentioned, the tight observation timeline was a highly salient source of anxiety for Math4All. Getting reliable, in particular, was expected to be an especially time-consuming phase of the observation process. Indeed, time to reliability was the first matter of concern to appear in our very first COG training.

The training—the first in a series of three—begins generically. Our COG trainer, Andrew, introduces himself, and we offer our own round of introductions, first those in the office—Leah, Ashley, Caitlin (another RA), and me—and then those calling in from other offices and from home. Leah gives Andrew a brief rundown of Math4All, and Andrew describes the training process will look like.

We begin with three days of training sessions with Andrew. Then we will go out and practice in classrooms not taking part in the Math4All study. At the same time, Ashley will fly out to for additional in-person training with COG to become our "anchor." Upon her return to GSE, she will "do the same thing" [paired reliability visits] with each member of the team. Each of us will be deemed reliable if and only if we can demonstrate our reliability with respect to our anchor, Ashley, on two consecutive occasions.

"The goal," Andrew repeats, "is that we're all rating things the same way, so we can try to keep everybody on the same page in how they're interpreting things in the classroom, and how they're using the system." Like a good seminar host, he pauses and asks, "Any questions at this point?" Leah, our project manager, chimes in: "I just have a question, everybody will be … reliable … within two reliability visits? Is that true?"

In listening back to my recording of this moment, I take note of Leah's halting delivery. As I quote her here, I use ellipses to mark distinct pauses before and after the word "reliable." The symmetrical pauses are themselves nested between the two bookends of "I just have a question" and "Is that true?" In the moment, and in re-listening, I am sure I know why Leah has asked this question in this way. What if getting reliable took too long? Or, even worse, what if we were not able to get reliable? The poetically hedged nature of Leah's question indexed, to me, her keeping a wary distance from the promise of a quick two-visit reliability.

Regardless of the accuracy of my reading, time was certainly a major concern for Leah, who was ultimately responsible for the on-schedule completion of project activities. Leah was more heavily invested in the success and longevity of the Math4All project as a condition of her continued employment stability and financial security than any of us RAs. Classroom observations had been promised in the same grant proposals that funded her full-time position with Math4All; and 90+ observations needed to be completed within a few months' time.

"I just have a question. Everybody will be … reliable … within two reliability visits? Is that true?" "That's the expectation," Andrew responds. "Okay," says Leah, and she continues to feel out the situation, "so it doesn't usually take anyone longer than that?"

Here I laugh a little, having just spent many months developing and getting reliable on a coding scheme for the Institute's Better Curriculums for Better Classrooms (BC2) project. As I understood it, and as I blithely assumed more experienced researchers like Leah must have understood it, reliability was a process which always took longer than planned. Not only that, but the COG system appeared to encompass a far more complex coding scheme than any I had ever used. From my experimentation with coding schemes for BC2, I supposed that the more complex the coding scheme, the harder it would be to get reliable. It would also be my first time coding

from a live observation, not from more persistent materials like digital transcripts or video. There was no telling what degree of difficulty the ephemerality of our object would add to the process.

Our trainer Andrew does not laugh. He responds earnestly to Leah: "So the hope is that you can [get reliable in two visits]. We add more if needed, but hopefully, um…" he pauses before settling on a definitive "That's the goal." He goes on to provide some context for the almost unbelievable expectation he is giving voice to: "The hope is that we can help you guys train this month, and we're hoping you can send us—when you go out to practice classrooms—send us at least your first um, uh, notes from that observation for us to look through and give you specific feedback."

COG will help scaffold our reliability, not only through training, but through the discussion of the notes we take during practice observations. Not only that but, "When Ashley comes here, we're going to show her how to check people's data so that we can help the rest of you guys make sure that you're doing it right—or at least doing it the same way." That is, the activities which make up classroom observation are textual practices—practices of breaking up observed phenomena into discrete notes and codes. COG's primary function in supporting reliability will be in regimenting our ability to turn experience into text, and to train one of our own to perform that same regimenting function.

Andrew concludes: "The hope is, if it takes two [reliability visits], if it takes three [visits], but that's the, that's the expectation." "Okay, sounds good," Leah says.

Risks of Reliability

Reliability is not just a site of anxiety for those of us on the Math4All observation team, it presents a potential site of concern for the COG team as well. We are certainly not the first outside group that they have trained on the system, and not even the first group that they have

trained remotely. (They already have enough experience to know that video chatting with so many trainees is not wise, hence our audio-only training set-up.)[2] However, anytime a new group is trained on the system, reliability re-emerges as a site of possible conflict.

In the worst case, a failure of trainees to get reliable could open up questions about the validity of the tool itself. One major reason Math4All chose the COG tool (besides time and money) was its validity. It had already undergone processes of formal validation. By showing that measurements taken by the tool were able to distinguish between classroom outcomes, or some similarly known measure, the tool was understood to be measuring what it claimed to measure. In the common explanatory metaphor, validity is concerned with the closeness of darts to a target, while reliability describes the closeness of the darts to each other.

The ability to bracket away validity—to say that it had already been taken care of—was one of main affordances of using an externally developed tool to collect data, as Math4All was not principally concerned with classroom observation as a research problem in and of itself. However, if the Math4All team was unable to get reliable, the question of validity might be re-

[2] This all occurred several years prior to the COVID-19 pandemic. Remote conferencing, especially for a training-type event, was not yet as highly conventionalized as it became once remote work became the norm for white collar professionals like academics. Instead, at best, one simply hoped that the host organization was sufficiently familiar with the conferencing software that it had chosen. In this case, COG was very well-versed in their conferencing platform, though I have no memory of what platform was used. It appeared to be blandly corporate enterprise software, likely provided as a matter of course by Empire University's IT department. In early 2017, Zoom was not yet ubiquitous in universities and among the general population, though businesses, universities, and frequent tele-conferencers were quickly migrating toward it in the years prior to COVID-19. (https://usefyi.com/zoom-history/, https://www.insidehighered.com/blogs/technology-and-learning/zoom-hot-higher-ed, https://www.forbes.com/sites/alexkonrad/2019/04/19/zoom-zoom-zoom-the-exclusive-inside-story-of-the-new-billionaire-behind-techs-hottest-ipo/)

opened. Reductively put, an inability to get reliable might be attributed to some fundamental problem with the tool itself.

On the other hand, an inability to get reliable might be attributed to the personal or collective incompetence of the Math4All trainees. The COG system is complicated, and requires us to attend to, note, and code a wide range of details concerning real-time phenomena, and to do so by typing, not on a keyboard, but on the undifferentiated glass surface of a tablet device. During one training, Caitlin, another graduate RA, expresses her reliance on procedural uniformity as a personal deficit: "It's so subjective. I'm so bad at that." (In fact, Caitlin's propensity for mechanical objectivity will make her the first of the team to get reliable.) Needless to say, the possibility of our own incompetence served as a looming threat to our identities as generally capable people and good researchers (cf. Sanders and Cuneo 2010).

When one Math4All observer was continually unable to get reliable, the Math4All team was left to adjudicate between these two explanations: Was it a problem with the observer, or was it a problem with the tool? Both explanations were entertained by various members of the Math4All team. Leah, in the position of greatest responsibility, was concerned that she might be blamed for choosing a bad tool; others expressed frustration that the lagging observer continued to insist that their coding choices accurately reflected what they knew to be true from their direct observation of the events, rather than deferring to collectively established procedure. Ultimately, as far as I know, the question of where the fault lay was only settled in practice, and was never explicitly addressed: The Math4All team moved forward with the COG system for collecting classroom data, and continued using the COG system the following year.

Practically speaking, if the validity of the tool had seriously been called into question, not only would we have had to scrap all the existing observational data, but we may have injured the

Math4All PIs' relationship with the COG PI, who was in fact doing us a big favor. Silence around the validity of the tool was essential to maintaining the forward movement of the project and the positive relationship between the Math4All and COG PIs. Silence around the one observer's inability to get reliable was similarly essential to maintaining the observer's professional reputation and personal self-esteem, the former of which might also redound upon the reputation of the Math4All PIs.

In fact, I received significant pushback from the Math4All team for including possible identifying information about this non-reliable individual in my talks on this subject. I have since attempted to put a thicker shroud around their identity here as a result. To be clear, I am not attempting to impute any actual lack of competence to this individual or the Math4All team, nor any actual lack of validity to the COG system. My point is merely that tool validity and personal/professional identity are possible arenas of contestation that failures to get reliable threaten to open up. I discuss the anticipatory anxieties around these issues only as factors conditioning the contingent unfolding of actual scientific practice.

In examining the travel of the COG system then, validity as a license to travel, is effectively an unquestionable property of the system, locked into place by the very conditions which necessitate and enable cooperation (we don't have time to make our own tool; we respect each other as colleagues). Validity may certainly be contestable on other fronts, for instance, in response to publications regarding the COG system, or even in backstage talk amongst researchers. However, in the case of practical cooperation between research groups—with collegial relations, publication, and future grant proposals on the line—the stakes suggest a tendency toward silence around validity. Indeed, the lack of validity talk among observers is why it only appears here in my discussion of reliability, despite the complementary nature of the two

(recall their continual appearance together in Figure 1, the NCLB definition of SBR, on page 18). Yet, as further elaboration on the training process will show, not all contestations opened up by reliability were addressed with silence.

The Scene

Our COG training was conducted over a series of three conference calls, allowing us to share an audio channel, access on-screen chat, and follow along with a presentation that Andrew shared from his screen. As such, as is commonly the case now in the era of COVID-19, we were welcome to participate from home if we so desired. One RA who had an hour-long car commute was very happy to take advantage of this option; but it was a rather disappointing development for me, as I felt the training would be less socially effervescent if a shared audio channel and slideshow were to be our only modes of encountering one another (experience suggested that the on-screen chat function would be rarely used, if ever, and this proved correct).

In an attempt to position myself in a more promising site for participant-observation then, I opted to go into the Math4All offices for training. I surmised that at least Leah and Ashley, being full-time on Math4All, would be there. As it so happened, Caitlin—another grad student RA—also decided to join, feeling that she might get more out of the training if she were with others.

So we were all four together during the training in Leah and Ashley's relatively small shared office space, making use of all four of the desktop computer stations which took up the majority of its square footage. Leah and I faced one wall, and Caitlin and Ashley faced the other, Ashley's back to me. A narrow aisle of thin industrial carpeting, illuminated by fluorescent lighting, separated our swivel desk chairs.

Leah's computer came to serve as our central speaker and microphone. Any time one of us wanted to add something to the conversation, we could swivel ourselves toward Leah's station and start talking. During breaks and after the training, we would swivel towards each other to chat about the session, eventually drifting into talk about our lives in general. Taylor, another RA working in an office down the hall, would sometimes come by and join us during our chats, leaning against the doorway of the small office.

In each of our training sessions, Andrew led us through different aspects of the COG coding scheme. After each session, we were assigned homework consisting of a few "practice scenarios." Each practice scenario contained an "event log" situated above a few rows of codes (e.g., Activity Type: Whole group, Small group, Individual). These rows reflected the types of codes we had discussed during our previous training sessions. We were to read over each practice scenario and choose the appropriate value(s) for each field. When we reconvened the next day, Andrew would run through how the COG team coded each scenario and open the floor for follow-up questions and discussion.

The Event Log

On the last day of in-office training, we begin by engaging in the now-familiar process of going over practice scenarios, how COG coded them, how we coded them, and why. The first practice scenario concerns the following event log[3]:

> During carpet time, the teacher asks, "Yesterday we talked about all kinds of signs, didn't we?" The students nod and say "yeah" or similar. "Does anyone remember any of the signs we talked about?" A few hands shoot up quickly. The teacher calls on a few of the students, and they say stop sign, speed limit, do not enter, yield, one way. The teacher writes the name of each on the board, and as she does, they talk about what colors the signs are.

[3] Text has been adapted to discourage identification.

> She draws shapes on the board around a few of the sign names. She asks the students to point at which shape has the most sides. All arms shoot up immediately to point at the octagon as a few students say out loud, "stop sign!" "Let's count how many sides a stop sign has," the teacher says, and she points to each side as the students count along with her. After they reach eight, she asks, "So how many sides did it have?" Most of the students say "eight!" though a few are pulling at the carpet and not looking at the board. They repeat this with a few more signs.
> The teacher then holds up a picture of a pedestrian crossing sign and asks, ''What does this sign mean?" A few children raise their hands, and she calls on one who says that it tells you to watch out for people crossing the street. "That's right," she responds. She then shows more pictures of street signs and the class talks about what each one means.

Because the COG system was a live observation system, not relying on audio and/or video recordings, the use of event logs, like the one above, was necessary to facilitate the assignment of the whole range of codes required to characterize each Event of Instruction (EOI), the COG system's unit of analysis. Observers were not expected to walk away from classrooms with complete records at the end of any (multi-hour) session. It was practically impossible for observers to make all the decisions necessary to discern and code EOIs in the midst of on-going activity while still paying attention to the contingencies of its unfolding.

Event log technology was, then, in part devised to support observer memory, manifesting as an open-ended text field amid a mobile app otherwise comprising uncompromising checkboxes, drop-down menus, and radio buttons. In the app, each EOI record contained such a field for the observer to type freeform notes, unconstrained by any coding scheme.[4] Our priority while in the classroom was to capture the details of our observations before they could be forgotten, hurriedly typing key phrases and jottings into the event log field in idiosyncratic shorthands.

[4] The only other text fields in the app took only integer values and times, e.g., "Number of students" or "Start time."

Beyond supplementing our memories, event logs were a central mediating technology for the reliable entextualization of observed phenomena into discrete EOIs with distinctive properties across coders. After the session ended, we would often drive to a nearby mall or commercial center, finding a coffee shop to sit in and clean up and fill out our event logs. In going over the details of each EOI again, we would not only figure out which codes we should select, but, at times, we also re-assessed our determination of EOI boundaries. Perhaps what we had noted as a single EOI was better captured as two EOIs. Indeed, much of the documentation attached to the first email we received from COG was concerned with setting forth rules for determining what should be noted as a single EOI, as multiple EOIs, or perhaps not an EOI at all.

That is to say, the cleaning up and filling out of an event log was not unconstrained altogether. This was not only because our notes were expected to capture details relevant to the COG coding scheme, but also because the event log would serve as a site for an second-round quality check of our observation records.

In the first case, we were expected to be familiar with the various codes involved in the COG coding scheme in order to make sure relevant details were included in our log. The numerical data was not difficult collect, and did not necessitate much use of the event log. For instance, we became habituated in quickly noting the number of students involved in each activity. However, the event log was essential for gathering details which spoke to the quality of teacher-student interactions or the degree of student engagement, for instance. As such, in training, we were advised to try and capture the exact language used in interactions. This was considered to be one of the best uses of the event log, the most help in reconstructing the event after the fact. It was conceded, of course, that it would not always be possible to get every word

down, depending on the speed or verbosity of the interaction. We should try, however, to capture what we could of the quality of the exchange, if we could not produce a verbatim account.

By the same logic, our event logs were to be narrative in form, capturing the unfolding of an interaction as a key index of its quality, e.g., how did the teacher respond when a student gave a wrong answer? Finally, we were also advised that rather than writing notes like "students were bored," to instead describe their behaviors, "X number of students raised their hands," or "X number of students were looking at the carpet." Indeed, in doing live observation, the strictures of the COG event log were fairly identical to common recommendations for ethnographic fieldwork, relying on inductive description and, certainly in the case of linguistic anthropology, an attention to discourse in interactional unfolding.

The description of visible behaviors, explicit language, and interactional turns in teacher-student and student-student interactions in the event log would serve as a helpful reference not only in our later coding of the event, but in the process of "data checking." We as observers were not to be the only readership of our event logs. A data checker would not only check the completeness of the observation record (every box ticked), but the correspondence of our codes with the narrative record of our event logs. In producing event logs then, we were not only creating a reference for our future selves, but advancing an evidentiary claim to a third party, the data checker. Observers, however, rarely came into adversarial relationships with data checkers. Rather, the data checker was another resource for coming to the appropriate normative coding of the event. Direct calls for consultation were even encouraged in event logs ("Because of ____ , I decided to code it X, but I'm not sure if it should be X or Y.")

The use of event logs like these in our training materials presented opportunities to align our own engagement with the event logs with COG's established textual practices in

precipitating event logs from live observations, and coding observations through discerning engagement with event logs. In this particular log, many characteristics of good event logs are present: direct quotation ("Yesterday we talked about all kinds of signs, didn't we?"), use of quantifiers in the description of student behaviors ("A few hands shoot up quickly" "most students say" "a few are pulling at the carpet"), and a nearly turn-by-turn narrative presentation. This modeling of good logging practice, in conjunction with explicit instructions about the event log genre, participated in the production of standard textual practices which peripheral trainees could emulate in their movement towards full participation in a COG community of textual practice in classroom observation (cf. Goodwin 1994, Lave and Wenger 1991). In doing so, we hoped our efforts in alignment would pay off in the calculation of high reliability scores.

