Technology-Based Personalization: Instructional Reform in Five Public Schools

David Nitkin

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2018

ABSTRACT

Technology-Based Personalization: Instructional Reform in Five Public Schools

David Nitkin

This dissertation addresses the question: *How does an attempt to redesign instructional delivery using technology-based personalization affect the technical core of teaching, learning, and student outcomes?* In recent years, many prominent educators, business leaders, and philanthropists have suggested that schools be redesigned to personalize students' learning experiences using technology. However, the justification for these reforms remains largely theoretical. Empirical research on technology-based personalization is sparse, and what little research does exist focuses predominantly on macro effects rather than the specific school-level, class-level, student-level, and lesson-level mechanisms that contribute to overall student achievement. The absence of research that pushes inside the "black box" of implementation is particularly problematic given a century of failed attempts to reform the technical core of instructional delivery, with symbolic reforms typically withering in the face of institutional resistance.

This study attempts to address that gap by examining the implementation of an innovative model for using technology-based personalization to deliver middle school math instruction. I draw upon theoretical tools from institutional theory, instructional improvement, and the history of educational reform to deepen our understanding of how technology-based personalization affects the role of students and teachers, the logistics of content delivery, and students' learning outcomes. Unlike previous studies in K-12 settings, which typically use summative assessments and virtual control groups to estimate aggregate effects on student learning, this study examines the relationships among a diverse set of lesson-level variables,

including instructional method, instructional content, group size and composition, teacher characteristics, student characteristics, and learning outcomes. In doing so, this study contributes to our understanding of the on-the-ground processes and mechanisms by which technology-based personalization affects (or does not affect) student learning.

Although the instructional model documented in this case study will remain anonymous, it is well known and respected among educators and philanthropists, and regarded as one of the most prominent and archetypical examples of technology-based personalization currently active in American schools. Using multiple methods, including novel applications of hierarchical linear modeling, cluster analysis, and heatmap data visualization, I explore: (a) the degree to which ground-level implementation of technology-based personalization represents an authentic departure from the traditional technology of schooling, and (b) the relationships among various elements of the model and student learning outcomes. I draw on longitudinal data from a full year of implementation in five schools, including the daily lesson assignments and assessment scores of 1,238 unique students supervised by 48 teachers.

This study supports four main findings: (a) the program succeeds in altering the technical core of instruction in several fundamental ways; (b) policy and logistical constraints limit the program's ability to reform the technical core of instruction to the degree that it aspires; (c) students who enter the program as already higher-performing are more successful on daily exit slips than students who enter the program with lower performance; and (d) the quantitative methods used in this paper represent useful and replicable tools for exploring the data produced by technology-based and personalized models.

Table of Contents

List of Tables and Charts

List of Figures

1. Introduction

Challenges Posed by Academic Diversity

Educators have consistently grappled with the challenge of meeting the varied academic needs of a diverse study body. The challenges posed by academic diversity are a consequence of three fundamental realities at the core of American public education: (1) the mandate that all students up to a certain age must attend school; (2) the desire for all students to obtain a uniform, baseline level of academic achievement; and (3) pre-existing economic, social, and cognitive disparities among the American public (Bidwell, 1965). In combination, these factors leave schools with the complex task of addressing the diverse academic and socio-emotional needs of all students, regardless of background, and ensuring that they meet the ever-increasing expectations of college, employers, and society at large.

Schools and districts have explored a variety of strategies for addressing this tension. Ability tracking and curricular differentiation have historically been two of the most common solutions, particularly in secondary schools (Lee & Ready, 2009). However, these strategies have increasingly been criticized for exacerbating divisions based on ability, race, class, ethnicity, and disability (Barr & Dreeben, 1983; Oakes, 1985). The rise of the standards and assessments policy regimes over the last thirty years has also reduced the popularity of curricular differentiation, which some have attacked as enabling lax standards that undermine achievement (Manna, 2011). However, the common alternative practice of organizing students into age-graded cohorts, irrespective of academic readiness, places the bulk of the "differentiation burden" upon classroom teachers. Specifically, variance in students' academic ability is far greater within classrooms than between classrooms in the same school or district, with some estimating that as

much as 62% of the variance in fifth-grade mathematics ability is situated within classrooms (Barr & Dreeben, 1983; Corno, 2008; Martinez, Schecther, & Borko, 2009).

The most prevalent classroom-level strategy for accommodating student diversity is ability grouping. Particularly common at the primary level, this technique sees teachers grouping students for instructional delivery based on the results of formal or informal assessments of academic readiness and ability (Pallas et al., 1994). Some researchers have gone so far as to describe the academic group, rather than the classroom, as the primary structure through which teachers deliver instruction to students (Barr & Dreeben, 1983). Indeed, the ability to accurately assess student learning and adjust instruction in real-time is one of the central tasks of teaching. According to Corno (2008), this type of differentiation is not a formal strategy or program, but instead what talented and experienced teachers learn to do naturally based on their accumulated teaching experiences. Talented teachers develop heuristic shortcuts that they use to customize and craft instruction to meet the needs of their students in real-time. This aligns with Bidwell's description of teachers as bridging the gap between the divergent skills of incoming students and the uniform academic outcomes expected by the bureaucratic schooling enterprise (Bidwell, 1965; Corno, 2008). However, this level of differentiation is difficult for teachers to execute effectively, and requires potentially unsustainable levels of pre-work and preparation (Beteille & Loeb, 2009; Carnoy & Levin, 1985; National Mathematics Advisory Panel, 2008).

Taken to its extreme, this suggests that the most effective mechanism for addressing the unique academic needs of learners would be to assign an individual tutor to each student (Bloom, 1984; VanLehn, 2011). These tutors could custom-tailor the instructional content to match each student's preexisting skills and knowledge. Individualized tutors could also uniquely tailor the method of instruction based on each student's preferences and proclivities, with teaching and

learning deliberately varied across text-based, oral, or visually oriented material to maximize each child's unique learning trajectory (Gardner, 2011). However, assigning an individual tutor to each student would obviously be cost-prohibitive using existing technologies. Moreover, this approach would neglect the fundamentally social nature of classroom life and could inhibit the development of students' interpersonal, collaboration, and communication skills.

The Promise of New Technologies to Address Academic Diversity

Although the potential of technology to supplement and even replace teacher-led instruction had been suggested long before the era of personal computers, significant improvements in information technology have led to a new round of calls for integrating technology and instruction (Cuban 1986; Tyack & Cuban, 1995; Wolf, 2010). Horn and Staker (2014) cite three rationales for technology-based instruction: (1) personalizing learning for each student; (2) providing all students with access to a wider array of high-quality content; and (3) controlling costs. Of the three, personalization is the most widely discussed and promoted. For example, billionaire Mark Zuckerberg recently announced personalized learning as a priority investment area of his newly minted Chan Zuckerberg Foundation, and the well-funded Bill & Melinda Gates Foundation, Michael and Susan Dell Foundation, and The Emerson Collective have also invested heavily in technology-based instructional models (Cavanagh, 2014; Herold, 2016a). We should not be surprised that philanthropists who made their fortune in the technology sector have proven eager advocates for technology-based solutions within the field of education, nor that the personal passions and predilections of these billionaires can have an outsized influence on education policy and practice (Ravitch, 2010).

An additional rationale for technology-assisted instruction is the exponential increase in student learning data that can be captured via technology-based learning platforms. These data

are much larger in volume than traditional education data, and also of a much finer grain-size, time-specific and inherently longitudinal, and naturally integrated with information on program delivery (Krumm et al., 2018; Natriello, 2012, 2013). This not only offers the potential of allowing technology-based systems to learn and improve over time, but is also a boon to researchers. For example, in a study of off-task behavior, Baker and Gowda (2010) found that the use of automated behavior detectors reduced by an order of magnitude the time needed to analyze student behavior data compared to traditional text replay analysis methods.

However, the research and development of personalized learning has been hampered by the lack of a consensus definition for what it actually means to be an "innovative" or "personalized" school. A recent EdWeek report suggested that "In the diverse and ever-changing world of educational technology, the term 'personalized learning' seems to be everywhere, though there is not yet a shared understanding of what it means" (Cavanagh, 2014). A consortium of prominent philanthropies, including the Bill & Melinda Gates Foundation, Eli & Edith Broad Foundation, and Michael & Susan Dell Foundation recently published the following "working definition of personalized learning:"

> Personalized learning seeks to accelerate student learning by tailoring the instructional environment—what, when, how and where students learn —to address the individual needs, skills and interests of each student. Students can take ownership of their own learning, while also developing deep, personal connections with each other, their teachers and other adults. Personalized learning includes [four elements]: (a) Learner Profiles - Each student has an up-to-date record of his/ her individual strengths, needs, motivations and goals; (b) Personal Learning Paths - All students are held to clear, high expectations, but each student follows a customized path that responds and adapts based on his/ her individual learning progress, motivations and goals; (c) Competency Based Progression - Each student's progress toward clearly-defined goals is continually assessed. A student advances and earns credit as soon as he/she demonstrates mastery; and (d) Flexible Learning Environments - Student needs drive the design of the learning environment. All operational elements—staffing plans, space utilization and time allocation—respond and adapt to support students in achieving their goals (Education Week, 2014; Pane et al., 2017).

In 2010, a symposium convened by the Software & Information Industry of America (SIIA), ASCD, and Council of Chief State School Officers published an alternate list of essential elements for personalized learning, including: (a) Flexible, Anytime/Everywhere Learning; (b) Redefine Teacher Role and Expand "Teacher"; (c) Project-Based, Authentic Learning; (d) Student Driven Learning Path, and (e) Mastery/Competency-Based Progression/Pace (Wolf, 2010). Although there is significant overlap between the definitions produced by the Gates Foundation and the SIIA symposium, there are also substantive differences in the role of the teacher and the prominence of project-based or authentic learning.

Further complicating matters, although the terms "blended learning" and "personalized learning" are often used interchangeably, they actually represent distinct but frequently overlapping constructs; a school may be blended without being personalized, or personalized without being blended (Picciano, 2014; Brodersen & Melluzzo, 2017). The Christensen Institute defines blended learning as "a formal education program in which a student learns: (a) at least in part through online learning, with some element of student control over time, place, path, and/or pace; (b) at least in part in a supervised brick-and-mortar location away from home; and (c) the methods along each student's learning path within a course or subject are connected to provide an integrated learning experience" (Horn & Staker, 2014). In addition, many have used the term "competency-based learning" synonymously with both personalized learning and blended learning, although both the Gates and SIIA definitions included competency-based advancement as only one element of the broader personalization concept (Horn, 2017). This profusion of models and definitions has made it difficult to assess the overall effectiveness of blended or personalized learning models writ large, or even define whether a model should count as blended or personalized at all (Pane et al., 2015).

However, despite these disagreements in how precisely to define new models, advocates of redesigning schools through technology-based personalization are united in their theory for how such models will support students. In their view, technology-based personalization will not only allow each student to engage with instruction that is matched to his or her unique aptitudes and interests, but also reduce costs, improve student outcomes, and expand access to often-scarce content like AP courses and foreign languages (Childress & Amrofell, 2016; Horn & Staker, 2014). Some also argue that it will make teaching a more sustainable and rewarding profession and reduce burnout by shifting some tedious instructional tasks away from teachers (Arnett, 2016; TNTP, 2014). In other words, if instructional content is increasingly delivered via technology, teachers will be able to focus on a more limited, sustainable, and rewarding set of tasks, such as building relationships with and motivating students. This may be particularly relevant in developing countries that may lack qualified teachers with domain-specific content knowledge (Muralidharan, Singh, & Ganimian, 2016).

Focus of the Dissertation

In this dissertation, I examine the relationship between technology-based personalization and student learning outcomes through a case study of an anonymous technology-based personalized program (referred to from here forward as "TBPP") in five public K-8 schools in a mid-sized urban district. TBPP, which is produced by a small non-profit organization, utilizes a technology-intensive personalized model in which an automated algorithm generates customized daily schedules for each teacher and student, including both specific learning objectives and formalized instructional tasks. These daily schedules are designed to maximize each student's progress towards mastery of the Common Core Math Standards, the ultimate goal of the program. At the end of each day's lesson, students take a short "exit slip" assessment, which is

automatically graded and used to update each student's personalized list of skills to learn throughout the year. This list is then used to generate the next day's personalized schedule for each student.

This study makes a significant contribution to the research literature by pushing inside the "black box" of personalized instruction to explore the specific school-level, class-level, student-level, and lesson-level mechanisms that contribute to overall student achievement. This includes an examination of the complex interactions among individual students, teachers, content, contexts, and learning methods. Although individual tutoring has long been understood as one of the most effective mechanisms for instructional delivery, much research on tutoring and small group instruction has focused on overall effects without attempting to explain the causal mechanisms by which the process works (Bloom, 1984; Corno, 2008; Snow & Swanson, 1992; VanLehn, 2011). This is also true of the literature on technology-based personalization, which has typically addressed the general effects of various models on student learning more heavily than the specific avenues through which student learning is produced (Barrow, Markman, & Rouse, 2007; Murphy et al., 2014; Pane et al., 2015, 2017; Wang & Woodworth, 2011; Wendt & Rice, 2013; Wenglinsky, 2005). Furthermore, much of the existing research on technology-based personalization in K-12 settings assumes that personalized models are being implemented as intended, but does not adequately explore the possibility that teachers or students may be buffering themselves from the attempted reform by continuing to act in ways that are typical of the traditional technology of schooling (Honig & Hatch, 2004; Tyack & Cuban, 1995). This dissertation has significant implications beyond the context of the TBPP program itself; a better understanding of the complex interactions among students, teachers, tasks, content, and

learning outcomes could have profound implications for all personalized learning models, as well as the wider phenomenon of classroom teaching and learning.

This study will utilize TBPP as a case study to explore the following research questions:

1. To what degree does the day-to-day, ground-level implementation of TBPP represent an authentic departure from the traditional technology of schooling? Conversely, to what degree are teachers and students engaging in symbolic reform while continuing to exercise traditional instructional patterns?
2. What are the relationships among various elements of the TBPP model and student outcomes?
    a. What is the association between variation in daily exit slip score and variation in instructional method, teacher characteristics, group size, and/or content? Do these relationships vary for different types of students?
    b. To what extent do daily content assignment or exit slip data predict end-of-year results on the PARCC and MAP assessments? Does this vary for different types of students?

In addition, this study will demonstrate the efficacy of several novel approaches to exploring the diverse, broad, and deep datasets produced by personalized learning programs. Although hierarchical linear modeling, cluster analysis, and data visualization heat maps have been applied effectively across a wide range of fields, this paper will represent one of the first times they have been applied to the daily instructional assignment and student outcome data generated by personalized learning program (Krumm et al., 2018). While the primary purpose of

this paper is not to break new methodological ground, it may nonetheless demonstrate a new and useful application of established statistical techniques to a type of data that is rapidly growing in volume and prominence. In 2016, Horn & Freeland Fisher described traditional education research as industrial in its assumptions of standardization at scale. Instead, they called for a new research model that explores personalized outcomes and harnesses the vastly richer data and enhanced analytic power created by recent technological advances. This paper will utilize TBPP as a case study to explore what this new research model could look like in practice.

2. Literature Review

Academic Diversity and Personalization

Context and historical trends. The necessity to differentiate instruction to meet pupils' unique needs has existed for as long as education itself. Corno (2008) cites references to educational differentiation in Chinese, Hebrew, and Roman texts dating back more than two millennia. For example, the Roman rhetorician and teacher Quintilian wrote during the reign of Domitian that:

> Some students are slack and need to be encouraged; others work better when given a freer rein. Some respond best when there is some threat or fear; others are paralyzed by it. Some apply themselves to the task over time, and learn best; others learn best by concentration and focus in a single burst of energy. (Quintilian, trans. 1921)

In addition to the above emphasis on differentiation by learning style, Quintilian also described the need for differentiation based on students' prior knowledge and abilities. He used the process of climbing a tree as a metaphor for the ascent to knowledge, with the teacher's role as helping each student climb to the branch just a little farther than the one he or she could reach unaided (Corno, 2008).

One of the earliest documented attempts to formally implement personalized learning in an American school district was the Pueblo Plan of the 1880s. The brainchild of Preston Search, superintendent of schools in Pueblo, Colorado, the Pueblo Plan rearranged the curriculum so that students could advance through material at their own pace. The distinction between grade levels was eliminated and teachers evaluated students based on how many units of study the student had completed rather than letter grades (Januszewski, 2001; Keefe & Jenkins, 2000; Tyack &

Cuban, 1995). Similar attempts at personalized models were implemented in St. Louis in the 1870s, Cambridge, Massachusetts in the 1890s, and Portland, Oregon in the 1900s (McDonald, 1915). While there is no consensus for why these models did not persist, one potential explanation is the degree to which each plan was associated with the charismatic superintendent who championed it. As the leaders who implemented the models moved on, school districts may have found it difficult to maintain their innovative structures in the face of isomorphic pressures from the broader institutional environment (DiMaggio & Powell, 1983; Meyer & Rowan 1977, 1978).

In 1916, John Dewey published a strong philosophical rationale for personalized models with his landmark "Democracy in Education." In this and other texts, Dewey argued that children should not be marched lockstep through a curriculum, but instead encouraged to nurture their own learning through self-guided exploration and discovery (Dewey, 1916). Although never fully implemented at scale in American schools, Dewey's ideas would provide much of the underpinning for pedagogical constructivism, a still-popular school of thought which suggests that students must authentically experience and engage with content in order to deeply understand it (Cohen, 1990; Wenglinsky, 2005).

Dewey's work, along with the work of Maria Montessori and other child-centered progressives, was also a strong influence on the Dalton Plan, a personalized model that generated intense interest among educators and the general public during the 1920s and 1930s. The Dalton Plan did away with self-contained classes, fixed times for discrete subjects, and annual promotions and retentions of students. Instead, students were empowered to negotiate monthly contracts with their teachers outlining both their minimum, mandatory tasks and additional opportunities for self-directed enrichment. Students moved at their own pace through the

curriculum and had significant latitude to choose their own content, peer collaborators, and physical workspaces. By 1932, nearly ten percent of American schools reported that they had implemented some version of the Dalton plan. However, this popularity would not prove durable; when a researcher attempted in 1949 to identify schools that still utilized the Dalton Plan, she found it in use at only a single site – the original Dalton School in Manhattan. Despite the early fanfare and publicity, the plan ultimately withered in the face of teachers objecting to the massive increase in paperwork, parents who worried about the plan's effect on student discipline, and students themselves, who sometimes complained that maintaining a personalized learning plan was solitary and boring compared to traditional classwork completed in the company of peers (Tyack & Cuban, 1995).

The theoretical justification for personalized learning was buttressed in 1978 when Harvard University Press published for the first time in English Lev Vygotsky's framework for differentiating content through each student's unique "zone of proximal development" (Vygotsky's work had been published in the Soviet Union in the 1920s and 1930s, but did not attract attention in the West until the late 1970s). In this model, the zone of proximal development serves the same role as Quintillian's next highest branch - just out of the student's independent grasp, but reachable with guided support from a teacher. Subsequently, Howard Gardner (2011) and others have produced significant research on the effects of differentiation based on "learning style," which can include instructional methods such as musical-rhythmic, visual-spatial, and verbal-linguistic. However, many others have disputed whether teaching students in their preferred learning style is associated with improved outcomes, or even whether distinct learning styles truly exist at all (Dembo & Howard, 2007; Paschler et al., 2008).

Historical challenges to personalized models. Although personalized instruction may be best-suited to meet the unique learning needs of each student, modern American schools were explicitly designed to promote standardization and uniformity. The structures that we associate with modern schooling, including single-teacher classrooms and age-grade cohorts, were charted and implemented by the "administrative progressives" at the turn of the 20th Century out of a desire to bring business-like rationality, hierarchy, and scientific management to the enterprise of education (Cuban, 1993; Tyack, 1991; Tyack & Cuban, 1995). Schools were designed to accomplish the dual goals of assimilating millions of young immigrants into a democratic American society while preparing all students to contribute to an industrial economy. Policies like the age-grade cohort allowed educators to impose a degree of uniformity across a large and heterogeneous group of students, while structures like the Carnegie unit and the division of knowledge into discrete subjects imposed a standardized bureaucratic structure across what had been a largely decentralized and incoherent educational enterprise (Bidwell, 1965). By implementing their vision for schooling at a moment when enrollments were rapidly expanding, the administrative progressives ensured that it would become embedded in regulation, legislation, and the public's collective vision of legitimate schooling (Tyack & Cuban, 1995; Tyack & Tobin, 1994).

This "one size fits all" design conflicts with the varying needs of a diverse student body. Bidwell (1965) provides one of the earliest and most effective analyses of how schools grapple with this tension. In his view, the age-grade cohort system combines teacher autonomy and the bureaucratic requirements for standardization, with advancement between grades roughly analogous to the examination of a product at different points on an assembly line. Bidwell argues that, "…the typical educational technology requires persisting interaction between an individual

teacher and his students. Such interaction permits the teacher to assess subtle variations in student performance and to adjust instructional methods accordingly, in a way which would not be possible were the student to move over very short periods of time from one teacher to another" (Bidwell, 1965). This means that the challenge of addressing the variability in student outcomes is vested in the classroom teacher, who is granted significant autonomy to modify instruction as he or she sees fit. However, a robust body of evidence suggests that this approach may be ineffective at scale, with teachers reporting significant levels of stress, overwork, and burnout while students are too often frustrated, unchallenged, and disengaged from classroom instruction (Beteille & Loeb, 2009; Carnoy & Levin, 1985; National Mathematics Advisory Panel, 2008).

Differentiation as a pedagogical strategy. Differentiated instruction, also known as adaptive teaching, is one of the most prominent classroom-level strategies for adjusting instruction to meet students' unique and dynamic needs (Tomlinson, 2001). Corno (2008) describes adaptive teaching as the real-time assessment and differentiation which experienced teachers utilize throughout instruction. In his words, "In teaching adaptively, teachers respond to learners as they work. Teachers read student signals to diagnose needs on the fly and tap previous experience with similar learners to respond productively" (p. 161). Ball et al. (2008) and Shulman (1987) suggest that teachers' ability to successfully engage in differentiated instruction is in large part determined by their pedagogical content knowledge, which includes the ability to diagnose student misunderstandings, generate appropriate models, and effectively explain complex and nuanced ideas. Troublingly, while pedagogical content knowledge may be significantly related to student achievement gains, incoming mathematics teachers' understanding of content is frequently thin and rule-bound (Ball, 1990; Hill et al., 2005).

While differentiated instruction is widely recognized as characteristic of good teaching, there is little evidence that teachers are capable of implementing it successfully at scale (National Mathematics Advisory Panel, 2008; Tomlinson et al., 2003). In one recent study, researchers provided teachers with extensive professional development and ongoing coaching on how to implement differentiation in their classrooms. However, three years later, they found no increase in the level of differentiation utilized by these teachers (Petrilli, 2012). Teachers themselves admit that they struggle to implement differentiation in their classrooms. In a 2008 national survey, more than eight in ten teachers said that differentiated instruction was "very" or "somewhat" difficult to implement (Farkas et al., 2008). Likewise, in a 2010 survey, a similar proportion of education school professors acknowledged that it is difficult to tailor instruction to match the individual needs of students on a daily basis (Farkas, 2010). In the words of one professor, "We are asking teachers to be more integrative, to be more focused on the interests of the children, to be more focused on individualizing…Yet we are still talking twenty five kids in a classroom and one teacher…We don't have homogeneous classrooms anymore and our teachers are still being treated as if everybody is homogeneous, so it doesn't work." Differentiation across a class of twenty to thirty students may simply be too difficult for the vast majority of teachers to execute effectively without adopting an unsustainable workload (Delisle, 2015).

Given that the challenge of implementing differentiated instruction increases as classes become larger and more diverse, the instructional form most conducive to utilizing it is individual tutoring. Tutoring delivers value through two separate mechanisms: (1) targeted instruction focused on the precise skills and content in the student's zone of proximal development, and (2) increased ability to motivate students through relationship-building and improving their attitudes towards the subject matter and themselves as learners (Snow &

Swanson, 1992). Indeed, Bloom (1984) used a randomized control trial to demonstrate that students participating in individual or small group tutoring typically performed two full standard deviations higher than students participating in traditional whole-class instruction. Interestingly, a third group of students who learned in a whole group setting using teacher-led adaptive techniques achieved results one standard deviation higher than the control, reaffirming the potential of individual adaptation and differentiation to support learning. In a meta-analysis of studies on human and automated tutors, VanLehn (2011) found a smaller effect size of 0.79 for human tutoring and 0.76 for automated tutoring, but reaffirmed the power of adaptive and competency-based instruction regardless of instructional method. Bloom also found tutoring and adaptive teaching to be highly correlated with student engagement, with students in the traditional classroom spending 65% of time on task, students in a large adaptive classroom spending 75% of time on task, and students engaged in tutoring spending 90%+ time on task.

The Use of Technology to Personalize Instruction

The theoretical rationale for technology-based personalization. Over the last decade, prominent figures from the business and philanthropic worlds have argued that new technologies offer the power to effectively deliver differentiated instruction to all students and significantly improve student outcomes. One of the most influential of these voices is that of Clayton Christensen, the Harvard Business School professor and coiner of the phrase "disruptive innovation" (Christensen, 2013). In Christensen's view, established market leaders rarely create fundamentally innovative products, since their past successes lock them into a business model and mindset aligned with their existing value proposition. Instead, innovation typically comes from "disruptive" entrepreneurs, often from outside the sector. These disruptors begin by offering alternative, inferior products to customers who are not currently served by or cannot

afford high-quality, mainstream products. However, in time, the disruptors use the revenue from those early adopters to refine and improve their products, eventually displacing the previously dominant players. One classic example is the process by which transistor radios replaced vacuum tube radios. Initially, the disruptive transistor radio was inferior to the established vacuum tube radio. However, American teenagers purchased them as a cheap alternative, and in time transistor technology improved to the point that they completely supplanted once dominant vacuum tube radios (Christensen, 2013). Although Christensen's theory of disruptive innovation is disputed by some, it remains popular among the Silicon Valley entrepreneurs and investors who are among the most ardent advocates for leveraging technology as a tool to improve K-12 instruction (Lepore, 2014).

In 2008, Christensen extended this argument to the field of education. He and his co-authors argued that the stagnant outcomes, century-old design, and lack of innovation in American public schools were typical of an industry ripe for "disruption." They identified online learning as an innovation that would supplant traditional brick-and-mortar schools, predicting that by 2019 half of all high school classes would be taught online (Christensen et al., 2008). Perhaps in part due to the inaccuracy of such predictions, technology advocates have recently shifted their focus from entirely virtual learning to blended learning, a model in which students spend part of their time learning from a teacher and part learning through technology (Horn & Johnson, 2012). The NewSchools Venture Fund recently suggested that $4 billion in strategic philanthropic investment could lead to 7% of schools transitioning to these types of "innovative models" over the next 10 years (Childress & Amrofell, 2016). While this is less ambitious than Christensen's initial estimate of half of all classes moving online, it still represents nearly 7,000

schools serving 3.5 million students, more than the total amount currently enrolled in all charter schools nationwide (Mead et al., 2015).

Advocates for new technology-based instructional models have been remarkably successful in attracting the attention of the popular press and well-heeled philanthropists. In 2015, Facebook founder Mark Zuckerberg announced that he would eventually give away 99 percent of his $45 billion fortune, with personalized learning as a priority investment area (Herold, 2016a). Zuckerberg outlined his bold philanthropic ambitions in a recent speech, stating, "Our hope over the next decade is to help upgrade a majority [of America's] schools to personalized learning and then start working globally as well… Giving a billion students a personalized education is a great thing to do" (Singer, 2017). Laurene Powell Jobs, the widow of billionaire Steve Jobs, recently made national headlines by donating $100 million to support high schools that adopt innovative, engaging approaches to learning, particularly for low-income and minority student populations (Gewertz, 2016). This philanthropic money is matched by increases in private investments in K-12 ed-tech companies, which grew from $77 million in 2010 to $537 million in 2015 (Childress & Amrofell, 2016). Although some have argued that technology companies making philanthropic investments in research and advocacy while simultaneously marketing for-profit educational software may pose a conflict of interest, such concerns are unlikely to slow the rising tide of enthusiasm for new and disruptive models (Ravitch, 2010).

The research evidence on technology-based personalization. The exuberance for technology-based personalization is not backed by a robust and conclusive body of empirical evidence on its effectiveness. A comprehensive meta-analysis of blended learning studies found only seven K-12 effect sizes from five high-quality studies, with two effect sizes favoring face to face instruction and five effect sizes favoring blended instruction (Means et al., 2010). Notably,

the authors found more than ten times as many studies examining blended or online instruction in higher education than in K-12 settings. Similarly, while a federal commission reported in 2008 that the use of instructional software has generally shown positive effects on mathematics achievement, it hedged that "Taken together, the available research is insufficient for identifying the factors that influence the effectiveness of instructional software under conventional circumstances" (National Mathematics Advisory Panel, 2008, p. xxiv). More recent studies have similarly found some positive effects, but the diversity of models, contexts, and methodologies make it difficult to draw sweeping conclusions about technology-based personalization as a whole (Brodersen & Melluzzo, 2017). For example, one study of blended learning in five charter networks found a wide array of instructional software and models in use, with a mixture of positive and negative effects. It also found that schools exhibited an eagerness to continually experiment with their models, meaning that even within a single school, the vision for blended learning was likely to change over time (Murphy at al., 2014a).

Several studies have attempted to address this issue by limiting their focus to specific programs or districts, but also reported uncertain estimates and mixed effects. Wendt & Rice (2013) found that the implementation of the online ST Math program produced positive results in some grades, but not others; Wang & Woodworth (2011) found that blended use of the Dreambox math program produced significant positive effects in overall mathematics achievement and a subtest score for measurement and geometry, but no effect on the subtests for problem solving, number sense, computation, or statistics and probability. The Center for Education Policy at Harvard University (2016) also found small positive effects for the use of Dreambox in a separate study. Murphy et al. (2014b) found that the implementation of Khan Academy varied so significantly within schools that it would be impractical to even attempt to

estimate a uniform effect on student achievement. In some cases, excitement over the newness of the personalized model may cause researchers or writers to overstate their effectiveness; the author of one white paper gushed that Summit Public Schools' blended model represents the future of learning, despite the fact that students' academic growth only marginally exceeded the national average in 2014-15 (Osborne, 2016).