Interestingly enough, our training in coding practice scenarios put us in a position much closer to data checkers than observers, especially, as we would soon find out, since the above event log was taken from an observation done by our trainer, Andrew.

Initial Disagreement

As Andrew runs us through the COG coding of the above scenario, we follow along on our printed or digital copies of the practice scenario. We silently mark up our papers or nod our heads as we move from code to code. It is only when Andrew opens up the floor for questions that a coding conflict appears which necessitates explanation. That is to say, many of our disagreements are noted silently, with a Math4All observer simply changing a code without feeling the need to question its logic. The Math4All team, however, will not allow one particular mis-alignment to be passed over as a mere mistake or oversight.

The field that emerges as the site of contestation is 'Cross-Cutting Content.' This field describes whether the teachers' instruction relates EOI content to other areas of school or life.

This type of code is fairly common in classroom observations, as US education researchers have taken up teachers' ability to integrate content across the curriculum and everyday life as an indication of pedagogical quality since at least the 1990s (see, for example, the entire October 1991 issue of *Educational Leadership* on the "integrated curriculum").

In the COG system, the Cross-Cutting Content field contains four values:

- Related to Life Outside of School
- Related to Other School Activities
- Related to Previous Lessons
- None

As will come up in the following conversation, "Related to Previous Lessons" applies only to math instruction, while "Related to Other School Activities" covers any other academic area of content integration. For the event log of record, Andrew reports that the COG team has marked only "Related to Life Outside of School."

No one says anything until he has concluded going through all the codes for this scenario, and opens the floor to requests for elaboration and clarification. Ashley, who is calling in from home on this day, offers that she has marked not only "Related to Life Outside of School" as COG did, but also "Related to Other School Activities" and "Related to Previous Lessons." Where COG has only selected the one option, Ashley has selected three.

As Ashley is speaking, her response is buttressed by a chorus of agreements from around me in the office and over the speaker: "Mhm," "Yeah," "Me too." When Ashley stops speaking, Caitlin adds "Same here," and Leah follows with a "Yeah, so do I."

I don't say anything, embarrassed, having coded Cross-Cutting Content so idiosyncratically that it matches neither the COG coding, nor Math4All's apparent consensus.

Despite my silence, Caitlin confirms out loud to Andrew, "So did all three people in the big office." Leah and Caitlin laugh good-naturedly, and I offer a sheepish chuckle. Andrew laughs a little as well, and begins to explain that "Related to Previous Lessons" must be math-related.

Questioning

Andrew's explanation of "Related to Previous Lessons" as math-only is taken in stride without comment (as corroborated later in the conversation). The Math4All team is now primarily interested in why the COG did not code the event as "Related to Other School Activities."

Leah begins this line of inquiry, "So can you give an example of what it would look like if it was, if it was 'to Other School Activities'?"

Andrew quickly responds, "Absolutely" but struggles for about 30 seconds to come up with an example on the fly (not an easy task), and instead re-clarifies, "You guys marked 'Other School Activities' or 'Previous Lessons'?"

Immediately, about four Math4All members across several locations say "Other School Activities" as if performing cascading parts in a choral round. Their lack of hesitation stands in stark contrast to Andrew's halting difficulty in improvising a hypothetical. He confirms, "'Other School Activities'? Okay." Another 20 seconds of hesitation precedes Andrew asking yet another question, this time turning to the practice scenario: "Were you more focused on the first paragraph or the second paragraph more? Like what was your thinking on that, what sort of stood out to you as the relationship [to Other School Activities]?"

It is worth noting that in repeatedly returning to questions to probe our thinking, Andrew is enacting a highly valorized style of pedagogy. To put it in perhaps reductive terms, the COG system gives high ratings to teachers who ask stimulating questions in response to student

misunderstandings. Such a strategy is taken as an indicator of high-quality instruction, and in our training, the Math4All members all seemed on board with the hierarchy of pedagogical moves they were being asked to rate, being more concerned about how to distinguish them in practice.

That Andrew is halting in his production of this pedagogical style is not surprising, as he is not a teacher by trade, but here a researcher taking on a teaching role as one responsibility among many. In describing his various pauses, hesitations, and so forth, I am only aiming to give readers a feeling of the real-time unfolding interaction, not making an argument about Andrew's competency in some way or another. However, the possibility of the COG team's incompetence will indeed come to loom over the training session that I am describing. It will be the discursive defusing of this impending negative possibility that will ultimately engender a stronger bond of trust between the two groups.

Turning to the Text

> Let us return to Andrew's question:
>
> "Were you more focused on the first paragraph or the second paragraph more? Like what was your thinking on that, what sort of stood out to you as the relationship [to Other School Activities]?"

In asking about the practice scenario as a text demarcated by paragraphs, Andrew turns our attention to our respective copies of the event log. None of these copies constitutes *the* practice scenario, though they are all instantiations of it. The practice scenario itself is an intersubjectively produced text which the event log is "about." It is an event (a lesson about signs) in intersubjective space which is both presupposed and entailed by the various print, digital, and sonic artifacts involved in talking about it.

As we will see, our intersubjectively held practice scenario can change without necessitating any changes to its artifactual representation in the multiple copies of the event log.

In fact, for the Math4All team to go from a strong disagreement about how the practice scenario should be coded, to coming to see it "in the same way" as Andrew/the COG team, no changes need to be made to our copies of the event logs. Rather, it was our textual practices in engaging with the event logs that required changing. What we imagined when we invoked "the practice scenario" would be transformed in-and-through the transformation of our textual practices, in this case, our practices of reading the event log.

> "Were you more focused on the first paragraph or the second paragraph more? Like what was your thinking on that, what sort of stood out to you as the relationship [to Other School Activities]?"

Leah responds, directly citing the language of the event log, "Well for me, it was like, 'Yesterday we talked about all kinds of signs, didn't we?' 'Does anyone remember any of the signs we talked about?'" Others mutter their agreement in the background. How else to say it is what it is? As an exercise, I urge the reader to attempt an explanation for why "Yesterday we talked about all kinds of signs, didn't we?" does not constitute evidence of a relationship to Other School Activities.

Rising Tension: "Back-and-Forth," "Grey Area"

Andrew only has time to acknowledge Leah's explanation with an "mhm" before I latch onto Leah's turn and set us on a topical digression, away from the resolution of the Other School Activities problem. While the topical disjuncture is rather unfortunate, what follows is interactionally continuous with the rest of the training situation, contributing to the slowly rising tension that has already begun emerging with Leah's inability to get a satisfactory explanation to why the patently obvious "Other School Activities" relationship is being ignored by COG.

I give a long explanation about why I did not code the scenario as "related to Life Outside of School." This was the only code COG assigned to the scenario, and was also part of

the Math4All consensus. I explain that I didn't think the lesson drew on students' personal experiences, and therefore didn't warrant the code. After I offer my explanation, Caitlin follows me in explaining her own coding choices, Life Outside of School and Other School Activities.

Though this movement into individual rationales takes us away from the disagreement about Other School Activities, it does lead Andrew into an explanation which heightens an already nascent air of dissatisfaction among the Math4All team. Andrew explains the "back-and-forth" conversation that took place within the COG team with respect to the Life Outside of School code, again, the only code the COG team assigned this scenario. Andrew concedes that it might not be a good code if the example is not directly drawing on students' personal experiences.

Andrew's admission of COG "back-and-forth" pings the Math4All team's already present concern about the difficulties of getting reliable. How could we be expected to get reliable on a system which doesn't seem to make sense, and which is now apparently being revealed as arbitrary even among authorities in its use? Given this revelation of COG's "back-and-forth," members of the Math4All team immediately respond with both overt and covert challenges to the coherence of COG's guidance.

Taylor identifies herself over the phone, "This is Taylor." She immediately names our collective concern in her first three words (emphasis added): "*For reliability purposes*, I was under the impression that it had to be very explicit, from our conversation yesterday. So today, I- from reading this, it didn't feel explicit enough." Taylor draws today's session into line with the previous trainings, in order to figure an apparent disjuncture between the two. In particular, Taylor is concerned about our textual practices around event logs. She has been assured that the genre conventions of event logs include their strict reportage of discrete, observable behaviors,

including speech behavior. Further, she has been assured that any coding activity based on an event log must be able to advance an evidentiary claim by pointing to specific words and phrases in the event log: "it had to be very explicit." These were the textual practices which would secure reliability, and if we were no longer collectively and strictly holding to those practices, what hope did we have of getting reliable?

Andrew agrees with Taylor. He says that the COG team had originally coded Cross-Cutting Content as "None," "for that exact reason, Taylor." He says that he was "doing probably what [he] shouldn't have been doing" by going back and reconsidering the relationship of road signs to knowledge about driving generally. "It's a little bit of grey area."

While Taylor and Andrew have their discussion, Caitlin whispers to Leah: "Didn't they [COG] originally code it 'related to Life Outside of School'? And now they're arguing against it?" As Taylor is overtly confronting Andrew about COG's inconsistency in applying their own criterion of explicitness, Caitlin and Leah covertly establish mutual dissatisfaction with the inconsistency of [Andrew as] COG's coding. When Andrew says "grey area," another marker of inconsistency, Leah and Caitlin shoot each other looks.

Andrew continues, leaning into Taylor's point about explicitness:

> "I would agree though that, um, I think the stronger- a stronger example of a relationship would be a lot more explicit, um, than this. And I would kind of gear towards, um, being more conservative than kind of opening it up and saying kind of any of these, you know, could be because it's implied. That's kind of where I lean, I would want an explicit connection to the math concept that they're looking at."

At the end of Andrew's turn, Leah sighs as she says, "So… so what's the conclusion?" At least a few of the members of the Math4All team have found the past ten minutes of discussion a little too tenuous. Rather than producing greater clarity in how to reliably code from event logs, today's training—the final formal training session—has so far seemed to only muddy the waters.

"Grey areas" do not seem to bode well for our getting reliable. But Andrew does not appear inclined to perform the authoritative, decisive stance that the Math4All team desires, and which Leah is explicitly requesting now: "So what's the conclusion?"

Andrew answers again with no attempt to disavow inconsistency, instead explicitly labelling his conclusion as a walk-back: "I would walk it back to None for this one." He continues, even qualifying this walk-back as a "leaning" above any more definitive stance, as it were: "That's kind of where I'm leaning, because that's how we had it coded originally." At this point, Andrew speaks at length on the limitations of event logs as a technology and their usefulness in data checking. I will elide this stretch for another time, as it is not taken up again in the rest of the interaction under analysis.

Return to Other School Activities

Andrew's long turn does not seem to clarify any of Math4All's concerns. If anything, it appears that the Math4All team is waiting to hear something that speaks directly to their specific coding concern. At the conclusion of Andrew's speech, Taylor immediately brings the conversation back to the initial concern brought up by Leah: "It's still not entirely clear to me why this is not a relationship to Other School Activities, could you talk a little bit more about that?"

I jump back in, echoing earlier requests for a scenario where Other School Activities would apply and an elaboration on the nature of explicit evidence: "Yeah, what would that actually look like, what in the notes would make you think it was related?" This time Andrew rather quickly produces a hypothetical scenario in which more math activity occurs in the discussion of signs. Taylor does not take this as an answer to her question: "Maybe my question

wasn't clear. Not about '[Related to] Previous Lessons' about math. It's clear that the activity was not explicitly about math. More so about 'to Other School Activities.'"

Leah follows up to clarify, reiterating her previous explanation: "Right, to me it's like very obvious that that should be what it is, because it's clearly saying 'remember how we talked about signs yesterday.' So to me, I don't understand why that isn't 'to Other School Activities'." Leah is almost laughing as she reads out the sentence from the scenario, presumably because the situation has reached the level of absurdity; the explicitness of the connection is so evidently transparent, and its dismissal so frustratingly opaque.

I then continue in my apparent mission to deny Leah a direct answer from Andrew: "Andrew, is what you're saying, because there's a whole paragraph about where they're talking about signs, and then she's like, 'let's count sides,' so it's not super related to what's happening in the notes [event log]?" Andrew agrees, and goes on to clarify what he is agreeing with.

What Andrew goes on to say re-orients Leah and the rest of the Math4All team so completely, that the tension of the preceding fifteen minutes fully dissolves, and, at least in my case, erases any conflict between COG and Math4All from memory. In fact, when I returned to my recording of this training over a year after the fact, I was startled to (re)experience the highly palpable tension of this conflict. I remembered being nervous about getting reliable more generally, but I had retained no recollection of any fundamental challenge to the COG system and training. I listened without memory of how the situation would resolve, or if it would resolve it all. It felt as if I was watching a television drama, unable to imagine how the writers would get the characters out of the seemingly unresolvable conflict they had set up.

Andrew's curative explanation centers around a feature of event logs that had escaped the Math4All team: An event log need not be co-terminous with an Event of Instruction; it could set

up the context for an EOI, describing activities preceding its start and following its conclusion. This was especially true in the case of "embedded" EOIs, EOIs that occurred in the midst of some other activity. In such a case, the event log should describe the containing event within which the EOI was Embedded. Critically, inclusion of surrounding activity in an event log would allow for better discernment as to whether the EOI should in fact be coded as Embedded.

Andrew:

> This is a scenario based off of something that I actually saw in a classroom, and I think, we have been discussing ever since, in terms of what's going on. And um, I think the—we agreed that it was an embedded activity, because the way that the math portion of it came out was sort of an aside, and the activity was not set up to be a math activity. So we thought we were getting some of that by coding it as 'embedded' um, but the teacher wasn't really going out of her way to, uh, to tie it all together.
>
> It was just kind of like, we're talking about signs, now here's a math thing. Alright let's go back to talking about signs. So it seemed embedded, but it didn't seem sort of, sort of connected as part of a complete idea, like, we're going to do the sign thing, now here's a little math thing that I'm going to embed, and then we're going to go back to doing a sign thing.
>
> Um, so it's, it's a little tricky. We're trying to differentiate those teachers who are really focused on, and explicitly trying to make strong connections between um, school subjects as opposed to a teacher who is embedding sort of like, we're going to throw a little math thing in here so we can check off these, um, this, this content requirement or something, and then we'll go back to whatever subject we were talking about. That's kind of what we were trying to differentiate here.

This explanation is an unexpected, if highly desired, revelation for members of the Math4All team. A stark shift in the affective atmosphere of the training takes place. The skepticism that pervaded the training over the course of the previous 15 minutes no longer infuses the turns that follow. Leah responds: "So I guess I had thought it was a math lesson, like I didn't even count it as embedded, I counted it as [a non-embedded type of EOI], so um, but I guess that wasn't clear to me. Now I can see it's sort of- she was- they're just kind of talking

about signs, and she's kind of throwing in a couple comments about math, is that what you're saying?"

Leah is coming to realize that her evidence of "Other School Activities," such as the teachers' utterances about "what we talked about yesterday," falls *out of the bounds* of what is considered "the math lesson" [=the EOI]. In substantiating an Embedded code, such utterances only become meaningful as pieces of an EOI's "context," no longer having any meaningfulness with respect to characteristics of the EOI itself, including the evaluation of any Cross-Cutting Content.

"Now I can see…" The work of producing agreement on this single code has taken over a quarter of an hour, and discussion of this single practice scenario has so far taken us 38 minutes into a planned 90-minute training. The Math4All team trusted the COG team enough to invest time and money into the use of their system, but that trust was suspended when we were unable to imagine a sensible refutation of what we perceived as an undeniably clear warrants for coding the event log in a particular way. We repeatedly requested an answering rationale, not only because we were unable to produce one ourselves, but because we all wished to see our trust in the COG system vindicated. Before we understood that not all of the language in the event log should be read as part of the EOI (or as Leah puts it, "math lesson"), our trust was suspended. In reaching intersubjective alignment with respect to what part of the practice scenario, of the event log, constituted the EOI, we were able to not only release the suspension of trust, but, as it happened, move through future conflicts with more confidence that they would be resolved.

Teachers and Texts

The preceding analysis was mostly about the "denotational text" of our training interaction, i.e., what it was "about," that being the practice scenario and what the practice

scenario itself was "about." The proceeding analysis will elaborate more extensively on the "interactional text" that unfolds in-and-through the unfolding of said denotational text. The earlier analysis has already recognized that the inculcation of a new textual practice, a new way of reading, of breaking an event log into meaningful parts (EOI vs non-EOI parts), has profound social consequentiality, moving actors into collective alignment, re-instating trust, setting the stage for formal reliability, and so on. Discussion of the interactional text of the event ("what's happening") offers an opportunity to illuminate the more overtly social aspects of what goes on in-and-through talking "about" something or other.

We can get into this discussion by way of this question: What if this "misunderstanding" had been cleared up earlier? If Leah and the rest of the Math4All team had realized the import of the Embedded code when it was first delivered in Andrew's initial run-through at the beginning of our training session?

Empirically, we cannot treat the timing of this intervention—at the end of a 15-minute stretch of building tension—as inconsequential. Certainly, we will never know what 'would have happened' otherwise, but we should at least be assured that if the Math4All team could simply (and carefully) read the manual and be able to reliably use the COG system, then the expense and effort of training would not be necessary to begin with. Given the COG team's prior experience proliferating their system, the training protocol would likely already have been scrapped long before they got to us, and we would have been directed immediately into reliability visits.

In fact, at the beginning of the training, the Math4All team had already learned about the Embedded code, and many of us, as Leah notes, had marked a failure to code the scenario as Embedded when Andrew first went through COG's coding at the beginning of the training session. However, this difference in coding did not appear remarkable to any of us, as none of us

made any remark. We simply trusted COG's coding, and changed our codes to Embedded simply because Andrew said so. (Of course, we did not have any idea in the moment how consequential this code would prove to be.)

Indeed, we trusted Andrew not simply as a representative of COG, but we looked to him as students might look to a teacher. In offering 'correct answers,' Andrew's interactional role, from the point of view of the Math4All team, was that of a teacher in a teacher-student relationship. We occupied the interactional position of his students, in line with the typical one-to-many configuration of "teaching." We treated Andrew not merely as a teacher, but a teacher of the old school: we expected him to have all 'the right answers,' his authority built on a bedrock of knowledge. Our growing frustration with Andrew arises when we perceive him to be uncertain or evasive in his answers (recall Leah sighing, "so what's the conclusion?").

Andrew, on the other hand, enacts an entirely different model of teaching—a more egalitarian, inquiry-based model—which does not require the teacher to perform a strong authoritative stance, and in fact, warns against it. As previously mentioned, Andrew's question-centered practice most resembles the progressive pedagogy that the COG system, as well as the COG and Math4All researchers, most value in evaluating the practice of classroom teachers. Further, his long explanations of how and why the COG team came to the decisions they did also most resembles the "talking to" and "teaching" models of intervention that scholars of research and materials use tend to recommend (see Remillard 1999; Honig 2012; Ball and Cohen 1996)

At issue in the dis-integration of our interaction is not a difference in interactional roles (teacher vs student), but a difference in our expectations as to how those roles should be enacted (teacher as source of knowledge vs. teacher as facilitator of knowledge-building). As we continue to butt heads on this latter matter in our continued interactional efforts, a baseline

shared collegiality begins to diverge into frustration on one end and something else on the other. (I have no idea as to Andrew's experience in all this; I only know that part of our continued and growing frustration was an apparent lack of urgency on Andrew's part in providing a definitive answer to meet our frustration.)

Given this divergence in the interactional text (student-teacher relationship), emerging alongside a divergence in the denotational text (the Other School Activities code), the Math4All team sought a re-unification of the interaction as a way to vindicate our trust in the COG group—that is, to justify our belief that they had created an observation system in line with what Math4All would have created, should we have been in the business of creating an observation system.

Crucially, what emerges as the source of re-unification is not a clarification of the Other School Activities code, or what the teacher means by "Yesterday we talked about all kinds of signs, didn't we?" Instead COG/Andrew's 'license to teach' is recovered through the shift in reading practice previously discussed. Indeed, reading is not just about deciphering the meaning of words, but, as Andrew's explanation of the Embedded code elaborates, being able to read an event log as containing an Embedded EOI *is the same thing as* being able to differentiate between types of teachers and the quality of their teaching practice. That is, textual practices like reading, not only do the work of constituting the groups which collectively practice them by setting a standard of membership (do *you* see what *we* see?), but the enactment of such practices populates the world with objects—types of teachers, types of teaching—in accordance with the interests of said group (cf. Street 1984, Collins in Silverstein and Urban 1996). In short, textual practices are both the result of, and a form of, ideological work.

The Promise of Objectivity

I asked in the last section, what if this 'misunderstanding' had been cleared up earlier, sooner, faster? Quite a lot was accomplished in the 15-minute timeframe of the conflict, and much of that accomplishment was the building up of suspended tension. The time contesting this single scenario was time well spent in our mission to get reliable. Skepticism has given way to confidence. The Math4All team no longer finds ourselves confronted with the inconceivable, and what's more, we have witnessed a miraculous transformation: the unthinkable made thinkable. From this point on, when Andrew uses phrases like "it's definitely up for interpretation," his equivocations no longer inspire the deep sighs or consternation of these past fifteen minutes. From the messy stuff of conflict, a foundation of trust has been set.

As Sanders and Cuneo (2010) have similarly observed, the negative emotionality experienced during the course of the training aided us in our journey towards reliability. The clicking into place of intersubjective alignment at the end was only all the more revelatory given the difficult beginnings of our conversation. If we could get here from there, perhaps we were not foolish to hope for reliability within two visits after all. I think of this hope as the feeling of objectivity: A forward-looking feeling of the promise and possibility of harmony, based on the resolution of a previously troubling dissonance. It arises as an apparent disjuncture is resolved as the 'discovery' of a 'pre-existing' continuity.

If we return our gaze to the event log, we can see that there were already paragraph breaks marking off the beginning and end of the embedded EOI from the rest of the text. In seeing this, Andrew's motivation in asking Leah about which paragraph she was focusing on also becomes clearer ("Were you more focused on the first paragraph or the second paragraph more?"). Andrew has formatted the event log such that the Embedded EOI is in fact embedded

between two separate paragraphs. These paragraph breaks, however, only become meaningful to the initiate in retrospect, though they were always present in the text-artifact.

This does not mean, however, that the Math4All team has been led to discover a pre-existing structure whose meaning was always there. Rather, knowledge about what the event log for an Embedded EOI looks like is reproduced in the reproduction of COG's textual practices within the Math4All team. Only in-and-by reproducing that textual practice could the Math4All team experience the paragraph structure of the event log as "being there" — as meaningful, effective, consequential.

Hope also projects (asymptotically) into the future the feeling that all disjunctures will be made recognizable as instances of continuity. In this chapter's example, hope is achieved by holding onto trust amid conflict. Given this case, we might also better understand why Andrew does not shy away from expressing uncertainty, nor evince any obvious signs of distress when repeatedly questioned by Leah and Taylor. Given his extensive experience developing and training people on the COG system, he understands that small adjustments may resolve dissonance into harmony. He is well practiced in his hope that reliability will arrive.

Science as Literary Practice

Latour and Woolgar (1979) discuss the scientific laboratory as a factory which produces reports, "a system of literary inscription" (p.53). Noting the preponderance of literary practices within the laboratory, their fictive observer/anthropologist comes to the following conclusion:

> By pursuing the notion of literary inscription, our observer has been able to pick his way through the labyrinth. […] The anthropologist feels vindicated in having retained his anthropological perspective in the face of the beguiling charms of his informants: they claimed merely to be scientists discovering facts; he doggedly argued that they were writers and readers in the business of being convinced and convincing others.

While, as a linguistic anthropologist, I take heart in Latour and Woolgar's appreciation of the centrality of language-in-use in scientific practice, I offer two comments on the conclusion they present.

First, that science is not a special domain of activity in its domination by literary practices. Language in its various uses has a special capacity for cultivating social activity. Notably, all this talk around a single practice scenario demonstrates the cultural capacity of language's reflexivity, its ability to foster the organic growth of social activity as new events of language-in-use are creatively and recursively generated from prior such events. Here we have a (1) conversation about (2) how to engage with a written artifact in (3) producing evaluative descriptions of (4) differentiable types of discursive events (e.g. a math lesson vs an aside), that is, *talk* about *how to read*, in order to *talk* about different *types of talk*.

Second, to take the position that science comprises numerous textual practices, including those primarily linguistic/literary, does not imply that it is any less real (cf. Latour 2004). From a pragmatic point of view, reality is an indispensable working hypothesis for all of us as we go about our lives (Peirce [1898] 1992), and everything we do in trying to investigate reality changes the reality we wish to investigate. As this chapter shows, talk about talk comes into being in the course of scientifically, politically, pedagogically meaningful and consequential activities: differentiating between types of teachers, rating the quality of instruction, designing a useful observation system, doing reliable data collection.

By establishing relationships of reliability in-and-by talking about talk, the COG and Math4All researchers did the work of maintaining the coherence of Research as a scientific community. Yet relationships of reliability do not merely organize Research from within by chaining together researchers. Rather, reliability lives in and organizes the entirety of Research-

Practice nexus. As with classroom observation, researchers' data collection does not occur solely in Research settings, but often in Practice settings; and even when it does occur in campus laboratories or the like, its participants are taken from Practice, being teachers, students, and so on.

Reliability practices organize the Research-Practice nexus in-and-by continually licensing Research's place within the relationship of knowing which has historically distinguished the two—Research knows Practice; Practice is known by Research. In-and-by "getting reliable" in classroom observations, coding interviews, and so on, we differentiated ourselves from practitioners in-and-through the very act which brought us together: researchers' [objective] study of practitioners [subjective] work, i.e., "*we* systematically study what *you* happen to do."

In-and-by an analysis of the centrality of textual practices for the production of reliability, we come to recognize textual practices like reading as non-straightforward, and instead inculcated through discursive practices of social reproduction. In doing so, we are better able to understand how once 'straightforward' evidence ("remember yesterday we talked about signs?") can recede into irrelevance as new straits forward come into view (Embedded-ness and its implications). More importantly, we can begin to articulate a model of knowledge circulation premised upon the observation that knowledge only 'travels' inasmuch as its conditions of production can be reproduced. That is, knowledge can only 'move' to the extent that its originating community can do the ideological work of reproducing its practices and values amongst the 'recipients' of that knowledge. If we can understand the work of disseminating knowledge to be an activity of social reproduction, just as dissemination is itself an activity of biological reproduction, then we can begin to understand both the failure of, and resistance towards, education reforms which 'merely' wish to impart knowledge.

Chapter Five: Adaptative Strategies

This chapter discusses what happens after the training sessions concluded, and we on the Math4All observation team were left to our own devices.

The previous chapter discussed how the travel of the COG system required not only a "valid" passport, but a "reliable" means of transportation. This system of transport was organized by way of the regimentation of the Math4All team's textual practices. The formal calculation of reliability between Andrew and Ashley in Empire City would then serve as a crucial site for grounding claims that the observation system the Math4All team was using was "the same" system that COG developed. However, in this case of "reliable use," which is presented in contrast to "faithful implementation," the COG team did not have any issue with us on the Math4All side making some changes to the system to better suit our needs.

In the course of our training sessions, the COG team emphasized that their tool should be modified to suit our needs rather than our needs modified to fit their tool. In adapting the tool, we continued to consult with the COG team as experts, though we did not always heed their warnings about the practical feasibility of our proposed adaptations.

Adaptations were necessary for a few reasons. First, the COG tool's original use was in the formative assessment of teacher instructional practices, as a means of supporting teachers' professional development. As such, teachers were told to prepare to engage in math instruction during their scheduled observations. Further, the tool had fairly high standards for what would count as an EOI—in particular, a relatively long minimum duration—because each EOI needed to be substantive enough to ground discussions with teachers about their instructional practices.

Math4All, on the other hand, intended to use the COG system to collect outcome measures about the amount of math instruction that took place with and without Math4All use. As such, we did not want to influence teachers' behaviors by telling them what we were there to observe. Instead, with the PIs' input, we simply set up a scheduling protocol which would provide as good a random sampling of observations as we could manage given school schedules, and did not discuss with the teachers that we would be looking for math-related activity for in our observations.[1] Given all this, the observation team was nervous that we would not end up recording any EOIs during our visits, as math instruction was expected to already be fairly rare in preschool settings.

Given this concern, Math4All developed a new kind of EOI—the "micro EOI"—which would allow us to capture shorter math-related activities. When we discussed this possibility with the COG team, they provided some words of warning that attempting to capture activity at that level might produce difficulties for observers, who may not be able to keep up with noting so many small events. Further, because shorter moments might be harder to catch, reliability with respect to the number of EOIs observed might be harder to achieve. However, the COG team showed no inclination to bar us from making this adaptation, only suggesting possible downsides to the adaptation.

The other reason for adapting the COG system concerned the existing codes around what math content was being addressed during instruction. The codes in the COG system had been developed to align with the Common Core math standards, but we were interested in teachers' engagement with the constructs operationalized by Math4All. As such, the content codes were

[1] Various Math4All staff did comment that the teachers would probably assume that we were concerned with math instruction given the nature of the Math4All intervention.

revised to reflect the skills that Math4All hoped to educate teachers in. Would teachers' use of Math4All involve learning about the math skills it assessed? Would becoming so educated then increase their likelihood of addressing those skills in their instructional practices?

The Math4All team was allowed to make these modifications to the COG system because we were trusted as fellow researchers. We were expected to share a common moral vision of not only what counted as good research practice, but also what counted as good teaching practice. For instance, in Andrew's explanation about the Embedded code, he describes how said code is a way of differentiating between types of teaching and types of teachers. In doing so, he does not need to explain to us which teacher is better, the teacher who is "really focused on, and explicitly trying to make strong connections between school subjects" or the teacher "who is sort of like, we're going to throw a little math thing in here so we can check off this content requirement." As members of the same Research circles, we are all able to quickly recognize the social personae that Andrew is mobilizing in describing the utility of the Embedded code, and to immediately clarify the confusion that plagued the earlier stretch of training.

Because of this moral bond of trust, COG was not concerned with letting us tinker with their system towards the realization of our own goals. The data that we would be producing would not be *for* COG, but for ourselves. Where I discuss the resolution of a training conflict as a vindication of trust—a re-affirmation that the decision to align ourselves with one another was a good one; COG's relinquishment of control, of supervision over Math4All's use and adaptation of their tool, marked the consummation of our trust relationship in extending the capacity for classroom observation across the division of labor within the Research community.[2]

[2] Such relationships are of course always contestable, so the question of trust could always be re-opened, for instance, by questioning Math4All's membership in the scientific community, or by

Discursive Technologies

In creating adaptations to the COG system, Math4All had created some new difficulties for itself. We were no longer able to rely on an established training apparatus with curated practice scenarios or the guidance of observers practiced in identifying "micro EOIs" or instruction that corresponded with our new content codes. We were now in the position of training ourselves on our own adaptations.

By the point that micro EOIs and new content codes had emerged as possible adaptations, almost all the members of the observation team had had the opportunity to go on practice observations. These were observation sessions scheduled in classrooms that were not enrolled in the Math4All study, but which would serve as sites for test driving the COG mobile app in a real classroom setting. This proved invaluable, not only for the hands-on experience of familiarizing ourselves with the app, but in providing fodder for our efforts in getting reliable on our modified system.

As mentioned in the previous chapter, we wanted to minimize the number of reliability visits necessary to achieve reliability. (Recall that reliability would be determined by a calculative comparison of observation records from two consecutive paired reliability visits.) Toward that end, we deployed a technique that has long been controversial with respect to its appropriateness and effectiveness in addressing the wide range of concerns it has been mobilized to address: we had a lot of meetings.

---

strategies of social distancing e.g., you acted against scientific norms in *your* adaptations of *my* tool.

In the course of these meetings, we further did not have the benefit of practice scenario event logs to jointly orient towards, and so we conducted our internal training with the assistance of another material technology: speech. While previous theorizations of cooperation have discussed the mediating role of artifacts, many tend to omit speech itself as a material artifact, instead tending to discuss more perduring artifacts, particularly in their commodity forms[3] (cf. Agha 2011). Like any text-artifact, that is, any object mediating socially consequential activities of meaning-making, in its mediation of such social activities, speech is simultaneously a material and semiotic phenomenon (cf. Keane 2003, Harkness 2013). Beyond the articulation of phonetic sounds, fluctuations in pitch, volume, tone—as will be described in the forthcoming example—index moments of material as well as semiotic transformation.