Many of these technology-based tutoring systems trace their design to computer-assisted instructional tools that were built at Carnegie Mellon University in the 1970s and 1980s (Murray, 1999; Yazdani, 1987). These systems began with simple branching trees of instructional content, but eventually expanded to include the ability to generate new questions based on pre-set mathematical operations and general teaching strategies. More recently, intelligent tutoring systems such as Reasoning Mind, ALEKS, and ASSISTments have been found to produce significant student gains in some contexts compared to traditional instructional models (Hardy, 2004; Koedinger et al., 1997). These tools have been discussed extensively in the learning analytics and educational data mining literature, and the underlying mathematical principles used to create them were essential in developing many of the technology-based personalization programs that are currently being used in classrooms across the country.

One of the most prominent and widely heralded attempts to implement technology-based personalization is Teach to One: Math (Childress & Amrofell, 2016; Horn & Staker, 2014). The initial pilot of the program was named one of the Top 50 Inventions of 2009 by TIME magazine, and it has since been covered favorably by The Washington Post, Education Week, and Forbes (Brown, 2012; Horn, 2013; Vander Ark, 2017). In 2016, Bill Gates dubbed Teach to One "the future of math," and his Bill & Melinda Gates Foundation is one of several prominent philanthropies that have invested tens of millions of dollars in the program (Newcomb, 2016).

However, despite its success in attracting philanthropic donations and positive attention from the popular press, the small amount of research conducted on Teach to One so far indicates a mixed and uncertain impact. One early study using a randomized control design found no effect on student learning, although the study's author acknowledges that this result is imprecise due to the study's very small sample size and several methodological issues that arose during implementation (Rockoff, 2015). The same study also collected evidence via surveys that indicated that teachers and administrators believed the program was effective, while students were initially skeptical, but came to accept the program in the second year of implementation. Two reports by researchers at Teachers College, Columbia University found gains that surpassed national norms, with the highest gains for students who started the year with the lowest academic ability. However, these reports were based on comparisons to national norms from the NWEA MAP assessment, an analytic approach that does not support robust conclusions. Not surprisingly, the authors cautioned that they were unable to make causal inferences based on the available data, and emphasized that the results were highly heterogeneous across schools, with some schools experiencing statistically significant negative growth (Ready, 2014; Ready et al., 2013). A recent study of Teach to One's first-year implementation in a mid-size urban district utilizing a more robust comparative interrupted time series approach found no significant effect across all grades combined ($p>.10$). However, the estimates varied somewhat across grade levels, with a marginally significant negative impact of Teach to One participation on student mathematics performance in fifth grade ($ES = -0.371$ *SD*s; $p<.10$) and no significant effects in sixth through eighth grades (Ready et al., 2017).

Many of the studies of K-12 blended learning that have been published are limited by the absence of robust comparison groups. For example, one widely cited study reported that sixty-

two public charter and district schools utilizing personalized approaches produced significantly improved student results in both math and reading, with larger gains for students who experienced personalized learning for the longest amount of time (Pane et al., 2015). However, the study employed "virtual comparison groups" generated by the Northwest Evaluation Association (NWEA) to generate these findings. Although the sample included schools across multiple districts and states, the comparability of the treatment and "comparison" schools remains unclear. In addition, every school in the study had previously applied for and been accepted into a competitive grant-making program, suggesting potential selection bias due to some common unmeasured characteristic associated with improved student learning, such as strong leadership or a cohesive instructional vision. Data and analytic methods used by other studies of blended learning, although strong in many respects, also raise questions about the equivalency of treatment and control groups (Center for Education Policy at Harvard University, 2016; Murphy et al., 2014a; Pane et al., 2017; Ready, 2014; Ready et al., 2013; Rockoff, 2015; Wenglinsky, 2005; Woodworth et al., 2015).

The didactic nature of many instructional technologies, combined with the multiple-choice and procedural format of most standardized assessments, can also create a bias in assessments that impedes effective evaluation. For example, recent studies of programs funded by the Next Generation Learning Challenges (NGLC), Charter School Growth Fund's Next Generation School Investments, and the Gates Foundation's Personalized Learning Pilots all measured student achievement using the Northwest Evaluation Association (NWEA)'s Measures of Academic Progress (MAP) assessment, a multiple-choice test that does not require collaboration, argumentation, or oral or written communication (Murphy et al, 2014a; Pane et al., 2015, 2017). If these technology-based instructional models only improved students' procedural

skills at the expense of a broader set of higher-order thinking skills, the MAP test and similar assessments would likely provide an overly optimistic assessment of learning.

The very small number of high quality studies that do exist show either no effect or a positive effect in a very narrow context, raising questions about their external validity. For example, while several studies have used rigorous randomized designs to find significant positive effects for computer-aided Algebra programs, there is no evidence that these results are replicable outside of that specific subject (Barrow, Markman, & Rouse, 2007; Pane et al., 2013). One of the most promising recent studies comes from an after-school program in urban India, which used a randomized experimental design and found that students using online learning software made significant gains in math and Hindi compared to a control group. However, it is unclear how well these results might translate to instruction within the school day or in an American context (Muralidharan, Singh, & Ganimian, 2016).

Perhaps the most conclusive finding in the research literature is the ineffectiveness of "virtual" models that deliver instruction entirely online without a face-to-face component. Means (2010) found no effect for online-only instructional models compared to face-to-face instruction, and a comprehensive study by the Center for Research on Educational Outcomes (CREDO) at Stanford found large and significant negative effects for online charter schools, with students making the equivalent of 72 fewer days of reading growth and 180 fewer days of math growth compared to demographically similar "twins" in traditional brick-and-mortar district schools (Woodworth et al., 2015). The National Education Policy Center has also published a series of reports indicating that the outcomes of students enrolled in virtual and online-only schools lag significantly behind those at traditional brick-and-mortar schools (Huerta et al., 2015; Miron et al., 2013; Molnar et al., 2014). Many virtual schools are more loosely regulated than their

traditionally structured counterparts, increasing the likelihood of fraud and abuse; a recent report by *Education Week* described in damning detail how only 55% of students enrolled in Colorado's largest online charter school logged into the school's instructional portal in a typical week, with a paltry 0.1% of students engaging with the school's online content for the recommended 20 hours per week or more (Herold, 2016b). Defenders of virtual schools suggest that they enroll more difficult-to-serve students or those that are already more likely to drop out at time of enrollment, although most published studies contain robust demographic controls that should account for such differences in student backgrounds.

Variability by context and student characteristics. As is often the case in attempted school reforms, technology's effect on teaching and learning is highly dependent on the specific details of the program and the context in which it is implemented (Cohen, Raudenbush, & Ball, 2003). Teachers and administrators consistently cite the variable quality of instructional software, unreliability of hardware, poor integration of data systems, and unavailability of internet bandwidth as key obstacles to successful implementation (Freeland & Hernandez, 2014; Hew & Brush, 2007; Murphy et al., 2014a; Pane et al., 2017). Teachers' unfamiliarity with software and a lack of quality professional development and coaching are also key barriers (Cuban, 1986; Hew & Brush, 2007; Murphy et al., 2014a). By increasing student autonomy, blended and personalized learning models also increase the importance of strong classroom management; there is evidence that American students are prone to engage in off-task behavior when using technology for instructional purposes (Baker and Gowda, 2010; Murphy et al., 2014a; Rodrigo, Baker, Ryan, & Rossi, 2013).

The effect of instructional technology on student outcomes is also likely to be dependent on the age of the participating students. Much of the innovation in personalized learning has

occurred in grades five and up, where students are perceived to be more capable of learning autonomously (Christensen et al., 2008). In contrast, many early childhood and elementary educators have been deeply skeptical of technology's ability to supplement or replace teachers (Cordes & Miller, 2000). Some suggest that reformers' "infatuation" with computers distracts from addressing young children's most pressing needs, which include strong bonds with caring adults, hands-on experiences with the physical world, and time for unstructured play. Others warn that computers pose a risk to students' physical health, including vision problems and obesity (Cordes & Miller, 2000).

Troublingly, some evidence indicates that technology-based programs may exacerbate existing race- and income-based inequalities (Philip & Olivares-Pasillas, 2016). Wenglinsky (2005) found that low-income and minority students are more likely to use technology for didactic, "drill and practice" instruction, which NAEP data show to be negatively associated with academic achievement. A separate study of the online Cognitive Tutor system for high school Geometry found that students in urban schools were significantly more likely to make careless errors and engage in off-task behavior than students in suburban and rural schools (Baker & Gowda, 2010). However, since this study included no control for traditional instruction, it is unclear whether the use of technology produced, mitigated, or is entirely unrelated to this gap. The potential for instructional differentiation to reinforce inequality can exist even in the absence of any digital technology; in his study of instructional adaptation in traditional classrooms, Corno (2008, p.166) found that "Some teachers form subgroups for differential treatment but… inadvertently lower standards and reduce opportunities for students whom they believe cannot do the work." Although forcing all students to move at the same pace through the same content may be inefficient for students whose abilities fall far below or above grade-level norms, it could also

have a leveling effect as all students receive consistent supports in pursuit of a common goal. In contrast, allowing students to move at their own pace could widen inequalities by allowing strong students to race ahead while inadvertently reducing expectations for students whom teachers or algorithms have determined cannot do grade-level work (Corno, 2008).

Competing Paradigms for Instructional Improvement

Investment in teacher capacity and skill. Many of the most prominent strategies for instructional improvement over the last twenty years are at best unrelated and at worst contradictory to technology-based approaches. For example, the bulk of the recent literature on school improvement has called for greater investment in teachers' capacity and skill rather than supplementing or supplanting them with new technologies (Ball & Cohen, 1999; Chetty at al., 2011; Elmore, 2010). Ball and Cohen (1999) describe teachers as the key instructional mediators and ultimate determinants of student learning. In their view, teachers' own opportunities to learn are perhaps the most crucial factor in improving students' academic outcomes. Many of the instructional techniques most heralded in the literature are also particularly difficult to deliver solely via technology, including those that ask students to analyze unfamiliar situations, invent mathematical procedures, and solve interdisciplinary problems in unpredictable contexts (Grouws & Cebulla, 2000). Some have suggested that the best teachers do not engage in any direct instruction or evaluation at all, but instead simply pose well-designed problems and ensure that each student's thinking is transparent to the rest of the class, tasks for which computers seem particularly poorly suited (Stein, 2001).

Teachers' ability to engage in these types of instruction is highly dependent on their pedagogical content knowledge. As first described by Shulman (1986), pedagogical content knowledge is the collection of skills and understandings required to successfully teach a specific

domain of content. For example, teaching a student to add fractions requires not only general pedagogical skills and the ability to correctly complete the relevant mathematical procedures, but also the deeper knowledge of fractions needed to assess subtle student misunderstandings, design effective models, and appropriately scaffold content for above- and below-grade level students (Ball, 2008). In recent years, Deborah Ball and others have exerted significant effort to assess the nature and effects of pedagogical content knowledge. Among their most prominent findings are that (1) many pre-service teachers lack pedagogical content knowledge and believe that mathematics is simply a series of rules to be memorized; (2) teachers' pedagogical content knowledge is at least somewhat domain specific, meaning that teachers can be more or less effective at teaching different content based on the depth of their knowledge of that content; and (3) teachers' pedagogical content knowledge in mathematics is significantly related to student achievement (Ball, 1997; Hill et al., 2005).

Emboldened by this research, many academics, policymakers, and educators have called for a reform agenda focused on building the pedagogical content knowledge and instructional capacity of teachers. For example, Ball and Cohen (1999) advocate for reconceiving professional development as an inquiry-based activity grounded in practice. Similarly, Richard Elmore has spent the latter half of his career advocating for practices such as instructional rounds, peer observation, and peer accountability designed to build teachers' instructional capacity through the long, hard work of collaborative inquiry (Elmore, 1996, 2006, 2010; Elmore & Birney, 1997). These researchers underscore the importance of teachers and students building both procedural fluency and deep, multifaceted understanding of content. They describe "knowing math" as not just getting the right answer or understanding relevant procedural rules, but also knowing why a rule is true and how it connects with other big mathematical ideas (Ball, 1990).

Critically, they emphasize the difficulty of accurately assessing what a student truly knows, or what "knowing" even means, and claim that robust pedagogical content knowledge centered in a talented human teacher is the best possible tool for the ongoing assessment of students' knowledge and the targeting of instruction to students' needs (Ball, 1990).

This paradigm for teaching, learning, and instructional improvement stands in stark contrast to the assumptions of many technology-based personalized models. Most instructional technology programs rely on algorithms that assume knowledge to be binary – students have either mastered a specific skill or not – and assess learning through automated multiple choice assessments (Arnett, 2016; New Classrooms, 2017; Rockoff, 2015). This overwhelming reliance on relatively unsophisticated assessment stands in opposition to Ball & Cohen's assertion that educators must "confront the inherent inconclusiveness and incompleteness of knowledge," (p. 17) as well as Stein's belief that short, multiple-choice assessments with clearly defined right and wrong answers completely preclude the kind of creative, student-generated discussion required for deep learning (Cohen & Ball, 1999; Stein, 2001). The work of Ball and others also suggests that programs like TBPP may inhibit teachers' effectiveness by requiring them to deliver instruction across a wide band of grade levels with minimal time for preparation, reducing their ability to build relevant pedagogical content knowledge (Ball et al., 2008). Finally, assigning different tasks and content to each student may make it difficult for teachers to collaboratively study student work, preventing the shared inquiry advocated by Elmore (1996).

Gaps in the Research Literature

Given the rapid pace of innovation in technology-based instructional models, we should not be surprised to find significant gaps in the research literature. First, while several studies have measured the overall effects of technology-based instruction, they have largely treated

instruction as a "black box," and have neglected to explore the specific mechanisms through which student learning is produced (Barrow, Markman, & Rouse, 2007; Wenglinsky, 2005). For example, a succession of recent research funded by the Bill & Melinda Gates Foundation cited the diversity of the models under study as a barrier to examining the ground-level mechanics of instructional delivery; the authors of the Gates studies argue that "Although [certain] core attributes are common among the schools in the study, there is considerable diversity in the details of the schools' instructional models because innovation was encouraged in the competitive grant programs they participated in. That is, the schools in this study are not adopting a single standardized model of personalized learning" (Pane et al., 2015, p.3). In short, while some research has documented the degree to which technology affects learning outcomes, there is a deficit of evidence on the specific classroom-level, group-level, and student-level avenues by which these effects are generated. More work is required to document how technology-based instructional models affect and are affected by student characteristics, the role of the teacher, and the location, context, and nature of instruction.

In parallel, the literature on technology-based instruction would benefit from additional studies of comprehensive reforms rather than supplemental or add-on programs. Many of the effect sizes currently documented in the literature are for after-school tutoring programs, out-of-class interventions, or other modes of instructional delivery positioned in addition to, rather than in the form of, core classroom instruction (Barrow, Markman, & Rouse, 2007; Muralidharan, Singh, & Ganimian, 2016; Pane el al., 2013). The few studies of comprehensive instructional programs that do exist generally fail to thoroughly document the precise nature of the model under study, leaving some ambiguity as to the key classroom-level differences between traditional, teacher-led models and the new, technology-enabled model (Murphy et al., 2014a;

Pane et al., 2015). The research literature would benefit from additional studies that examine how comprehensive models of technology-assisted instruction affect and are affected by the technical core of schooling, including the role of the teacher, the role of the student, the organization of instruction, and quantifiable measures of academic achievement (Elmore, 2010).

This is particularly relevant given the rich body of literature on the difficulty of enacting reforms that meaningfully impact the technical core of schooling (Bidwell, 1965; Carnoy & Levin, 1985; Cohen, 1990; Cuban, 1986, 1990, 1993; Elmore, 1996, 2010; Tyack & Cuban, 1995). The past century of American education is littered with abundant examples of well-funded interventions that are promising in theory, but fail to meaningfully affect the core interactions among teachers, students, and content. Schools and teachers have a well-documented propensity to adopt reforms only symbolically or partially while buffering classroom practice from meaningful and enduring change. Indeed, when researchers from the Rand Corporation visited the classrooms of forty schools attempting to implement personalized learning between 2012 and 2015, they found that none of the schools were as radically different from traditional schools as theory had predicted (Pane et al., 2017). These historical and recent examples illustrate the acute need to look beyond the macro-level effect sizes reported in the existing literature on technology-based personalization in order to document how technology-based personalization truly affects teaching and learning (Cohen, 1990; Cohen & Barnes, 1993; Elmore, 1996; Honig & Hatch, 2004; McLaughlin, 1987; Spillane, Reiser, & Reimer, 2002; Weick, 1976).

3. Theoretical Framework

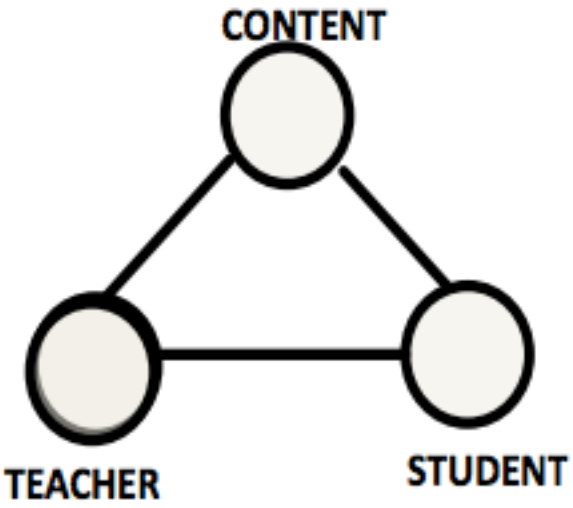
This paper examines how an attempt to redesign instructional delivery using technology-based personalization affects the technical core of schooling and student outcomes. Although TBPP is only one of many instructional models currently attempting to operationalize technology-based personalization in schools, it is typical of the movement as a whole in its utilization of individualized learning pathways, dynamic and homogeneous groupings, and digital technology as an evaluator, sorter, and instructor of students. In this paper, I draw upon theories of New Institutionalism, institutional isomorphism, and instructional reform to evaluate the effectiveness of TBPP – and by implication, technology-based personalization writ large - in substantively altering the technical core of schooling and enhancing student outcomes.

The Traditional Technology of Schooling

The instructional core. TBPP is one of many current attempts to reform the technical core of schooling through technology-based personalization. Its proponents hope to replace today's industrial education model, which assumes standardization at scale, with a post-industrial model that assumes personalization and differentiation. The technical core, also known as the "instructional core," is the fundamental level at which teaching and learning occurs. In simplest terms, it is defined as the interaction of teacher and student in the presence of content (Elmore, Fiarman, & Teitel, 2009). Given the pivotal role of the instructional core in determining student learning outcomes, any attempted reform is effective only insofar as it influences one of its three central pillars; a reform must alter the level of content, teachers' knowledge and skill, and/or student engagement in order to impact student learning outcomes (Cohen, 1990; Cohen, Raudenbush, & Ball, 2003; Elmore, 1996, 2010; Hess, 1999; McDonnell & Elmore, 1987). Some have gone so far as to declare that "if you can't see it in the core, it's not there," effectively

declaring all reforms to be meaningless if they do not affect the instructional core (Elmore, Fiarman, & Teitel, 2009).

Figure 1: The Instructional Core



Despite numerous efforts at reform over the last century, the technology of schooling has remained stubbornly consistent (Bidwell, 1965; Carnoy & Levin, 1985; Cohen, 1990; Cuban, 1986, 1990, 1993; Elmore, 1996, 2010). In this traditional model, teachers act as presenters of knowledge and students as passive recipients. Physically, groups of twenty to thirty students are oriented towards the "front" of the room, where a single teacher presents information through various media. All of the students in the class study the same content at the same time; when differentiation occurs, it comes in the form of scaffolds to help struggling students access content rather than a differentiation of the content itself. Although students may spend time working in groups or individually, they typically do so under a high degree of supervision and in pursuit of a learning target that is shared by the entire class. The overriding metaphor is that of the industrial

assembly line, with batches of students exposed to a uniform set of content for a fixed period of time, assessed, then advanced to the next set of content. New technological advances have been seamlessly integrated into this process without disrupting its fundamental contours, as radio, overhead projectors, and video supplement the teacher as the top-down mechanism for delivering knowledge to students.

The durability of the traditional technology of schooling. The historical record and a wide body of research literature provide ample reason for skepticism of technology's power to substantially affect the instructional core. Radio, television, and the personal computer were each heralded as potentially revolutionary educational tools in their time, but each failed to fundamentally change the technical core of teaching and learning (Cuban, 1986). The work of Tyack & Cuban (1995) provides a compelling narrative of how and why these and similar reform efforts have failed in the past. In their view, the most durable efforts to reform schools have typically involved either cosmetic changes or "add-ons" that leave the technical core of student, teacher, and content unaffected. For example, while reforms such as adding kindergarten grades or reducing class sizes require additional resources, they do not demand a substantial change in educator practice. In contrast, reforms targeted at the fundamental "grammar of schooling," like the Dalton Schools and the Eight-Year Plan, failed to gain widespread popular support and eventually withered (Cuban, 1990; Tyack & Cuban, 1995).

The resiliency of the traditional technology of schooling within the instructional core can be traced to several fundamental root causes. The first is a product of timing, whereby the policies favored during the massive educational expansion of 1880 to 1930 became "baked into" the fundamental logic of schooling. (Bidwell, 1965; Tyack, 1991; Tyack & Cuban, 1995; Tyack & Tobin, 1994). For example, a sprawling body of legislation and regulation has codified many

of the traditional structures of schooling, including the organization of students into age-grade cohorts, the division of knowledge into discrete subjects like social studies and science, and the importance of the Carnegie Unit in eligibility for graduation. Many of these structures spurred the creation of built-in constituencies with strong incentives to maintain the status quo, including teachers unions, vendors, professional associations, and postsecondary institutions. School buildings themselves also represent a physical codification of the traditional technical core of schooling, with an "egg crate" architectural design that divides space into roughly identical classrooms fit for twenty to thirty students each (Tyack & Cuban, 1995). Indeed, one of the primary obstacles in implementing personalized models like TBPP is finding or building spaces inside of traditional school buildings in which instruction can be delivered to one hundred students simultaneously (Pane et al., 2017)

A second fundamental barrier to reform is the decentralized and fragmented political control of American education, which has generally impeded consistent change initiatives across district and state lines (Cohen & Bhatt, 2012). This problem is exacerbated by competing and ambiguous goals for the educational enterprise itself, which have often prevented coherent policymaking (Carnoy & Levin, 1985; Labaree, 1997). Schools have been asked at varying times to prioritize the competing ideals of democratic equality, social efficiency, and social mobility, producing an incoherence that inhibits their ability to successfully accomplish any one of the three. Cuban (1990) describes how dominant social groups have often chosen to assign social and political problems to schools rather than attack them head-on, which would create more conflict and dislocation. For example, during the Civil Rights era, schools were tasked with promoting racial integration and social justice, while twenty years later during an era of globalization and international business competition they were assigned the task of promoting

economic competitiveness through rigor and skill development (Cross, 2004). These whipsawing priorities prevented coordinated and systemic effort to advance either achievement or equity, instead producing additional layers of bureaucracy and incoherence (Elmore, 1993). This conflict between the competing goals is one of many irresolvable tensions that inhibits coherent attempts to reform the technical core (Stone, 2002).

These tensions have also combined to prevent the creation of a cohesive and widely utilized technical body of knowledge regarding teaching and learning that could serve as a rallying cry for reorganization and instructional improvement (Cohen & Bhatt, 2012; Cohen, Raudenbush, & Ball, 2003; Labaree, 1992). Prior to becoming teachers, all educators spent decades as students themselves in classrooms that were organized according to the traditional technology of schooling; this gives them a schema and predisposition towards current methods rather than the blank slate enjoyed by entrants into other professions (Cuban, 1993). Preparing teachers to implement a new instructional model would require time for collaboration, feedback, and knowledge formation, which currently does not exist in most schools (Spillane, 2005; Spillane, Reiser, & Reimer, 2002). Unlike many other professions in which collaboration is the norm, the organization of schools into isolated classrooms deprives teachers of the opportunity to share best practices and learn from one another as professionals, further inhibiting reform (Elmore, 2010).

Institutional Barriers to Reform

Decoupling and technology/task misalignment. The factors described above have contributed to schools' adopting a loosely coupled structure in which reforms are adopted symbolically while the technical core remains largely untouched (Meyer and Rowan, 1977, 1978; Weick, 1976). In contrast to traditional Weberian notions of hierarchical authority,

bureaucratic control, and rational behavior, loosely coupled organizations gain legitimacy through the adoption of the myths, rituals, and ceremonies of the broader environment (DiMaggio & Powell, 1983; McLaughlin, 1987; Scott & Davis, 2007). This allows them to represent themselves as "legitimate" in the eyes of various stakeholders while abstaining from the difficult and uncertain work of improving the technical core. In a survey of 57 districts, Hess (1999) found widespread evidence that superintendents, confronted with the challenges of a difficult-to-access technical core, a lack of widely accepted goals and measurements, and competing pressures from multiple stakeholders, responded by adopting a variety of symbolic and divergent reforms that were ultimately ineffective in improving instruction. A cynic might wonder if recent calls to integrate technology and instruction may result in similarly symbolic reforms, with superintendents eager to claim the legitimacy and resources gained by adopting personalized learning models while experienced teachers and principals assume that "this too shall pass."

The foundational literature on organizational theory provides additional reasons to doubt reformers' optimism. This research suggests that an organization's efficiency will be maximized when management style, technology, task, and environment are all in alignment. For example, in simple and predictable environments, efficiency can be maximized through automation and top-down decision-making, while in complex and unpredictable environments, person-centered technologies and distributed decision-making will maximize efficiency (Burns & Stalker, 1961). Simple environments are those in which procedures are simple, stable, or homogeneous, while complex environments are those in which procedures are unique, unknown, or shifting (Henderson & Nutt, 1978). Similarly, Van de Ven & Delbecq (1974) classify task complexity

according to two independent dimensions: task difficulty and task variability. The lower the variability and complexity of a task, the more prone it is to automation.

Traditionally, K-12 classrooms have been described as exhibiting the high degrees of variability and unpredictability that characterize complex environments, meaning that they are a poor fit for mechanization (Bidwell, 1965; Corno, 2008; Fullan, 1996). Some technology advocates suggest that recent advances in the technologies for assessing students, analyzing data, and delivering instruction offer the potential to change this calculus, allowing some of teachers' traditional tasks to be re-classified as "simple" rather than "complex," thus enabling automation (Arnett, 2016). For example, giving students' feedback on procedural math skills like multiplication fluency does fit neatly in Van de Ven & Delbecq's description of a simple task that "[possesses] a known procedure that specifies the sequence of steps to be followed in performing the task]," (p.183) suggesting that shifting this type of instruction from human-based to technology-based systems could improve efficiency. However, it remains unclear what percentage of teachers' work might fall within this category, or how schools of education, labor unions, or the public might resist such a radical reconception of the role of a teacher (TNTP, 2014). In addition, there is some evidence that technology-based learning models serve to make instruction more didactic and procedural, which may not be conducive to teaching the broad set of complex cognitive and social skills required for success in the 21st Century (Murphy et al., 2014a; Wenglinsky, 2005).

Institutional isomorphism. An additional constraining force on instructional reform is isomorphic pressure to retain the forms and practices of the broader institutional environment (DiMaggio & Powell, 1983). Isomorphism encourages organizations to adopt practices not because of their technical efficiency, but instead because they provide legitimacy in the eyes of

powerful stakeholders (Meyer & Rowan 1977, 1978). Typically these stakeholders rest outside the organization itself and encourage conformity with preexisting structures and norms. The result is that each unit in an environment, such as a school in a district or a district in a state, comes to resemble all other units, regardless of the technical efficiency of the dominant organizational processes.

Isomorphic pressure can take several forms. The most direct of these is coercive isomorphism, which is produced by direct pressure such as a government mandate or conditional revenue. For example, criterion-based, state-mandated assessments of student achievement may be interpreted as a form of coercive isomorphism, since they threaten schools with sanctions or closure if they do not prepare students to express their knowledge of specific content in a mandated format (Hyslop & Mead, 2015). The high-stakes nature of these tests may discourage educators from teaching higher- or lower-level skills that will not appear on state tests, even when those skills are within the zone of proximal development for individual students. Indeed, a survey of 62 public charter and district schools implementing technology-based models indicated that students' ability to work at their own pace was limited by a perceived need to emphasize grade-level content and prepare for standardized tests (Pane et al., 2015, 2017). Although a new generation of technology-based assessments such as the SMARTER Balanced and PARCC assessments offer the possibility of assessing a broader range of skills, and the recently passed Every Student Succeeds Act (ESSA)'s loosening of federal control over state-level assessment and accountability provides space for further innovation, the majority of state accountability systems still rely only on assessments of grade-level standards (Clarke-Midura & Dede, 2010; Klein, 2016). These high-stakes assessments, and the normative rewards and sanctions associated

with them, represent a powerful form of coercive isomorphism that constrains schools to familiar forms such as age-grade cohorts and standardized instruction.

Organizations can also be influenced by normative isomorphism in which the individuality of units is constrained not through direct pressure, but instead through the imposition of professional and organizational norms. These norms are often transmitted through professional associations, certification requirements, and popular conceptions of strong or appropriate practice (DiMaggio & Powell, 1983). Normative isomorphism is particularly relevant for instructional models that attempt to leverage technology-based personalization to rethink the role of the teacher. While reformers may be excited for the increased efficiency produced by asking lower-paid aides or paraprofessionals to supervise students learning directly from technology, parents and the public may be strongly attached to the popular conception of students being taught by certified and experienced teachers. This may be true even if non-certified teachers were demonstrated to produce equivalent student learning gains when supervising a technology-based model; the popular conception of the teacher as the dominant mediator and mastermind of the learning process holds a symbolic resonance that may be difficult to dislodge.