Re-Creating the Classroom

The observation team meets in a small conference room at the Golden School of Education. The GSE is housed in one of the older buildings on campus, with an interior that looks to have last been renovated sometime in the 1970s. This room is half the size of a typical seminar room. A small oval conference table sits atop industrially carpeted floor, leaving just enough room for us to sit around it. A projector hangs from the ceiling, pointed at a whiteboard on one wall, above which a retractable screen hangs. Leah, Ashley, Caitlin, myself, and three

[3] For example, Star 2010, in clarifying if well-known entities like the American flag are boundary objects, rightly points out that the determination depends on if such flags are being used in the course of some working arrangement. She then points the reader toward the manufacture and marketing of actual American flags as places to find such working arrangements. My point here is that in understanding how the American flag mediates social organization, looking at invocations of the flag in speech (including, but not limited to, instances of the phrase "the American flag") should not be overlooked as 'immaterial.' In fact, such discursive materializations of the flag do a great deal of social work which cannot be captured by following only the life course of American flag commodities.

others are in attendance. Taylor is not here, choosing to attend a meeting for the other project she is a part of. There is some consternation around this fact, as it could be argued that our meeting has much more practical bearing on Taylor's responsibilities than the other. Indeed, in this chapter I hope to show the value of being "in the room" for the task of getting reliable.

This meeting is one in a series of meetings that Leah has called to go over the revised list of content codes. Content codes are used to describe the mathematical content of EOIs, from Counting 1–10 and Counting 11–20 to Identifying Shapes. Leah has her computer open and is projecting the list of codes on the wall. In each meeting, we have been working our way through portions of the list, clarifying our collective understanding of the difference between "analyzing" and "extending" patterns, between what a lesson about "cardinality" looks like versus a lesson about "counting." The schedule of meetings as a whole can feel tedious and obligatory, but in practice, each meeting typically involves some amount of charged back-and-forth between observers as we attempt to establish alignment in our textual practices of coding, how each of these codes should be used in effectively breaking up classroom phenomena into meaningful pieces. This chapter will show how this alignment of our textual practices was done through the co-participatory work of discursively producing and acting on classroom scenes in intersubjective space.

The Classroom Scene

Where previously we had all been students within a hierarchical teacher-student relationship between COG and Math4All, now certain members of the Math4All team have come to more frequently take on the teacher role with respect to observation. Leah, the most academically credentialed of all of us on the team, had been with Math4All the longest, and was well attuned to the constructs of the intervention. As such, she led discussions around

distinguishing between content codes. Meanwhile, Ashley had at this point already taken her trip to Empire City and established reliability with Andrew, becoming our reliability anchor. As such, she became our resident authority on observational procedure. Even Leah turned to Ashley on questions of proper protocol. With Ashley and Leah most prominently occupying the teaching positions within our group, the rest of us once more took on the student role.

During meetings, Leah would project the list of content codes on the wall, and lead the rest of us through explanations of each code, prompting discussions about how to spot these academic constructs in action in the classroom context, among other things. One of the most frequently used techniques in coming to mutual understandings of reliable coding practice did not rely on the use of videos, as in Chapters Two and Three, nor on the use of printed/digital documents as in the previous chapter. Rather, we spoke classrooms into existence, and co-operated upon them, renewing and transforming classroom scenes toward the disambiguation of our collective textual practices in coding.

Figure 2 is a transcript of such an interaction, in which Ashley gives a description of classroom scene in order to address how to code an EOI when its goal changes mid-event. Leah then asks a question, and Ashley reproduces the scene.

My contention in presenting this example is that what Ashley is doing by speaking is not transmitting her inner thoughts via sound waves into the minds of the rest of the observation team, but creating a text in intersubjective space which functions much like "the practice scenario," or "a solicited event of bear production" did in the previous chapters. In the cases of the practice scenario and the solicited event of bear production, their artifactual bases remained the same—that is, the inscriptions of the event log and the audio-visual phenomena of the video never changed—while their effective meanings-in-context transformed as they were encountered

and re-encountered over the course of an unfolding event of interaction. In this situation, however, given the ephemerality of speech, we will see that the evolving meaning of the invoked classroom scenario very quickly motivates changes in its artifactual renewal in speech.

Ashley

If they were like
"We're going to count to 10"
"1 2 3... 10"
and then some kid was like
"No, let's count to 15"
and she's like
"Oh, you guys want to keep going on?"
and they count to 15,
you wouldn't say it was `Count 1 through 10`,
you'd say `[Count] 11-20`
because they took it one step further than the initial goal.

Leah

Yeah, so you do count the higher one,
even if it wasn't the intended goal.

Ashley

Because the intended goal the way she stated it was
"Let's count from 1 to 10"
but then the kid was like
"Let's be even better than we thought
we were going to be and count to fifteen!"

Figure 2: Ashley and Leah have an exchange about an imagined classroom event

Bringing the Classroom to Life

In bringing her narrated classroom to life, Ashley must make it as available for our direct observation as the practice scenario or the bear scene. In order to do so, she makes abundant use of techniques of direct speech, that is, she speaks through the characters whose existence she is asserting, from their point of view. Each time she does this, she uses the locutive marker "like," commonly used in American English to index direct quotation.

01 If they were *like*
02 "We're going to count to 10"
03 "1 2 3... 10"
04 and then some kid was *like*
05 "No, let's count to 15"
06 and she's *like*
07 "Oh, you guys want to keep going on?"
08 and **they count to 15,**

Figure 3: Ashley's use of direct and indirect speech

In Figure 3, I show direct speech underlined and indirect speech in bold for the first eight lines of the transcript, the lines in which Ashley first invokes the classroom scene. Indirect speech, unlike direct speech, makes a claim of 'second-hand' reportage. Ashley's heavy use of direct speech works to bring the classroom scene to life in the meeting room, making it available to us for direct observation—the activity we are supposed to be getting reliable on. She performs the classroom scene directly 'as it happened,' even taking on the first act of counting as an elided instance of direct speech (line 03). The second time the tedious act of counting appears (line 08), she defers to indirect speech.

In the room, not captured by the transcript, Ashley also changes her vocal quality whenever she ventriloquizes her narrated characters, pitching it higher and/or changing the timbre, adding a melodic prosody to her phrases that is absent from her strictly narrative mode.

This adds a distinctive vibrancy to the characters whose speech she is both presencing and animating.

The scene in intersubjective space then can be roughly represented as in Figure 4, with Ashley's utterances marked at the top and the narrated event demarcated within boxes. Note that Ashley's utterances occur in the meeting room, which is not depicted below, while the instances of direct speech occur both within the meeting room and within the intersubjective scene.

Tracking the pronominal deictics (they, we, let's, she), as in Figure 5, further contributes to the effect of direct speech, where third-person pronouns act to assert the existence of a character in the scene—"she", the teacher, "they," the students—while first- and second-person pronouns produce the effect of those characters speaking to each other, in line 02 for instance, "we" is the teacher speaking about herself and her students, not Ashley speaking about herself and the other observers in the room with her.

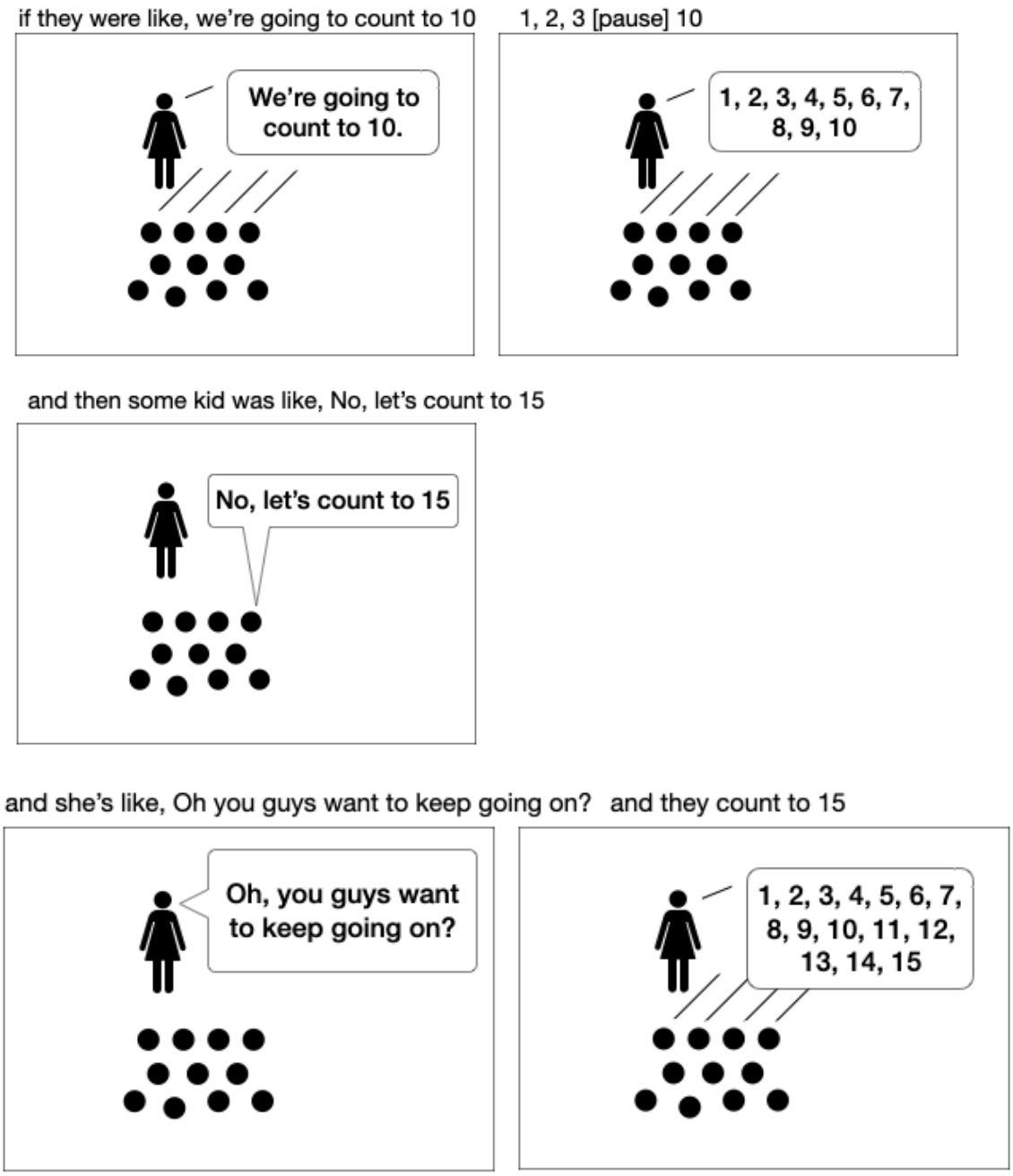

Figure 4: The Narration and the Narrated Event

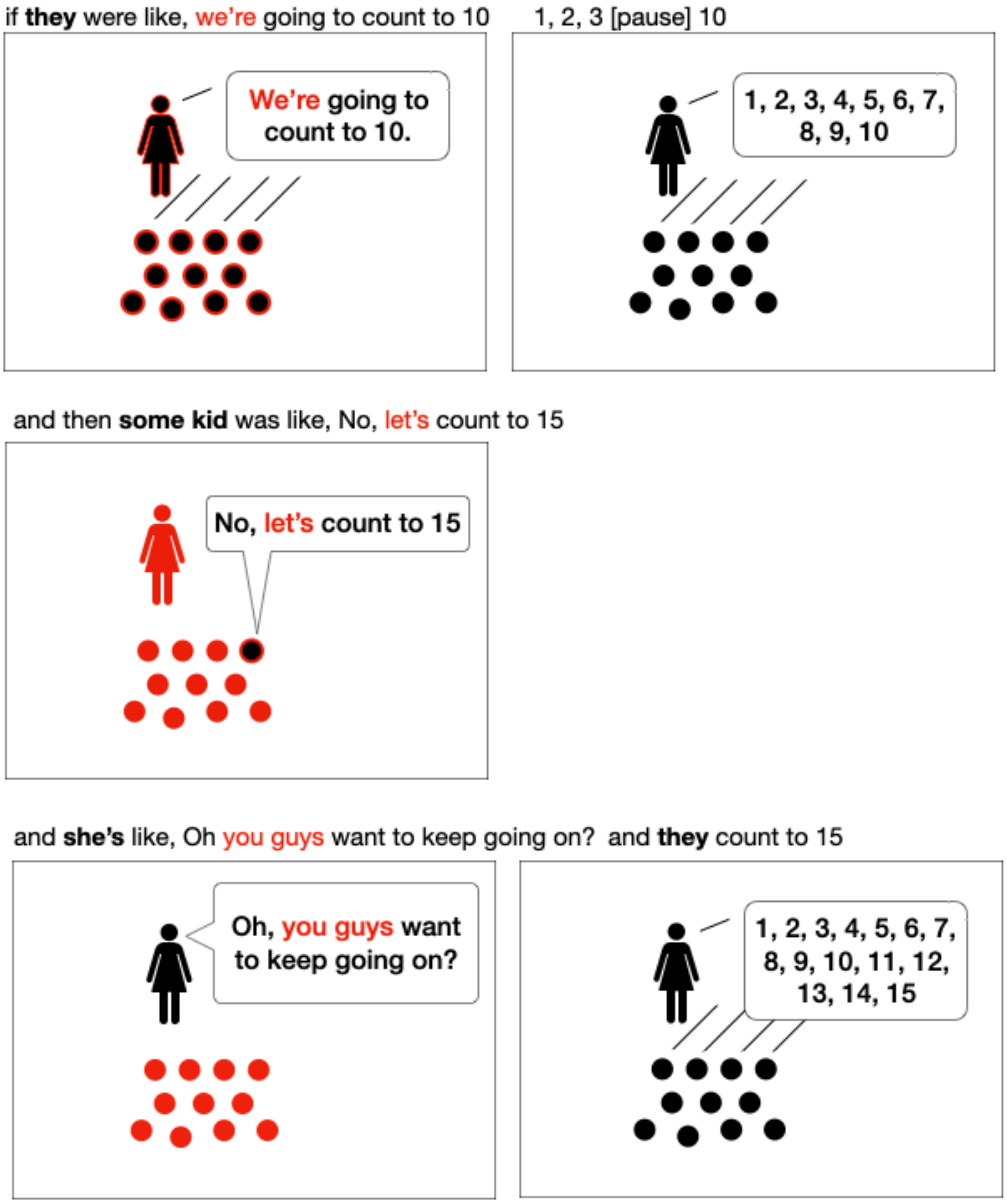

Figure 5: Use of third-person versus first- and second-person pronouns.

Third-person pronouns and their referents are marked in black, while first- and second-person pronouns and their referents are marked in red. If a referent is referred to both by the narration and within the narrated event ("they" and "we" in the first frame), the referent takes on the color of the pronoun which occurs first, and then takes on the second color as a border.

It is likely that readers, just like the people in the room with Ashley, are able to recognize direct speech without this technical breakdown of the patterns which produce its presencing effect, the effect of bringing the classroom into the meeting room. Ashley's direct speech is so self-evident that in my transcript I have allowed myself to include quotation marks around it, as is conventional in other literary forms. Because we share in these cultural forms with Ashley, this technical explanation may feel unnecessary, and even pretentious. Yet, I engage in this evidentiary procedure nonetheless, as would be expected in the identification of interactional patterns in the analysis of less familiar cultural forms. It is worth remembering that the effects of language-in-use are not transparently decodable from grammatical form alone, but are culturally, interactionally, contextually mediated. One day, should this ever be read again, even the average American academic reader might find our taken-for-granted forms strange. Finally, in calling attention to the use of pronouns, we have already begun to lay the groundwork for understanding the brilliance of Ashley's next few lines.

Bringing "You" Into the Scene

So far then, Ashley has created a virtual classroom scene towards which all of us in the meeting can orient. Modeled in this way, our reliability—our alignment with our anchor, Ashley—can be understood as emerging from the similarity of our alignment toward this virtual classroom scene. In directing us to the proper way of aligning towards this scene, Ashley then introduces the figure "you." This "you" appears to be a classroom observer, who is correctly selecting the code Counting 11–20 over Counting 1–10. We can depict this, as in Figure 6, as a narrated event of coding, in which the previous classroom scene is nested.

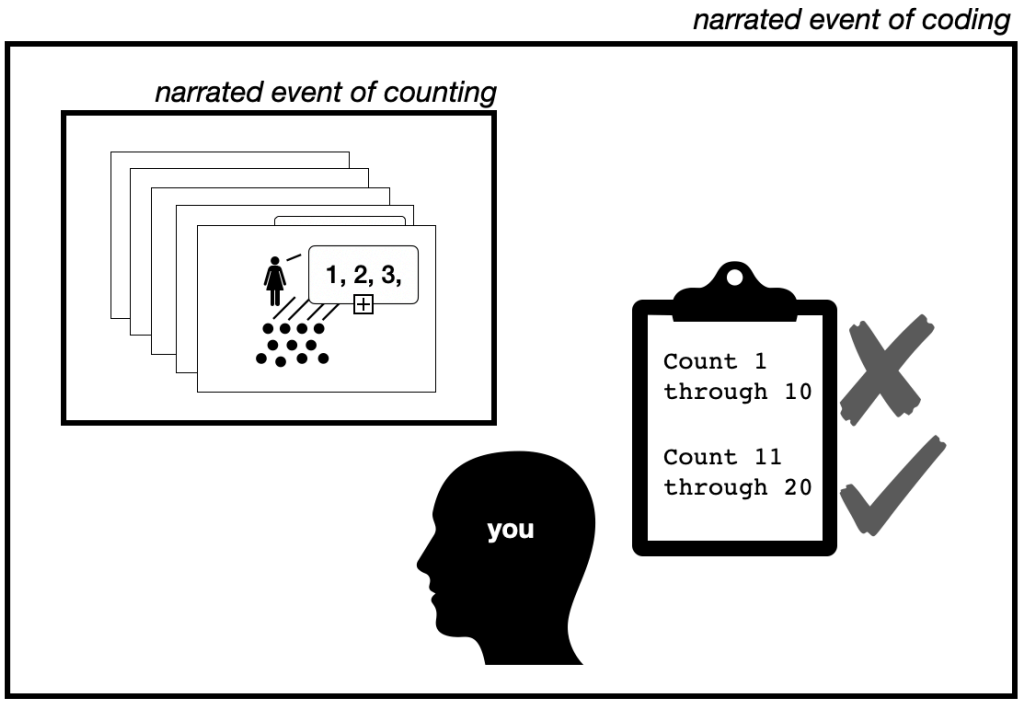

Figure 6: The narrated classroom scene nested within yet another narrated event of coding, in which "you" are the observer/coder.

But "you" is not simply any label for this narrated observer. The word "you," like the pronouns discussed in the earlier example, is a deictic, a linguistic unit whose reference depends upon its context of use. Other deictics include words like "here" or "there," "this" or "that," "come" and "go."[4] In conversation, I can only know who the word "you" points to depending on who says it and how they say it. What is remarkable in this situation is that the word "you" is simultaneously evaluable with respect to the narrating event of coding AND the "narrating

[4] Linguistic anthropologists have sometimes called deictics "shifters," and deictics have featured heavily in theorizing indexicality as the mechanism by which language has social effects (Silverstein 1976). Wortham (2005) performs an analysis similar to mine, tracking the deictic "you" in describing how repeated student recruitment into certain roles within classroom examples mediates student identity formation.

event," the event in which Ashley is speaking to a group of observers about an observer coding a classroom scene. Because she is speaking *to* us, we, of the narrating event, are also the "you" of the narrated event.

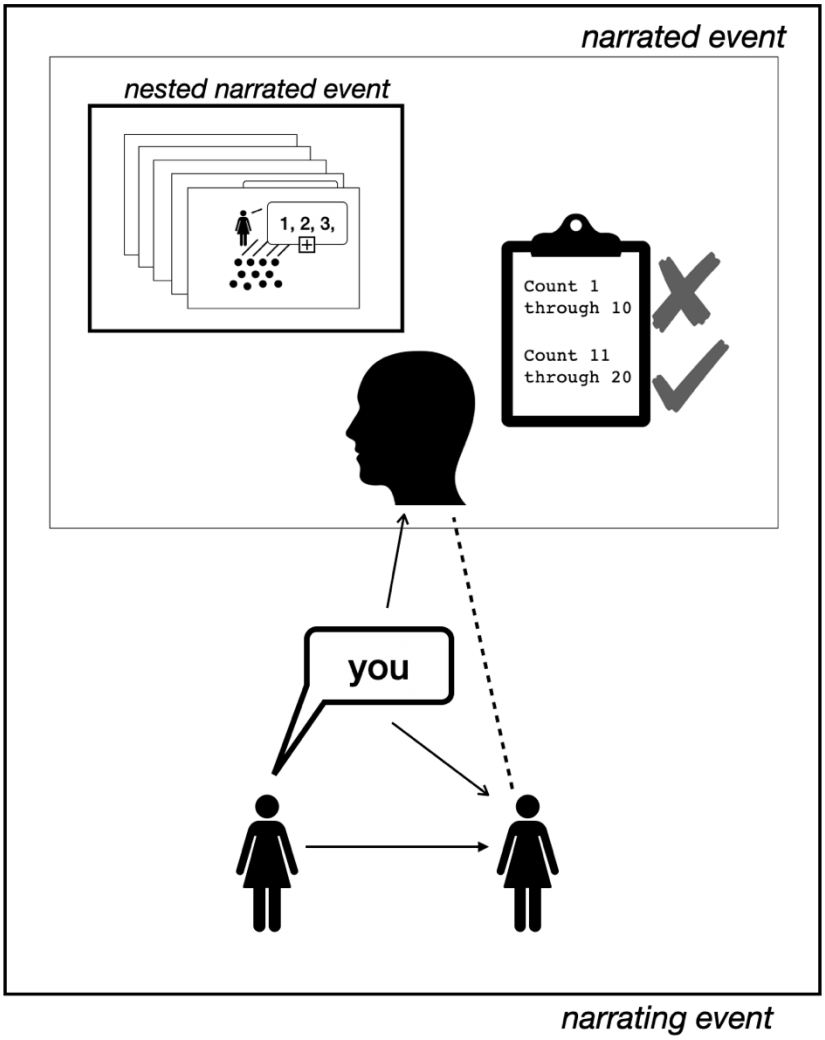

Figure 7: "You" anchoring the narrated event to the narrating event.

As shown in Figure 7, Ashley's use of "you" crosses the boundary between narrated and narrating events, both bringing her interlocutors (the other observers-in-training) into the classroom scene as the narrated observer, and indexically anchoring the classroom scene within

the here-and-now of the observation meeting, as the thing which we, being part of the narrated event of coding, are observing. Not only that, but in line 11, Ashley tells us what each of us, as the narrated classroom observer, should be thinking about the classroom scene when we make our coding decision: “they took it one step further than the initial goal.”

Once More with Intention

Ashley has produced a complex, unfolding mediator in the intersubjective space of the meeting. Even within this short stretch, Ashley herself, in lines 09–11, is able to reflexively point to and comment on the event projected by her own prior utterances. Indeed, as her speech unfolds, each moment becomes available context for the next moment, and we can understand

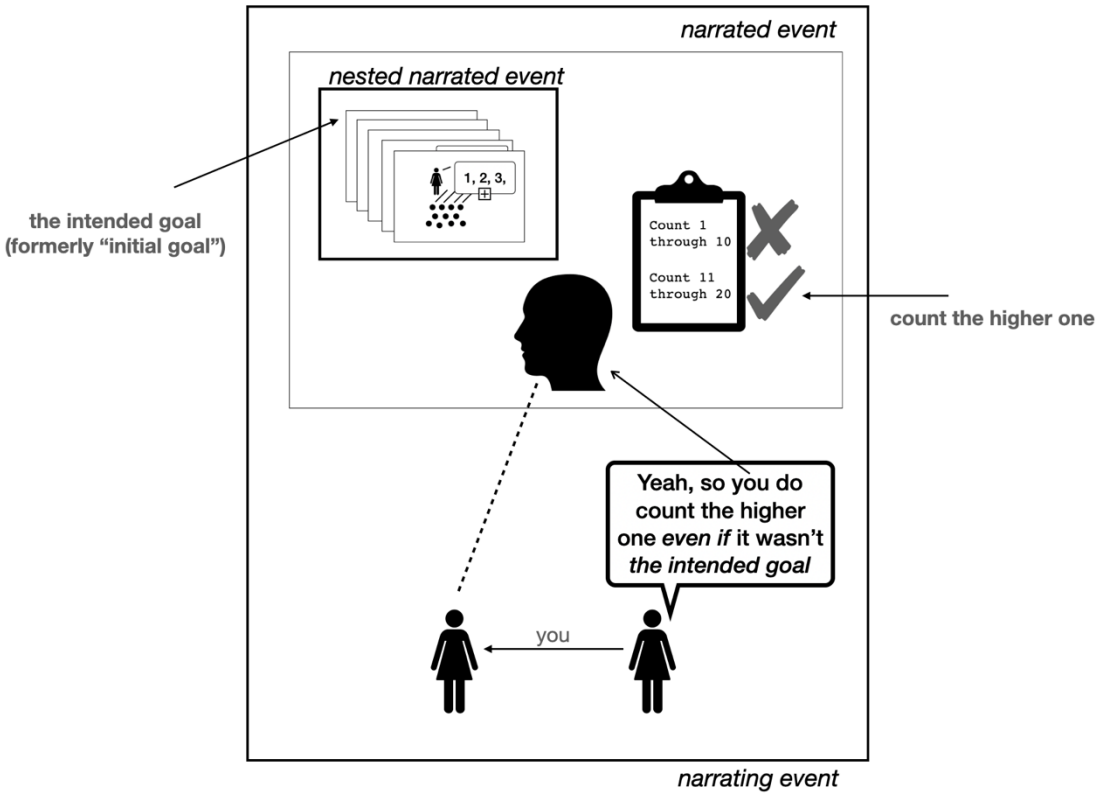

Figure 8: Leah introduces “the intended goal”

her to be in meta-discursive interaction with her own utterances as they unfold over time. She might as well have produced a video of a classroom scene and played it for us, as the PD team did for the teachers in Chapter Two. The material ephemerality of her speech makes little difference in this respect. The narrated events are intersubjectively available to both present and future Ashley, and as we will see, anyone else who cares to take them up in their own conversational contribution, as Leah does in Figure 8.

Leah's contribution is a re-telling of the narrated event of coding, now in the timeless present ("you do" vs Ashley's "you would"), in which "you," the observer, make note of a fresh concept: intentionality. Again, we can understand Leah's turn as Leah pointing back at the events conjured by Ashley's utterances and working on them herself, much like she would add a comment in a digital document, or mark up the margins of a paper manuscript. Again, the ephemerality of the events' initial materialization in Ashley's speech does not prevent Leah from taking them up, in a move of simultaneous reproduction and transformation.

Now, of course, Leah's contributions are also available in intersubjective space. Ashley takes advantage of this, taking up Leah's "intended goal" and using it to re-tell the original narrated event, now injected with markers of intention.

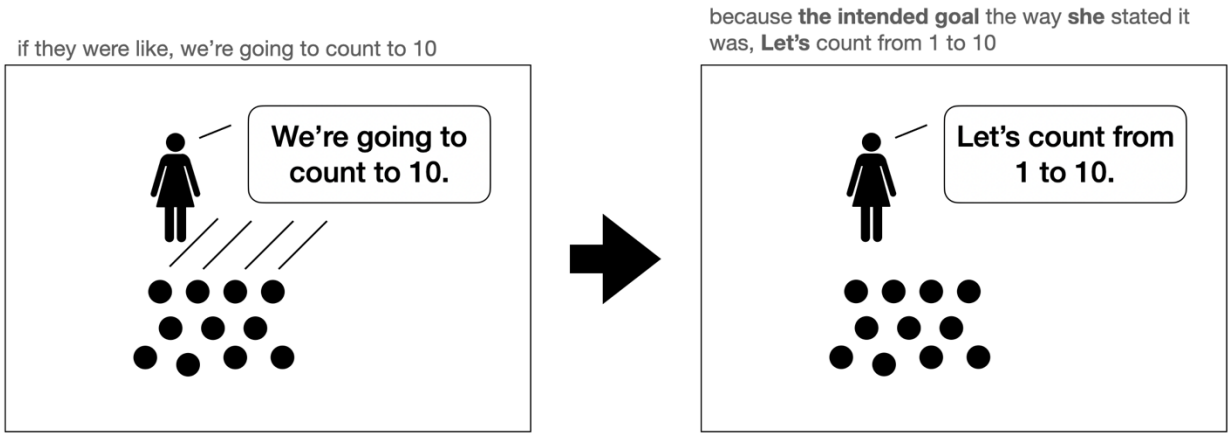

Figure 9: The original telling (left) re-told with intention (right).

As Figure 9 calls out in bold, Ashley incorporates intentionality in a number of ways. First, in line 12, Ashley re-names counting to ten as "the intended goal" rather than "the initial goal" as she referred to it in line 10. Second, rather than use the ambiguous "they" of line 01, Ashley uses the pronoun "she" to explicitly call out the teacher as an intentional agent. Third, Ashley shifts from the indicative mood of "We're going to count to 10" into the hortative mood of "Let's count from 1 to 10," indicating an active request. This is the same hortative mood that Ashley originally attributed to "some kid" in line 05, as he exhorts his classmates to count to 15.

Breaking the Fourth Wall

In the final scene of her re-telling, Ashley modulates her voice up to a high child-like pitch, and in a breathless rallying cry, intones "Let's be even better than we thought we were going to be and count to fifteen!" I use an exclamation mark in the transcript to mark the fevered pitch of her delivery; but perhaps more notable is the combination of direct and indirect speech evident in this cry (lines 17–18). In re-working and re-playing the narrated event for the pedagogical purpose of illustrating normative coding practice to Leah and the other observers, Ashley merges the interpretive stance of the narrated coding event (line 11 "they took it one step further than the initial goal"), into the direct speech element of the narrated classroom event (line 05 "Let's count to 15"). The resulting effect is akin to a theatrical breaking of the fourth wall.

The fourth wall is a term used to describe the virtual divide between audience and actors in a stage play, separating the events within the world of the play (our narrated event) and the events of the world in which we watch the play as a play (our narrating event). "Breaking the fourth wall" is when events occur within the play that reflexively call attention to its play-ness. As a phrase, "breaking the fourth wall" is most often used to describe instances of actors

speaking to the audience, a group of people whose existence they should not be aware of. However, any number of things may happen to suggest that the actors know that they are in a play, or that the play was written by writers who know this is a play of a particular genre, and so on.

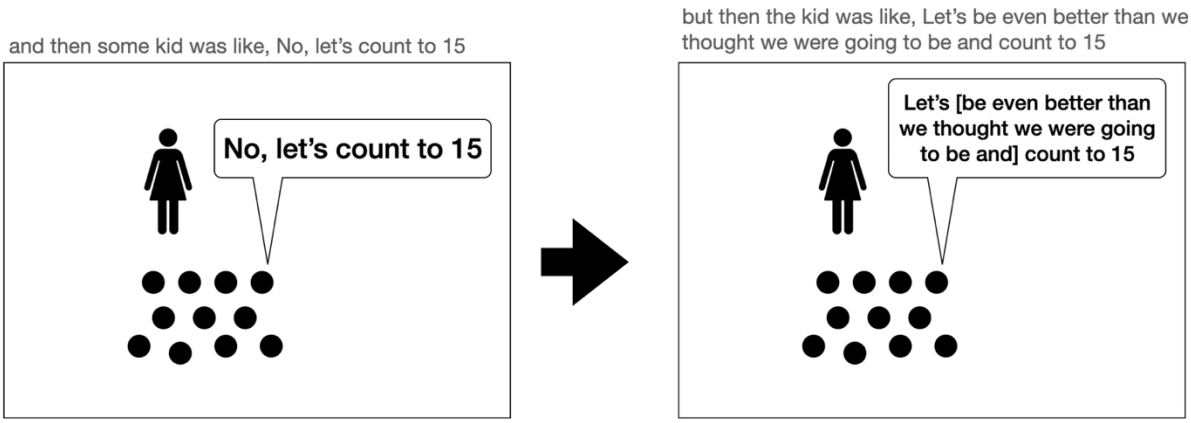

Figure 10: Original telling (left) to breaking the fourth wall (right)

In Ashley's narrated event, the bold student breaks the fourth wall by not only making his original demand, "Let's… count to fifteen!" but also describing the nature of his request in terms directly relevant to his heretofore unacknowledged audience, Ashley's addressees in the narrating event: "Let's be even better than we thought we were going to be…!" The whole of this short exchange is brought to bear in effecting this lesson: that when "you," the observer, hear the interruption of line 05, it should sound to you exactly like the declaration in lines 17–18.

The Reality of Intersubjective Space

In the preceding analysis, I argue that a discursive mediator unfolds in intersubjective space in the form of virtual nested narrated events, and that it is functionally equivalent—if not

more dynamically responsive—to more perduring text-artifacts like videos and other documents, as we have seen in the previous chapters.

One might wonder, to what extent the classroom scene, the seemingly most crucial "shared" object by all the individuals in the room, is actually shared? How do we even know that my interpretation of the scene is equivalent to the interpretations of the various people in the room? Is it really appropriate to analyze it as a "shared" mediator? Wouldn't it be more accurate for the diagrams to show each person imagining their own version of the narrated event in their personal cognitive domain, rather than orienting toward some central narrated event in intersubjective space? All of these questions point to the same concern with the analysis: How important is it that the scene be 'the same' for each person in the room?