A recent RAND study paints a vivid picture of these isomorphic forces in forty schools attempting to implement personalized learning programs (Pane et al., 2017). These schools reported that their most significant barriers in implementing technology-based personalization included the difficulty of explaining competency-based grading systems to parents and who were accustomed to traditional A-F grades. Most schools were actually forced to convert their competency-based grades into A-F scores for state-level reporting and college applications, a powerful example of isomorphic pressure forcing a reversion to traditional practices. The study

also found that charter schools tended to display more extensive implementation of many aspects of personalized learning, while traditional district schools tended to look more similar to the national sample of schools. This finding is not surprising, given charter schools' independence from many of the bureaucratic, regulatory, and union-related pressures that constrain district schools (Huerta & Zuckerman, 2009).

Isomorphic pressures are strongest in fields in which goals and technologies are ambiguous, organizations are highly dependent upon limited sources for resources, and there are powerful professional organizations: all apt descriptors of the field of education. In particular, the difficulty of setting and measuring meaningful educational goals encourages the adoption of symbolic rather than technical indicators of success, further encouraging isomorphism. Hess (1999) outlines several root causes for the difficulty of measuring goals in urban school districts, including: (a) heterogeneous student groups; (b) lack of universal, widely accepted assessments of learning; (c) disagreement over purpose of schooling; (d) rapid leadership turnover; and (e) social dysfunctions in urban areas that make it hard to disentangle the effects of school quality and poverty. These factors all inhibit educators' ability to reform and improve the technical core.

Buffering, symbolic adoption, and street-level bureaucracy. Schools and teachers adopt a variety of strategies to protect themselves in the face of multiple, competing demands and a constant churn of symbolic reform. These rarely include outright defiance, which is politically dangerous and may result in restricted access to valuable resources. Much more common is a practice of buffering, or strategically engaging with external demands in limited ways (Honig & Hatch, 2004; Tyack & Cuban, 1995). This can include limiting interactions with reform agents, ignoring negative feedback, or negotiating to shape the terms of compliance. Similarly, actors may choose to add peripheral structures without altering the technical core of

teaching and learning (Meyer & Rowan, 1977; 1978). For example, when tasked with reducing achievement gaps or promoting racial integration, districts may choose to create Offices of Civil Rights or hire Chief Diversity Officers, powerful symbols of compliance. However, these actions are likely to have little or no direct effect on the technical core of instruction. In another example, a recent study on the effects of high school graduation credit requirement reforms found that schools responded to higher standards by changing their criteria for awarding diplomas, but then awarding diplomas to a higher number of students who did not meet the requirements. They symbolically adopted the reform by implementing the new mandate, but failed to meaningfully change their practice at the student level (Carlson & Planty, 2012).

A wide body of research literature demonstrates the prevalence of cooption, symbolic compliance, and non-compliance in the loosely coupled world of education policy (Cohen, 1990; Cohen & Barnes, 1993; Elmore, 1996; McLaughlin, 1987; Spillane, Reiser, & Reimer, 2002; Weick, 1976). In the words of Richard Elmore (1996), "The closer an innovation gets to the core of schooling, the less likely it is that it will influence teaching and learning on a large scale… innovations that are distant from the core will be more readily adopted." This supposition is aligned with a separate literature on street-level bureaucracy and institutional innovation, which describes how service workers with substantial discretion over the execution of their work, such as teachers, police officers, and health workers, frequently buffer or ignore top-down directives (Lipsky, 1971; Lipsky, 1980; Weatherly & Lipsky, 1977). This is particularly true when these street-level bureaucrats are faced with inadequate resources, frequent challenges to authority from involuntary clients, and contradictory or ambiguous goals and expectations. This literature suggests that teachers tasked with implementing technology-based personalization may adopt it superficially or symbolically while continuing to engage their traditional teaching methods and

waiting for reformist policymakers to be replaced or lose interest. In words that could easily be applied to today's schools, a New York teacher from the 1930s expressed her fatigue with the seemingly endless cycle of symbolic reform: "Last year it was the socialized recitation, or the Gary Plan, or dramatization or correlation; this year it is motivation, silent reading, or the Dalton Plan. Each is taken up in turn, indiscriminately adopted, presently elbowed out to make room for the next newcomer; and yet we are not saved. The old problems remain" (Tyack & Cuban, 1995).

Description of the Reform Studied in this Dissertation: TBPP

TBPP is one of the most prominent new models attempting to reform the instructional core through technology-based personalization that redesigns classroom instruction in an attempt to match each student with the specific content that will best support his or her academic growth. The key design features of the program have remained largely consistent since its inception in the late 2000s. The learning environment is reorganized into one large room containing between four and eight adult instructors and approximately 100 students, frequently including students from multiple grade levels. Upon entering the room, students open personal laptop computers, log into the TBPP online portal, and consult their personal "learning lists," which tell them what they will be learning that day and how they will learn it. At the end of each day, students take a short, multiple choice "exit slip" to determine their mastery of that day's content. The program then uses the exit slip results to update the student's individual learner profile and to determine each student's assignment for the next day.

A TBPP lesson is designed to take a student approximately 35 minutes to complete. Students experience two lessons back to back each day, typically addressing the same skill, followed by the day's exit slip, which they have ten minutes to complete. Lessons are grouped

into “rounds” that last between two and three weeks before culminating in a “learning list wrap-up” assessment that evaluates the student’s mastery of the five to seven skills assigned to him or her for that round. If students exit slips and learning list demo demonstrate that they have mastered the content assigned to them for that round, they will be assigned more advanced content for the next round. If a student does not master a particular skill, he or she will typically be assigned to continue working on that skill in a subsequent round.

Each thirty-five minute TBPP lesson utilizes one of seven different instructional methods. In the Online Instruction (OI) and Online Practice (OP) methods, students work independently on digital content that they access through the online TBPP portal on their personal laptops. OI introduces students to new content, whereas OP provides practice opportunities with content to which students have already been introduced. The Paper Practice (PP) method also sees students working independently using either online or traditional paper/pencil content. In the Large Group (LG) and Small Group (SG) methods, students work in groups of two to six to solve mathematical problems addressing a shared skill. Students in the OI, OP, PP, LG, and SG methods are supervised by adults as they work, but these adults could be either certified math teachers (CMTs) or Teacher Assistants (TAs), who are not certified to teach math. The Teacher Instruction (TI) method is most similar to typical instruction, with CMTs guiding groups of six to thirty students through a shared mathematical concept. Finally, in the Long Term Projects (LTP) method, students work with the same peer group and CMT over multiple sessions to solve a complex, real-world problem. This day-to-day consistency makes the LTP method different than all other methods in which new groups are generated dynamically each day. The TBPP algorithm intentionally assigns each student to a balance of methods, and at any given time, different students in the TBPP classroom will be simultaneously utilizing each of these methods. This

means that a typical TBPP classroom will simultaneously feature some students learning independently using the OI, OP, or IR method, some students working in small groups in the LG or SG method, and some students learning from teachers in the TI or LTP method.

Although the TBPP program is designed only for students in Grades 5-8, the skills available for instruction include content ranging from early elementary school to Algebra. The skills map itself was created by TBPP staff, and has not to my knowledge been validated by outside researchers or content experts. Many of TBPP's curricular materials have been sourced from established content providers and software publishers, while others have been created entirely by TBPP staff. However, although TBPP provides instructional resources to use for all methods, teachers are allowed to customize them or use different materials of their own design if they choose.

Typical student and teacher experience. A typical student – we will call him Joseph – begins his daily TBPP experience by walking into a large, open learning space that is approximately the size of four traditional classrooms. Upon entering the learning space, Joseph will retrieve his personal laptop, log onto the TBPP portal, and check his personal schedule for the day. This schedule will be composed of two lessons, each utilizing a separate instructional method. For example, Joseph may have been introduced to the skill of multiplying decimals yesterday in a TI, but did not demonstrate mastery on yesterday's exit slip. As a result, the TBPP algorithm today assigns him to spend thirty-five minutes practicing decimal multiplication in an LG with four other students who also need to master this skill, then assigns him to a PP where he will work on the skill independently using online content from Pearson that he accesses via his computer. After seventy minutes, Joseph opens his personalized exit slip through the TBPP portal and attempts to answer five multiple choice questions on multiplying decimals. Joseph is

delighted to see that he has gotten four out of the five questions correct, which TBPP interprets as indicating mastery. Joseph logs off and closes his computer, knowing that he is now ready to move on to the more complex skills, such as dividing decimals, that TBPP will present to him in the next round. Figure 2 below provides an overview of Joseph's schedule on the typical day that I just described.

Figure 2: Sample Student Daily Schedule

Joseph Johnson 8th Grade *Tuesday*	
Concept: Understanding Percents: parts per 100 **Target Skill:** Multiplying decimals	
Session 1 Large Group (LG)	9:10am
Session 2 Paper Practice (PP)	9:45am
Exit Slip Exit slip	10:20am
Joseph leaves for his next class	10:30am

We can also understand the program through the eyes of a typical teacher - call her Ms. Jackson – who begins her daily TBPP experience the afternoon before instruction is scheduled to occur. At 4pm she opens her computer, logs onto the TBPP portal, and examines her schedule

for the next day (it is worth noting that since the next day's instructional assignments are based in part on the current day's results, it is impossible for her to begin preparing until today's exit slips are completed). Each day, Ms. Jackson teaches two separate eighty-minute periods; she has seventh and eighth grade students in the morning, then fifth and sixth grade students in the afternoon. Through the portal, she sees that her morning period will begin with a TI on adding fractions at 9:10am, then continue with an LG at 9:45am. For each lesson, she is able to see the individual students she will be teaching, as well as their assessment history, including how many lessons they have each previously experienced on the skills she'll be teaching. She is also able to download a lesson plan and related instructional materials to help her teach the skill. Because Ms. Jackson is a veteran teacher, she chooses to reuse one of her old lesson plans to teach the TI, but she likes the materials that TBPP provides for the LG, so she prints a set to use with her students the next morning.

At 9:10am, Ms. Jackson stands at the door to greet students as they enter the learning space. She quickly takes attendance, then moves to the section of the room where she will be teaching her TI. As she teaches her lesson, a nearby group of seventh and eighth grade students works on Algebra content in a SG station, while another group of students, headphones perched atop their heads, works independently in a OI station while supervised by a TA. After the TI ends, Ms. Jackson transitions to a separate part of the room, where she supervises a new group of students as they work together in an LG. She spends the period circulating from student to student, keeping them each on task and addressing misconceptions. As she circulates, she also uses her laptop to assign each student grades for "participation" and "effort;" these low-stakes grades do not affect their progress through TBPP skills, but can be reflected on the report cards

that go home to parents. At 10:20am, she urges students to wrap up their work and begin their exit slips, and at 10:30am she sends them out the door and on to their next classes.

TBPP as an attempt to reform the instructional core. The TBPP model diverges from the traditional technology of schooling in several significant ways, including the role of teachers and students, the design of the physical space, the assignment of instructional content, and teachers' decision-making latitude.

Chart 1: Traditional Technology of Schooling vs. TBPP Model

Technical Element	Traditional Technology	TBPP Model
Role of teachers and students	Teachers are active presenters of knowledge and students are passive recipients.	Students learn from teachers, computers, and each other.
	Teachers use formal and informal assessment to understand each student's progress.	Exit slips and the TBPP portal give teachers and students a shared understanding of progress.
	Isolated teachers work with a fixed group of twenty to thirty students for a full year.	Teams of teachers share responsibility for the learning of approximately 100 students. Teachers work with unique subgroups each day.
Physical space	One teacher commands a space filled with between twenty and thirty students. Students are typically oriented towards the "front" of the room.	Multiple teachers share a common space filled with approximately one hundred students. Space is flexible and dynamic.
Instructional content	All students in the class study the same content at the same time.	Students typically work on different content than one another and can move at different paces through the content depending on how quickly they achieve mastery.
Teacher decision-making	High levels of teacher discretion on how to group students and deliver instruction.	Automated algorithms determine what content to deliver, how to group students, and how to deliver instruction for the OI, OP, and PP methods.

TBPP embodies all four of the elements included in the Gates Foundation "working definition" of personalized learning: (1) individual learner profiles; (2) personal learning paths; (3) competency-based progression, and (4) flexible learning environments. However, it is also important to note what the program does not do. For example, it does not use data to evaluate which methods might be most effective for each student, as might be suggested by Gardner (2011). Instead, method assignments are motivated by the desire to expose all students equally to all methods, as well as logistical convenience given how many students in a classroom need to work on each skill in a given day. Similarly, TBPP does not collect or use data on which students have been more or less successful on exit slips when they worked in a group together in the past. Finally, once a skill is assigned, TBPP does not provide more scaffolding for less able learners or less scaffolding for more capable learners, as recommended by Snow & Swanson (1992); the instructional content for a given skill is the same for all students.

Although TBPP is an archetypal example of the kind of technology-based personalization envisioned by the Gates Foundation, the design of the program encompasses several decisions and assumptions that are not necessarily inherent to the use of instructional technology in general (Scardamalia & Bereiter, 2001). First, the design of the OI and OP methods represents a deliberate choice to use technology to deliver instruction in short, discrete bursts rather than to facilitate long-term investigation of authentic real-world problems (while the LTP method does address authentic, real-world problems, that method is facilitated by teachers, not technology). Second, TBPP's use of technology as its primary mechanism for assessing student understanding means that the program is limited to evaluating the narrow range of skills and knowledge that technology can assess without human support. In particular, the short, multiple-choice format of exit slips means that they are more likely to consider basic procedural skills than complex skills

related to theoretical understanding or evaluation. In combination, these decisions about where and how to leverage technology may serve to make TBPP instruction narrower and more didactic than traditional classroom instruction. Indeed, there is evidence that similarly designed technology-based models have had precisely this effect (Murphy et al., 2014a; Wenglinsky, 2005).

The design of TBPP's proprietary skill network and algorithm also reflects specific epistemological choices about the nature of learning and knowledge. For example, it assumes that learning occurs through identifiable pathways, and that linear and dependent relationships can be drawn from one skill to another. Similarly, it assumes knowledge to be binary, and that students can be categorized according to whether or not they have mastered individual, discrete mathematical skills. The skill network and algorithm encompass neither the possibility of partial mastery nor the idea that knowledge may be context-dependent; instead, a student is assumed to have either mastered a skill or not with no room for additional nuance.

Finally, it is worth highlighting the intentional design choice for the TBPP algorithm to personalize instructional assessments based entirely on the assumed levels of mastery within the skills network. The algorithm does not attempt to match students to methods in which they have been more successful in the past, nor does it pair them with teachers or peers with whom they have experienced past success. Indeed, the program's creators intentionally designed the algorithm to provide all students an equitable span of experiences across methods, teachers, and peers. However, these built-in program features mean that the algorithm does not automatically "learn" or improve its instructional assignments over time, other than to adjust to students' dynamic positions in the skill network. TBPP is designed to personalize based only on students'

individual content mastery, not their preferred learning method, teacher, peer group, or any other instructional element.

4. Data and Research Methods

Description of Data

This study leverages a diverse set of quantitative data from five public K-8 schools in a mid-size urban district during the 2015-16 academic year, when all five schools were in the first year of implementation of the TBPP program. I have combined data from two sources: (a) daily programmatic data collected by the non-profit that manages TBPP, including detailed daily lesson assignments and exit slip scores for all students participating in the program within the five schools; and (b) students' demographic data and scores on the Fall 2015 MAP assessment, Spring 2016 MAP assessment, and state-mandated PARCC math assessments from Spring 2016.

This study was completed in conjunction with a larger, four-year study which explores TBPP's causal impact on student mathematics performance and analyzes TBPP's implementation processes. The demographic data, MAP results, and PARCC results were collected by the research team for the larger project, of which I was a member. However, the inclusion of daily programmatic data is unique to my study.

Daily programmatic data. One of the embedded features of the TBPP program is the ability to collect detailed daily programmatic data. These data include linked lessons and exit slips for each student at a daily level, allowing me to associate specific instructional experiences with student outcomes. The daily lesson data is highly detailed, including information on the method, content, teacher, curricular materials, date, and time of day. Exit slips are multiple-choice format and machine-scored. The vast majority of exit slips contain five questions, but some contain four or six questions. A student must answer at least 75% of questions correctly in order to "pass" and advance in his or her TBPP skills progression.

The questions on each exit slip are drawn from a library of content-specific items written

by content experts in the employ of the non-profit organization that produces TBPP. This non-profit organization claims to test the validity and reliability of these items themselves using rigorous and mathematically sound procedures. However, I did not have access to detailed descriptions of the process that they use for validating their items, nor did I have access to the item-level data that I would need to test the validity and reliability of the items myself. Accordingly, while this study assumes that the exit slip assessments are psychometrically valid, I was unable to conclusively determine that this is the case.

In 2015-16, TBPP recorded data for 247,560 instructional events and 170,075 linked exit slips from 1,238 unique students and 48 teachers across the five schools participating in this study. These 170,075 linked exit slips reflect double counting of exit slips on days in which back-to-back instructional events for a single student addressed a common skill; only 123,776 unique exit slips were actually administered. The instructional events included seven distinct methods, with the role of the teacher varying depending on the method. In independent methods such as Online Instruction (OI), Online Practice (OP), and Paper Practice (PP), the role of the teacher is to ensure students remain on task and to support individual students with content as needed. In collaborative methods such as Large Group (LG) and Small Group (SG), the role of the teacher is to act as a guide and facilitator of student-led groups. Adult-led methods such as Teacher Instruction (TI) and Long Term Projects (LTP) are similar to traditional classroom instruction in which a teacher organizes instruction, delivers new content, actively checks for understanding, responds to student misunderstanding, and facilitates guided and Paper Practice.

Two types of teachers participate in TBPP: certified math teachers (CMTs) and Teacher Assistants (TAs) who are not certified to teach math. These TAs may include special education teachers, English as a second language specialists, or teachers certified in other content areas,

such as social studies. My models utilize a dummy variable to indicate whether teachers are CMTs or TAs (CMT=1, TA=0). While CMTs were assigned to oversee all instructional methods, TAs were only assigned to OI, OP, PP, LG, and SG methods (i.e. not TI or LTP methods). 27 CMTs and 21 TAs delivered instruction in the five schools included in this study.

Because the grade-level of a skill assigned for instruction can be either above, on, or below the typical grade level of a student, I generated an additional variable to reflect the difference between the grade level of the instructional content and the grade level of the student engaged in instruction. For example, lessons delivered to a 6th grade student featuring 4th, 5th, or 6th grade content would be coded as -2, -1, or 0, respectively. I also employed a series of dummy-coded variables in my analyses to reflect the method of instruction (OP, PP, LG, SG, LTP, and TI compared to OI). Table 1 below reflects the total number of instructional events for each method, as well as total number of linked exit slips (see Missing Data section below for discussion of the gap between instructional events and exit slips).

Table 1: Instructional Events and Exit Slips per Method

	Instructional Events (*n*=247,560)	Exit Slips (*n*=170,075)
Independent-Led Methods		
Online Instruction (OI)	61,211	51,809
Online Practice (OP)	31,154	26,104
Paper Practice (PP)	31,675	26,172
Student-Led Methods		
Large Group (LG)	12,975	11,132
Small Group (SG)	18,729	15,305
Adult-Led Methods		
Teacher Instruction (TI)	38,636	32,567
Long Term Projects (LTP)	53,180	6,986

TBPP organizes students into within-school classes that participate in instruction at the

same time and in the same location. These classes include all of a school's students in one or more grade levels, meaning that they are far larger than a traditional class; in the five schools included in this study, class size ranged from 82 to 128, with a median of 107.5. Each of these classes is typically served by between four and eight adults, with one adult assigned to teach or supervise the discrete learning tasks occurring in each section of the room. This means that although the class sizes are significantly larger than in a traditional model, teacher to student ratios are roughly similar.

Demographic data, MAP results, and PARCC results. The district's student population is predominantly low-income and black and/or Hispanic, and the demographics of the five schools under study are representative of the district as a whole. I employ a series of dummy-coded measures to account for students' demographic characteristics in my analyses, including indicators for gender (female=1, male=0), limited English proficiency (LEP) and special education (SPED) status (yes=1, no=0), separate indicators of free- and reduced-price lunch status (yes=1, no=0, compared to fully paid lunch status), grade (fifth, sixth, and eighth, compared to seventh), and race/ethnicity (black, Hispanic, and Asian/Pacific Islander students, compared to whites).

My data also include the Spring 2016 PARCC score and the Fall and Spring MAP scores from the 2015-16 academic year. The PARCC (Partnership for Assessment of Readiness for College and Careers) assessment is administered annually in compliance with federal testing mandates by a consortium of eight states and the District of Columbia. The assessment is aligned to the Common Core State Standards and is given to all students in grades 3 through 8 in both ELA and math. PARCC is criterion-based, meaning that all students are assessed using a common set of grade-level questions; their responses to those questions are used to place them in

one of five performance levels, with the top two levels representing proficiency.

The MAP (Measures of Academic Progress) assessment is produced by the Northwest Evaluation Association (NWEA), a national organization that provides assessments, professional development, and research for schools. The MAP assessment is available in ELA and Math for students in grades K through 11. In contrast to the PARCC exam, the MAP assessment utilizes a Rasch measurement model, meaning that students are measured on a continuous scale ranging from kindergarten to the high school level skills. The MAP assessment is also computer-adaptive, meaning that it differentiates the questions presented to each student depending on how that student performed on earlier questions. For this study, I z-scored (standardized) MAP scores within each grade, allowing "apples to apples" comparison of MAP data across multiple grade levels (Howell, 2002).

Missing Data

I am fortunate to have complete data on all independent variables, including students' demographic data and the daily instructional assignment for each student. The completeness of the instructional data is a product of the TBPP model itself, since it naturally creates a complete record of the instructional experience for each student every day as a byproduct of designing and assigning that experience.

However, while my dataset contains complete information on all independent variables, some instructional events lack data for the dependent, exit slip variable. The most prominent reason for missing exit slip data is related to the unique design of the LTP method. Unlike most methods, which are discrete, one-day instructional events, the LTP method engages students in a complex, real-world task that takes multiple days to complete. Because these LTPs unfold over more than a week, TBPP only assigns an exit slip for approximately one in seven LTP lessons

(usually on day two or three of the task). Accordingly, I removed from my analytic sample the 46,194 LTP lessons that are not paired with exit slips, leaving a total of 201,366 instructional events and 170,075 exit slips linked to 1,238 students and 48 teachers. After removing the LTP lessons without exit slips, 6,986 LTP lessons remained in the dataset. Because this omission is due to the design of the TBPP model, rather than missing data that should have been included in the data file but is absent, it does not raise any serious analytic or conceptual concerns.

After the unmatched LTP lessons were removed, 31,291 of the remaining 201,366 instructional events lacked corresponding exit slip data. There are several reasons why an instructional event could lack a linked exit slip, including timing issues (i.e. student runs out of time to complete the exit slip), technology issues, behavior issues, a fire drill, or a partial absence/early pickup. Given that the reasons for a missed exit slip are many and unknowable, the exclusion of these exit slips is unlikely to bias the analytic outcomes, and the direction of any potential bias is uncertain. However, I tested for the possibility of bias by calculating for each student a "percent of exit slips missing" variable, then using ordinary least squares regression (OLS) to search for relationships between student-level exit slip completion and any measured student characteristic, including school, grade level, gender, race/ethnicity, free- and reduced-price lunch status, limited English proficiency, special education status, and Fall 2015 MAP score. Tables 2 below indicates statistically significant relationships between the percentage of exit slips complete per student and that student's school and grade, with effect sizes ranging from -.072*** at School 3 to .050*** at School 2 (I used School 1 and Grade 7 as reference categories). While these differences are meaningful, the lack of variance in exit slip score across schools and grades means that they are unlikely to significantly bias the results. There are also statistically significant relationships between exit slip completion and several other time-

invariant demographic indicators in Model 3, but these relationships are explained away when MAP scores are included in Model 4.

Table 2: Predictors of Exit Slip Completion Per Student

	Model 1	Model 2	Model 3	Model 4
School 2[a]	.081***	.083***	.080***	.050***
School 3	-.047**	-.046**	-.034*	-.072***
School 4	.069***	.071***	.074***	.047***
School 5	-.001	.001	-.006	-.022*
Grade 5[b]		-.011	-.010	-.026*
Grade 6		.005	.005	-.019*
Grade 8		-.022~	-.024*	-.034***
Female[c]			-.011	-.004
Black[d]			-.008	.005
Hispanic			.007	.005
Asian			-.089*	.042
Free lunch[e]			.030*	-.002
Reduced lunch			.013	.013
LEP[f]			-.039**	.003
SPED[g]			-.040**	-.019
Fall MAP Math[h]				.008
R-squared	.086	.091	.114	.194

~p<.10
* p<.05
** p<.01
***p<.001

[a] School 1 is used as a reference category
[b] Grade 7 is used as a reference category
[c] Male is used as a reference category
[d] White is used as a reference category
[e] Paid lunch is used as a reference category
[f] Not limited English proficiency is used as a reference category
[g] Not special education is used as a reference category
[h] MAP scores are standardized (z-scored) within each grade level

Table 3: Exit Slip Completion Per Student Per School

	Students	Mean	Standard Deviation	Min	Max
School 1	267	.796	.217	0	.967
School 2	230	.878	.085	0	.970
School 3	220	.749	.127	0	.904
School 4	243	.865	.128	.064	.988
School 5	278	.795	.156	0	.922

Table 4: Exit Slip Completion Per Student Per Grade

	Students	Mean	Standard Deviation	Min	Max
Grade 5	175	.829	.141	0	.972
Grade 6	361	.826	.141	0	.988
Grade 7	357	.823	.176	0	.971
Grade 8	345	.796	.164	0	.950

I also generated basic summary statistics to explore whether the exit slip completion rate varied across methods, teacher types, or content levels. Table 5 indicates that across the six non-LTP TBPP methods, exit slip completeness ranged from 82% to 86% (LTP completeness in the dataset was 100%, since LTP lessons without exit slips had been previously excluded). Across the two teacher types, exit slip completeness ranged from 84% to 86%, and across the six potential content levels, exit slip completeness ranged from 80% to 86%. Exit slips were slightly more likely to be complete for lessons on or below a student's grade level than for lessons above a student's grade level. However, these associations are small enough in magnitude that they are unlikely to be large enough to significantly bias the results.

Table 5: Percentage of Exit Slips Complete by Method, Teacher Type, and Content Level

	Instructional Events	Exit Slips	Exit Slip Completeness
Method			
Online Instruction (OI)	61,211	51,809	85%
Online Practice (OP)	31,154	26,104	84%
Large Group (LG)	12,975	11,132	86%
Small Group (SG)	18,729	15,305	82%
Teacher Instruction (TI)	38,636	32,567	84%
Long Term Projects (LTP)	6,986	6,986	100%
Paper Practice (PP)	31,675	26,172	83%
Teacher Type			
Certified Math Teacher (CMT)	66,221	56,669	86%
Teacher Assistant (TA)	135,145	113,406	84%
Content Gap			
Three grades below student (-3)	7,530	6,464	86%
Two grades below student (-2)	28,929	24,209	86%
One grade below student (-1)	44,907	38,784	86%
At student's grade level (0)	83,500	71,196	85%
One grade above student (+1)	31,052	24,894	80%
Two grades above student (+2)	597	487	82%

Finally, I evaluated the percentage of lessons with completed exit slips per student (see Chart 2 in the Appendix). This analysis revealed that 45 students, representing 3.6% of the total number of students, completed fewer than 50% of their exit slips. I chose to eliminate these 45 students and their associated 1,085 lessons and 329 exit slips from the dataset. Among these students, the median number of lessons was 9, indicating that most participated in TBPP for less than one week. The relatively low number of eliminated students and lessons should reduce the likelihood of analytic concerns. After removing these 45 students, the final analytic dataset contained 200,281 instructional events and 169,746 exit slips linked to 1,193 students and 48 teachers.

Of the 169,746 total instructional events, 92,414 are "paired" lessons, meaning that they occur back to back on the same day with another lesson addressing the same instructional content. An additional 77,332 of the instructional events are "stand-alone" lessons, meaning that they are the only instructional event paired with a particular exit slip. As with the LTP issue described in the Missing Data section above, the presence of stand-alone lessons is a feature of the TBPP model rather than a case of problematic missing lesson data. Stand-alone lessons are produced when students are assigned to spend half of the TBPP period engaged in a LTP lesson or meeting with their "homeroom" group, both of which occur relatively frequently.

Tests for Normality of Data

I generated histograms to evaluate the normality of my outcome variable (standardized exit slip score) and each of my continuous predictor variables (average group MAP score, content gap of instruction, standardized student MAP score, and centered group size). This analysis indicated that the data were normally distributed (see Charts 4 through 9 in the Appendix). This is particularly important in the case of the outcome variable, where the distribution reveals enough variance to be able to conduct meaningful analyses; had almost all students earned the same exit slip score each day, it would have been very difficult to draw meaningful conclusions about the relationship between the time-variant instructional variables and daily learning outcomes. The distribution of group sizes is slightly non-normal, but the data is close enough to normal to allow for meaningful analysis and interpretation.

I also evaluated the distribution of methods within each skill. Were easier or more difficult skills taught using some but not all methods, it could have biased the estimates obtained in my quantitative analyses. In Table 6 below, each row represents a discrete skill, and each column represents a method. The cells are populated with the total number of lessons within the

dataset that utilized that skill and method. Finally, I transformed the table into a heatmap by assigning each cell a color based on the number of lessons it represents; cells with fewer lessons are colored red, and cells with more lessons are colored green.

Table 6 demonstrates sufficient variability of instructional methods within each skill to obtain meaningful results from quantitative analysis. This is true at all grade levels. This analysis also suggests several other interesting features of the data. First, it reveals the normality of the distribution by content level, with more lessons delivered for skills falling in Grades 5, 6, and 7 of the TBPP's proprietary skills map then above or below those grades. Second, it reveals that at least one OI lesson was used to teach every single skill, with every other method assigned to only address a subset of the total pool of skills. Finally, it indicates that LTP lessons in particular are not evenly distributed across all skills; only 27% of the 288 total skills have at least one associated LTP lesson, compared to between 80% and 100% for the other six methods. Because LTP lessons are time-intensive to teach and labor-intensive to create, it is likely that LTP lessons have been assigned to only the most important, foundational, or high-leverage skills.