There is no point of view from which we can neutrally evaluate the congruency of each individual's conception of the classroom scene to the others'. Further, we cannot simply assume that any individual conceptions are somehow inscribed within clearly bounded thought or brain activity such that they could be de-contextualized from the interactive contexts within which said conceptions emerged and unfolded. However, these caveats do not allow us to simply assume congruency. On the contrary, I follow statistical thought in assuming the ubiquity of variation; and similarly, I expect there to be some shape to that variation (i.e., it is non-random). In describing the production of intersubjective alignment then, I am describing the by-degrees reduction of the range of interpretive variation, along axes of evaluation that have been selected for as meaningful within the interaction (e.g., the interlocutors care about who brings up what counting goal and in what order, but they do not care what color the students' shirts are).

All that said, we must take recourse to pragmatism in response to the above-stated objection to the analytic representation of the classroom scene as an intersubjective mediator.

Pragmatism emphasizes that the reality of any object is equivalent to its consequentiality. Under this logic, we can treat the scene as unfolding in intersubjective space if the co-participants within the interaction *act as if* each participant is, or should be, orienting to 'the same' scene. This is a pragmatic orientation to the phenomenological experience of intersubjectivity as real and consequential over and above any God's-eye evaluation of the sameness of each person's private imaginings.

Is it possible that there were people in the room who did not speak up who had an entirely different picture of the classroom scene in their minds and simply could not then make sense of what was happening? Or that there were people present who simply were not engaged with the discussion at all? Yes and yes. Confusion was readily and repeatedly brought up by all members of the reliability team (excepting the single member of the team who was regularly inactive during meetings, instead focusing upon other work while physically present[5]). None of the observation team members appeared to have any reticence about making a felt loss of intersubjectivity known, when they no longer seemed to be "on the same page" as others.[6] Not only that, but confusion about what was happening in any given scene was usually the cause of the kind of re-playing and re-framing of scenes that we saw in the example. That is, even though the example is one in which existing agreement (or non-existent disagreement) is built up, it might as well have proceeded with Leah expressing confusion or disagreement over what the

[5] It is possible that my presence as an observer affected this behavior, but I cannot say either way.

[6] The only sanction against discussion during observation meetings was a recurring reminder from Leah that we only had so much time, and that we could debate any single point "for hours." This was usually issued after extensive conversation on any one point of confusion had already taken place, sometimes allowing Leah to move the conversation forward to the next topic.

intended goal was, and Ashley reproducing the final scene in the same manner toward the same pedagogical end.

Perhaps paradoxically, disagreement is generally the cause of meta-discursive work on some mediator in intersubjective space, such as 'the conversation we've been having,' or 'the example that you brought up.' In fact, disagreement and confusion are only possible given participants' working hypothesis of a shared intersubjective space. Where else might a point of confusion or disagreement be found? If we do not act as if there is some singular thing which we can both orient towards, then what could we possibly agree or disagree about? Disagreement, private or public, is an excellent index of the feeling of a shared reality which intersubjectivity names. Only if we presume that we live in 'the same' reality, can we meaningfully disagree; only if we presume we are talking about 'the same' classroom scene can we disagree over its proper coding.

So, our analysis does not require everyone to share an objectively identical orientation (if such a thing were possible or could be ascertained) towards a mediator to proceed. It is enough that individuals proceed in their interactions as if they should be, or could be, talking about the same thing, even if they disagree on what that thing is or how to talk about it.

Conclusion

In adapting the COG tool toward their own use, the Math4All team resorts to the adaptive technology of talking during meetings in order to work towards inter-observer reliability without COG's established support infrastructure. Ashley's invocation of and reflexive commentary on an imagined classroom scene assists the Math4All observers in calibrating their textual practices, sometimes with the help of characters in the scene themselves.

Methodologically, this analysis shows how discourse—language-in-use—can point to and build on itself as it unfolds over time, demonstrating how language's capacity for reflexivity—that language can be used to talk about language—allows it to be an ideal medium for human culture, just as a nutrient broth mixed in agar serves as an ideal medium for cell culture. In this case, the building up of a series of rehearsals of a classroom scene acts as an unfolding mediator against which each participant (including those non-speaking) can orient themselves, and in doing so, calibrate their alignment with Ashley, and approach formal reliability.

For the linguistic anthropologist, the analysis of communication is not about determining what people really meant, or reckoning the degree of fidelity between what was intended and what was received. Instead, communication is taken to be the activity of (re)producing cultural concepts and social relationships in interaction.

Conclusion

In this work I have tried to establish that even objective means of creating and sharing knowledge are not asocial or apolitical "technical" affairs. They rely on textual practices which emerge from the political projects of various groups; the discursive, interactional regimentation of such practices—the work of communication—is the activity of social reproduction and social differentiation.

But beyond their ideological character, which is after all, not special in the least, strategies of objectivity, which prominent strains of education research have turned to in their projects of knowledge delivery, tend to hinder themselves by straining the very relationships required for their success. In particular, the work of fidelity does not work to communicate shared interests between researchers and teachers, even, as we see in Chapters One and Two, when intermediaries directly set this alignment of interests[1] as their goal. Rather, the constraining action of fidelity ends up communicating teachers' place outside of the domain of science, experiment, knowledge production. That is, communicating fidelity was simultaneously a form of ex-communication.

Divisions of Labor

In tracking teachers' and researchers' trajectories from peripheral to full participation, I have argued that the full participation of teachers under fidelity does not much resemble the full participation of researchers under reliability. Even in their full participation, teachers do not gain the ability to produce their own knowledge, but remain tied to researchers' projects of knowledge production. I take the cases of fidelity and reliability to illustrate two different

[1] See "translation" in Latour (1993)

possibilities for organizing the division of labor in the Research-Practice nexus, that is, the coordination and distribution of the types of work entailed by a scientific system of education. I describe two important points of contrast below.

The first point of contrast concerns the distribution of textual practice between trainers and trainees at the point of full participation. In the case of reliability, at the point of full participation, a set of the Math4All researchers and a set of the COG researchers were engaging in the same practices of classroom observation, this claim of sameness grounded in their demonstrations of reliability. In the case of fidelity, at the point of full participation, the Math4All teachers were engaged in textual practices which were unique to their station.

The PD team staff were not in the business of assessing children on Math4All. Though they were also adept at the textual practices they were training the teachers in; their primary occupation was support, not assessment. Further up the ladder, it was unclear if Math4All researchers with no teacher support function had ever seen the Math4All assessment in its full artifactual glory—what we described as "the box"—as it was being distributed and used during the efficacy study.[2] This is not to say the PIs and other researchers were not conceptually familiar with the tasks themselves. They would talk about select tasks in presentations, and often described unique and clever design features which improved task validity. However, I do not know of any occasions in which they would have encountered or handled The Box or its contents during the course of my fieldwork.

[2] The Box had been assembled mostly on the "psych side" with help from additional undergraduate / postgrad RAs working across several of the Math4All PIs' projects. I was very enthusiastic about the large amount of mindless labor involved in binder assembly (cutting shapes out of paper, sliding sheets into plastic sleeves, and so on) and so volunteered to do as much of this work as I could. This personal proclivity towards mechanical work was regarded as mildly bizarre by some and highly relatable by others.

The second point of contrast concerns the level of autonomy or discretionary judgment afforded to former trainees at the point of full participation, illuminating differences in the entailments of reliability and fidelity as trusting and non-trusting modes of organization, respectively.

With reliability, the Math4All team, at the end of their training period, were able to both modify the observation tool and independently collect data in service of their own research agenda. That is, at the point of full participation, when training activities had effectively ended, an additional capacity for classroom observation was extended across a division of epistemic labor organized by research agendas. The COG team would continue to develop their observation tool and use it to study the quality of classroom instruction; the Math4All team would continue to develop their assessment and use classroom observation to collect outcome data. Having met a professional standard of reliability, both were authorized to conduct classroom observations in a certifiably objective manner in pursuit of their own research aims.

That the COG team and Math4All team had entered into a relationship which was expected to end in this way bespeaks a fundamental level of trust premised on their mutual recognition of each other as being part of the same scientific community. They were each part of a division of labor within education research similar to the societal divisions of labor discussed by Shapin (1994) in citing Putnam's "division of linguistic labor" and Rorty's notion of science as a "model of human solidarity" (p.23). Within the COG-Math4All relationship of reliability, this division of labor remained even after training, such that even at the point of full participation, the Math4All team did not have the standing to go around training people in classroom observation. COG group retained their relatively exclusive status as a "site of emanation" for the practice of classroom observation (Silverstein 2013). However, we on the

Math4All team were trusted to do observations without COG supervision, and our statisticians were trusted to analyze the resulting data without any involvement from COG.

In discussing the role of trust in the division of labor, I am not only describing the practically necessary situation in which I can go about functioning in relative ignorance of most subjects because I trust that an array of by-degrees accessible someones does hold that knowledge; rather I am trying to describe different ways of organizing cooperative activity given the presence or absence of trust. I take Math4All and COG's parting of ways as a consummation of the trust that brought them together to begin with. This parting of ways is starkly revealing about the constraints of the distrusting fidelity relationship in relation to the flexibility of the trusting reliability relationship.

In the fidelity case, the intended use of the Math4All assessment required users to remain tethered to the Math4All group. When teachers participating in the efficacy study produced student data, that data was not theirs to analyze, to turn into knowledge. Rather, it was to be sent to the Math4All research team (via website) in order to undergo statistical transformation into student knowledge. In fact, in the course of being trained in administration, the Math4All teacher-participants were warned off of any attempts to use their observations and scores as a basis for producing student knowledge with respect to math abilities. Rather, any hypotheses they came to in observing students' behavior in the course of the assessment itself would have to be confirmed or disconfirmed by the statistical machinery of Math4All itself.

Unlike the situation between COG and Math4All, Math4All did not trust teachers to produce good knowledge. The fidelity relationship operationalizes the distrusting relationship between researchers and teachers: researchers do not view teachers as fellow members of their

moral community. Even if they possessed the same scientific values (i.e., objectivity), they lacked the training to enact those values through appropriate practices of knowledge production.

In this case then, rather than extend the capacity for knowledge production to collective actors with independent agendas, the fidelity model extends the domain of Math4All researchers' knowledge production by employing teachers as technicians in the most mechanical sense of the word. This division of labor is not defined by differentiated aims as in the reliability case, but differentiated roles in service of a single aim. That aim is determined by the party whose intentions set the standard for fidelity, in this case, the Math4All researchers. Lacking the flexibility of use which characterized Math4All's use of the COG observation system, implementation figures a rigid assembly line of knowledge manufacture, divided among workers specializing in particular tasks. When analyzed in this way, the producer-consumer / producer-user metaphor which has often been used to characterize the role relation between researchers and teachers does not quite hold. Teachers are not only consumers or end-users, but indispensable actors in the manufacturing process.

Even when the student data gets transformed into student knowledge and returned to teachers—and this knowledge is sincerely intended to support teachers—another series of workshops will take place, to help teachers understand how to faithfully interpret that knowledge as meaningful. And then another series of workshops will follow in order to guide teachers in faithfully using those interpretations to guide their instruction. That is to say, the regimentation of teachers' textual practices not only enables the faithful production of data to feed researchers' knowledge production, but the subsequent delivery of that knowledge as "useful" is also premised on the further regimentation of teachers' textual practices. Knowledge and data 'travel' are not dependent on the qualities of the knowledge or data itself, but on the successful

cultivation of teachers as a well-behaved class of doers, faithfully responding to the demands of the thinking class.[3]

In reliability, knowledge travels along lines of reproduction—Math4All observers learn to engage in the same textual practices as COG observers; in fidelity, knowledge travels along lines of stratification—each strata learns the textual practices suited to their station.

Objections

One might object: Is this not a perfectly acceptable division of labor? Is it unacceptable merely because we can identify a hierarchical structure of exclusion? Doesn't any division of labor logically necessitate exclusivity in order to effect efficiency through specialization? And finally, is this not a highly uncharitable mischaracterization of the motivations of those who work in Research?

I will respond in reverse order.

First, in describing how fidelity produces Practice as an object of Research, I am positing social phenomena which are empirically continuous with the best intentions of the actors involved with them. Research's treatment of Practice as a site of data extraction does not depend on the existence of researchers actively hellbent on turning schools into data mills for their own glorification. This would require that consequences be constrained by intentions (a premise similar to the logic of fidelity) (cf. Dewey 1922, p. 231). My analysis is premised on the opposite

[3] An assembly line–style division of labor also existed within the Math4All research team itself, as the individuals on the observation team were by and large not the same individuals who were doing the statistical analysis, and when there was a demand for greater data collection capacity, many lower-ranked research assistants were conscripted to man the effort. That is to say, we can observe the fractal recursion (Irvine and Gal 2000) of the thinking/doing divide within the thinking fraction itself; and we will also observe its recursion within the doing fraction in the differentiation of groups of participating teachers as more thinker-like (the GC group) and more doer-like (the others).

assumption; that consequences are not mediated by intention, but context. If action is entextualized intention, consequences are contextualized action.

I believe that the great majority of education researchers invested in SBR-style objectivity, fidelity, and evidence, deeply and sincerely believe that they are taking the best, most ethical, course of action possible, given their considered assessment of the world they live in and their place within it. It is *because of*, and not *in spite of*, their sincere efforts to improve American education through scalable, evidence-based interventions that the events I have spent this work analyzing have taken place at all.

Second, is the fidelity model bad simply because it is hierarchical? Aren't researchers better suited to do knowledge work, and wouldn't teachers be better off given the knowledge researchers generate?

As historians of education have demonstrated, researchers' authority in education is not a straightforward consequence of the goodness of the knowledge they produce, but, as Chapter One elaborates, the result of convergent political interests under an ideology of objectivity. In fact, the very existence of the Research-Practice Gap concept calls into question what "good" should mean in education research, if not "useful." But, as Raudenbush comments following an astute reading of Porter (2003): "Could historically specific conditions have generated a logic of inquiry that transcends those conditions to be of much more general use, a logic that might be useful in other historical worlds?" (p.275).

How useful is the abstract knowledge produced by objective methods? As Burkhardt and Schoenfeld (2003) write: "[G]eneral theories are weak, providing only general guidance for design; nonetheless they receive the lion's share of attention in the research literature" (p.10). The production of evidence, even about a specific intervention, is a form of general theory,

because “the” intervention itself is always an abstracted general construct (a text), though it may attach itself to procedural and material specificities (text-artifacts). Once “the” intervention hits the ground, re-contextualization must occur, and the intervention must either change, adapting to the new environment, and/or the environment must adapt to it.

In the COG reliability case, both of processes of adaptation were evident, as the observation group became entrained in new textual practices, and as they made their own modifications to the tool given the role the COG system would play within the Math4All context. However, fidelity requires the maintenance of the integrity of “the” Math4All intervention, in order to calculate the efficacy *of the intervention itself*, demanding that only the environment change (the teachers, the teachers’ schedules, the organization of space in the classroom for administration, etc.) to adapt itself to Math4All. The successful scaling up of Math4All then, demands a homogenization of its destination environments in order to travel, especially the homogenization of teachers and their textual practices.

I argue that the activities which produce homogeneity under fidelity foreclose the usefulness of educational interventions by imposing a distrusting relationship between Research and Practice. The demands of fidelity in the service of objective evidence require the mechanization of teachers’ work, or at the very least, for Practice to act as Research thinks appropriate. Dewey [1916] (1997) provides what is an apt description of this situation:

> Much is said about scientific management of work. It is a narrow view which restricts the science which secures efficiency of operation to movements of the muscle [read: fidelity of implementation]. The chief opportunity for science is the discovery of the relations of a man to his work—including his relations to others who take part—which will enlist his intelligent interest in what he is doing. Efficiency in production often demands division of labor. But it is reduced to a mechanical routine unless workers see the technical, intellectual, and social relationships involved in what they do, and engage in their work because of the motivation furnished by such perceptions. The tendency to reduce such things as

> efficiency of activity and scientific management to purely technical externals is evidence of the one-sided stimulation of thought given to those in control of industry—those who supply its aim. Because of their lack of all-around and well-balanced social interest, there is not sufficient stimulus for attention to the human factors and relationships in industry. (p.85)

As Chapter Three showed, the PD team's workshop planning was driven by a desire to engage teachers' "intelligent interest" in Math4All, and reveal to them the "technical, intellectual, and social relationships" involved in what they do. However, in the actual workshops, these motivations were undercut by the requirement to secure fidelity. That is to say, the maintenance of interventions as singular, coherent entities, whose efficacy or effectiveness may be measured, runs counter to the ability to recruit teachers as participants within a scientific system of education so imagined.

However, as the Golden Child group shows us, there are teachers who are more willing and able to see themselves as participants within the projects of academia. Indeed, the educational political projects of the research university have always involved attempts to produce a teacher workforce which is more amenable to academic control (Murphy 1990, Lagemann 2000). For example, historian Lagemann calls attention to the efforts of University of Chicago President William Rainey Harper toward raising educational requirements for K–12 teachers in order to reduce the number of working-class, immigrant, community-based, union-friendly teachers standing in the way of his ambitions for the University as a site of educational authority (2000, p.13).[4]

In the present-day, the K–12 public school teaching workforce is overwhelmingly White (79%) and well-educated (over half hold post-baccalaureate degrees) (NCES 2021), but early childhood educators are both less White and significantly less educated (Paschall 2020). As such,

[4] He did not succeed on this front.

policies requiring higher levels of education for early childhood educators would be a boon for interventions like Math4All, permitting easier travel (scaling up) among a more homogeneously White and affluent class of preschool teachers.

Alternatives

While the routinization of teachers' work is troubling in and of itself, the squandering of their scientific potential is perhaps even more troubling for the project of creating a scientific system of education, especially one that holds democratic aspirations. Charles Sanders Peirce, in his 1898 lectures, proclaimed the first, and sole, rule of reason: **Do not block the way of inquiry** (1992, p.178, emphasis in original). Counter this precept, SBR-style interventions attempt to distribute knowledge without distributing knowledge production, to distribute scientific knowledge without distributing scientific inquiry. The clearest alternative to this pipeline system of scientific inquiry is then the democratization of scientific practice, not a bridging of the Research-Practice Gap, but an erasure of the boundary defining Research as distinct from Practice.

Numerous researchers have proposed incorporating teachers into the research process as scientific agents, and provided models through which this could happen, with more extensive knowledge of teachers, schools, and so on, than I have ever possessed. Among those proposals include Dewey's famous "laboratory school,"[5] practitioner research,[6] design experiment,[7] some varieties of research-practice partnerships,[8] professional learning communities,[9] and more. My

[5] No longer functioning according to Dewey's vision, it now serves the well-heeled children of University of Chicago faculty.
[6] Cochran-Smith & Lytle 2009
[7] Cobb et al. 2003
[8] Coburn and Stein 2010
[9] Stoll et al. 2006

work here has not been to investigate these models. I can only hope that the practitioners involved will be afforded the same environmental supports that academic researchers demand for their own projects. Toward the development of such alternative models of scientific education, I hope to model in this study a form of inquiry which attends to the consequentiality of means over the vindication of ends, the unfolding of the present and its range of outcomes over the correspondence of intention and some circumscribed effect.

Appendix A: Critiques of the Human Capital Approaches to the Achievement Gap

The Achievement Gap has been recognized by a wide variety of actors, from a wide range of interested points of view. Crucially, the Achievement Gap does not signify the same type of problem or same type of solution from all such points of view. Below I note three major lines of critique against the human capital view that the Achievement Gap is fundamentally driven by gaps in skills development, or to say the same thing, that the Achievement Gap can be ameliorated through skills-based interventions.

First, stratification is a (condemnable) feature not a bug. This line of critique holds that the human capital approach fundamentally misunderstands the societal function of schooling, in taking it to be a site of skills development, whether cognitive or non-cognitive. Schools are actually sites for the reproduction of social stratification in accordance with dominant (e.g., anti-Black, patriarchal, capitalist) regimes of value.[1] Both the (e)valuation of skills and the benchmarks of achievement are dynamically adjustable in service of this stratifying function.[2] That is, the same practices may not be recognized as skills when practiced by different populations; and the most valuable skills and credentials will always be kept out of reach for those in disadvantaged populations, by way of gatekeeping and/or educational inflation.[3] Generally, in this view, schooling for public good begins with a fundamental revaluation of the aims of education.

[1] Bourdieu and Passeron [1977] (1990) and Meyer (1977) are usually cited as the major scholarly promulgators of this view.
[2] See, for instance, Puckett and Rafalow (2021) on the "negotiated" categorization of technological skills in education.
[3] See Labaree (1997) on the "credentials race."

Second, sociological and economic scholarship indicates that educational parity does not translate into socioeconomic parity. In fact, in order to achieve the same levels of socioeconomic status, Black families must accumulate more schooling and credentials than White families.[4] Furthermore, evidence shows that anti-Black discrimination actually intensifies at higher levels of educational attainment.[5] Generally, in this view, the solution to problems of racial disparity involve reckoning with the historical and on-going conditions of that inequality through direct legislative redress, such as a program of reparations for Black Americans (e.g., Darity and Mullens 2020).

Third, "gap" discourses around skills take up culturally White practices as the neutral standard against which minority practices are then legible as deficient. Focusing on "skills development" among racialized minority populations bespeaks a point of view from which minority populations are deficient with respect to White majority, and a failure to recognize (1) that the White standard is not neutral but works to devalue and/or erase (2) minority communities' existing skills.[6] Generally, in this view, the appearance of gaps should incite investigations of the subject position from which such gaps are constructed. For instance, Rosa and Flores (2015) argue for the problematization of White listening practices, rather than racialized discursive practices.

[4] See the work of Patrick L. Mason and William Mangino cited in Darity and Mullens (2020), Chapter 2, footnote 21.
[5] Again, as cited in Darity and Mullens (2020), see Tomaskovic-Devey, Thomas, and Johnson (2005) "Race and the Accumulation of Human Capital: A Theoretical Model and Fixed-Effects Application," American Journal of Sociology 111, no. 1 (July 2005): 58–89.
[6] See Avineri et al (2015) for linguistic anthropological critiques of the Language Gap.

Appendix B: Chapter Two Full Transcript

Leah: I mean one thing you could do is like, show a video of a kid doing it, and have them be in little groups and talk about like what they think it means- what they think this means about what the child understands

Rose: Or even like practice scoring videos- like watch a video, practice scoring, based on what you see this kid doing, […]
Barbara: I like the idea of at least- maybe not for the very first one, but for some of them, have that be their first things, then have like, what do you notice about this task, what questions do you have?

Rose: Exactly, so you're actively engaging them instead of like, this is the task, now what. You know, like you're doing the opposite

Barbara: I think they're gonna [unintelligible]

Leah (interrupting): Barbara, what idea did you like? The um having them score it?

Barbara: I think having them, before we tell them much of anything, or maybe there's like an overview, some way we use the manual text, I don't know, that before we talk too much at them, we actually have them score something using a video, and then you know, talk to each other, and then use that to [unintelligible] what questions did you have about this task, what did you notice about it, what you know what I mean? I think we need to engage them.

Anna: Or even if you want to take a step back and ask them, what kind of mathematical thinking do you see here? Right, like, you don't even tell them what skill this is for, just like what do you notice this kid doing- especially for something like the [redacted] task, show them that and be like, what sorts of things is the kid saying, what is he doing, what kinds of thinking do you see happening here? So that they're sort of thinking about the kid and what they're getting out of watching the kid do this, than this is the right way to score this
Rose: Well then, that's giving a great- I like that in terms of motivating like turning back to the manual, like, let's read, let's see- then you're kind of primed for like, ooo I wonder what _is_ going- like, this is what I think is going on

Leah: Right, Yeah, I mean so I think, the problem is that involves finding good videos

Rose: I think [unintelligible] video for this daylong video

Leah: I think it might be hard to find this perfect video that shows this perfect thing that we're thinking of

Ashley: I will tell you right now it is very hard to find perfect videos, because I've watched all of them

Rose: Well I don't know what perfect would be

Barbara: What has to be perfect about it?

Leah: Not, not, not perfect, but this child talking about [math concept] in this [math skill] task, like I didn't see that very much and so, uh, can you remember any kids doing that?

Gina: Not really but I mean

Leah: I mean, so like, we just have to use what we found, basically

Rose: I don't think you need anything special for this particular idea, that Anna and I shared, with the analogies, I don't think you need anything special at all, you need a clear video of that task where, and hop- and hopefully from the first assessment, where you can actually see what the kid is pointing to, and so it doesn't matter if the kid is actually saying- so teachers actually have the score sheet [waving paper] for that particular task, they're practicing scoring, cos that's going to be something, that they're, that they need to do, and then, after it's done, then you have this conversation about whatever, I mean, the kid doesn't have to answer correctly, you just have to be able to see what the kid is doing for this particular task. Surely you've got a video that does that, otherwise, why are we videotaping?

Leah: So, right and I, and I honestly, I think having them score it is fine, I'd almost rather them not be sitting there focusing on the score sheet and just have them watch it, and then um, ... because there's a lot of rich things happening and then, and the scoring I think, _is_ pretty straightforward, um, I mean we could have them score it afterwards

Barbara: I think it's less about the scoring and more about their active engagement with the video, rather than passive engagement with the video, I think that's why, so it's not so much that the scoring is hard, it's that we have to have ways to help them not be passive learners

Leah (overlapping with "learners"): So we could have some notes, like write notes about what you noticed

Anna: Another way to do it, if we do it the way that Rose suggested which is like, have them score it as they're watching it so they kind of get used to doing the scoring, but also say like, what kinds of things would you put in your notes for this kid

Leah: Exactly

Anna: So there's a notes column

Leah: Yeah

Anna: You might say, what are the kinds of things that were really interesting to you that you might want to follow up with in instruction, or you might want to see later if another kid, so like, what kind of thing- cos the notes column is all that rich mathematical thinking we're hoping they're seeing, right? So that's a way that we can link it back to the assessment

Rose: Right

Anna: And it's honestly

Barbara: My only concern- I like that, but my only concern is when I'm doing this assessment I actually find I don't have any time to do that- [unintelligible] to think that this is a reasonable- but Leah can I go back a little quick step, and just say- ask- like when I'm doing a PD the first thing I usually do is spell out what the goals are for the- for bringing them there, I mean not for them, but even for myself, and I realize that that's something I'm a little- I'm not sure if the main point of the ten o' clock to forty-five [time block] is... so that they are in some way proficient with administering the task or if it's more about um, kind of general learning about the topic and I know it's a little bit of both but I think it might be helpful to us to prioritize that, articulate it in some way
Leah: Yeah, I agree, I mean, I think we- In some way we need to talk about yeah, thinking more deeply about children's thinking about children's thinking in that area, rather than necessarily doing the task correctly, um,

Barbara: But I think we're going to have to prioritize- I don't think you can have people think deeply in that way about one two three four five six seven eight [counting the number of skills] and I don't think they warrant the same [unintelligible] thinking

Works Cited

Abbott, Andrew Delano. 1988. *The System of Professions: An Essay on the Division of Expert Labor*. Chicago: University of Chicago Press.

AERA. 2006. "Standards for Reporting on Empirical Social Science Research in AERA Publications." *Educational Researcher* 35 (6): 33–40.

AERA. 2015. "AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs: Approved by AERA Council, June 2015." *Educational Researcher* 44 (8): 448–52.

Agha, Asif. 2011. "Meet Mediatization." *Language & Communication* 31 (3): 163–70.

Akrich, Madeleine. 1992. "The De-Scription of Technical Objects." In *Shaping Technology/Building Society: Studies in Sociotechnical Change*, edited by Wiebe E. Bijker and John Law, 205–24. Inside Technology. Cambridge, Mass: MIT Press.

Andrejevic, Mark. 2010. "Reading the Surface: Body Language and Surveillance." *Culture Unbound* 2 (1): 15–36.

Apple, Michael W. 1986. *Teachers and Texts: A Political Economy of Class and Gender Relations in Education*. New York: Routledge & Kegan Paul.

Au, Wayne. 2007. "High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis." *Educational Researcher* 36 (5): 258–67.

Avineri, Netta, Eric Johnson, Shirley Brice-Heath, Teresa McCarty, Elinor Ochs, Tamar Kremer-Sadlik, Susan Blum, et al. 2015. "Invited Forum: Bridging the 'Language Gap.'" *Journal of Linguistic Anthropology* 25 (1): 66–86.

Ball, Deborah Loewenberg, and David K. Cohen. 1996. "Reform by the Book: What Is: Or Might Be: The Role of Curriculum Materials in Teacher Learning and Instructional Reform?" *Educational Researcher* 25 (9): 6.

Becker, Gary S. 1993. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. 3rd ed. Chicago: University of Chicago Press.

Bijker, Wiebe E., Thomas P. Hughes, Trevor J. Pinch, and Universiteit Twente, eds. 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, Mass: MIT Press.

Bourdieu, Pierre, and Jean Claude Passeron. 1977. *Reproduction in Education, Society, and Culture*. 1990 ed. Theory, Culture & Society. London ; Newbury Park, Calif: Sage in association with Theory, Culture & Society, Dept. of Administrative and Social Studies, Teesside Polytechnic.

Bowker, Geoffrey C, and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences (Inside Technology)*. Cambridge, Mass: MIT Press.

Bryk, Anthony S, and Barbara L Schneider. 2004. *Trust in Schools: A Core Resource for Improvement*. New York: Russell Sage Foundation.

Burkhardt, Hugh, and Alan H. Schoenfeld. 2003. "Improving Educational Research: Toward a More Useful, More Influential, and Better-Funded Enterprise." *Educational Researcher* 32 (9): 3–14.

Carless, David. 2009. "Trust, Distrust and Their Impact on Assessment Reform." *Assessment & Evaluation in Higher Education* 34 (1): 79–89.

Carr, E. Summerson. 2010. "Enactments of Expertise." *Annual Review of Anthropology* 39 (1): 17–32.

Carr, E. Summerson, and Michael Lempert, eds. 2016. *Scale: Discourse and Dimensions of Social Life*. Oakland, California: University of California Press.

Century, Jeanne, and Amy Cassata. 2016. "Implementation Research: Finding Common Ground on What, How, Why, Where, and Who." *Review of Research in Education* 40 (1): 169–215.

Coburn, Cynthia E., and Mary Kay Stein, eds. 2010. *Research and Practice in Education: Building Alliances, Bridging the Divide*. Lanham, MD: Rowman & Littlefield Publishers.

Coburn, Cynthia E., Judith Toure, and Mika Yamashita. 2009. "Evidence, Interpretation, and Persuasion: Instructional Decision Making at the District Central Office." *Teachers College Record* 111 (4): 1115–61.

Coburn, Cynthia E., and Joan E. Talbert. 2006. "Conceptions of Evidence Use in School Districts: Mapping the Terrain." *American Journal of Education* 112 (4): 469–95.

Cohen, David K., Donald J. Peurach, Joshua L. Glazer, Karen E. Gates, and Simona Goldin, eds. 2014. *Improvement by Design: The Promise of Better Schools*. Chicago, Ill; London: The University of Chicago Press.

Coleman, James S. 1990. *Foundations of Social Theory*. Cambridge, Mass: Belknap Press of Harvard University Press.

Cuban, Larry. 1984. *How Teachers Taught: Constancy and Change in American Classrooms, 1890-1980*. Research on Teaching Monograph Series. New York: Longman.

Darity, William A., and A. Kirsten Mullen. 2020. *From Here to Equality: Reparations for Black Americans in the Twenty-First Century*. Chapel Hill, NC: The University of North Carolina Press.

Daston, Lorraine, and Peter Galison. 2007. *Objectivity*. New York: Zone Books.

Davis, Elizabeth A., and Joseph S. Krajcik. 2005. "Designing Educative Curriculum Materials to Promote Teacher Learning." *Educational Researcher* 34 (3): 3–14.

Dehue, Trudy. 1997. "Deception, Efficiency, and Random Groups: Psychology and the Gradual Origination of the Random Group Design." *Isis* 88 (4): 653–73.

Dewey, John. 1916. *Democracy and Education: An Introduction to the Philosophy of Education*. New York: Free Press.

Dewey, John. 1922. *Human Nature and Conduct: An Introduction to Social Psychology*. New York, NY: The Modern Library. <https://www.gutenberg.org/files/41386/41386-h/41386-h.htm>

[DOE] US Department of Education. 2002. Transcript of the conference on "The Use of Scientifically Based Research in Education." Washington D.C., February 2002. <https://www.govinfo.gov/content/pkg/ERIC-ED466791/pdf/ERIC-ED466791.pdf>

Eisenhart, Margaret, and Lisa Towne. 2003. "Contestation and Change in National Policy on 'Scientifically Based' Education Research." *Educational Researcher* 32 (7): 31–38.

Evans, Ronald W. 2004. *The Social Studies Wars: What Should We Teach the Children?* New York: Teachers College Press.

Feuer, Michael J., Lisa Towne, and Richard J. Shavelson. 2002. "Scientific Culture and Educational Research." *Educational Researcher* 31 (8): 4–14.

Forsythe, Diana E. 1999. "Ethics and Politics of Studying up in Technoscience." *Anthropology of Work Review* 20 (1): 6–11.