Table 6: Distribution of Instructional Methods within Skills

	Skill	OI	OP	LG	PP	TI	SG	LTP
Ungraded	110	20	1	1	22	0		
	127	23	11		9			
	135	55	45	0	12	41	21	
	155	23	11		4			
	159	111	26	16	33	40	11	
	163	22	10		4	7	6	
	168	57			33	5	6	
	188	122		10	86	53	32	
	191	5	5		2		3	
	205	37	0	3	20	4	3	
	207	115	33	113	81	250	65	
	234	53	34	42	53	133	63	
	241	1			2			
	244	53	15		47	5	12	
	251	11	3	3	0	7	3	
	255	80	64	42	42	133	30	
	260	42	32	6	24	53	44	
	266	56	41	3	15			
	285	60	45	0	33	10	11	
	291	10	3		3			
	305	35	32	3	10	35	30	
	314	42	30		7	5		
	323	101	67	11	43	15	6	
	333	20	10		12			
	340	1	2		1			
	341	1	1					
	347	2						
	354	41	46	6	20		6	
	356	6	0		6			
	360	1	1					
	374	35	14	13	47	62	13	
	378	7	4		7			
	379	65	17	10	41	23	12	
	388	2			2			
	566	60	14	10	54	80	60	
Grade 2	195	2	2		2			
	500	10	4	2	16			
	502	13	3		10			
Grade 3	147	34	30	21	42	12	14	
	145	63	26	3	57		0	
	240	110	30	6	30		3	
	256	112	52		110	2	13	
	272	43	17		60		3	
	287	35	3		37	5	3	
	530	67	10		40	4		
	579	250	170	63	67	202	62	24
	581	30	4	4	22		3	
Grade 4	100	135	121	24	30	71	35	
	105	455	245	24	250	132	102	
	110	546	557	127	240	336	207	
	115	70	7		62	22		
	125	302	325	72	105	132	120	51
	142	547	152	53	467	220	85	
	153	105	82	32	102	115	47	
	165	230	140	5	82	61	33	
	171	104	77	0	23	5	3	
	172	325	227	55	105	204	157	
	185	352	205	42	146	104	110	
	185	502	153	33	525	174	111	120
	240	130	170	2	80	5	4	
	242	554	333	151	243	275	140	
	320	525	207	101	343	242	122	
	250	220	43	117	242	135	50	
	280	172	83	14	52	53	11	
	295	334	33	141	340	302	123	
	330	334	215	3	137	100	35	12
	515	215	61	47	231	151	83	
	520	114	52		75		6	11
	525	151	72	3	102	40	57	
	575	307	175	70	155	231	115	
	585	155	101	25	40	103	45	

	Skill	OI	OP	LG	PP	TI	SG	LTP
Grade 5	107	379	173	95	151	85	48	47
	112	299	111	26	95	44	29	168
	124	136	84	47	79	91	92	
	126	354	224	11	165	82	58	57
	184	364	251	27	144	58	93	188
	187	479	176	199	399	442	285	24
	192	458	329	29	77	199	88	98
	196	246	368	79	262	351	192	198
	229	185	116	42	189	157	79	
	229	285	285	81	128	289	184	
	239	552	166	189	229	488	248	41
	257	314	241	48	111	184	99	68
	274	357	289	74	245	288	156	48
	277	488	352	85	128	247	78	99
	282	389	284	92	99	147	56	185
	289	384	214	51	98	252	129	179
	301	416	176	156	368	449	197	
	331	389	339	58	288	326	189	69
	384	158	185	28	84	65	68	
	385	259	91	49	199	199	99	
	386	382	189	58	382	196	121	
	387	195	99	21	199	111	98	
	416	158	28	46	151	99	42	
	417	189	82		87	5	6	185
	418	688	19	158	548	388	288	145
	419	199	94	52	128	188	81	
	428	178	147	74	197	141	44	
	521	98	48	18	91	185	98	
	527	421	114	88	242	442	146	
	531	86	91	11	116	99	41	
	533	365	114	17	244	29	99	74
	534	429	199	28	155	64	51	
	535	356	58	97	391	189	98	99
	577	354	251	46	125	97	82	24
	578	279	158	9	97	94	42	
	583	398	286	96	199	199	199	
Grade 6	105	349	99	52	258	155	128	47
	116	277	146	11	185	16	14	
	117	491	269	91	264	76	182	
	120	265	229	91	185	187	85	128
	122	224	168	159	74	352	79	
	125	287	282	45	121	75	64	
	130	342	285	94	191	176	95	
	140	258	151	99	174	388	97	58
	141	491	247	69	128	196	91	52
	170	118	186	85	78	341	41	
	177	894	42	222	364	651	352	47
	186	544	259	65	268	285	156	69
	209	641	289	69	228	328	184	298
	209	358	284	67	166	225	59	
	219	545	328	178	282	452	167	99
	216	649	179	191	472	484	257	
	222	589	186	198	228	262	164	387
	225	388	282	146	271	265	199	
	226	352	124	22	199	186	116	276
	228	188	99	94	96	219	42	
	233	499	298	16	199	199	85	27
	245	172	187	46	78	78	98	
	247	659	384	167	242	391	156	
	264	285	158	25	28	48	19	
	265	491	398	58	82	166	68	
	298	756	577	392	185	582	298	196
	292	397	295	127	149	249	185	16
	299	445	259	69	186	184	85	
	302	227	126	37	124	76	48	
	304	485	278	79	294	186	97	
	317	275	146	99	165	77	68	
	318	546	267	189	298	446	141	
	478	167	156	82	127	198	98	
	487	225	196	77	157	284	68	
	499	669	461	165	298	385	294	382
	508	176	284	69	98	149	85	29
	529	187	2	17	179	28	69	
	524	191	178	56	84	82	119	
	532	246	485	147	428	541	318	42
	536	69	64	9	59	26	18	4
	537	188	5	9	98	96	27	
	549	86	49	9	47	4	29	
	567	526	314	245	249	499	97	198
	588	432	241	128	218	446	226	194

	Skill	OI	OP	LG	PP	TI	SG	LTP
Grade 7	106	484	257	66	188	125	84	
	113	246	182	78	118	96	52	78
	155	587	92	158	469	426	248	59
	144	156	161	74	75	116	73	42
	152	488	261	92	294	356	165	76
	154	135	92	99	54	136	188	61
	156	52	28		17		8	
	161	349	132	87	138	271	85	
	166	128	52	51	38	281	82	44
	174	255	159	66	53	358	146	29
	179	427	172	58	182	295	128	152
	181	58	15	18	72	47	38	
	182	357	284	72	148	297	148	224
	198	255	127	53	95	276	128	24
	199	218	129	143	66	365	83	
	211	285	28	81	282	285	116	
	215	429	1	118	364	345	151	129
	219	155	38	56	73	254	64	
	230	115	74	11	72	91	91	93
	231	586	229	177	429	371	181	
	249	154	65	38	216	168	186	
	258	357	252	45	184	91	181	116
	271	428	279	61	54	112	93	
	279	648	345	288	351	779	251	42
	294	481	94	148	382	429	144	
	295	247	168	46	116	78	44	
	315	128	96	17	45	58	29	
	383	366	178	91	182	135	98	
	388	68	47	26	93	71	48	113
	391	384	214	96	211	668	166	
	392	185	91	95	46	81	95	
	393	64	27	43	73	123	26	
	418	356	76	188	235	269	118	98
	511	88	61		93	68	43	
	522	115	95	15	53	112	68	
	526	117	69	22	98	29	96	
	541	383	282	93	92	336	141	197
	546	187	194	42	183	174	183	
	584	169	197	38	5	125	95	
Grade 8	111	482	284	111	117	342	98	16
	114	2	1					
	125	288		114	268	226	77	44
	145	327	126	27	29	97	63	228
	146	174	186	16	45	93	56	
	158	158	61	92	78	21	91	93
	167	322	355	94	138	311	284	
	173	62	22	13		18	12	
	178	145	186	13	56	26	28	147
	188	255	116	25	124	74	28	13
	194	7			5			
	204	255	155	98	93	78	96	
	205	1						
	221	4	5		1			
	224	248	96	18	43	26	14	
	232	45	43		17		9	
	236	171	183		67	26	24	
	240	488	98	57	251	361	247	
	242	29	29	9	7			
	249	18	8	5	17	7	6	
	251	78	52	6	51	27	12	
	262	154	79	26	93	92	93	
	269	122	18	7	88	44	94	
	270	95	98		6			
	275	28	5		5			
	288	2	1					
	294	6	4		1			
	308	12	5					
	310	118	88	63	44	218	56	
	313	241	123	5	57	28	26	
	315	1			1			
	322	118	75	27	16	188	6	161
	325	365	255	222	121	454	148	56
	332	38	28		16		9	
	335	123	122	18	75	77	65	
	444	31	25	18	24	23	18	
	503	112	97	15	87	25	15	
	512	1			1			
	513	283	114	43	115	83	38	
	514	33	4		67		22	
	538	232	144	72	112	151	63	
	542	46	28	17	38	14	6	
	544	51	51	4	23	34	38	
	552	88	53	16	34	84	43	

	Skill	OI	OP	LG	PP	TI	SG	LTP
Grade 9	297	196	84	29	55	88	95	293
	259	168	59		41	25	28	
	267	163	65	94	97	208	45	
	279	22	8		41	29	15	
	285	248	92	46	79	158	97	25
	308	61	95	6	91	6	16	
	319	152	186	12	61	18	24	
	328	245		48	187	286	86	
	324	78	15	95	98	98	29	
	326	95	4	2	97	15	6	
	329	11			9			
	331	244	148	84	48	259	189	
	333	58	22	8	29	4	15	
	334	18	9		2		2	
	336	18	2		29			
	337	333	283	189	86	364	182	
	342	189	98	67	164	172	187	
	343	187	95	54	153	214	146	58
	345	158	22	22	199	91	48	
	346	5	5		6			
	349	91	66	9	57	16	25	
	352	4						
	358	95	29	8	16	6	18	
	363	86	91	11	96	11	18	
	364	11	17		11		9	
	365	196	82	14	66	75	95	98
	366	95	66	46	86	178	79	
	371	282	178	95	63	243	148	
	372	55	58	76	18	249	172	
	381	75		5	66	95	48	
	384	15		5	11			
	385	7	2		4		2	
	547	229	98	18	94	44	98	158
	548	303	271	95	163	176	82	
	549	151	152	48	116	158	74	
	550	186	94	46	208	229	144	
	551	125	183	95	93	124	65	
	553	6	2		4		9	
	554	157	29	45	123	201	69	154
	555	68	15	25	75	78	96	
	556	96	9	16	79	45	26	
	557	47	14	17	29	64	46	
	558	47	4		44	15	28	
	559	98	15		14		6	
	560	16	9		24		9	
	561	193	18	88	176	153	64	
	562	4	1		4			
	563	58	93	28	17	44	6	
	564	53	6	97	42	92	18	
	565	45	1		93	7	16	

Quantitative Methods

I used two primary quantitative techniques to explore my research questions. The first is a hierarchical linear model (HLM) that nests lessons within students (Raudenbush & Bryk, 2002; Woltman et al., 2012). The second is hierarchical cluster analysis paired with clustergram heatmap data visualizations (Bowers, 2007, 2010; Eisen at al. 1998; Lee, et al., 2016; van'tVeer, 2002).

Overview of hierarchical linear modeling (HLM). Hierarchical linear modeling is a statistical technique for examining the relationships among cases that exist in nested structures. For example, a study of voting behavior may focus on voters in different states. In this case, there are two levels of analysis – votes and states – with the first voter level nested within the second state level. In an educational context, researchers may seek to explore the relationships between several curricula and the mathematics achievement of students nested within classrooms, which are in turn nested within schools. These types of nested structures are relatively common in social science research (Raudenbush & Bryk, 2002; Means et al., 2010; Murphy et al, 2014a; Ready & Wright, 2011; Singer & Willett, 2003; Woltman et al., 2012; Wood et al., 2017).

The use of a multi-level model offers several advantages for this study. First, it enabled me to explore the proportion of variance in student outcomes at the lesson, student, class, and school levels, directly addressing my research question as to what degree variation in TBPP's daily program implementation is related to variation in student outcomes. Second, it provided more accurate standard errors than traditional OLS, which would erroneously assume independent responses across lessons without taking into account covariance based on student characteristics, class-level factors, or school-level factors. Third, it allowed me to model the

effect of time, which is important in a dataset with multiple longitudinal data points for each participant. Finally, it enabled me to model cross-level effects and explore whether the relationship between lesson-level variables and student outcomes differs based on the types of students engaged in instruction (Raudenbush & Bryk, 2002; Singer & Willett, 2003; Woltman et al., 2012).

Fitting the model. I utilized a two-level hierarchical linear model to explore the relationships between various elements of the TBPP model and students' outcomes, as measured by standardized (z-scored) daily exit slips. My model utilized adaptive centering with random effects (Raudenbush, 2009). This means that lesson-level effects are group-mean centered within students, and that each student's daily instructional data is compared to the average of his or her data over the course of the year. This approach is particularly useful when exploring a program such as TBPP which features complex interdependencies among students' academic performance and the nature of the instruction assigned to them each day.

Although I initially intended to utilize a four-level model that nested instructional events within students within classes within schools, when I fit a one-way random-affects ANOVA model to partition the variance in exit slip scores, it revealed that less than 1% of the variance in exit slip scores lay across schools and classes, 12% of the variance lay across students in the same class, and 88% of the variance lay across lessons completed by the same student. Accordingly, I eliminated the level-3 and level-4 models and proceeded with a two-level model featuring lessons nested within students. My Level-1 model includes several time-variant instructional variables, including number of exposures to the skill, the number of rounds in which the skill had been taught to the student, the number of exposures the student has had to TBPP, instructional method, teacher classification, group size, average fall MAP score of the

group, and the content gap, which is defined as the difference between the grade level of the skill assigned for instruction and the grade level of the student engaged in instruction. My level-2 model features time-invariant student-level demographic variables, including beginning-of-year achievement on the NWEA MAP assessment.

I tested for two types of interactions across variables. First, my level-1 model included interactions between the instructional method and other lesson-level variables to test whether the effects of the teacher, group size, group mean MAP score, or content level varied depending on the instructional method. Second, I used a "slopes-as-outcomes" approach to test for cross-level interactions between the time-variant instructional variables in my level-1 model and the time-invariant variables in my level-2 model (Seltzer, 1995). These cross-level models tested for interactions between level-1 instructional variables and each student's standardized score on the Fall 2015 NWEA MAP assessment. While a typical instructional model would likely feature very strong correlations between a student's score on a baseline academic assessment such as MAP and subsequent daily academic assessments results, this effect was minimized given TBPP's use of baseline academic data to calibrate the difficulty of each student's daily instructional content. I tested for the within-level and across-level interactions separately; in other words, I did not test for interactions between the intra-level-1 interaction terms and the level-2 variables.

I describe my multi-level model below:

Level – 1: $Y_{ij} = \pi_{0j} + \pi_{1j}(SKILLCOUNT_{ij}) + \pi_{2j}(ROUNDCOUNT_{ij}) + \pi_{3j}(TBPPCOUNT_{ij}) + \pi_{4j}(METHOD_{ij}) + \pi_{5j}(TEACHER_{ij}) + \pi_{6j}(GROUPSIZE_{ij}) + \pi_{7j}(GROUPMEANMAP_{ij}) + \pi_{8j}(CONTENT_{ij}) + \pi_{9j}(METHOD_{ij} * TBPPCOUNT_{ij}) + \pi_{10j}(METHOD_{ij} * TEACHER_{ij}) + \pi_{11j}(METHOD_{ij} * GROUPSIZE_{ij}) + \pi_{12j}(METHOD_{ij} * GROUPMEANMAP_{ij}) + e_{ij}$

Level – 2: $\pi_{pj} = \beta_{p0} + \beta_{p1}(X_j) + \beta_{p2}(FALLMAP_j) + r_{pj}$

where: Y_{ij} = the exit slip score for lesson *i* delivered to student *j*, standardized (z-scored)

π_{0j} = the mean exit slip score for student *j*

SKILLCOUNT = the total number of lessons in which this student has received instruction on this skill, up to and including this lesson

ROUNDCOUNT = the total rounds in which this student has received instruction on this skill, up to and including this lesson

TBPPCOUNT = the total number of lessons in which this student has received instruction via TBPP, up to and including this lesson, centered and divided by ten

METHOD = the instructional method for the lesson

TEACHER = a dummy indicator for whether the teacher is a CMT or a TA

GROUPSIZE = the total number of students whose instruction is simultaneously supervised or led by the same teacher, centered

GROUPMEANMAP = the mean standardized MAP score for the group of students whose instruction is simultaneously supervised or led by the same teacher

CONTENT = the instructional content level, coded as described previously

FALLMAP = student *j*'s Fall score on the NWEA MAP assessment, standardized (z-scored)

X_j = a vector of the student demographic variables for student *j*

e_{ij} = the residual, unexplained variance associated with lesson *i,* assumed to be normally distributed with a mean of zero and a variance of σ^2

r_{pj} = the residual, unexplained variance associated with student *j,* assumed to be normally distributed with a mean of zero and a variance of σ^2

My within-student (Level 1) model estimates the extent to which variance in exit slip scores is associated with specific elements of a daily TBPP lesson. My student-level (Level 2) model then describes exit slip scores as a function of student characteristics. My final analysis entails a ‘‘slopes-as-outcomes’’ approach, which allowed me to ascertain whether the relationship between TBPP's instructional delivery (i.e. Level 1 variables) and exit slip performance varies for different types of students (i.e. Level 2 variables) (Seltzer, 1995). All analysis was conducted using Stata 15 software.

Because the data is longitudinal over the course of the year, it is important to include indicators for time within the model so as not to violate the statistical assumption of

independence of observations (Raudenbush & Bryk, 2002; Singer & Willett, 2003). I model the longitudinal nature of the data using three variables: SKILLCOUNT, ROUNDCOUNT, and TBPPCOUNT. The SKILLCOUNT variable reflects the number of lessons in which the student has received instruction on the skill, including the current lesson. Similarly, the ROUNDCOUNT variable reflects the number of rounds in which the student has received instruction on the skill, including the current round. Finally, the TBPPCOUNT variable reflects the student's total number of exposures to the TBPP model as a whole, centered and divided by ten. This is particularly relevant given some previously published indicators that students become more familiar with and successful in programs like TBPP over time (Murphy et al, 2014a; Ready, et al. 2017; Rockoff, 2015). I also test the interaction between TBPPCOUNT and METHOD to explore whether this "familiarity" effect differs between more traditional methods, like TI, and methods with a steeper learning curve, like OI.

Although reciprocal causation, also known as endogeneity, can create interpretative difficulties when studying time-varying predictors, there is a very low risk of reciprocal causation in this study (Singer & Willet, 2003). The time-varying predictors in my model (e.g. group size, lesson content, etc.) cannot be influenced by student participants within a single day's lesson, as they are determined by the external process of the TBPP algorithm. Although the exit slip score is coded in the data set as contemporaneous with the other variables, the other time-dependent variables are determined by the TBPP algorithm in advance of the instruction that culminates in the exit slip; there is no avenue for a day's exit slip to retroactively influence the time-dependent variables for that day.

Overview of cluster analysis. I utilized cluster analysis to explore the relationships between yearlong student outcomes and longitudinal patterns in both content assignments and

exit slip outcomes. Cluster analysis is a descriptive data mining procedure for uncovering latent groupings within unstructured data (Jain, Murty, & Flynn, 1999; Romesburg, 1984). It has sometimes been described as a form of "quantitative phenomenology" due to its ability to display detailed and rich patterns of data within and across individual cases (Bowers, et al., 2017). There are two types of cluster analysis: structured analysis, in which the researcher presupposes certain assumptions about the character of the groups, and unstructured analysis, in which the nature of the groups is determined by the structure of the data itself (Bowers, 2007; Eisen et al., 1998; Lee, et al., 2016). I chose to utilize unstructured analysis due to the paucity of extant literature on technology-based personalization which could provide guidelines regarding the structure of the data (Murphy et al, 2014a; Pane et al., 2015; Wang & Woodworth, 2011; Wendt & Rice 2013). Although there are many types of unstructured cluster analysis, I chose to utilize hierarchical cluster analysis with an average linkages clustering algorithm due to its ability to efficiently uncover underlying structures within large datasets (Bowers, 2007, 2010; Eisen & DeHoon, 2002; Eisen et al., 1998; Jain, Murty, & Flynn, 1999; Jorion et al., 2018; Romesburg, 1984; van'tVeer et al., 2002).

The combined use of cluster analysis and clustergram heatmaps presents several features that are well suited to this project. First, they do not rely upon the typical assumptions associated with OLS regarding multicollinearity, heteroskedasticity, and case independence, making them particularly useful when exploring educational datasets that are highly interdependent and nested (Bowers, 2007; Howell, 2002). Second, as Bowers (2007) describes, they retain the granularity of the data rather than aggregating to the mean and reporting a generalized trend. This is especially valuable when studying topics with an underdeveloped base of literature, such as technology-based personalization. Horn and Freeland Fisher (2016) suggest that while the bulk

of education research has historically investigated which interventions are most likely to work on average for a typical student, future research should instead probe deeper to chart predictably effective paths for individual students or types of students. The deep, broad, and diverse data produced by innovative technology-based learning models may unlock expansive new frontiers for educational research similar to the way that the mapping of the human genome sparked a revolution in medical research. Accordingly, it is appropriate to explore whether the statistical techniques that have proven so powerful in the field of bioinformatics may hold the same promise in educational contexts.

For this analysis, I grouped the students according to the similarity of the pattern of their standardized exit slip scores across the year. In addition, since exit slip scores are directly associated with each day's assigned content, and content assignment is in turn determined by each student's unique progression through the TBPP skills map, I conducted a separate cluster analysis using the content gap of assigned lessons as the relevant set of data upon which to cluster (content gaps were coded using the procedure described in the Data section above). In other words, I conducted the cluster analysis twice – once with students grouped according to similarity in the pattern of their exit slip scores, and a second time according to similarity in the pattern of the content levels assigned to them by the TBPP algorithm.

After completing the cluster analyses, I utilized several visualization techniques to aid analysis and make the results more easily comprehensible. First, I drew cluster trees, which are sometimes also known as dendrograms (Eisen et al., 1998; Romesburg, 1984). Cluster trees use lines to link cases and clusters of cases based on their similarity to one another. The algorithm places cases and clusters closest to those with which they are most similar, enabling the reader to use the length of the connecting line as a proxy for the quantitative similarity of the underlying

data. I also used a form of heatmap known as an clustergram to visualize the data. First pioneered in the field of bioinformatics, clustergrams represent the variables of interest with blocks of color, aiding the human eye in quickly and efficiently detecting patterns across cases (Bowers, 2007, 2010; Eisen et al., 1998; Lee, et al., 2016; Jorion, et al., 2018; van'tVeer et al., 2002). A clustergram typically displays cases as rows and data categories as columns. For my analysis, rows represent students and columns represent methods, days of instruction, or months of instruction. Each individual data point is represented by a color that reflects its value. Accordingly, the clustergram enable us to visualize the complete learning trajectory of each student longitudinally over the course of the year. Cluster analysis and heatmap visualization were completed using RStudio 1.0.143 software, with support from code written by Bowers & Zhao (2018) and developed through the support of the National Science Foundation under grant no. 1546653.

Clustergrams also enable the linking of dichotomous outcome variables to individual cases. In the bioinformatics literature, this technique is used to explore whether groups of genes are associated with the appearance of certain tumors, facilitating the development of diagnostic methods and treatments (Eisen et al., 1998; van'tVeer et al., 2002). Within the field of education, variables like high school completion and ACT attempts have been used as dichotomous outcomes (Bowers, 2007, 2010). For this study, my clustergrams will include three variables of interest: (1) a students' score on the Fall 2015 NWEA MAP math assessment; (2) a student's proficiency level on the Spring 2016 PARCC math assessment, and (3) a dichotomous variable reflecting whether a student met the "typical growth" norm published by NWEA for the period between Fall 2015 and Spring 2016. These analyses will directly address my second research question: what are the relationships among various elements of the TBPP model and student

outcomes, and in particular, do what extent do daily content assignment or exit slip data predict end-of-year results on the PARCC and MAP assessments? They will also enable me to explore whether the results differ for clusters of students, including latent groups that may not be identifiable based on available indicators (e.g. gender, LEP status, etc.)

The list below summarizes my analytic process, drawing heavily from Romesburg (1984) and Bowers (2007):

1. Convert clustering variables (i.e. exit slip or content assignment) onto a standardized scale
2. Create a resemblance matrix by calculating a distance measure between every case
3. Combine the two most similar cases into a cluster
4. Recalculate the resemblance matrix
5. Iterate over steps 3 and 4 until all of the cases are clustered into one cluster, e.g. n-1 times
6. Rearrange the order of the cases on the basis of their similarity according to the results of step 5
7. Draw the dendrogram
8. Draw the clustergram
9. Interpret the clusters

The clustering algorithm begins by matching the most similar cases based on the similarity of their respective data. These two cases are then redefined as a cluster, and the resemblance matrix is recalculated with the new cluster serving as a case. This process continues iteratively, with cases grouped into larger and larger clusters, until the clustering algorithm defines all cases as belonging to a single cluster encompassing the entire population of cases. This requires *n*-1 iterations, with n representing the total number of student cases. The clustering

process does not change the underlying data for each case, but instead reorganizes them so that similar cases are grouped together.

The process described above reflects two specific analytic decisions: the choice of hierarchical clustering as a clustering method and the use of average linkages as a distance measure. Below, I briefly describe the literature on the available alternate options and the rationale for my analytic choices.

Choice of clustering algorithm. I chose to utilize a hierarchical clustering method over the two most prominent alternatives, K-means clustering and self-organizing maps (Eisen & DeHoon, 2002; Jaskowiak, Campello, & Costa, 2014). The primary disadvantage of K-means clustering is that this technique requires the supposition of a pre-set number of clusters prior to initiating the clustering algorithm. Since there is no reason based on the literature or theory to assume *a priori* a specific number of clusters, any choice would be arbitrary and could interfere with obtaining the most accurate results (Eisen & DeHoon, 2002; Jain, Murty, & Flynn, 1999). One alternative option could be to utilize principal component analysis to identify a number of clusters that represent a significant portion of data, then apply k-means clustering for the classification (Ding & He, 2004). However, there is evidence that the principal components that contain most of the variation in the data do not necessarily capture most of the cluster structure, and clustering with principal components does not necessarily improve cluster quality (Yeung & Ruzzo, 2000).

Self-organizing maps, which were invented by Teuvo Kohonen in the early 1980s, are a technique for mapping high-dimensional vectors onto a smaller dimensional space (Eisen & DeHoon, 2002; Mangiameli, P., Chen, S. K., & West, D. 1996). One advantage of self-organizing maps compared to K-means clustering is that self-organizing maps do not require any

prior knowledge about the structure of the data. However, while self-organizing maps are well suited to high-dimensional input spaces like data on the structure of the human brain, my data requires clustering only according to the exit slip score or content gap. Accordingly, self-organizing maps would have been a poor choice for my data, which is poorly aligned with the type of continuous, high-dimensional input space for which self-organizing maps are typically utilized.

Choice of linking and distance methods. Even within the family of hierarchical clustering methods, there are several linking methods and distance measures from which to select (Costa, Carvalho, & de Souto, 2002; Jaskowiak, Campello, & Costa, 2014; Romesburg, 1984). I elected to utilize an average linkages method, which defines the similarity between any two clusters as the arithmetic average of the similarities between the objects in one cluster and the objects in the other (Romesburg, 1984). This method offers several advantages over the alternative single linkage and complete linkages methods. First, it is robust to missing data (Bowers, 2007, 2010). Second, it incorporates the full range of data from each case rather than only the most similar or dissimilar measure, making it a good fit for a research question that seeks to explore the full yearlong experience for each student. Finally, average linkages is widely used within the literature, and Romesburg (1984) suggests it as the preferred hierarchical clustering method (Eisen et al., 1998; Bowers, 2007).

In contrast, the alternate hierarchical clustering methods all offer serious drawbacks in their applicability to this data and research question. Whereas the average linkages method encompasses all corresponding objects within each cluster, the single linkage and complete linkage methods calculate the distance based on only the smallest or largest distance among cases, respectively (Jaskowiak, Campello, & Costa, 2014; Romesburg, 1984). Because the

multiple-choice format of TBPP's exit slip constrains students' daily outcomes to a relatively small number of possible values, these methods are likely to overestimate the similarity among cases. Centroid clustering represents a final alternative method, but the literature recommends against using this technique in combination with Pearson's correlations, since the differences in normalization of data vectors can produce strange situations in which distances decrease as we move up the cluster tree (Eisen & DeHoon, 2002).

I chose to utilize uncentered Pearson's correlations as a distance measure. One prominent alternative measure is Euclidian distance, which simply calculates the direct distance between the measures (Bowers, 2007; Jaskowiak, Campello, & Costa, 2014; Romesburg, 1984). While Euclidian distance is widely used in the literature, it does not work well for data that is not normalized, such as the data on the grade-level gap of assigned content that I include in my analyses (Eisen & DeHoon, 2002; Jain, Murty, & Flynn, 1999). In addition, in an empirical test of accuracy using yeast data, Costa, Carvalho, & de Souto (2002) found that Euclidian distance had the lowest accuracy in three out of four tested datasets, and was not demonstrably superior in the fourth. An alternative form of Euclidian distance is Cityblock or Manhattan distance, which calculates the sum of the distance along each dimension rather than the shortest distance overall (Eisen & DeHoon, 2002; Jaskowiak, Campello, & Costa, 2014). However, since this method is a variation of Euclidian distance, it suffers from many of the same shortcomings.

Spearkman's Rank and Kendall's Tau represent alternative techniques for calculating distance. These methods reduce the effects of outliers by converting data into ranks rather than calculating distance based on actual value, and are often used when analyzing ordinal data (Eisen & DeHoon, 2002; Howell, 2002; Romesburg, 1984). However, a visual examination of the

distribution of exit slip and content assignment data suggests that outliers will not be an issue, and ranking the data will not make sense given how many repeated values the dataset features.

Accordingly, I will calculate distance measures between cases using uncentered Pearson correlations, defined as:

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_x^{(0)}}\right)\left(\frac{y_i}{\sigma_y^{(0)}}\right)$$

where

$$\sigma_x^{(0)} = \sqrt{\sum_{i=1}^{n}\frac{(x_i)^2}{n}}$$

and

$$\sigma_y^{(0)} = \sqrt{\sum_{i=1}^{n}\frac{(y_i)^2}{n}}$$

In contrast to traditional Pearson correlations, in which each data point is subtracted from the case mean as part of the calculation, the uncentered correlation formulas above assume the mean for each case to be zero. This is important in situations in which two vectors have the same shape, but are separated by a constant value – for example, two students whose exit slip scores improved at the same rate over the course of the year, but began at a different starting point. In such a scenario, a traditional centered Pearson correlation would produce a correlation coefficient of 1, indicating that these two cases are identical, but an uncentered correlation method would helpfully distinguish between them (Bowers, 2007, 2010; Eisen & DeHoon, 2002; Eisen et al., 1998; van'tVeer et al., 2002).

Although the preceding section evaluated the relative strengths and shortcomings of various clustering methods and distance measures, it should be noted that no technique has been demonstrated to be universally superior to all others (Bowers, 2007; Costa, Carvalho, & de

Souto, 2002; Eisen & DeHoon, 2002; Eisen et al., 1998; Jain, Murty, & Flynn, 1999; Jaskowiak, Campello, & Costa, 2014; Romesburg, 1984). Instead, the choice of analytic techniques is highly dependent on context, data structure, and research question. There is not yet a robust literature on the application of clustering techniques to the student-level, daily data produced by personalized learning programs; this study may represent a first step towards building the more developed base of evidence that could be applied to this type of data in future studies.

5. Results: Hierarchical Linear Modeling

Lesson-Level Results

I utilized several models to examine the relationship between lesson-level predictors and standardized exit slip scores. The first of these, Model 1 (see Table 7 below), included as predictors the total number of lessons in which the student has studied the skill, the total number of rounds in which the student has studied the skill, the total number of TBPP lessons completed by the student since the start of the school year (centered and divided by ten), and dummy indicators representing the method of instruction, with OI as the uncoded comparison group.

Model 1 suggests that exit slip scores are .014*** standard deviations lower for each additional lesson in which a student is exposed to a skill and .021** standard deviations lower for each additional round in which a student is exposed to a skill. This implies that some students may become stuck on particular skills and have a hard time becoming "unstuck," even after repeated lessons. However, these results may also be influenced by survivorship bias, as students who pass exit slips are automatically excluded from the pool of students exposed to that skill an additional time. Model 1 also suggests that exit slips scores are .014*** standard deviations lower for every ten lessons in which a student participates in TBPP, regardless of how many times he or she has been exposed to that skill. In addition, the results for the method dummy variables suggest statistically significant positive effects for the OP (.035***) and LTP (.079***) methods compared to the OI reference category, but statistically significant negative effects for the LG (-.051***), SG (-.035***) and TI (-.016*) methods. However, although these results are statistically significant, their magnitude is quite small (Cohen 1988, 1992).

Table 7: Multi-level Regression on Standardized Exit Slip Results with Level-1 Interactions

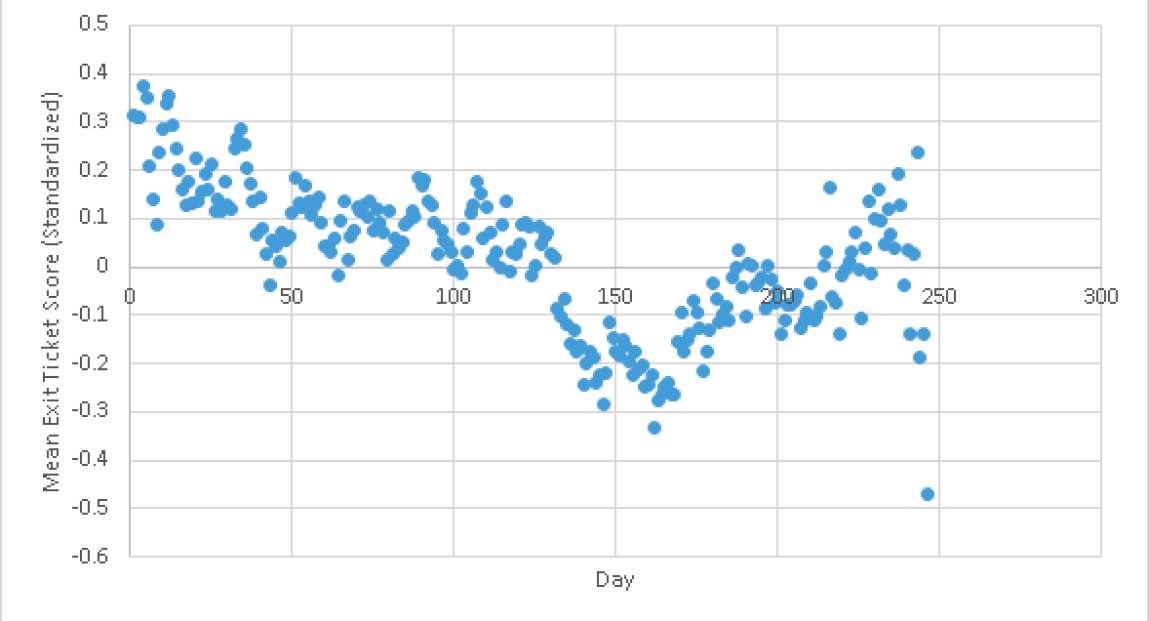
	Model 1 (n=169,745)	Model 2 (n=169,745)	Model 3 (n=169,745)	Model 4 (n=169,745)	Model 5 (n=169,745)	Model 6 (n=169,745)
Lesson (Level 1)						
Skill exposures	-.014***[a]	-.007***	-.007***	-.007***	-.007***	-.007***
Round exposures	-.021**	-.027***	-.026***	-.027***	-.028***	-.027***
TBPP exposures[b]	-.014***	-.006***	-.008***	-.006***	-.006***	-.006***
OP[c]	.035***	.034***	.034***	.012	.025**	.035***
PP	-.012	-.017*	-.016*	-.022~	-.015~	-.017*
LG	-.051***	-.007	-.005	-.033*	.050**	-.005
SG	-.035***	-.006	-.005	-.020~	.002	-.004
LTP	.079***	.141***	.142***	.148***	.140***	.137***
TI	-.016*	.048***	.049***	.054***	-.020~	.046***
Math teacher		-.016**	-.015*	-.033***	-.017**	-.015*
Group size[d]		.001	.001	.001	.002*	.005
Group MAP mean[e]		-.022***	-.025***	-.021***	-.021***	-.044***
-3 content gap		.486***	.485***	.485***	.483***	.484***
-2 content gap		.414***	.415***	.414***	.412***	.414***
-1 content gap		.234***	.234***	.234***	.231***	.234***
+1 content gap		-.116***	-.117***	-.116***	-.116***	-.116***
+2 content gap		-.433***	-.429***	-.434***	-.433***	-.430***
N/A content gap		.057***	.054***	.056***	.059***	.057***
Lesson-Level Interactions						
OP * TBPP exposures			.000			
PP * TBPP exposures			.000			
LG* TBPP exposures			.004**			
SG * TBPP exposures			.004**			
LTP * TBPP exposures			.002			
TI * TBPP exposures			.005***			
OP * MAteacher				.035*		
PP * MAteacher				.010		
LG* MAteacher				.067**		
SG * MAteacher				.031~		
OP * groupsize					.003*	
PP * groupsize					.000	
LG* groupsize					.009***	
SG * groupsize					.001	
LTP * groupsize					.002	
TI * groupsize					-.010***	
OP *Group MAP mean						-.009
PP * Group MAP mean						.008
LG* Group MAP mean						.086***
SG *Group MAP mean						.065**
LTP*Group MAP mean						.086**
TI * Group MAP mean						.014
Constant	.040**	-.091***	-.093***	-.081***	-.090***	-.090***
Student (Level 2)						
Random effect	.119	.154	.155	.154	.154	.155
Residual	.874	.847	.847	.847	.846	.847

~p<.10. * p<.05. ** p<.01. ***p<.001

[a] Outcome is standardized (M = 0, SD = 1); [b] Centered and divided by ten; [c] All methods are compared to OI
[d] Group size is centered; [e] Measure is the mean of MAP scores standardized within each grade (M = 0, SD = 1)

The schools' implementation of PARCC "test prep" in the spring complicates the interpretation of the effect of TBPP exposures. As indicated by Chart 2 below, students' mean exit slip scores decline significantly around Day 131 of implementation, which is the point in the year when the school district required that TBPP's algorithm be modified to assign all students to "on grade-level" content every day in order to prepare them for the high-stakes, state-mandated PARCC assessment. This is in contrast to TBPP's typical practice of assigning students to content that it determines to be in their zone of proximal development, which is typically below grade level.

Chart 2: Standardized exit slip results over time



To account for this change, I re-ran Model 1 using only the data from the first 131 days of the academic year, prior to the discontinuity introduced by test prep. This parallel analysis indicated no effect for TBPP exposures, a marginally significant negative effect for skill exposures (-.004~), and a significant negative effect for round exposures (-.036***). Although the findings for the first 131 days are slightly different than those for the full year, neither

analysis supports the existence of an increase in outcomes over time as students and teachers become more familiar with the new system. Instead, outcomes appear to remain consistent or decline slightly over time.

In Model 2, I add several additional lesson-level variables, including the teacher type (CMT compared to TA), centered group size, standardized group Fall MAP math mean score, and a dummy variable representing the gap between the student's grade level and the grade level of the lesson content. The findings indicate a statistically significant negative effect for CMTs compared to TAs (-.016**), although it is important to note that this dataset only allows the comparison of CMTs to TAs for the OI, OP, PP, LG, and SG methods, since the TBPP algorithm does not assign TAs to lead the TI or LTP methods. Model 2 indicates no statistically significant relationship between group size and exit slip score, and a negative relationship between group MAP mean and exit slip score (-.022***). This suggests that each standard deviation increase in the mean Fall Math MAP score of an instructional group is associated with a .022*** standard deviation decrease in the exit slip score of the students in that group. The method effects are in some cases slightly different in Model 2 than in Model 1; Table 8 below summarizes the effects in each model. In Model 2, there is a statistically significant positive relationship between exit slip scores and the OP (.034***), LTP (.141***) and TI (.048***) methods and a negative relationship for the PP method (-.017*) compared to the OI reference category. Again, however, the magnitude of these effects is quite small, with the exception of the positive effect for the LTP method.

Table 8: Method Estimates for Model 1 vs. Model 2

	Model 1	Model 2
Independent-Led Methods		
Online Instruction (OI)	N/A – Reference Category	
Online Practice (OP)	.035***	.034***
Paper Practice (PP)	-.012	-.017*
Student-Led Methods		
Large Group (LG)	-.051***	-.007
Small Group (SG)	-.035***	-.006
Adult-Led Methods		
Teacher Instruction (TI)	-.016*	.048***
Long Term Projects (LTP)	.079***	.141***

The largest overall effects within Model 2 are for the content level dummy variables, which represent the difference between the grade level of the instructional content and the student's grade level. There were very large and statistically significant positive effects for instructional content below a student's grade level, and very large and statistically significant negative effects for content above a student's grade level (-3=.486***; -2=.414***; -1= .234***; +1=-.116***; +2=-.433***). As a test for robustness, I also re-ran Model 2 using a different methodology for calculating the match between student and content level; rather than compare the instructional content to the student's grade level, I instead compared it to the grade level associated with that student's Fall Math MAP score (NWEA, 2015). This parallel analysis also indicated a statistically significant effect of -.117*** for each grade that the instructional content exceeded the student's MAP level. In other words, each level that the lesson's content exceeded the student's zone of proximal development as measured by MAP was associated a .117 standard deviation decrease in exit slip score. This means that students also performed better on content that was below their zone of proximal development.

Although it is not surprising that students would perform worse when tested on content above their zone of proximal development, what is surprising is that TBPP would produce these types of mismatches in the first place. After all, TBPP is specifically designed to eliminate student/content mismatches by implementing personalized instructional pathways for each student. However, I found that prior to Day 120 of instruction, 19.9% of all lessons addressed content that was more than one standard deviation above the student's Fall MAP level and 24.8% of lessons addressed content that was more than one standard deviation below the student's Fall MAP level[12]. In other words, nearly half of all lessons addressed content that was either far above or far below the zones of proximal development suggested by students' beginning-of-year assessments.

There are at least two potential explanations for this consistent pattern of mismatches. The first is that Fall MAP score is an imprecise estimate of a student's true academic level, especially as the year progresses and her or his abilities develop. In other words, as each student learns new math content and is matched to more challenging lessons through TBPP's ongoing analysis of daily assessment data, his or her Fall MAP score will quickly become outdated, creating an apparent mismatch. This explanation is partially supported by the fact that students were more likely to be matched with content that appeared too difficult than too easy, which would be consistent with the students' abilities growing beyond their beginning-of-year levels. However, this fails to explain the 24.8% of lessons that were assigned below the zone of proximal development suggested by students' Fall math MAP scores.

---

[1] I calculated grade levels associated with NWEA MAP scores by identifying the RIT score associated each grade level's Fall 2015 NWEA MAP math norm, then calculating non-overlapping RIT bands for each grade centered around each grade's norm

[2] Data for the full year is likely heavily influenced by the advent of test prep, with 30.8% of lessons addressing content more than one standard deviation above a student's Fall 2015 MAP level and 12.0% addressing content more than one standard deviation below.

An alternate explanation for mismatches is the logistical difficulty of generating a "right-fit" assignment for each student every day. Although the TBPP algorithm can assign students in the OI, OP, and PP methods to work on any content at any time, the TI, LTP, LG, and SG methods all require multiple students ready to work on the same content simultaneously. Accordingly, TBPP's scheduler may be forced to routinely place some students in groups focused on content that is either too low or too high. This problem is likely exacerbated by the algorithm's commitment to exposing each student equally to each method, regardless of her skill level or the skill level of her peers. However, the fact that content mismatches are approximately equally likely to appear within each method, as indicated by Table 9 below, suggests that the difficulty of creating groups for the TI, LTP, LG, and SG methods is not the sole cause of mismatches.

Table 9: Lesson Level Vs. Students' Fall Math MAP Levels – First 120 Days of TBPP

	TI	OI	OP	LG	PP	SG	LTP
% of lessons **above** student's grade level	22%	18%	17%	24%	20%	23%	22%
% of lessons **at or near** student's grade level	57%	54%	56%	55%	53%	57%	55%
% of lessons **below** student's grade level	20%	28%	27%	21%	27%	21%	23%
Total % of lessons **above or below** grade level	43%	46%	44%	45%	47%	43%	45%

Lesson-Level Interactions

Model 3 retains all of the lesson-level variables from Model 2, but adds interaction terms between each method and the total number of exposures to the TBPP model. This enables an examination of whether the relationship between exit slip scores and familiarity with the TBPP

model varies based on method[3]. In this model, the estimates for the method effects (e.g. OP, PP, etc.) represent the relationship between that method and exit slip score when the centered TBPP exposure value is set to zero. When the value for TBPP exposures differs from zero, its relationship with exit slip scores is described by the combination of the coefficient for TBPP exposures and the coefficient for TBPP interactions.

For example, when the value for centered TBPP exposures is zero, the effect size for a student in a TI is .049***. However, when the value for centered TBPP exposures is 1, then the combined effect size is .046. This value is obtained by combining the .049 value for the TI, the .005*** value for TI * TBPP exposures, and the -.008*** value for TBPP exposures. Similarly, when the value for TBPP exposures is 2, then the combined effect size is .043, representing a combination of the .049*** value for the TI effect, the .010 value for the "TI * TBPP exposures" effect (.005*** times two), and -.016*** for the TBPP effect (-.008 times two). As the pattern described above indicates, there is a net effect of -.003*** within the TI method for each point of increase in the TBPP exposures variable (-.008*** minus .005***). This indicates that for every ten additional TBPP lessons, a student's mean exit slip in the TI method is on average decreased by .003 standard deviations.

Looking across all of the interaction terms, we see a range from .000 to .005***. Because all of these values are smaller in magnitude than the -.008*** value for TBPP exposures alone, we can infer that there is a negative relationship between TBPP exposures and exit slip scores for all methods. However, the magnitude of that relationship varies across methods; it is largest for OP (-.008***), PP (-.008***), and OI (-.008***), but smallest for TI (-.003***). Although the

[3] These interaction terms consider total exposures to TBPP across all methods, not just the method with which the interaction is applied. In other words, "LG * TBPP exposures" reflects the effect of an LG lesson given the total number of all TBPP lessons, not just the total number of LG lessons.

magnitude of these effects is small, they can add up rapidly, given that values for TBPP exposures ranges from -10.9 to 13.6. In other words, a typical student's average exit slip score on the OI method is .080 standard deviations higher on average on the first day than the hundredth day of TBPP instruction.

Model 4 retains all of the lesson-level variables from Model 2, but adds interaction terms between each method and the teacher type. Interpreting the effect sizes using the same method as in Model 3, we see that the effect of a CMT is .002*** in the OP method, -.055*** in the PP method, .001*** in the LG method, and -.022*** in the SG method.

Table 10: Combining Effects for Model 4

	CMT effect	Method effect	Interaction effect	Combined effect
OP	-0.033***	0.012[4]	0.035*	0.002***
PP	-0.033***	-0.022~	.010	-0.055***
LG	-0.033***	-0.033*	0.067**	0.001***
SG	-0.033***	-0.020~	0.031~	-0.022***

These results are difficult to interpret. Given the similarity between the LG and SG methods; it is not clear why we should see a negative effect for a CMT in the SG method, but no effect for the LG method. Similarly, it is not immediately apparent why there should be a negative effect for a CMT in the PP method, but no effect in the OP method, which is structurally quite similar. Again, however, these effects are small in magnitude, and may represent an artifact of the very high statistical power of the model rather than meaningful variation in the program's effectiveness.

Model 5 tests for interaction effects between method and group size. It indicates statistically significant effects for group size within four methods: OI (.002*), OP (.005*), LG

[4] Non-significant effects are not included in the combined effect

(.011***), and TI (-.008***). Again, the effects for the OP, LG, and TI methods were calculated by combining the group size effect and the interaction effects. The negative relationship between group size and student achievement in the TI method is in keeping with previous literature on class size effects (Krueger & Whitmore, 2001; Mosteller, 1995). The positive interaction term between group size and the LG method can potentially be explained by LG's collaborative structure. In a collaboration-based method like LG, larger groups may be more likely to contain at least one student who understands the content well enough to explain it to others, who can translate the task into Spanish for a LEP peer, or with whom a middle schooler will have a close relationship that enables them to work productively. This means that collaboration-based methods may exhibit positive network effects as group size grows.

Finally, Model 6 examines the interaction between the method and the overall ability of the instructional group, as indicated by the standardized group Fall MAP math mean score. Notably, the three methods that feature the highest degree of student-to-student interaction have significant positive interaction effects: LG (.042***), SG (.021**), and LTP (.042***). This means that participating in a group with higher-performing students was positively associated with exit slip outcomes for these methods, while participating in a group with lower-performing students was negatively associated with exit slip outcomes. This is strong evidence of a peer effect, in which students benefit from proximity to higher-performing students when they are in methods that allow for significant peer interaction. There was no statistically significant interaction effect between group MAP mean and the OP, PP, OI, or TI methods, suggesting that any positive effects of group MAP mean do not extend to methods in which students are not spending a significant amount of time interacting with peers. Instead, the effect of group MAP mean in those methods is -044***, suggesting that students in higher-performing OP, PP, OI,

and TI groups scored lower on average on exit slips, all else being equal. One potential explanation for this effect is that higher-performing students are more likely to advocate for themselves by raising their hand to demand the teacher's attention, reducing the level of support available for other students in their OI, PP, OI, or TI group. Another explanation could be that teachers prepare more extensively or exert more effort when supervising a method with lower-performing students.

Student-Level Results

In Model 7 (see Table 11 below), I retain the lesson-level variables from Model 2, then introduce several time-invariant Level 2 variables, including fall MAP score, gender, race/ethnicity, free/reduced price lunch status, grade level, LEP status, and special education status. Of all these variables, only Fall math MAP was associated to a statistically significant degree with standardized exit slip outcome. The effect was .090***, suggesting that each standard deviation increase in a student's Fall MAP math score was associated with a .090 increase in that student's exit slip score. Conversely, it also suggests that each standard deviation decrease in a student's Fall MAP math score was associated with a .090 standard deviation decrease in that student's exit slip score. The interpretation of this result depends on what construct one assumes Fall Math MAP to represent. To the degree that it represents mathematics skill, the results suggest that the TBPP algorithm may not be setting ambitious enough targets for high-performing students. This would be aligned with my previously discussed findings regarding the prevalence of content level mismatch, which indicates that TBPP routinely matches high-performing students with content that is not rigorous enough to be in their zone of proximal development. In contrast, were one to assume that Fall Math MAP score were correlated with a broader set of abilities, then the result would suggest that higher-scoring

students were more likely to be successful on exit slips for reasons not related to mathematical skill - perhaps due to a greater ability to adapt to the innovative nature of the TBPP model, or increased ability to stay focused. Finally, since the MAP assessment is taken on a computer, the Fall Math MAP score may be assumed to represent facility with technology, which would certainly be relevant to TBPP's technology-heavy instructional model.

To evaluate these possibilities, I re-ran Model 7 with Fall Reading MAP in place of Fall Math MAP. I found that Fall Reading MAP was related to exit slip outcome (.044***), but with approximately half the effect size of Fall Math MAP. This lends credence to the theory that general academic ability, diligence, or facility with technology may be positively associated with exit slip performance, since those two competencies, but not math ability, were assessed by the MAP Reading assessment. However, the difference between the magnitude of the MAP Math and MAP Reading effects suggests that there is also some relationship between prior mathematical ability and exit slip performance.

In Model 8, I evaluated the interaction between a student's math MAP score and method, finding positive interactions for all methods except PP. The effects were largest for the LTP (.252***) and TI (.171***) methods.[5] Because the TBPP algorithm attempts to create homogeneous groups by placing students with peers who are at their same level, the positive effects across all methods could be interpreted as indicative of TBPP tracking students in ways that exacerbate existing inequalities. In other words, if high-performing students are more likely to grouped with other high-performing students, and low-performing students are more likely to be grouped with low-performing students, and students perform better when placed with higher-

[5] Again, I derived these values by combining the values for Fall MAP and the interaction effects. For example, the effect of a one standard deviation increase in Fall MAP score for a student in the LTP method is equal to .252***, which I calculating by combining the FallMAP effect of .071*** and the "LTP * FallMAP" effect of .181***.

performing peers, then participation in TBPP would widen gaps between the mathematical ability of the highest and lowest performing students. In addition, the fact that these effects seem strongest in the LG, SG, LTP, and TI methods is further evidence for the existence of peer effects, given that these four methods all involve significantly more student-to-student interaction than the PP, OI, or OP methods.

Model 9 examines the interaction between teacher type and a student's Fall MAP score, finding a marginally significant positive interaction (.010~) and a negative overall effect for a math teacher (.020*). This may indicate that higher performing students are more likely to benefit from the unique pedagogical content knowledge of MAs, or that lower performing students are more likely to benefit from the specialized special education or LEP skills possessed by TRs. However, given the marginal significance and very small effect size, I would caution against reading too much into this finding. Finally, in Model 10 I examined the relationship between Fall MAP score and group size. I found that for every standard deviation that a student's Fall MAP score increased, the association between exit slip outcomes and group size decreased by .001*** standard deviations. This suggests that larger group sizes are likely to be more harmful to the performance of lower performing students than higher performing students. Again, however, the effect size is quite small in magnitude.

Table 11: Regression on Standardized Exit Slip Results with Cross-level Interactions

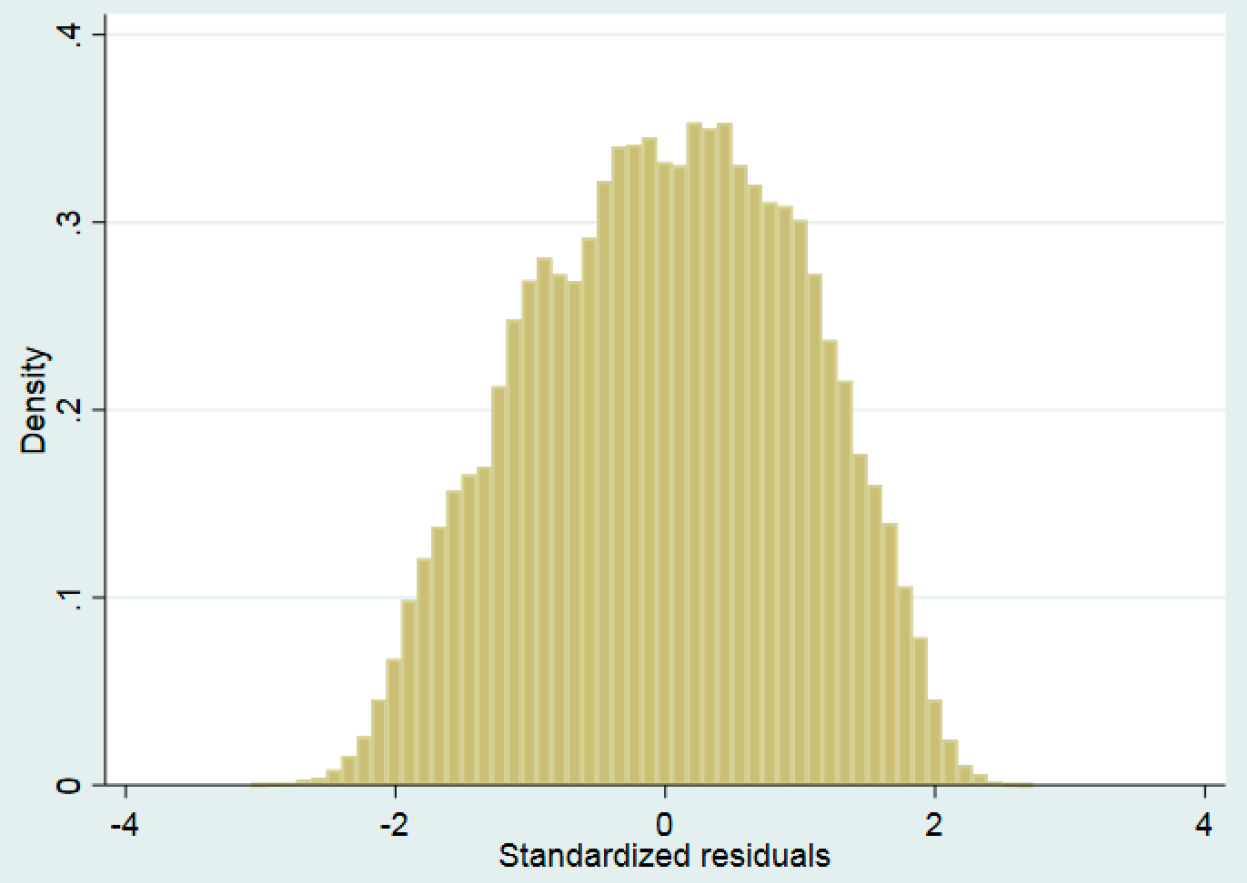
	Model 7 (n=153,062)	Model 8 (n=153,062)	Model 9 (n=153,062)	Model 10 (n=153,062)
Lesson (Level 1)				
Skill exposures	-.007***	-.008***	-.007***	-.007***
Round exposures	-.025***	-.028***	-.025***	-.025***
TBPP exposures	-.006***	-.007***	-.006***	-.006***
OP	.031***	.033***	.032***	.032***
PP	-.017*	-.018*	-.017*	-.017*
LG	-.006	-.003	-.006	-.005
SG	-.012	-.009	-.013	-.012
LTP	.149***	.148***	.150***	.149***
TI	.050***	.051***	.050***	.051***
Math teacher	-.021**	-.016*	-.020*	-.020*
Group size	.001~	.001	.001	.001
Group MAP mean	-.024***	-.051***	-.024***	-.032***
-3 content gap	.514***	.495***	.514***	.511***
-2 content gap	.427***	.420***	.427***	.426***
-1 content gap	.237***	.236***	.237***	.237***
+1 content gap	-.117***	-.114***	-.117***	-.117***
+2 content gap	-.453***	-.435***	-.453***	-.451***
N/A content gap	.059***	.065***	.059***	.060***
Constant	-.059***	-.055***	-.057***	-.057***
Student (Level 2)[a]				
FallMAP	.090***	.071***	.085***	.090***
Female	.007			
Black	.000			
Hispanic	.000			
Asian	.044			
Free lunch	.000			
Reduced lunch	.023			
LEP	.026			
SPED	.000			
Cross-Level Interactions				
OP * FallMAP		.020*		
PP * FallMAP		.006		
LG* FallMAP		.068***		
SG * FallMAP		.065***		
LTP * FallMAP		.181***		
TI * FallMAP		.100***		
Math teacher * FallMAP			.010~	
Group Size * FallMAP				-.002***
Random effect	.065	.076	.075	.075
Residual	.847	.847	.847	.847

~p<.10. * p<.05. ** p<.01. ***p<.001

[a] The model would not converge in Stata when grade level dummies were included alongside other Level 2 variables. However, I tested them in a separate model and found them not to be significant.

Checking for heteroskedasticity. After estimating Model 2, I checked for heteroskedasticity by obtaining the predicted value (Y-hat), the residuals, and the standardized residuals. A histogram of the standardized residuals reveals a normal distribution. I also generated a correlation table containing the squared residual, absolute value of the residual, and all independent and dependent variables. According to these results, the highest level of correlation for the squared residual is with the standardized exit slip score (-.110), followed by TBPP exposures (.061), skill exposures (-.019), round exposures (-.012), the PP method (-.011), group mean MAP (-.009), and the LG method (.004),

Chart 3: Distribution of Standardized Residuals



Summary of results from hierarchical linear modeling. In the preceding chapter I utilized several models to examine the relationship between standardized exit slip scores and multiple lesson-level variables, within-lesson interactions, and cross-level interactions. These analyses produced several interesting findings. These included the lack of an upward trend in performance that could be associated with students' and teachers' growing familiarity with TBPP, higher exit slip scores for the OP, TI, and particularly LTP method than for other methods, and a negative and significant effect size for math-certified MAs compared to TRs. I also found significant peer effects for the LG, SG, and LTP methods, as well as a positive association between each student's MAP math and exit slip scores, with the largest effects within the LTP, TI, SG, and LG methods. There was also a smaller, but still statistically significant relationship between MAP reading and exit slip scores, suggesting some unmeasured student characteristic other than mathematical ability that is associated with both MAP performance and daily exit slip performance.