Gal, Susan. 2015. "Politics of Translation." *Annual Review of Anthropology* 44 (1): 225–40.

Gal, Susan, and Judith T. Irvine. 2019. *Signs of Difference: Language and Ideology in Social Life*. Cambridge: Cambridge University Press.

Garfinkel, Harold. 1963. "A Conception of and Experiments with 'Trust' as a Condition of Concerted Actions." In *Motivation and Social Interaction: Cognitive Approaches*, edited by O.J. Harvey, 187–238. New York: Ronald Press.

Gieryn, Thomas F. 1983. "Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists." *American Sociological Review* 48 (6): 781.

Glenn, David. 2010. "Anthropologists Debate Whether 'Science' Is a Part of Their Mission." *The Chronicle of Higher Education*, November 30, 2010. <https://www.chronicle.com/article/Anthropologists-Debate-Whether/125571>

Goodwin, Charles. 1994. “Professional Vision.” *American Anthropologist* 96 (3): 606–33.

Groen, Mark. 2012. “NCLB--the Educational Accountability Paradigm in Historical Perspective.” *American Educational History Journal* 39 (1/2): 1–14.

Hall, Stuart. 1973. “Encoding and Decoding in the Television Discourse.” Paper presented the Council of Europe Colloquy on Training in the Critical Reading of Televisual Language, University of Leicester, September 1973.

Harkness, Nicholas. 2013. *Songs of Seoul: An Ethnography of Voice and Voicing in Christian South Korea*. Berkeley: University of California Press.

Heckman, James J. 2006. “Skill Formation and the Economics of Investing in Disadvantaged Children.” *Science* 312 (5782): 1900–1902.

Hempenstall, Kerry. 1997. “The Whole Language-Phonics Controversy: An Historical Perspective.” *Educational Psychology* 17 (4): 399–418.

Honig, Meredith I. 2012. “District Central Office Leadership as Teaching: How Central Office Administrators Support Principals’ Development as Instructional Leaders.” *Educational Administration Quarterly* 48 (4): 733–74.

Honig, Meredith I., and Cynthia Coburn. 2008. “Evidence-Based Decision Making in School District Central Offices: Toward a Policy and Research Agenda.” *Educational Policy* 22 (4): 578–608.

Honig, Meredith I., Nitya Venkateswaran, Patricia McNeil, and Jenee Myers Twitchell. 2014. “Leaders’ Use of Research for Fundamental Change in School District Central Offices: Processes and Challenges.” In *Using Research Evidence in Education*, edited by Kara S. Finnigan and Alan J. Daly, 33–52. Cham: Springer International Publishing.

IES and NSF Joint Committee. 2013. “Common Guidelines for Education Research and Development.” Institute of Education Sciences, US Department of Education and the National Science Foundation. <https://ies.ed.gov/pdf/CommonGuidelines.pdf>

Irvine, Judith T., and Susan Gal. 2000. “Language Ideology and Linguistic Differentiation.” In *Regimes of Language: Ideologies, Polities, and Identities*, edited by Paul V. Kroskrity. School of American Research Advanced Seminar Series. Santa Fe, NM: Oxford: School of American Research Press; J. Currey.

Jasanoff, Sheila, ed. 2010. *States of Knowledge: The Co-Production of Science and Social Order*. Transferred to digital print. International Library of Sociology. London: Routledge.

Johnson, Jim. 1988. “Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer.” *Social Problems* 35 (3): 298–310.

Jonçich, Geraldine. 1968. *Edward L. Thorndike: The Sane Positivist*. Middletown, Conn: Wesleyan University Press.

Kaestle, Carl F. 1993. "The Awful Reputation of Education Research." *Educational Researcher* 22 (1): 23–31.

Keane, Webb. 2003. "Semiotics and the Social Analysis of Material Things." *Language & Communication* 23 (3–4): 409–25.

Keane, Webb. 2018. "On Semiotic Ideology." *Signs and Society* 6 (1): 64–87.

Keller, Evelyn Fox. 1978. "Gender and Science." *Psychoanalysis and Contemporary Thought* 1 (3): 409–33.

Kennedy, Mary. 2005. "Sources Of Improvements in Teaching." In *Inside Teaching: How Classroom Life Undermines Reform*, 201–24. Cambridge, Mass: Harvard University Press.

Kliebard, Herbert M. (1987) 2004. *The Struggle for the American Curriculum, 1893-1958*. 3rd ed. New York, NY: RoutledgeFalmer.

Labaree, David F. 1997. *How to Succeed in School without Really Learning: The Credentials Race in American Education*. New Haven, Conn: Yale University Press.

Labaree, David F. 2010a. "How Dewey Lost: The Victory of David Snedden and Social Efficiency in the Reform of American Education." In *Pragmatism and Modernities*, edited by D Ttrohler, T Schlog, and F Osterwalder, 163–88. Sense Publishers. <https://web.stanford.edu/~dlabaree/publications/How_Dewey_Lost.pdf>

Labaree, David F. 2010b. "The Lure of Statistics for Educational Researchers." In *Educational Research: The Ethics and Aesthetics of Statistics*, edited by Paul Smeyers and Marc Depaepe, 13–25. Educational Research. Dordrecht: Springer Netherlands.

Ladson-Billings, Gloria. 2006. "From the Achievement Gap to the Education Debt: Understanding Achievement in US Schools." *Educational Researcher* 35 (7): 3–12.

Lagemann, Ellen Condliffe. 2000. *An Elusive Science: The Troubling History of Education Research*. Chicago, Ill; London: University of Chicago Press.

Lampland, Martha, and Susan Leigh Star, eds. 2009. *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Ithaca: Cornell University Press.

Latour, Bruno. 1993. *The Pasteurization of France*. Cambridge, Mass.: Harvard Univ. Press.

Latour, Bruno. 2004. "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry* 30: 225–48.

Latour, Bruno, and Steve Woolgar. 1979. *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills: Sage Publications.

Lave, Jean, and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Learning in Doing. Cambridge and New York: Cambridge University Press.

Louis, Karen Seashore. 2007. "Trust and Improvement in Schools." *Journal of Educational Change* 8 (1): 1–24.

Meens, David E, and Kenneth R Howe. 2015. "NCLB and Its Wake: Bad News for Democracy." *Teachers College Record*, 45.

Merton, Robert K. 1973. "The Normative Structure of Science." In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267–78. Chicago and London: University of Chicago Press.

Mertz, Elizabeth. 2007. *The Language of Law School: Learning to "Think like a Lawyer."* Oxford and New York: Oxford University Press.

Meyer, John W. 1977. "The Effects of Education as an Institution." *American Journal of Sociology* 83 (1): 55–77.

Murphy, Marjorie. 1990. *Blackboard Unions the AFT and the NEA, 1900-1980*. Ithaca: Cornell University Press.

Nader, Laura. 1972. "Up the Anthropologist: Perspectives Gained from Studying Up." Educational Resources Information Center, Record #ED065375. <https://eric.ed.gov/?id=ED065375>

National Center for Education Statistics. 2021. "Characteristics of Public School Teachers." The Condition of Education. Institute of Education Sciences. <https://nces.ed.gov/programs/coe/pdf/2021/clr_508c.pdf>

National Research Council (U.S.), Richard A. Duschl, Heidi A. Schweingruber, and Andrew W. Shouse, eds. 2007. *Taking Science to School: Learning and Teaching Science in Grades K–8*. Washington, DC: National Academies Press.

National Research Council (U.S.), Richard J. Shavelson, and Lisa Towne, eds. 2002. *Scientific Research in Education*. Washington, DC: National Academy Press.

[NCLB] U.S. Congress. 2002. *No Child Left Behind*. <https://www.congress.gov/bill/107th-congress/house-bill/1/text/pl>

Nelson, Adam R. 2016. "The Elementary and Secondary Education Act at Fifty: A Changing Federal Role in American Education." *History of Education Quarterly* 56 (2): 358–61.

O'Donnell, Carol L. 2008. "Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K–12 Curriculum Intervention Research." *Review of Educational Research* 78 (1): 33–84.

Oudshoorn, Nelly, and Trevor Pinch, eds. 2003. *How Users Matter: The Co-Construction of Users and Technologies*. Inside Technology. Cambridge, Mass: MIT Press.

Paschall, Katherine, Rebecca Madill, and Tamara Halle. 2020. "Professional Characteristics of the Early Care and Education Workforce: Descriptions by Race, Ethnicity, Languages Spoken, and Nativity Status." OPRE Report #2020-107. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Peirce, Charles S. 1898. *Reasoning and the Logic of Things: The Cambridge Conferences Lectures of 1898*. Edited by Kenneth Laine Ketner. Cambridge, Mass: Harvard University Press.

Penuel, William R., Caitlin C. Farrell, Anna-Ruth Allen, Yukie Toyama, and Cynthia E. Coburn. 2018. "What Research District Leaders Find Useful." *Educational Policy* 32 (4): 540–68.

Phillips, D. C. 2000. *Constructivism in Education: Opinions and Second Opinions on Controversial Issues. Ninety-Ninth Yearbook of the National Society for the Study of Education*. Yearbook of the National Society for the Study of Education. Chicago, Ill: University of Chicago Press.

Pinch, Trevor. 1993. "'Testing - One, Two, Three ... Testing!': Toward a Sociology of Testing." *Science, Technology & Human Values* 18 (1): 25–41.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Porter, Theodore M. 2003. "Measurement, Objectivity, and Trust." *Measurement* 1 (4): 241–55.

Porter Theodore M. 2008. "The Objective Self." Edited by Lorraine Daston and Peter Galison. *Victorian Studies* 50 (4): 641–47.

Puckett, Cassidy, and Matthew H Rafalow. 2021. "From 'Impact' to 'Negotiation': Educational Technologies and Inequality." In *The Oxford Handbook of Sociology and Digital Media*, edited by Deana A. Rohlinger and Sarah Sobieraj. Oxford University Press.

Raudenbush, Stephen W. 2002. "Identifying Scientifically-Based Research in Education." Remarks prepared for the working group conference on the Use of Scientifically Based Research in Education, Washington D.C., February 2002. Accessed from < https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.152.117&rep=rep1&type=pdf >.

Raudenbush, Stephen W. 2003. "Comments on 'Measurement, Objectivity, and Trust' by Theodore M. Porter." *Measurement* 1 (4): 274–78.

Remillard, Janine T. 1999. "Curriculum Materials in Mathematics Education Reform: A Framework for Examining Teachers' Curriculum Development." *Curriculum Inquiry* 29 (3): 315–42.

Remillard, Janine T. 2000. "Can Curriculum Materials Support Teachers' Learning? Two Fourth-Grade Teachers' Use of a New Mathematics Text." *The Elementary School Journal*, 331–50.

Remillard, Janine T. 2011. "Modes of Engagement: Understanding Teachers' Transactions with Mathematics Curriculum Resources." In *From Text to "Lived" Resources*, edited by Ghislaine Gueudet, Birgit Pepin, and Luc Trouche, 105–22. Dordrecht: Springer Netherlands.

Remillard, Janine T. 2016. "How to Partner with Your Curriculum." *Educational Leadership* 74 (2): 34–38.

Remillard, Janine T., and Luke Reinke. 2012. "Complicating Scripted Curriculum: Can Scripts Be Educative for Teachers?" Presented at the 2012 Annual Meeting of the American Educational Research Association, Vancouver, BC, April 2012. <http://www.gse.upenn.edu/icubit/sites/gse.upenn.edu.icubit/files/RemillardReinkeAERA2012.pdf>

Rosenbaum, Paul R. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science* 14 (3): 259–304.

Rosen, Lisa Stefanie. 2000. "Calculating Concerns: The Politics of *representation in California's 'Math Wars.'" Doctoral Dissertation, San Diego: University of California, San Diego. 304587357. ProQuest Dissertations and Theses Global.

Rudolph, John L. 2002. *Scientists in the Classroom the Cold War Reconstruction of American Science Education*. New York: Palgave. http://public.eblib.com/choice/publicfullrecord.aspx?p=735883.

Sanders, Carrie B., and Carl J. Cuneo. 2010. "Social Reliability in Qualitative Team Research." *Sociology* 44 (2): 325–43.

Schieffelin, Bambi B., Kathryn Ann Woolard, and Paul V. Kroskrity, eds. 1998. *Language Ideologies: Practice and Theory*. Oxford Studies in Anthropological Linguistics 16. New York: Oxford University Press.

Schneider, Barbara L., and Venessa A. Keesler. 2007. "School Reform 2007: Transforming Education into a Scientific Enterprise." *Annual Review of Sociology* 33 (1): 197–217.

Schrag, Peter. 2001. “The New School Wars: How Outcome-Based Education Blew Up.” The American Prospect. November 19, 2001. <https://prospect.org/api/content/c9d24ec7-f529-5a0b-8e79-b676f473da94/>

Shapin, Steven. 1994. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. Science and Its Conceptual Foundations. Chicago: University of Chicago Press.

Silverstein, Michael. 1976. “Shifters, Linguistic Categories, and Cultural Description.” In *Meaning in Anthropology*, edited by Keith H. Basso, 11–55. School of American Research Advanced Seminar Series. Albuquerque: University of New Mexico Press.

Silverstein, Michael. 1979. “Language structure and linguistic ideology.” In *The elements: A parasession on linguistic units and levels*, edited by Paul R. Clyne, William F. Hanks, and Carol L. Hofbauer, 193–247. Chicago: Chicago Linguistic Society.

Silverstein, Michael. 1981. “The limits of awareness.” *Sociolinguistic Working Paper No. 84*. Austin, TX: Southwest Educational Development Laboratory.

Silverstein, Michael. 1985. “Language and the Culture of Gender: At the Intersection of Structure, Usage, and Ideology.” In *Semiotic Mediation: Sociocultural and Psychological Perspectives*, edited by Elizabeth Mertz and Richard J. Parmentier, 219–57. Language, Thought, and Culture: Advances in the Study of Cognition. Academic Press, Inc.

Silverstein, Michael. 1996. “Encountering Language and Languages of Encounter in North American Ethnohistory.” *Journal of Linguistic Anthropology* 6 (2): 126–44.

Silverstein, Michael. 1997. “The Improvisational Performance of Culture in Realtime Discursive Practice.” In *Creativity in Performance*, edited by R. Keith Sawyer, 265–312. Greenwich, CT: Ablex Publishing Corporation.

Silverstein, Michael. 2003. “Indexical Order and the Dialectics of Sociolinguistic Life.” *Language & Communication* 23 (3–4): 193–229.

Silverstein, Michael, and Greg Urban, eds. 1996. *Natural Histories of Discourse*. Chicago: University of Chicago Press.

Sroufe, Gerald E. 1997. “Improving the ‘Awful Reputation’ of Education Research.” *Educational Researcher* 26 (7): 26–28.

St. Pierre, Elizabeth Adams. 2004. “Refusing Alternatives: A Science of Contestation.” *Qualitative Inquiry* 10 (1): 130–39.

St. Pierre, Elizabeth Adams. 2006. “Scientifically Based Research in Education: Epistemology and Ethics.” *Adult Education Quarterly* 56 (4): 239–66.

Star, Susan Leigh. 2010. “This Is Not a Boundary Object: Reflections on the Origin of a Concept.” *Science, Technology, & Human Values* 35 (5): 601–17.

Star, Susan Leigh, and James R. Griesemer. 1989. "Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19 (3): 387–420.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, Mass: Belknap Press of Harvard University Press.

Street, Brian V. 1984. *Literacy in Theory and Practice*. Cambridge Studies in Oral and Literate Culture 9. Cambridge [Cambridgeshire]; New York: Cambridge University Press.

Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge Cultural Social Studies. Cambridge, UK; New York, NY: Cambridge University Press.

Tomlinson, Stephen. 1997. "Edward Lee Thorndike and John Dewey on the Science of Education." *Oxford Review of Education* 23 (3): 365–83.

Tseng, Vivian. 2012. "The Uses of Research in Policy and Practice." *Social Policy Report* 26 (2): 1–24.

Tyack, David B. 1974. *The One Best System: A History of American Urban Education*. Cambridge, Mass: Harvard University Press.

Weiss, Carol H. 1980. "Knowledge Creep and Decision Accretion." *Knowledge* 1 (3): 381–404.

Woolgar, Steve. 1991. "Configuring the User: The Case of Usability Trials." In *A Sociology of Monsters: Essays on Power, Technology and Domination*, edited by John Law, 58–99. Sociological Review Monograph 38. London: Routledge.

Wortham, Stanton. 2006. *Learning Identity: The Joint Emergence of Social Identification and Academic Learning*. Cambridge; New York, NY: Cambridge University Press.