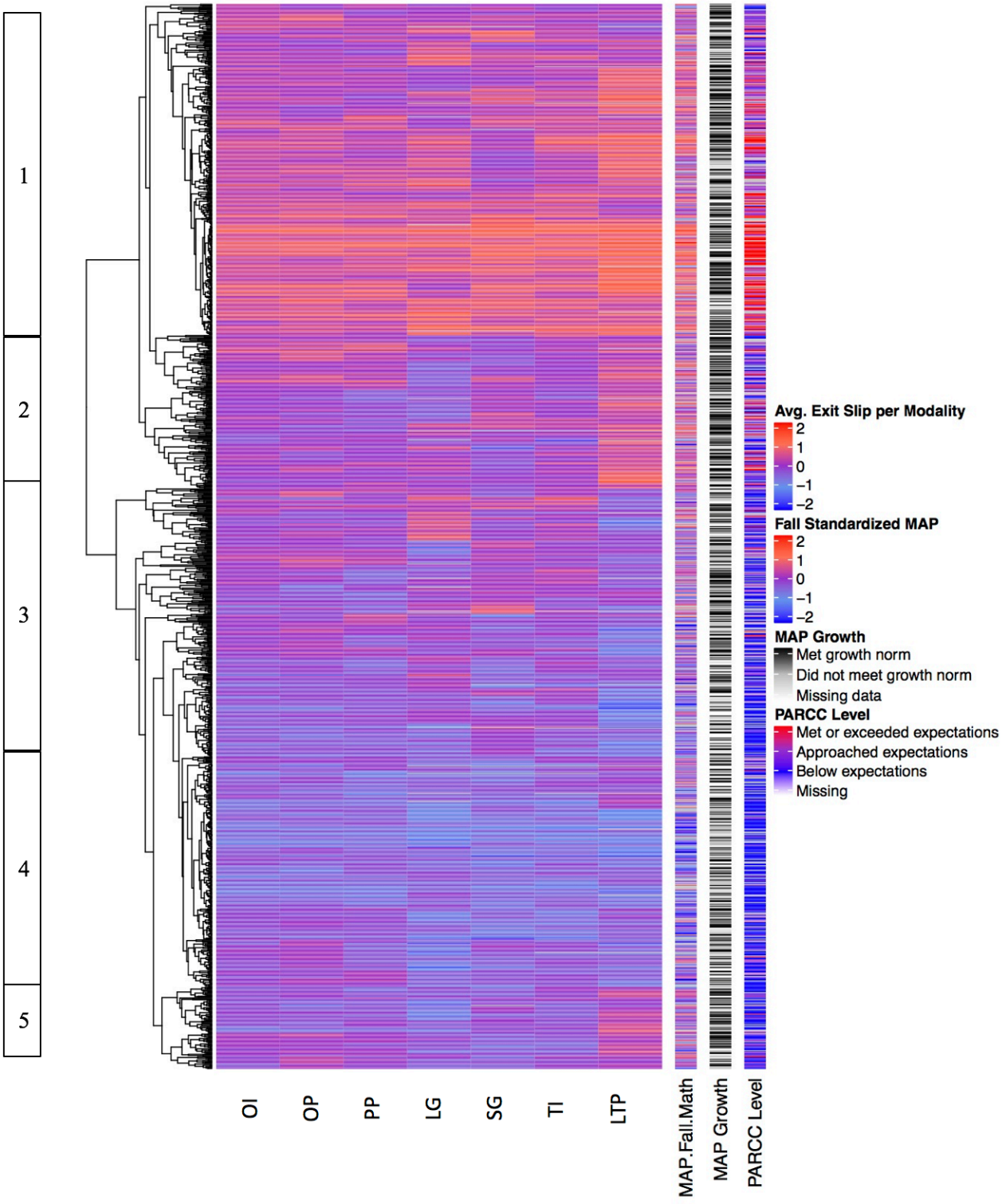
The variable with the largest magnitude relationship to exit slip scores is the content level of instruction, with students performing better on content below their grade level and worse on content above their grade level. Although it is not surprising that students would perform worse when tested on content above their zone of proximal development, it is surprising is that TBPP would produce these types of mismatches in the first place, given that TBPP is specifically designed to eliminate student/content mismatches by implementing personalized instructional pathways for each student.

6. Results: Hierarchical Cluster Analysis

Exploration of Data for Each Instructional Method

I generated several cluster analysis heatmaps to examine the relationships among exit slip results, content assignments, and year-long academic outcomes. The first of these, represented by Figure 3 below, displays the mean standardized exit slip score for each student (rows) disaggregated by the seven instructional methods utilized by TBPP (columns). Mean standardized exit slip scores are represented by color blocks, with blue representing the bottom of the scale, red representing the top of the scale, and purple representing the population mean. The similarity or dissimilarity of the patterns of exit slip outcomes is represented on the far left of the heatmap by the dendrogram, or cluster tree, with longer horizontal lines indicating dissimilar patterns and shorter lines indicating similar patterns. The three columns on the right of the heatmap indicate each student's standardized score on the Fall 2015 MAP math assessment, growth from the Fall 2015 to Spring 2016 MAP math assessment, and performance on the Spring 2016 PARCC math assessment. These columns enable comparison between students' exit slip patterns and their baseline mathematical ability prior to entering TBPP, growth in mathematical skills over the course of a year of participating in TBPP, and mathematical ability after a year of participation in TBPP, respectively. Although this study is unique in the application of cluster analysis and heatmaps to daily student assessment data, the overall approach is heavily informed by previous examples in the educational literature (Bowers, 2007, 2010, Bowers, et al., 2016; Lee, et al., 2016).

Figure 3: Standardized exit slip scores disaggregated by instructional method



This analysis yields several interesting findings. The first is the high level of correlation among exit slip performance, Fall 2015 MAP score, and Spring 2016 PARCC level. This is evident in the general consistency of the horizontal color bands, with blue, purple and red appearing synchronized across the three measures. This consistency indicates that students who enter TBPP with a higher mathematical ability are more likely to succeed on daily exit slips and also more likely to end the year proficient in grade-level mathematics content, as assessed by PARCC. For example, Students in Cluster 1 score well on daily exit slips, Fall 2015 MAP, and Spring 2016 PARCC, while students in Cluster 4 have lower scores on all three measures. The correlation between beginning-of-year and end-of-year mathematics performance is not surprising, given the well-document difficulty of disrupting entrenched student achievement gaps. What is surprising, however, is that these measures should also be correlated with daily exit slip performance. TBPP is designed to match each student with daily content at his or her precise zone of proximal development, which should make every student equally likely to master that day's exit slip, regardless of his or her starting level. Figure 3 may suggest that high-performing students are routinely matched with "too-easy" content and low-performing students with "too-hard" content. Alternately, it may indicate that there is some quality possessed by higher performing students beyond simple mathematical ability, such as socio-emotional skills or ability to learn, that makes them more likely to succeed on each day's exit slip.

There also appear to be small positive correlations between MAP growth and daily exit slip performance, Fall 2015 MAP performance, and Spring 2016 PARCC performance. The correlation between year-long MAP growth and daily exit slip data can be interpreted as evidence that daily exit slips are a useful measure of student learning, with higher performance on daily lessons associated with increased annual growth. However, the correlation between Fall

2015 MAP performance and annual MAP growth may also be evidence that TBPP provides inequitable experiences and outcomes for students who enter the program with different ability levels, as higher-performing students taking advantage of the program's autonomy to race ahead while lower-performing students languish or slip through the cracks.

Figure 3 also suggests a high degree of correlation in students' performance across all methods. In other words, students in Cluster 1 are generally successful on exit slips in all methods, while students in Cluster 4 are generally unsuccessful in all methods. However, there are some exceptions to this rule. For example, the students in Clusters 2 and 5 appear to be more successful in the LTP method than in other methods, while the students in Cluster 3 appear less successful in LTP than in other methods, such as TI and LG. Within-student differences in exit slip performance across methods would be in keeping with the theory of multiple intelligences that informed the creation of many personalized learning programs (Gardner, 2011; Horn & Staker, 2014). However, it is worth noting that TBPP is designed to expose all students to all methods with equivalent levels of frequency, not to adjust each student's method exposure based on her or his past performance.

The format of the analysis represented by Figure 4 below is similar to that of Figure 3, with the exception that the heatmap data represents the mean difference between each student's grade level and the grade level of the instructional content assigned to him or her within that method rather than mean exit slip performance. For example, lessons delivered to a 6^{th} grade student featuring 4^{th}, 5^{th}, or 6^{th} grade content would be coded as -2, -1, or 0, respectively. This enables an examination of the pattern of content assignment for each student within each method, as well as the relationships between that pattern of content assignments and Fall 2015 MAP performance, year-long MAP growth, and Spring 2016 PARCC performance.

Figure 4: Content levels of instruction disaggregated by instructional method

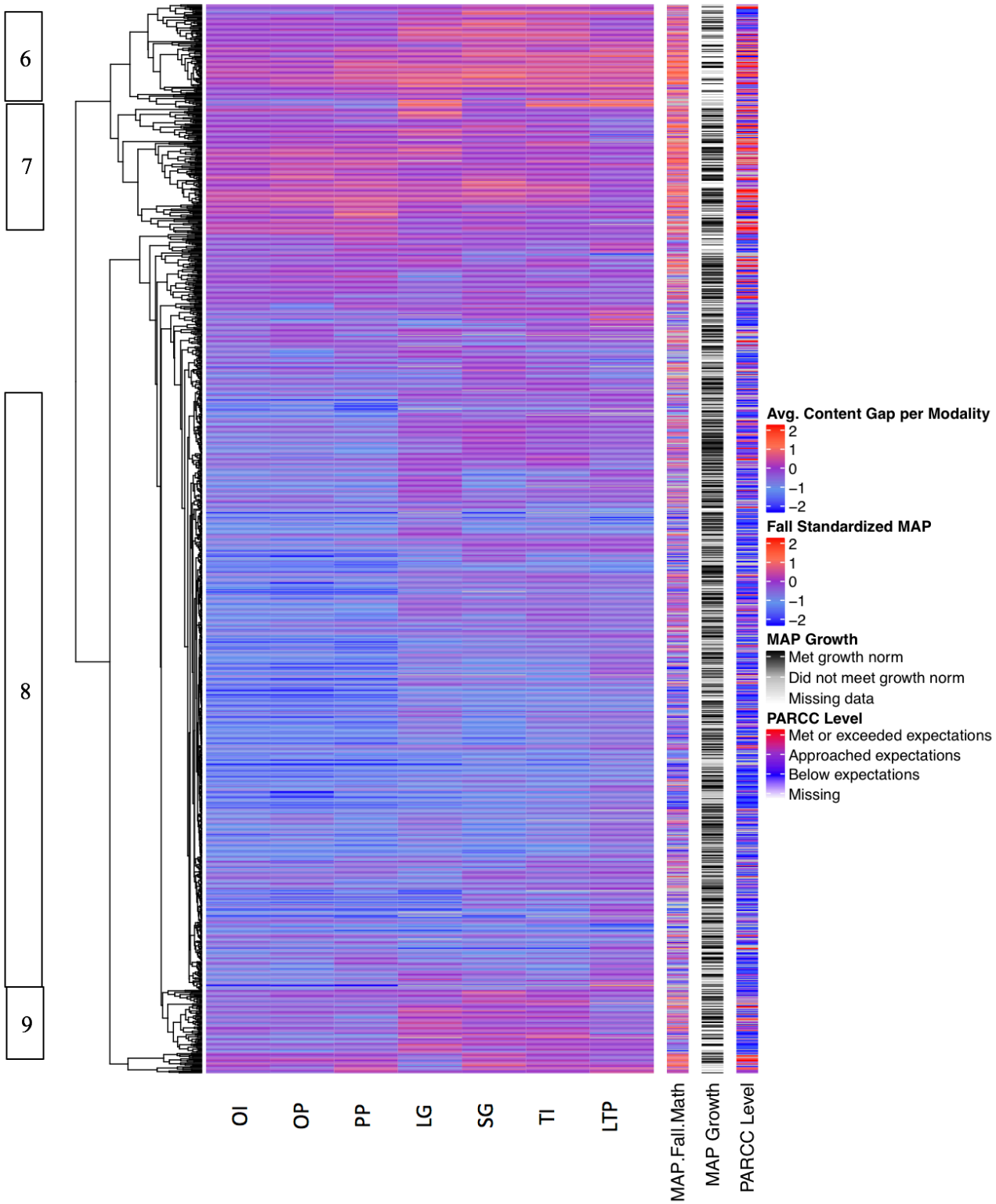


Figure 4 indicates several interesting data trends. First, the higher frequency of blue than red within the heatmap indicates that more instructional content is assigned below students' grade levels than above their grade levels. Second, the heatmap indicates a high degree of correlation between students' Fall 2015 MAP scores and the level of the content assigned to them; for example, students in cluster 6 generally performed above the mean on Fall 2015 MAP and were assigned above-grade level content, which is indicated by red shading on both measures, whereas students in cluster 8 were both more likely to perform below the mean on Fall 2015 MAP and to be assigned below-grade level content, which is indicated by blue shading on both measures. This would be in keeping with the theory of action for TBPP, which uses Fall 2015 MAP data to assign "just right" content to each student.

The heatmap also indicates a higher frequency of below-grade level content assignments within the OI, OP, and PP methods than within the LG, SG, TI, and LTP methods, especially for students in cluster 8. In contrast, some of the higher performing clusters, such as cluster 7 and cluster 9, display a higher frequency of below-grade level assignments within the LTP method than other methods. This may reflect the logistical challenge of generating a "right-fit" assignment for each student every day. Although the TBPP algorithm can assign students in the OI, OP, and PP methods to work on any content at any time, the TI, LTP, LG, and SG methods all require multiple students ready to work on the same content simultaneously. Accordingly, TBPP's scheduler may be forced to routinely place lower-performing students in groups focused on content that is too high (e.g. Cluster 8) or higher-performing students in groups focused on content that is too low (e.g. Cluster 7).

Longitudinal Exploration of Data

I also generated several heatmaps to examine longitudinal patterns of data across the duration of the academic year. In Figure 5 below, the heatmap displays standardized exit slip score for each student (rows) for each of 165 instructional days ranging from September 24, 2015 to June 20, 2016 (columns). As with Figures 3 and 4, standardized exit slip scores are represented by color blocks, with blue representing the bottom of the scale, red representing the top of the scale, and purple representing the population mean. The other elements of Figure 5, including the cluster trees, Fall 2015 MAP math data, MAP growth data, and Spring 2016 PARCC data are also generated and displayed in the same manner as in the previous analyses[6].

Figure 5 below reveals several interesting data features. First is the presence of several distinct clusters of students. Progressing from the top of the heatmap to the bottom, the students in Cluster 10 appear to have been generally successful on exit slips at the start of the year, but to have experienced declines in performance as the year progressed. This may be related to the implementation of PARCC test prep around Day 130 of instruction, which is indicated by the vertical bar labeled "14;" in other words, these students may have been successful when matched with below-grade level content at the start of the year, but struggled when the launch of test prep forced them to work exclusively with grade level content. In contrast, the students in Cluster 11 continued to experience significant success across the entire year, while the students in Cluster 12 struggled across the entire year. It is worth noting, however, that the vast majority of students

[6] In order to maximize the function of the clustering algorithm, I removed from the dataset the 120 students with the highest degree of missingness in lessons completed. This left 1073 students in the dataset for Figures 5, 6, 7, and 8 compared to 1193 for Figures 3 and 4. I tested for the possibility of bias by comparing the demographic indicators of the eliminated and non-eliminated students, including school, grade level, gender, race/ethnicity, FRPL status, IEP status, and LEP status. I found that the eliminated students were generally similar to non-eliminated students according to those indicators.

appear to have experienced both high and low scores on exit slips, indicating that each individual's performance could vary greatly from day to day. In contrast to the method-based heatmaps in Figures 3 and 4, there does not appear to be a high degree of correlation between year-long outcomes and the clusters of longitudinal data.

Figure 5 also contains several distinct vertical bands in which significant amounts of data appear to be missing. These bands occur throughout the year, but are most common in the months of March, April, and May. This may be associated with the implementation of test prep during this portion of the year. For example, teachers may have had students "take a break" from using TBPP so that they could take practice tests or otherwise prepare themselves for PARCC. Data also appears more likely to be missing in June, when students may be more likely to engage in non-instructional activities like field trips or end-of-year celebrations. Interestingly, the vertical bands appear to be roughly consistent across all clusters, suggesting that patterns in exit slip scores were more determinative for the clustering algorithm than patterns of exit slip missingness.

Figure 5: Standardized exit slip scores displayed longitudinally

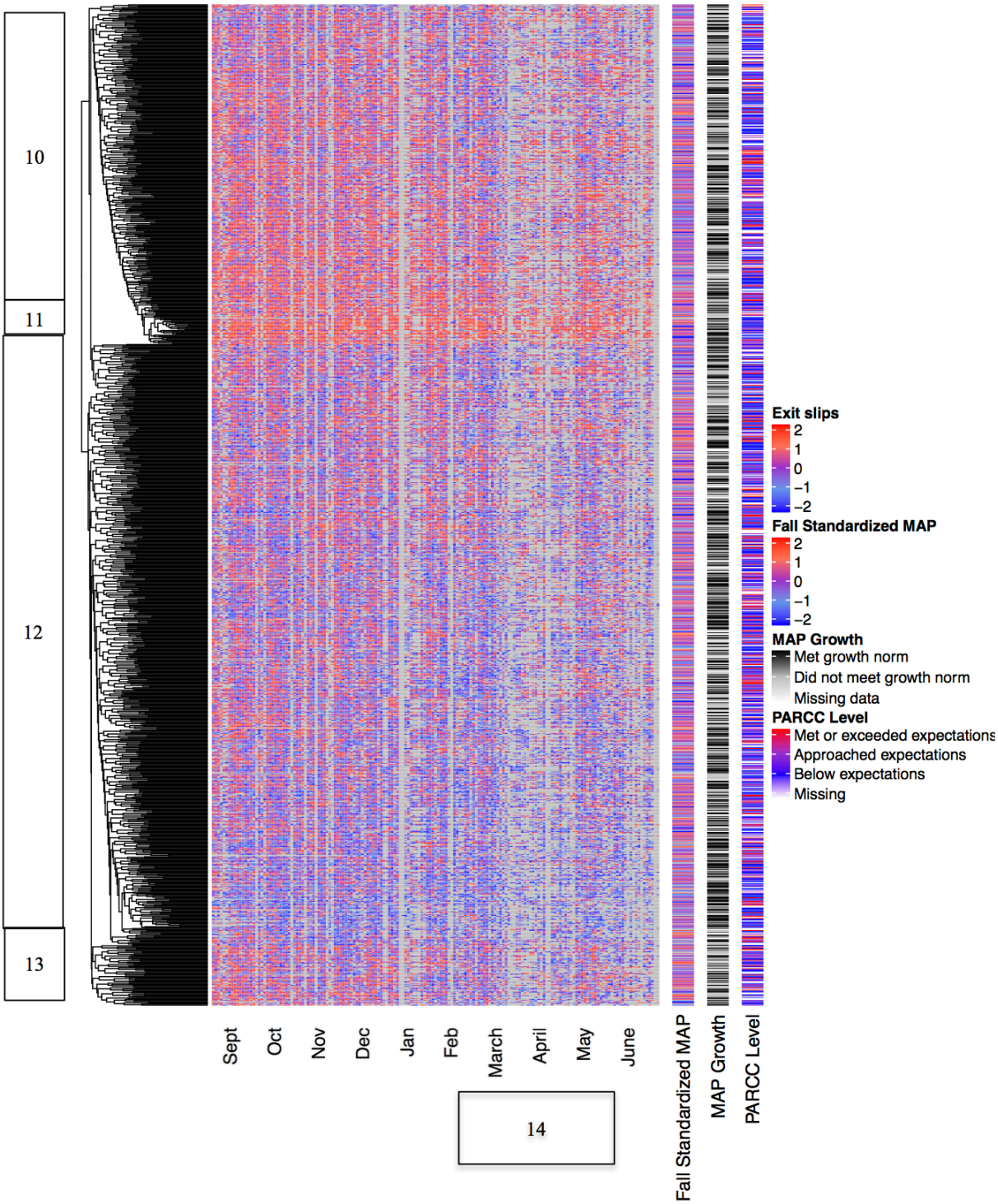


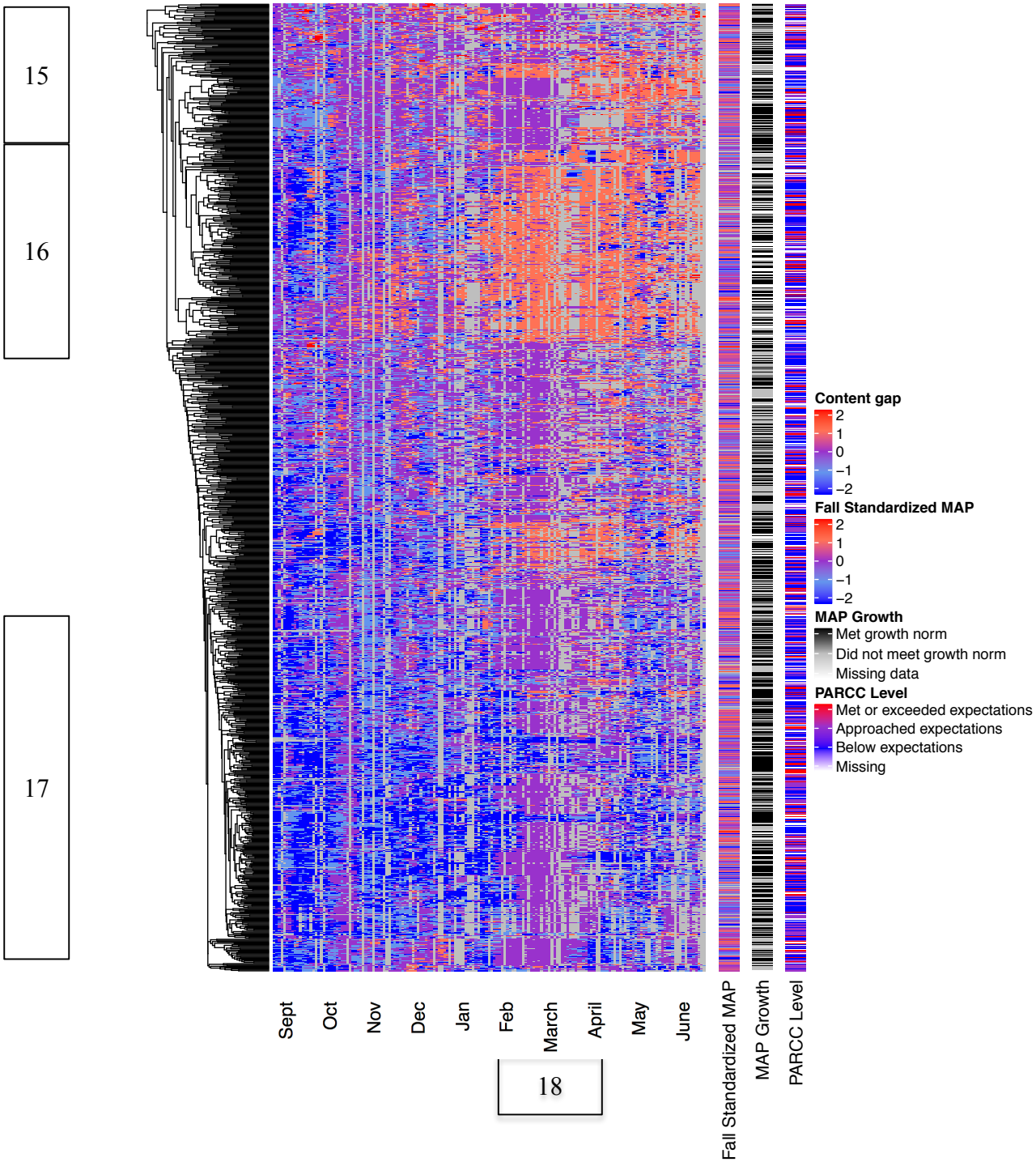
Figure 6 below also displays data longitudinally, but rather than standardized exit slip scores, the heatmap is clustered based on the content level of instruction. As in Figure 4, the content level is calculated as the gap between the content level assigned for the lesson and the student's grade level, enabling apples to apples comparisons across grade levels. In this heatmap, the color red is associated with content that is assigned above the student's grade level, while the color blue is associated with content that is assigned below the student's grade level.

Like Figure 5, the clustergram in Figure 6 contains several distinct clusters of students. Cluster 15 represents students who spent most of the year working with on-grade level content, but moved to mostly above-grade level content in the final third of the year. Fittingly, the MAP growth data indicates that these students were slightly more likely to meet their annual MAP growth than was the student population as a whole. The students in Cluster 16 began the year working with mostly below-grade level content, but were assigned above-grade level content once test prep began, and for the most part continued to work with above-grade level content for the remainder of the year. In contrast, the students in Cluster 17 began the year working with below-grade level content, shifted to on-grade level content for test prep, then reverted to below-grade level content once test prep was complete. Interestingly, a significant number of the students in Cluster 17 appear to have performed above the mean on the Fall 2015 MAP assessment and also met or exceeded expectations on the Spring 2016 PARCC assessment. This raises the question of why they were so consistently assigned below-grade level content throughout the year.

Figure 6 also contains a very clear marker for the period when test prep began, which I have labeled as vertical cluster 18. During this period, almost all students were assigned content that was on or above their grade level. This is evidence of how the policy constraint of high-

stakes testing may have forced an unorthodox implementation of TBPP by requiring students to engage in on- or above-grade level content even if it is above their zone of proximal development.

Figure 6: Content levels of instruction displayed longitudinally



Longitudinal exploration grouped by month. I also conducted a second set of longitudinal analyses with exit slip scores and content levels aggregated by month rather than displayed individually for each day. Figures 7 and 8 represent the results of those analyses.

Figure 7: Standardized exit slip scores displayed longitudinally with monthly groupings

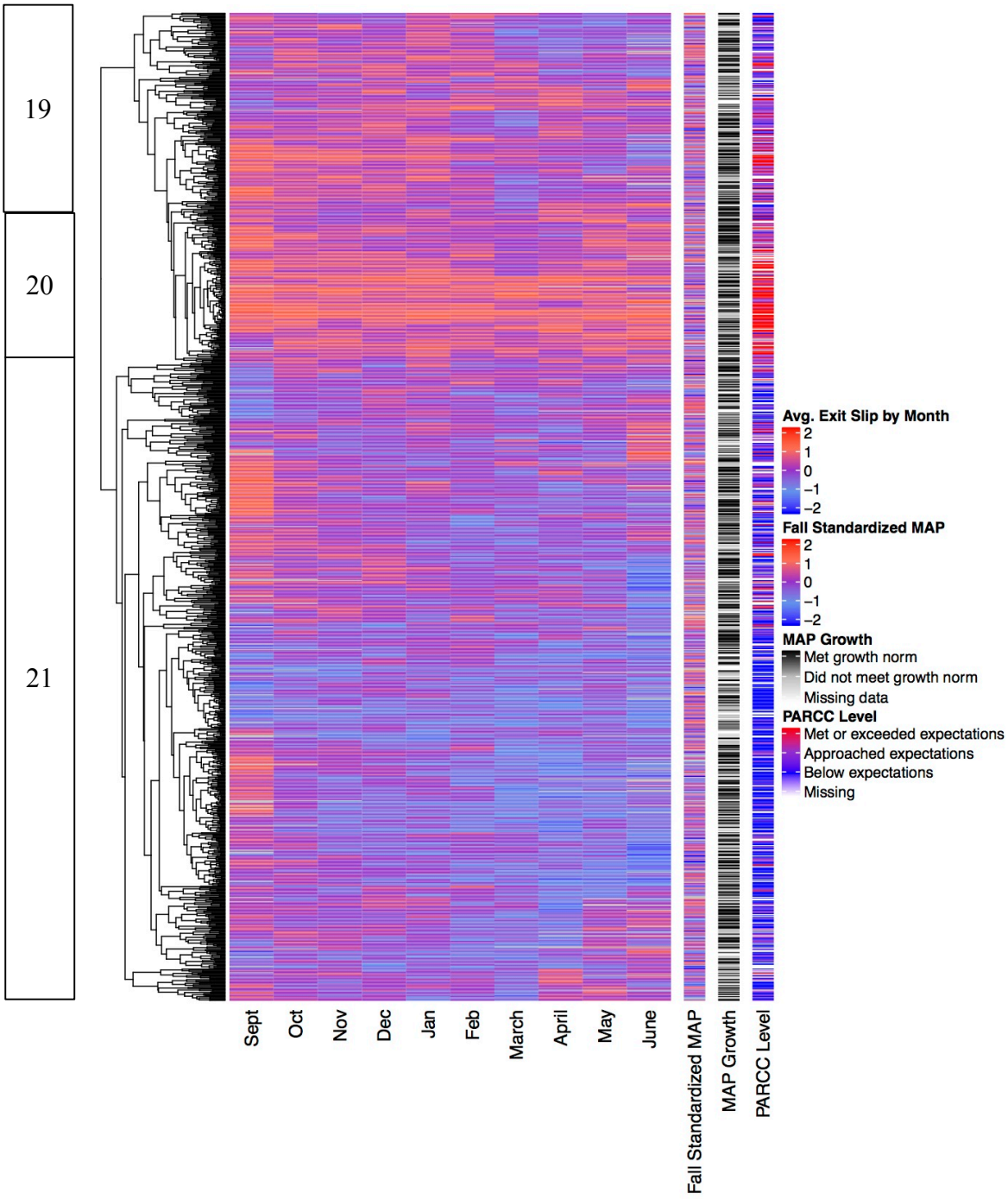
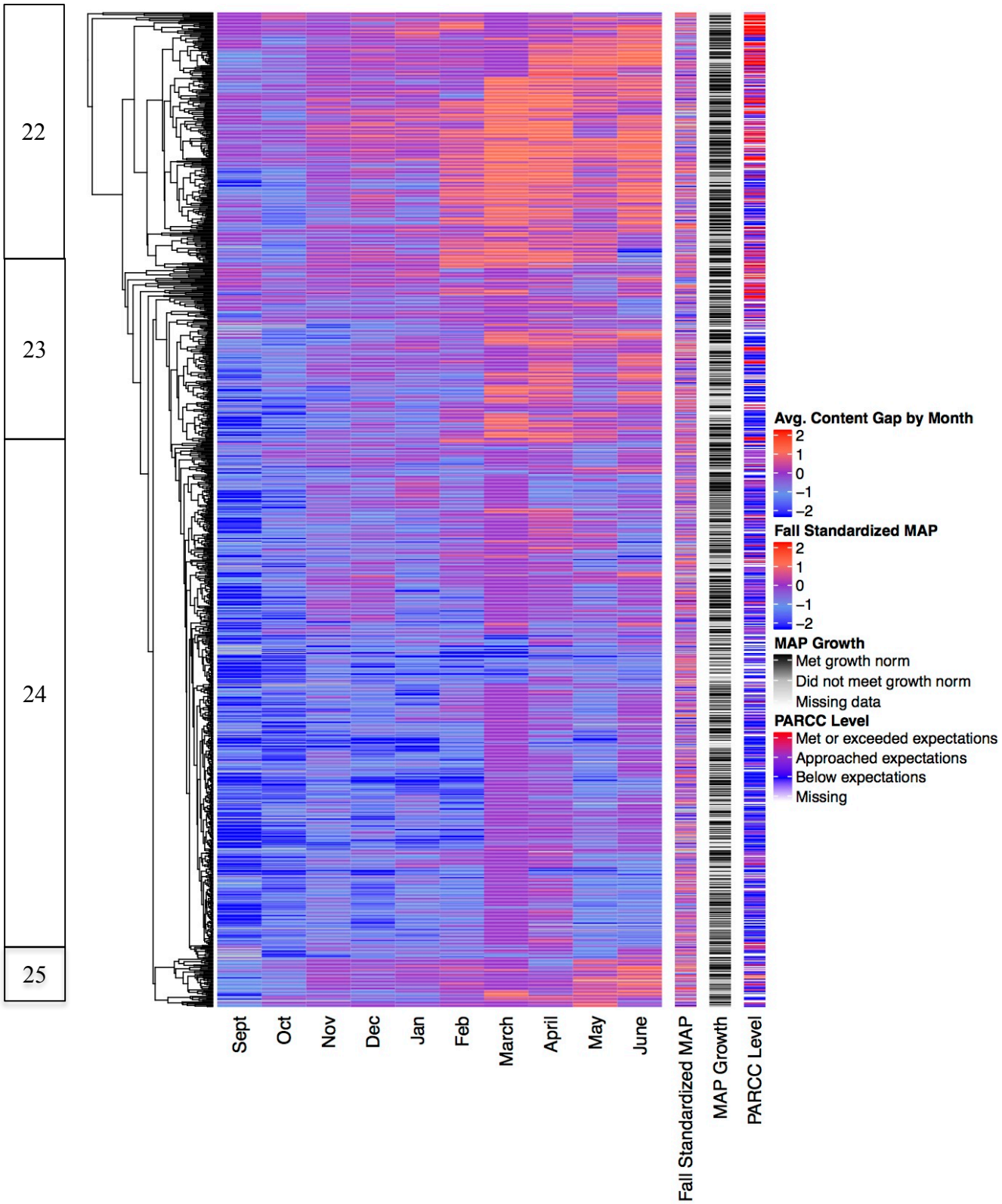


Figure 8: Content levels of instruction displayed longitudinally



Several interesting trends are apparent in Figures 7 and 8. First, the decision to aggregate the data by month rather than fully disaggregate for each day appears to improve the function of the clustering algorithm and support the generation of clearer and more distinct clusters. This is apparent in the longer horizontal lines in the dendrograms of Figures 7 and 8 compared to Figures 5 and 6, indicating a greater degree of similarity among the cases within each cluster. It is also apparent in the tighter correlation between the heatmap data and the PARCC data in both Figures 7 and 8. This relationship across multiple types of data suggests meaningful differences in the characteristics of students within each cluster. The decision to aggregate the data by month in Figures 7 and 8 also eliminates the "blotchiness" created by missing data in Figures 5 and 6, making the heatmaps easier to read and more visually accessible.

The clusters of students in Figures 7 and 8 are similar to those found in Figures 5 and 6, but more distinctly demarcated. Students in cluster 19 began the year with high exit slip scores, but their performance gradually declined, perhaps in tandem with the assignment of increasingly challenging content. Students in cluster 20 experienced the highest exit slip scores across the year, while students in cluster 21 experienced relatively low exit slip scores in every month but September. The correlation between exit slip scores and PARCC performance in all three clusters reinforces the finding that exit slip scores are a useful measure of student learning. They are also a striking display of the reality that different groups of students appear to have widely divergent experiences with TBPP. For the students in cluster 20, experience with TBPP seems associated with significant success, as indicated by high average exit slip performance every month. Not surprisingly, these consistently high-performing students are also the most likely to be proficient on the end-of-year PARCC assessment. However, the students in Cluster 21 have very different experience with TBPP. They typically score lower on their exit slips, and are also

much less likely to pass the PARCC assessment. Their relatively poor exit slip performance exists in spite of the fact that TBPP is nominally designed to match each student to content at their "just right" zone of proximal development. This suggests either that TBPP is matching these students with content that is too difficult for them, or that there is some factor other than the difficulty of the content that makes it more difficult for them to succeed on exit slips than the students in Cluster 20.

Figure 7 is also notable for the lack of any evidence of increasing performance over time as students and teachers gain familiarity with the program (Ready, et al. 2017; Rockoff, 2015). This is consistent with my HLM findings, which similarly found no increase in student performance as the year progressed. It may be that this improvement would occur in the second year of implementation. However, an examination of a second year of data unfortunately falls outside of the scope of this study.

Figure 8 also features several distinct clusters of students. Individuals in cluster 22 began the year with content on or below grade level, but experienced rapid increases in the level of content assigned to them. Students in this cluster were most likely to pass the PARCC math assessment, and also appear most likely to achieve their MAP math growth targets. Students in clusters 23 and 25 also experienced some growth, but their content assignments did not rise as quickly or as high as the students in cluster 22. In contrast, the students in cluster 24 were assigned below grade-level content all year long, with the exception of March and April, where the effects of test prep on content assignment are clearly apparent. Students in cluster 24 were much more likely to fail the PARCC math assessment than all other students. Unlike in Figures 3 and 4, there is a relatively low level of correlation between Fall MAP Math scores and the other measures reflected in the heatmap.

A final interesting feature of the data in Figure 8 is the significantly lower level of content assigned in September compared to the rest of the year. This suggests that the TBPP algorithm may intentionally begin the year by assigning all students below grade-level content to backfill missing skills or to boost their confidence with a new learning system. The very low level of content assigned in September is likely the root cause of the relatively high exit slip scores during that month in Figure 7.

Summary of results from hierarchical cluster analysis and heatmaps. I generated several cluster analysis heatmaps to examine the relationships among exit slip results, content assignments, and year-long academic outcomes. All analyses grouped students into several distinct clusters, affirming both the heterogeneity of student experiences within the program and the overall usefulness of these analytical techniques when studying the data produced by technology-based learning models. The heatmaps suggest a high level of correlation among exit slip performance, Fall 2015 MAP score, and Spring 2016 PARCC level. The correlation between beginning-of-year and end-of-year mathematics performance is not surprising, given the well-document difficulty of disrupting entrenched student achievement gaps. What is surprising, however, is that these measures should also be correlated with daily exit slip performance. TBPP is designed to match each student with daily content at his or her precise zone of proximal development, which should make every student equally likely to master that day's exit slip, regardless of his or her starting level. However, this did not prove to be the case.

The heatmaps suggest several additional interesting findings. The first of these is a high degree of correlation in students' performance across all methods, with the exception that many students appeared more likely to be successful in LTPs than in other methods (this is aligned with the results from HLM). Additional findings include a higher frequency of below-grade level

content assignments within the OI, OP, and PP methods than within the LG, SG, TI, and LTP methods, especially for lower-performing students; the existence of a period in March, April, and May when students are not matched with below-grade-level content and are more likely to exhibit missing data; and an approximately month-long period at the start of the year when almost all students are assigned content far below their grade levels.

7. Discussion

Effectively differentiating instruction is one of the most fundamental challenges of public education. Historically, the American education system has addressed this dilemma in two ways. The first is through tracking, in which students of different levels are sorted into homogeneous classrooms within schools. However, tracking has been criticized in recent decades for reinforcing inequalities based in race, ethnicity, and class, and it has recently fallen out of favor (Barr & Dreeben, 1983; Lee & Ready, 2009; Oakes, 1985). The second, more common strategy for addressing diverse student needs is classroom-level ability grouping in which teachers are given broad discretion to informally and formally assess students, organize them for instruction, and customize the content or pedagogical techniques used for each group (Barr & Dreeben, 1983; Bidwell, 1965; Corno, 2008; Martinez, Schecther, & Borko, 2009; Pallas et al., 1994). Despite its prevalence in American schools, ability grouping has at least two significant shortcomings. The first is the tension between the varied needs of a diverse student body and policy mandates that all students meet a common, minimum level of proficiency; these mandates have become increasingly explicit and consequential over the last twenty years (Hyslop & Mead, 2015; Manna, 2011). The second is the significant demand that ability grouping places upon the time, energy, and skill of classroom teachers. The work of continually assessing and regrouping is incredibly difficult, and can require that teachers plan multiple lessons for every day of instruction. There is evidence that the challenges of differentiating instruction are a significant contributor to teacher burnout and attrition (Arnett, 2016; Beteille & Loeb, 2009; Carnoy & Levin, 1985; National Mathematics Advisory Panel, 2008; TNTP, 2014).

Some educators, policymakers, and philanthropists have recently argued that new technologies offer the potential to more effectively support teachers in delivering differentiated

instruction that meets the unique needs of every learner (Cavanagh, 2014; Herold, 2016a; Horn & Staker, 2014). However, while some prior research has explored the overall effects of these programs, very little work has been done to describe the ground-level reality of how the behavior of students and teachers affects, and is affected by, these programs in the context of daily instruction. My research addresses this gap in the research literature by examining the implementation of a technology-based personalized learning program in five schools to better understand the complex relationships among school-level, class-level, student-level, and lesson-level factors and both daily and annual student learning outcomes. I examine these relationships using a variety quantitative methods, including hierarchical linear modeling, cluster analysis, and data visualization heatmaps.

I also examine the degree to which the day-to-day, ground-level implementation of TBPP represent an authentic departure from the traditional technology of schooling. Traditional forms of instructional delivery have proven exceedingly difficult to disrupt over the last hundred years, with successive waves of reform typically crashing on the rocks of entrenched organizational norms before receding with little or no trace (Carnoy & Levin, 1985; Cohen, 1990; Cuban, 1986, 1990, 1993; Elmore, 1996, 2010; Tyack and Cuban, 1995). My research examines the prospect that technology-based personalization may represent a divergence from this historical pattern, or whether teachers and students are merely engaging in symbolic reform while continuing to exercise traditional instructional patterns.

Findings

This study supports four main findings: (a) TBPP succeeds in altering the technical core of instruction in several fundamental ways; (b) policy and logistical constraints limit TBPP's ability to reform the technical core of instruction to the degree that it aspires; (c) students who

enter the program as already higher-performing are more successful on daily exit slips than students who enter the program with lower performance; and (d) the quantitative methods used in this paper represent useful and replicable tools for exploring the data produced by technology-based and personalized models.

Meaningful reform of the instructional core. I find that TBPP succeeds in altering the technical core of instruction in several meaningful ways. For example, whereas in traditionally organized instruction teachers work with a common group of students for an entire year or semester, in TBPP teachers work with multiple distinct, non-repeating groups of students each day. Whereas in traditionally organized instruction teachers are expected to teach a single, clearly defined scope of content, in TBPP teachers are expected to teach a wide array of content ranging from 2nd grade to high school math. Finally, while traditionally organized instruction is characterized by a high degree of teacher control over instructional content and method, teachers in TBPP have no ability to influence the assignment of students, content, or instructional methods. These are very significant changes. The literature on instructional reform is littered with failed reforms that only glancingly or symbolically alter the fundamental interactions among students, teachers, and content (Cohen, 1990; Cuban, 1986; Elmore, 2006, 2010; Honig & Hatch, 2004; Tyack & Cuban, 1995). TBPP appears to have succeeded where they failed in authentically altering the technical core of teaching and learning.

However, while TBPP succeeds in reducing the teacher's role as the ultimate arbiter and mediator of knowledge, it does not shift that power towards students, as has been encouraged by some proponents of technology-based personalization (Childress & Amrofell, 2016). Instead, TBPP shifts power from the teacher to the algorithm while leaving students relatively powerless to determine the course of instruction. This represents a significant divergence from earlier

reforms such as the Dalton plan, which relied heavily on student choice, and a direct repudiation of the learning models advanced by theorists such as John Dewey and Maria Montessori. Rather than direct their own learning, students in TBPP have their learning directed by an algorithm. Rather than choose their own goals through self-guided exploration and discovery, students are pushed toward a common, uniform level for excellence through instructional assignments that are dictated by their results on multiple-choice assessments of procedural skill. While ardent advocates of technology-based personalization argue that it will empower students to "choose your instructional method," the practical reality is that TBPP affords neither teachers nor students the ability to choose the content or methods in which teaching and learning will occur.

This shift in agency from teacher to algorithm manifests itself in the data in several interesting ways. The first is the small but statistically significant negative effective size for certified math teachers (CMTs) compared to teacher assistants (TAs). While this comparison is only possible for the OI, OP, PP, LG, and SG methods, it still comes as a bit of a surprise; one would intuitively assume that even in these methods, students would benefit from being supervised by math teachers with deep knowledge of the content under study. However, the data indicates that this is not the case, and that students perform equally well or better when supervised by teachers with lesser mathematical training and ability. This may reflect the limitations that TBPP places on teachers' ability to build and exercise relevant pedagogical content knowledge. A TBPP teacher may be told at 4:30pm that their instructional load for the next day will include an LG on 4th grade fractions, a TI on 9th grade algebra, another TI on 5th grade geometry, and a OI in which students are studying fifteen unique skills ranging from 3rd through 8th grade. It is extremely unlikely that the teacher would have the time to refresh their understanding of all of these concepts, prepare for the most common student misunderstandings,

and proactively plan for how to respond to each of them. Faced with such a dizzying array of content, trained CMTs may behave similarly to TAs with little or no experience in math instruction. Previous research has found that that pedagogical content knowledge is both domain-specific and associated with student achievement (Ball, 1990, 1997; Ball et al., 2008; Hill et al., 2005; Shulman, 1987). By forcing CMTs to teach multiple subjects with little preparation, TBPP may be negating the relevance of their pedagogical content knowledge in ways that suppress student outcomes.

This finding is very much in keeping with the theory of action suggested by many proponents of technology-based personalization (Arnett, 2016; Christensen, 2013; Christensen, et al. 2008; Horn & Staker, 2014). Many of these individuals suggest that schools would become more efficient and students better served were the role of the teacher more effectively differentiated, and individuals with varying levels of skill hired to engage in custom-tailored tasks at an efficient cost. For example, were TBPP to utilize CMTs exclusively for long-term LTP instruction while using lower-paid aides to supervise the OI, OP, and PP methods, it might produce equivalent or improved instructional outcomes at lower cost compared to traditional instructional models. Indeed, this is the exact approach utilized by instructional models like the Summit Personalized Learning Platform, which splits students' time between complex, long-term, real-world projects supervised by content-expert teachers and computer-based practice of basic skills while supervised by lower-paid instructional assistants (Osborne, 2016).

It is also interesting to note the significantly higher exit slip outcomes on the LTP method than for all other methods. The LTP method is unique in that it is the only method in which the teacher, students, and content remain consistent for more than one day. By the time that students take their exit slip on the second or third day of the LTP, teachers will have had several days to

build or strengthen relationships, informally assess students' knowledge of the material, and adjust instruction in response. In all other methods, teachers have only a single, thirty-five minute period to address the entire skill. LTP is also the only method in which teachers have the ability to deliver instruction, review data, then come back the next day to address specific misconceptions or target individually struggling students. It is striking that students appear to be most successful in the method in which teachers are best able to engage in these traditional instructional tasks.

However, the fact that the LTP method is associated with higher student outcomes than other methods does not necessarily suggest that traditional instruction is universally superior to technology-based, personalized models like TBPP. While the LTP method is more similar to traditional instruction than the other six TBPP methods, it features significantly smaller group sizes and far greater student homogeneity than a typical classroom, neither of which would be possible outside the context of the larger TBPP model. Any comparisons with traditional instruction are only suggestive and circumstantial; future studies would need to directly compare data from TBPP and traditional classrooms prior to drawing any firm conclusions about comparative effects.

Policy and logistical constraints. My second major finding is that policy and logistical constraints limit TBPP's ability to reform the technical core of instruction to the degree that it aspires. This finding manifests itself most clearly in the data related to test prep in February, March, and April. While TBPP's intention is that students engage only with content at their unique zone of proximal development, the clear purple vertical bar in the month of March in the longitudinal heatmap in Figure 8 (p. 110) demonstrates that many students are pushed to work with grade-level content around the time of PARCC testing. In addition, the pattern of students'

exit slip results suggests that this push toward grade-level content is associated with a decrease in student outcomes. The decision to focus on grade-level standards during the spring is not an inherent part of TBPP's design; on the contrary, it was imposed unwillingly upon the non-profit that manages TBPP by school and district administrators who feared the consequences of low PARCC scores. The policy constraints posed by high-stakes standardized testing clearly inhibit the ability of TBPP to function as intended during these spring months, a finding that is in keeping with other examples in the literature (Hyslop & Mead, 2015, Murphy et al, 2014a). In addition, the higher incidence of missing exit slips in March, April, and May could suggest intermittent implementation of TBPP due to teachers replacing TBPP instruction with practice tests, test prep workbooks, or other activities specifically designed to maximize performance on the PARCC assessment. In other words, the policy constraint posed by high-stakes testing may not only be incentivizing schools to reduce the personalization of content for part of the year, but also to partially abandon the use of TBPP altogether.

This is a powerful example of coercive isomorphism (DiMaggio & Powell, 1983; Meyer & Rowan 1977, 1978). The imposition of government-mandated assessments of student achievement, paired with the threat of sanctions or school closure in the case of low results, creates a powerful incentive for educators to abandon TBPP's model of skill-based differentiation and instead confront all students with the common set of grade-level standards that will appear on the PARCC exam (Hyslop & Mead, 2015; Pane et al., 2015; 2017). In other words, while TBPP may succeed in authentically reforming the technical core of instruction during most of the year, that reform seems to revert to mere symbolism during the window of time when the pressures of test-based accountability are most acute.

Logistical constraints may also inhibit the ability of TBPP to fully personalize content to students. As indicated by Table 9 (p. 88), nearly half of instructional assignments fell outside of the zone of proximal development suggested by students' performance on the Fall MAP math assessment. The heatmap of content levels within each instructional method in Figure 4 (p. 102) suggests that it may be easier to match students with far-below grade level content in the OI, OP, and PP methods than in the other four methods. This could be attributable to the fact that the OI, OP, and PP methods do not require any other students to be simultaneously working on the same skill. In contrast, assigning a student to LG, SG, TI, or LTP typically requires between five and fifteen other students who are also ready to be matched to the same skill. To give a practical example, if only two students need practice with a specific 5th grade geometry skill, it is logistically impossible for them to ever work on that skill in a TI, LTP, LG, or SG, since there will not be enough peer students to work on it with them. Even with more than one hundred students in a class, it may simply be impractical to match every student with his or her ideally leveled content every day. This logistical constraint likely inhibits the ability of TBPP to offer the fully personalized experience that it aspires to create.

In addition, the relatively high prevalence of purple coloring for the LTP method in Figure 4 suggests that it may be particularly difficult to match students with content in their zone of proximal development for LTP lessons. This is true for both low- and high-performing students. Students in the high-performing cluster 7 were mostly assigned above-grade level content in the first six methods, but their LTP assignments were more likely to be colored purple, indicating that they worked on comparatively lower-level skills within LTP lessons. Conversely, lower-performing students in cluster 8 also exhibit a mismatch between the coloring of their content assignments for LTP lessons compared to the other six methods, but in the opposite

direction, with non- LTP lessons predominantly colored blue for "below-grade-level" while LTP lessons feature a higher prevalence of purple coloring. The LTP method's multi-day nature likely makes it particularly difficult to generate groups of students who all need the same above- or below-grade level skill for an extended period of study. Again, the uniqueness of the LTP method is apparent; the fact that it is most similar to traditional forms of instruction means that it also least reflects the radical personalization at the heart of the TBPP model.

Unequal experiences and outcomes. My third major finding is that students who enter the program as already higher-performing are more successful on daily exit slips than students who enter the program with lower performance. This is apparent in the HLM results, which display a statistically significant positive relationship between Fall MAP math score and exit slip outcome, suggesting that students who enter the year with more mathematical ability are more likely to be successful on exit slips each day. It is also apparent in the heatmaps, which group students into clear clusters based on their performance on exit slips. The students who consistently perform higher on daily exit slips are also more likely to pass the PARCC exam, and vice versa. While this finding is to be expected in a typical instructional model, it is unexpected within TBPP, which is designed to match each student with content at his or her unique zone of proximal development; if every student is working on content that is at the exact right difficulty for him or her, then they should all be equally likely to be successful, regardless of their mathematical skills at the start of the year.

One potential root cause for this data trend could be the presence of significant peer effects in the LG, SG, and LTP methods, indicating that students typically score higher on exit slips when they are assigned to work alongside higher-performing students. This peer effect is particularly meaningful given that these are the three methods in which students have the most

opportunities to interact with other students in the course of learning. Given that the TBPP algorithm is explicitly designed to organize students into homogeneous groups, TBPP could be understood as a form of tracking that accelerates higher-performing students while denying lower-performing students the opportunity to learn from more mathematically capable peers (Barr & Dreeben, 1983; Lee & Ready, 2009; Philip & Olivares-Pasillas, 2016; Wenglinsky, 2005). The interaction effects between a student's Fall MAP score and the LG, SG, TI, and LTP methods lend further support for this theory. So does the heatmap in Figure 3 (p. 99), which seems to suggest that higher-performing students perform particularly well on exit slips when working within the LTP method.

A second root cause of the inequality in outcomes could be that the significant autonomy afforded to students by TBPP increases the importance of non-cognitive skills like motivation and grit, which could be more commonly found among higher-performing than lower-performing students. This would be in keeping with some of the extant literature related to on-task behavior in personalized learning environments, as well as the broader literature on non-cognitive skills and "success at school" factors in general (Baker and Gowda, 2010; Bowers, 2007, 2011; Brookhart et al., 2016; Duckworth, 2007; Murphy et al., 2014a; Rodrigo, Baker, Ryan, & Rossi, 2013). In other words, a student who is more diligent or cares more about education may score higher on the Fall MAP math exam, but may also be more motivated to work hard in student-directed methods regardless of his or her mathematical skill. The relevance of non-cognitive skills is supported by the fact that Fall MAP ELA scores are associated with exit slip performance with a statistically significant effect size roughly half that of Fall MAP math. This indicates that there is some underlying construct other than mathematical ability that is assessed by the MAP test and associated with exit slip performance.

Regardless of the cause, there does seem to be evidence of a "Matthew Effect" associated with TBPP in which students who enter the program as higher-performing experience greater daily success than students who enter the program as lower-performing (Merton, 1988). The Matthew Effect derives its name from a biblical verse from the Book of Matthew stating that "For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away." While the daily exit slip performance of students participating in TBPP notably does not appear to be associated with racial/ethnic background, gender, disability status, or English language learner status, there are significant differences based on students' Fall Math MAP scores, and to a lesser extent their Fall Reading MAP scores. I do not have access to data from a control group that would allow me to draw conclusions about whether TBPP increases inequality compared to traditional instruction. However, my findings do suggest that a rigorous program evaluation using a sophisticated method for causal inference such as comparative interrupted time series could be a fruitful avenue for future research (Bloom, 2003; Shadish, Cook, & Campbell, 2002).

In interpreting this finding, I should stipulate that inequality is not necessarily an unabashed evil if it is caused primarily by accelerating the growth of high-performing students. One of the key arguments in favor of technology-based personalization is that it allows curious, diligent, and intelligent students to race ahead and meet their full potential rather than languish bored in a class that moves too slowly for them. One could imagine a scenario in which TBPP promotes the growth of high-performing students in a way that expands inequality across students while having only very small negative effects on low-performing students, or even no negative effect at all. This is a classic example of the kind of value-laden trade-off that is endemic to both education and the social sciences more broadly (Labaree, 1997; Carnoy &

Levin, 1985; Stone, 2002). How should we weigh the importance of individual achievement vs. collective achievement, autonomy vs. equality, or high-performers meeting their full potential vs. low performers not being left behind? While quantitative analyses can provide useful evidence for evaluating the magnitude and direction of these trade-offs, the solutions will always involve philosophical questions that cannot be resolved through statistical analysis alone.

Usefulness of data methods. My fourth and final major finding is the overall usefulness of the methodological approaches used in my dissertation for exploring the broad, deep, and diverse data produced by personalized learning programs. The relationships between daily exit slip scores and end-of-year outcomes on the PARCC and NWEA MAP assessments suggest that exit slips are a useful measure of student learning, and that they are worthy of consideration for similar research in the future. Furthermore, the coherence and comprehensibility of my results suggest that exploring the relationships among diverse instructional variables and daily exit slip data through hierarchical linear modeling and hierarchical cluster analysis can yield meaningful insights into how the complex interactions among teachers, students, and content relate to variations in student learning.

This has significant implications even outside the context of technology-based personalization. Educational research is often limited by the difficulty of precisely associating instructional inputs with meaningful outcome measures; graduation rates and standardized test scores are the most commonly used metrics, but the fact that they are only gathered annually means that it can be difficult to disentangle causality among the myriad of complex factors that affect student learning. In contrast, personalized learning programs offer the ability to generate datasets in which daily student outcomes are integrated with program delivery in a way seldom encountered in educational research (Krumm et al., 2018; Natriello, 2012, 2013). The growing

prevalence of technology-based instructional models means that the pace of creation for these types of datasets is likely to accelerate in the future. This paper presents several innovative applications of established statistical techniques that would meaningfully aid researchers in exploring these new and very valuable datasets (Horn & Freeland Fisher, 2016; Natriello, 2012, 2013).

While hierarchical linear modeling is relatively common within educational research, it is most typically used to nest students within classes, classes within schools, or both (Raudenbush & Bryk, 2002; Means et al., 2010; Murphy et al, 2014a; Ready & Wright, 2011; Singer & Willett, 2003; Woltman et al., 2012; Wood et al., 2017). Nesting lessons within students represents a relatively novel application of this familiar quantitative technique. My use of hierarchical linear modeling in this paper is most similar to previous studies utilizing longitudinal data, since these studies also involve nesting multiple cases within individuals. However, the use of hierarchical linear modeling to explore the associations among multiple instructional variables and daily student outcomes represents an extension of this work to a new context (Singer & Willett, 2003).

Although some educational research over the last ten years has utilized hierarchical cluster analysis and data visualization heat maps, these techniques have not been commonly combined when studying schools or students (Bowers, 2007, 2010; Bowers, et al., 2016; Krumm et al., 2018). However, my dissertation reinforces previously published studies that argue for the usefulness of these descriptive techniques when exploring educational data. Hierarchical cluster analysis and heatmap data visualization offer several distinct advantages in comparison to regression-based statistical analyses. First, they perform well when exploring data that are multicollinear, interdependent, and nested, as is frequently the case in educational contexts.

Second, their visual nature makes them highly accessible to teachers, administrators, and policymakers; these techniques are sometimes described as "quantitative phenomenology" because they allow a rich description of individuals patterned in a way that enables us to see relationships and test hypotheses (Bowers, et al., 2017). Finally, their ability to reveal nuances across individual cases or groups of students makes them uniquely well suited to exploring data produced by personalized programs. Technology-based personalization offers the prospect of providing unique educational experiences custom-fitted to the needs of individual children. It seems fitting to analyze data from these programs using a technique that enables disaggregation at the student level rather than assuming a common effect, as is the case with regression analysis. The ability of data visualization heatmaps to enable a quick assessment of distributions, outliers, and clusters makes it a strong fit for studying educational models that are explicitly designed to create customized and non-standard student experiences.

Issues and Limitations

We should be mindful of several important limitations when interpreting the results of this study. The first is the significant diversity among instructional models utilizing technology-based personalization, which may limit the applicability of these findings to other contexts. The rapid pace of innovation among blended and personalized models means that there can be significant diversity in experiences across models, or even across schools or classrooms utilizing the same model (Brodersen & Melluzzo, 2017; Cavanagh, 2014; Horn & Staker, 2014; Murphy at al., 2014a; Pane et al., 2015; Picciano, 2014). While TBPP is typically described as one of the archetypical current examples of technology-based personalization, several of its key features are relatively unique, including the use of automated algorithms to make daily scheduling decisions for teachers, the inclusion of multiple instructional methods, and the highly comprehensive

nature of the program. It is unclear to what degree the findings from this study may be applicable to different technology-based personalized models, such as the Summit Personalized Learning Platform or the lab rotation model utilized by Rocketship schools (Childress & Amrofell, 2016; Horn & Staker, 2014; Osborne, 2016).

A second threat to the external validity of these findings is that they encompass only a single district and a single year of data. If there were some factor that made this district unique, or some reason that the 2015-16 academic year were different than a typical academic year, it could provide a bias that would reduce our ability to generalize these findings across other contexts. Of particular concern is the fact that I studied TBPP in its first year of implementation in this district. The radical differences between TBPP and traditionally organized instruction could create a significant learning curve in the first year of implementation, meaning that the 2015-16 academic year might not be representative of a typical year for the program. While there is some evidence for this type of "implementation dip" in the literature, it is worth noting that student performance did not seem to improve over the course of the 2015-16 academic year as students and teachers gained familiarity with the model (Murphy et al, 2014a; Rockoff, 2015).

A third issue was my decision to exclude the lessons for which exit slip data was missing, which may have created a bias that could interfere with the validity of the findings. While statistical comparisons of excluded and non-excluded lessons do not indicate any significant concerns, the varied and unknown causes for missing exit slips are still worth noting. This is also true of my decision to exclude the 3.6% of students who completed fewer than half of their exit slips. In contrast, my ability to draw upon data from all students enrolled in the program rather than taking a representative sample should increase the validity of the study, as well as increase the overall power of the statistical analyses.

Finally, the short, multiple-choice format of exit slips means that they are more likely to evaluate procedural and didactic skills than more complex skills related to theoretical understanding or evaluation. This may create a bias in my results if TBPP's ability to build those deeper skills were uncorrelated with its ability to build the procedural skills that exit slips are designed to assess. While this represents a limitation in my study, it may also represent a limitation within the TBPP model itself. The TBPP algorithm uses exit slips and the NWEA MAP math assessment as proxies for learning. However, both of these assessments are composed entirely of multiple choice questions, and students are not required to engage in collaboration, argumentation, or oral or written communication to complete them. Researchers like Richard Elmore and Deborah Ball have argued throughout their careers that "knowing math" means more than just getting the right answer or understanding relevant procedural rules, but also knowing why a rule is true and how it connects with other big mathematical ideas. Unfortunately, the assessment measures used by TBPP may be inadequate to fully assess those essential competencies.

Implications for Future Research

This study suggests several valuable avenues for future research. The most straightforward of these is to broaden my dataset to include data from the implementation of TBPP in other districts, or within this same district across multiple years. Expanding the scope of the research in this way could help to address some of the concerns related to external validity that arise when studying a program in only a single specific context. Similarly, it would be very useful to apply the analytic techniques from this dissertation to other programs utilizing technology-based personalization. Because the data from other programs is probably structured differently, it seems unlikely that the data could or should be pooled. However, it would be very

useful to apply similar analytic techniques and research questions to data produced by alternate technology-based personalized programs in order to explore whether the key findings from this study are also true in those contexts.

A second avenue for future research could be to complement the quantitative research utilized in this study with qualitative research, including classroom observation and interviews with teachers or students. I suggest in this paper that TBPP may accelerate inequality by enabling motivated or high-performing students to race ahead of their lower-performing peers; interviews with those students could help confirm or refute those findings. Similarly, observing lower-performing students when working within methods that provide a high degree of autonomy could illuminate whether their comparatively low performance on exit slips is attributable to off-task behavior or authentic struggles with math content. In addition, my theory that TBPP inhibits teachers' ability to build and exercise pedagogical content knowledge could be confirmed or undermined by interviewing teachers about their experience with the program, and particularly by asking whether they feel that TBPP affects their ability to effectively prepare for instruction when compared to more traditional forms of schooling. Given that TBPP relieves teachers from needing to complete many of the traditional tasks of teaching, such assessing and grouping students, it could also be useful to explore teachers' perceptions of how TBPP affects their workload and the overall sustainability of their jobs. It would also be very interesting to collect hard data on the attrition of teachers utilizing TBPP compared to traditional instruction to explore whether reducing teachers' scope of work affects their likelihood of departing from the profession.

Future research could also more deeply explore the relationships among student learning and various teacher characteristics, including teachers' pedagogical content knowledge, past

experiences, and training. For example, it would be very interesting to explore whether a former 5th grade teacher is more effective when teaching 5th grade skills than 9th grade Algebra content on which he or she has never been formally trained. Were this intuitive finding to be borne out in the data, it would provide further evidence for the importance of pedagogical content knowledge as a key determinant of students' learning outcomes. Similarly, it would be useful to explore whether there are consistent differences across teachers in their effectiveness when teaching particular mathematical content areas, such as algebra, geometry, or ratios and proportions, and whether those differences were correlated with teacher experience or interest. Were the data to indicate significant teacher-level variety in effectiveness based on content area, it would suggest that TBPP would be more effective were its algorithm to take into account teachers' unique abilities when generating daily instructional assignments.

Finally, future research could attempt to apply additional methods from the fields of learning analytics and educational data mining to confirm, add nuance to, or expand upon the findings from this study. The last two decades have witnessed an explosion in new techniques for exploring "big data," many of which have been applied to the field of education through the parallel fields of Educational Data Mining and Learning Analytics (Agasisti & Bowers, 2017; Bowers, 2017; Siemens & Baker, 2012). Although techniques like Bayesian Knowledge Tracing, Correlation Mining, Association Rule Mining, and Sequential Pattern Mining are beyond the purview of this paper, they and techniques like them could represent a useful extension of the work that I have undertaken in this dissertation (Baker, 2015; San Pedro et al, 2013; Snow et al., 2016).

Policymakers and researchers are eager to explore the outcomes from instructional models utilizing technology-based personalization. However, they risk missing important data

trends if they limit their exploration to end-of-year outcomes on state-mandated standardized assessments. This paper demonstrates the usefulness of also investigating the student- and lesson-level factors that affect learning at a daily level. Continuing this avenue of research may generate insights into not only technology-based personalization, but the phenomenon of teaching and learning more broadly.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and Student Achievement in Chicago Public High Schools. Journal of Labor Economics 24(1): 95-135.

Agasisti, T., Bowers, A.J. (2017) Data Analytics and Decision-Making in Education: Towards the Educational Data Scientist as a Key Actor in Schools and Higher Education Institutions. In Johnes, G., Johnes, J., Agasisti, T., López-Torres, L. (Eds.) Handbook of Contemporary Education Economics (184-210). Cheltenham, UK: Edward Elgar Publishing. http://www.e-elgar.com/shop/handbook-of-contemporary-education-economics.

Allison, P. D. (2002). Missing data. Thousand Oaks, CA: Sage.

Arnett, T. (2016). *Teaching in the Machine Age: How innovation can make bad teachers good and good teachers better*. Christensen Institute.

Baker, R. S. (2015). Big data and education. *Online EdX course.* http://www.columbia.edu/~rsb2162/bigdataeducation.html

Baker, R. S. J. d., & Gowda, S.M. (2010). An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools. *Proceedings of the 3rd International Conference on Educational Data Mining*, 11-20.

Ball, D. L. (1990). The Mathematical Understandings That Prospective Teachers Bring to Teacher Education. *The Elementary School Journal*, *90*(4), 449–466.

Ball, D. L. (1997). From the General to the Particular: Knowing Our Own Students as Learners of Mathematics. *Mathematics Teacher*, *90*(9), 732–737.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. *Teaching as the Learning Profession: Handbook of Policy and Practice*, *1*, 3–22.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education*, *59*(5), 389–407.

Barr, R. & Dreeben, R. (1983). *How schools work*. Chicago: University of Chicago Press.

Barrow, L., Markman, L., and Rouse, C. (2007). Technology's Edge: The Educational Benefits of Computer Aided Instruction. NBER Draft.

Beteille, T. & Loeb, S. (2009) "Teacher Quality and Teacher Labor Markets," In G. Sykes, B. Schneider, & D. N. Plank, eds., *Handbook of Education Policy Research* (New York: Routledge), pp. 596-612.

Bidwell, Charles. (1965). The school as a formal organization. In James G. March (Ed.), *Handbook of Organizations* (pp. 972-1018). Chicago: Rand McNally.

Bloom, B. S. (1984). "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." Educational Researcher 13(6): 4-16.

Bloom, H.S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27*(1), 3-49.

Brodersen, R. M., & Melluzzo, D. (2017). Summary of Research on Online and Blended Learning Programs That Offer Differentiated Learning Options. REL 2017-228. *Regional Educational Laboratory Central*.

Bowers, A. J. (2007). Grades and Data Driven Decision Making: Issues of Variance and Student Patterns. *Online Submission*.

Bowers, A. J. (2010). Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping out and Hierarchical Cluster Analysis. *Practical Assessment, Research & Evaluation*, *15*(7).

Bowers, A.J. (2011) What's in a Grade? The Multidimensional Nature of What Teacher Assigned Grades Assess in High School. *Educational Research & Evaluation*, 17(3), 141-159. doi: 10.1080/13803611.2011.597112 (Preprint available)

Bowers, Alex (2015). Using Big Data to Investigate Longitudinal Education Outcomes through Visual Analytics. National Science Foundation. Grant #1546652

Bowers, A.J. (2017) Quantitative Research Methods Training in Education Leadership and Administration Preparation Programs as Disciplined Inquiry for Building School Improvement Capacity. *Journal of Research on Leadership Education,* 12(1), p. 72-96. http://doi.org/10.1177/1942775116659462 (Preprint available)

Bowers, A. J., et al. (2016). "Building a Data Analytics Partnership to Inform School Leadership Evidence-Based Improvement Cycles." *Annual meeting of the American Educational Research Association, Washington, DC*.

Bowers, A.J., Blitz, M., Modeste, M., Salisbury, J., Halverson, R. (2017) How Leaders Agree with Teachers in Schools on Measures of Leadership Practice: A Two-Level Latent Class Analysis of the Comprehensive Assessment of Leadership for Learning. *Teachers College Record*, 119(4). http://www.tcrecord.org/Content.asp?ContentId=21677

Bowers A.J., Zhao, Y., (2018) Cluster analysis heatmap R code. Personal communication.

Brookhart, S., Guskey, T., Bowers, A.J., McMillan, J. Smith, L. Smith, J., Welsh, M. (2016) One Hundred Years of Grading Research: Meaning and Value in the Most Common

Educational Measure. *Review of Educational Research*, 86(4), p. 803-848 http://doi.org/10.3102/0034654316672069.

Brown, E. (2012, October 14). D.C. students test "Teach to One" learning system. *Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/dc-students-test-teach-to-one-learning-system/2012/10/14/9f945470-149b-11e2-be82-c3411b7680a9_story.html?utm_term=.6201a9e1e17e

Burns, Tom, & Stalker, G. M. (1961). *The management of innovation.* London: Tavistock Publications. Chapter 6: Mechanistic and organic systems of management (pgs. 96-125).

Carnoy, M., & Levin, H. (1985). *Schooling and work in the democratic state*. Stanford University Press.

Cavanagh, Sean (2014, October 20). What is 'Personalized Learning'? Educators Seek Clarity. *Education Week*. Available at http://www.edweek.org/ew/articles/2014/10/22/09pl-overview.h34.html

Center for Education Policy at Harvard University. (2016). DreamBox Learning Achievement Growth in the Howard County Public School System and Rocketship Education.

Center for Research on Education Outcomes (2015), *Urban Charter School Study Report on 41 Regions*. Stanford, CA.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (No. w17699). National Bureau of Economic Research.

Childress, S., & Amrofell, M. (2016). *Reimagining Learning: A Big Bet on the Future of American Education*. NewSchools Venture Fund.

Christensen, Clayton (2013). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Christensen, C. M., Horn, M. B., & Johnson, C. W. (2008). *Disrupting class : how disruptive innovation will change the way the world learns*. New York: McGraw-Hill.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates.

Cohen, D. K. (1990). A Revolution in One Classroom: The Case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, *12*(3), 311–329.

Cohen, D. K., & Bhatt, M. P. (2012). The importance of infrastructure development to high-quality literacy instruction. *The Future of Children, 22*(2), 117-138.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119-142.

Cordes, C. & Miller, E. (Eds.) (2000). *Fool's Gold: A critical look at computers in childhood.* College Park, MD: Alliance for Childhood.

Corno, L. (2008). On Teaching Adaptively. *Educational Psychologist*, *43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Costa, I. G., de Carvalho, F. A., & de Souto, M. C. (2002). Comparative study on proximity indices for cluster analysis of gene expression time series. *Journal of Intelligent & Fuzzy Systems*, *13*(2-4), 133-142.

Creswell, John W., and Vicki L. Plano Clark. 2011. *Designing and Conducting Mixed Methods Research*. SAGE.

Cross, C. T. (2004). Political education: National policy comes of age. New York, NY: Teachers College Press.

Cuban, Larry (1986). Teachers and machines: The classroom use of technology since 1920. New York: Teachers College Press.

Cuban, L. (1990). Reforming Again, Again, and Again. *Educational Researcher*, *19*(1), 3–13.

Cuban, L. (1993). How teachers taught: Constancy and change in American classrooms, 1890-1990 (2nd ed.). New York: Teachers College Press.

Delisle, James R. (2015, January 6). Differentiation Doesn't Work. *Education Week*. Available at https://www.edweek.org/ew/articles/2015/01/07/differentiation-doesnt-work.html

Dembo, M. H., & Howard, K. (2007). Advice about the use of learning styles: A major myth in education. *Journal of college reading and learning*, *37*(2), 101-109.

Dewey, J. (1916). Democracy and Education: An Introduction to the Philosophy of Education.

Diamond, J. B. (2007). Where the Rubber Meets the Road: Rethinking the Connection Between High-Stakes Testing Policy and Classroom Instruction. *Sociology of Education, 80*(4), 285-313.

DiMaggio, P.J., & Powell, W.W. (1983). "The iron cage revisited: Institutional isomoprphism and collective rationality in organizational fields." *American Sociological Review, 48*, 147-160.

Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning* (p. 29). ACM.

Dobbie, W., & Fryer, R. G., Jr. (2011). Getting beneath the Veil of Effective Schools: Evidence from New York City. NBER Working Paper No. 17632. *National Bureau of Economic Research U6*.

Duckworth, Angela L., et al. "Grit: perseverance and passion for long-term goals." *Journal of personality and social psychology* 92.6 (2007): 1087.

Education Week. (2014, October 22). Personalized Learning: A Working Definition. Retreived from http://www.edweek.org/ew/collections/personalized-learning-special-report-2014/a-working-definition.html.

Eisen, M. B., & DeHoon, M. (2002). Cluster 3.0 manual. Palo Alto, CA: Stanford University.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95, 14863-14868.

Elmore, R. (1993). "School Decentralization: Who gains? Who loses?" pp. 33-54 in Decentralization and School Improvement

Elmore, R. F. (1996). "Getting to scale with good educational practice." *Harvard Educational Review, 66*(1), 1-26.

Elmore, R. & Birney, D. (1997). "Investing in teacher learning." National Commission on Teaching and America's Future.

Elmore, R. (2006). The problem of capacity in the (re)design of educational accountability systems. Paper presented at the conference, Examining America's Commitment to Closing Achievement Gaps, Teachers College, Columbia University, November, 2006.

Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). Instructional rounds in education.

Elmore, Richard, (2010). "Leading the Instructional Core: an interview with Richard Elmore". Ontario Ministry of Education. In Conversations. Summer, Volume 11, Issue 3. Retrieved from http://www.edu.gov.on.ca/eng/policyfunding/leadership/Summer2010.pdf

Farkas, S., & Duffett, A. (2010). Cracks in the Ivory Tower? The Views of Education Professors Circa 2010. *Thomas B. Fordham Institute*.

Farkas, S., Duffett, A., & Lovelace, T. (2008). High-achieving students in the era of NCLB. *Washington, DC: Thomas B. Fordham Institute. Retreived from: http://www. edexcellence. net/publications*, 735-738.

Freeland, Julie & Hernandez, Alex (2014). Schools and Software: What's Now and What's Next. The Clayton Christensen Institute for Disruptive Innovation.

Gardner, H.,(2011). *Frames of mind: The theory of multiple intelligences* (3;3rd; ed.). New York: Basic Books.

Gewertz, C. (2016, September 14). 10 High School Redesign Projects Win $100 Million in "XQ Super School" Contest. *Education Week*.

Graham, J. W. (2009). Missing data analysis: Making it work in the real word. Annual Review of Psychology, 60, 549–576.

Grouws, D. A., & Cebulla, K. (2000). *Improving student achievement in mathematics*. International Academy of Education.

Hardy, M. E. (2004). Use and evaluation of the ALEKS interactive tutoring system. *Journal of Computing Sciences in Colleges*, *19*(4), 342-347.

Henderson, John D., and Paul C. Nutt. 1978. "On the design of planning information systems." Academy of Management Review, 3: 774-785

Herold, Benjamin. (2016a, March 9). Facebook's Zuckerberg to Bet Big on Personalized Learning Education Week.

Herold, B. (2016b, Novembe 3). A Virtual Mess: Inside Colorado's Largest Online Charter School. Education Week.

Hess, F. M., & Brookings Institution, W., DC. (1999). Spinning wheels: The politics of urban school reform. Washington, D.C: Brookings Institution Press.

Hew, K.F., & Brush, T. (2007). Integrating technology into K-12 teaching and learning: Current knowledge gaps and recommendations for future research. Educational Technology Research and Development, (55). 223-252.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hollands, F. M. (2003). The impact *of computer use on the individualization of students' learning experiences in public middle school science classrooms*

Horn, M. (2013, November 21). Teach To One Earns Promising Marks In Math Learning. *Forbes*. Retrieved from http://www.forbes.com/sites/michaelhorn/2013/11/21/teach-to-one-earns-promising-marks-in-math-learning/#5ca867b7250a

Horn, M. B., & Staker, H. (2014). *Blended: Using disruptive innovation to improve schools*. John Wiley & Sons.

Horn, M. B., & Freeland Fisher, J. (2016). *A blueprint for breakthroughs: Federally funded education research in 2016 and beyond*. Clayton Christensen Institute for Disruptive Innovation. Retrieved from http://www.christenseninstitute.org/publications/a-blueprint-for-breakthroughs/

Horn, M. B. (2017, January 3). Finding "Personalized Learning" and Other Edtech Buzzwords on the Gartner Hype Cycle. *EdSurge News*.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

L. Huerta & A. Zuckerman, "An Institutional Theory Analysis of Charter Schools: Addressing Institutional Challenges to Scale" Peabody Journal of Education Vol. 84, Iss. 3, 2009

Huerta, L., Shafer, S. R., Barbour, M. K., Miron, G., & Gulosino, C. (2015). Virtual Schools in the US 2015: Politics, Performance, Policy, and Research Evidence. *National Education Policy Center*.

Hyslop, Anne & Mead, Sara (2015). A Path to the Future: Creating Accountability for Personalized Learning

Jacob, A., & McGovern, K. (2015). *The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development*. TNTP.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Januszewski, A. (2001). *Educational technology: The development of a concept*. Libraries Unlimited.

Jaskowiak, P. A., Campello, R. J., & Costa, I. G. (2014). On the selection of appropriate distances for gene expression data clustering. *BMC bioinformatics*, *15*(2), S2.

Jorion, N., Roberts, J., Bowers, A.J., Tissenbaum, M., Lyons, L., Kuma, V., Berland, M. (2018) Uncovering Patterns in Constructionist Collaborative Learning Activities via Cluster Analysis of Museum Exhibit Log Files. A paper presented at the annual meeting of the American Educational Research Association (AERA), New York, NY: April 2018.

Keefe, J. W., & Jenkins, J. M. (2000). *Personalized instruction: Changing classroom practice*. Eye on Education.

Klein, A. (2016, November 28). Final ESSA Accountability Rules Boost State Flexibility in Key Areas. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2016/11/ed_dept_releases_final_account.html

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, *111*(468), 1-28.

Krumm, A., Means, B., & Bienkowski, M. (2018). *Learning Analytics Goes to School: A Collaborative Approach to Improving Education*. Routledge.

Labaree, D.F. (1997) Public Goods, Private Goods: The American Struggle over Educational Goals. *American Educational Research Journal, 34*(1), 39-81

Lee, V. E., & Ready, D. D. (2009). U.S. High School Curriculum: Three Phases of Contemporary Research and Reform. *The Future of Children*, *19*(1), 135–156.

Lee, J., Recker, M., Bowers, A.J., Yuan, M. (2016). Hierarchical Cluster Analysis Heatmaps and Pattern Analysis: An Approach for Visualizing Learning Management System Interaction Data. A poster presented at the annual International Conference on Educational Data Mining (EDM), Raleigh, NC: June 2016.

Lepore, J. (2014). The disruption machine: what the gospel of innovation gets wrong The New Yorker 23 June.

Lewis-Beck, M. 1980. *Applied Regression: An Introduction.* Beverly Hills: Sage.

Lincoln, Yvonna S., and Egon G. Guba. *Naturalistic inquiry*. Vol. 75. Sage, 1985.

Mangiameli, P., Chen, S. K., & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, *93*(2), 402-417.

Manna, Paul, Collision Course: Federal Education Policy Meets State and Local Realities, CQ Press, 2011

Martinez, J. F., Stecher, B., & Borko, H. (2009). Classroom Assessment Practices, Teacher Judgments, and Student Achievement in Mathematics: Evidence from the ECLS. *Educational Assessment*, (14), 78–102.

Matthews, L. E. (2003). Babies Overboard! The Complexities of Incorporating Culturally Relevant Teaching into Mathematics Instruction. *Educational Studies in Mathematics*, *53*(1), 61–82.

McDonald, R. A. F. (1915). *Adjustment of school organization to various population groups* (No. 75). Teachers college, Columbia university.

McDonnell, L., & Elmore, R. (1987). "Getting the job done: Alternative policy instruments." *Educational Evaluation and Policy Analysis, 9*(2), 139-152.

Mead, Sara, Mitchel, Ashley LiBetti, and Rotherham, Andrew J. (2015) *The State of the Charter School Movement*. Bellwether Education Partners.

Means, B., Toyama,Y., Murphy. R., Bakia, M., and Jones, K. (2010). Evaluation of evidence based practices in online learning: A meta-analysis and review of online-learning studies. Washington, D.C.: U.S. Department of Education.

Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *isis*, *79*(4), 606-623.

Meyer & Rowan (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*.

Meyer, J. W., & Rowan, B. (1978). "The structure of educational organizations." In M. W. Meyer (Ed.), *Environments and organizations* (pp. 78-109). San Francisco: Jossey-Bass.

Miller, L., Gross, B., Maas, T., Hernandez, J., & Lu, A. (2016). *Financing Personalized Learning: What Can We Learn from First-Generation Adopters?* Center on Reinventing Public Education.

Miron, G., Horvitz, B., Gulosino, C., Huerta, L., Rice, J. K., Shafer, S. R., & Cuban, L. (2013). Virtual Schools in the US 2013: Politics, Performance, Policy, and Research Evidence. *National Education Policy Center*.

Molnar, A., Huerta, L., Rice, J. K., Shafer, S. R., Barbour, M. K., & Miron, G. (2014). Virtual Schools in the US 2014: Politics, Performance. *National Education Policy Center*.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The future of children*, 113-127.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2016). *Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India.*

Murphy, E., Snow, E., Mislevy, J., Gallagher, L., Krumm, A., & Wei, Xin (2014a). Blended Learning Report, *Michael & Susan Dell Foundation*

Murphy, R., Gallagher, L., Krumm, A., Mislevy, J., & Hafter, A. (2014b). *Research on the Use of Khan Academy in Schools*. SRI Education.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education (IJAIED)*, *10*, 98-129.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. US Department of Education.

Natriello, Gary (2012). "Adaptive Educational Technologies and Educational Research: Opportunities, Analyses, and Infrastructure Needs." Background Paper Prepared for the National Academy of Education

Natriello, Gary (2013). "Adaptive Educational Technologies: Tools for Learning and for Learning About Learning." National Academy of Education, Washington, DC

New Classrooms (2017). "How it Works." Retrieved from https://www.newclassrooms.org/how-it-works/

Newcomb, T. (2016, June 16). Teach to One: Inside the Personalized Learning Program That Bill Gates Calls the "Future of Math." Retrieved from https://www.the74million.org/article/teach-to-one-inside-the-personalized-learning-program-that-bill-gates-calls-the-future-of-math

Northwest Evaluation Association (2015). "2015 NWEA Measures of Academic Progress Normative Data." Retrieved from https://www.nwea.org/content/uploads/2015/06/2015-MAP-Normative-Data-AUG15.pdf

Oakes, J. (2005). *Keeping track: how schools structure inequality* (Vol. 2nd). New Haven, Conn: Yale University Press.

Osborne, D. (2016). *Schools of the Future: California's Summit Public Schools*. Washington, DC: Progressive Policy Institute.

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stluka, M. F. (1994). Ability-Group Effects: Instructional, Social, or Institutional? *Sociology of Education*, *67*(1), 27–46.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., Karam, R., Daugherty, R., & Phillips, A. (2013). Does an Algebra Course with Tutoring Software Improve Student Learning? Santa Monica, CA: Rand Corporation. Retrieved from http://www.rand.org/pubs/research_briefs/RB9746.html

Pane, John F., Elizabeth D. Steiner, Matthew D. Baird and Laura S. Hamilton. (2015). Continued Progress: Promising Evidence on Personalized Learning. Santa Monica, CA: RAND Corporation, http://www.rand.org/pubs/research_reports/RR1365.html.

Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). Informing Progress. Santa Monica, CA: Rand Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR2042.html

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological science in the public interest*, *9*(3), 105-119.

Petrilli, M. J. (2012). *The diverse schools dilemma: A parent's guide to socioeconomically mixed public schools*. Thomas B. Fordham Institute.

Philip, Thomas & Olivares-Pasillas, Maria C., (2016). Learning Technologies and Educational Equity: Charting Alternatives to the Troubling Pattern of Big Promises with Dismal Results. *Teachers College Record*, ID Number: 21616

Picciano, A. G. (2014). Big data and learning analytics in blended learning environments: Benefits and concerns. *IJIMAI*, *2*(7), 35-43.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternate to the fixed effects model for studying time-varying treatments in school settings. *Education Finance and Policy*, *4*(4), 468-491.

Ravitch, D. (2010). The Death and Life of the Great American School System: How Testing and Choice are Undermining Education. New York: Basic Books.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal*, *48*(2), 335-360.

Ready, D., Meier, E., Horton, D., Mineo, C., & Pike, J. Y. (2013). Student Mathematics Performance in Year One Implementation of Teach to One: Math: Center for Technology & School Change, Teachers College Columbia University.

Ready, D., (2014). Student Mathematics Performance in the First Two Years of Teach to One: Math, Teachers College Columbia University.

Ready, D., Conn, K., Nitkin, D., & Shalev, S. (2017). Year-One Impact Results from the i3 Implementation of Teach to One: Math. A paper presented at the annual meeting of the American Education Research Association (AERA), San Antonio, TX: April 2017

Rockoff, J. (2015). *Evaluation Report on the School of One i3 Expansion*. New York, Columbia University Business School.

Rodrigo, M.M. T., Baker, Ryan S. J. D, & Rossi, L. (2013) Student off-task behavior in computer-based learning in the Philippines; Comparison to prior research in the USA. Teachers College Record, 115(10), 1.

Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.

San Pedro, M. O., Baker, R. S. J. d., Bowers, A., & Heffernan, N. (2013). Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. Presented at the The 6th International Conference on Educational Data Mining, Memphis, Tennessee.

Scardamalia, M. & Bereiter, C. (2001). Getting real about 21st century education. *The Journal of Educational Change (2)*

Seltzer, M. H. (1995). Furthering our understanding of the effects of educational programs via a slopes-as-outcomes framework. *Educational Evaluation and Policy Analysis*, *17*(3), 295-304.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Siemens, G., & d Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.

Singer, Natasha (2017, June 6). The Silicon Valley Billionaires Remaking America's Schools. *The New York Times*. Available at https://www.nytimes.com/2017/06/06/technology/tech-billionaires-education-zuckerberg-facebook-hastings.html?_r=0

Snow, R., & Swanson, J. (1992). Instructional Psychology: Aptitude, Adaptation, And Assessment. *Annual Review of Psychology*, *43*(1), 583–626.

Snow, E., Krumm, A., Bowers, A. J., Podkul, T., & Feng, M. (2016). Quantifying How Students Use an Online Learning System: A Focus on Transitions and Performance. Presented at the International Conference on Educational Data Mining (EDM), Raleigh, NC.

Stein, M. K. (2001). Take Time for Action: Mathematical Argumentation: Putting Umph into Classroom Diccussions. *Mathematics Teaching in the Middle School*, *7*(2), 110–112.

Stone, D. (2002). Policy paradox: The art of political decision making, revised edition. *London and New York, NY: WW Norton and Company*.

TNTP. (2014). Reimagining teaching in a blended classroom [Working paper]. Brooklyn, NY: Author. Retrieved August 31, 2016 from http://tntp.org/assets/documents/TNTP_Blended_Learning_WorkingPaper_2014.pdf

Tomlinson, C. A. (2001). *How to differentiate instruction in mixed-ability classrooms*. ASCD.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., ... & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted*, *27*(2-3), 119-145.

Tyack (1991). Public School Reform: Policy Talk and Institutional Practice, *American Journal of Education*

Tyack, D.B. & Tobin, W. (1994, Autumn). The "Grammar" of Schooling: Why Has It Been So Hard to Change? (in Social and Institutional Analysis; School Reform)

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

Van de Ven, A., and A. Delbecq. "A Task Contingent Model of Work Unit Structures," Administrative Science Quarterly, Vol. 19, No. 2 (June 1974), 183-197.

VanLehn, Kurt, "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems," *Educational Psychologist*, Vol. 46, No. 4, 2011, pp. 197–221.

Vander Ark, T. (2017, January 11). Teach to One: Inventing the Future of Math Learning. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/on_innovation/2017/01/getting_smart_podcast_teach_to_one_inventing_the_future_of_math_learning.html

van'tVeer, L. J., Dai, H., vandeVijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530-536.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: the development of higher psychological processes*. Cambridge, Mass: Harvard University Press.

Wang, H., & Woodworth, K. (2011). *Evaluation of Rocketship Education's Use of DreamBox Learning's Online Mathematics Program*. Menlo Park, California: SRI International.

Wendt, S., & Rice, J. (2013). *Evaluation of ST Math in the Los Angeles Unified School District.* San Francisco: WestEd.

Weick, K. (1976). Educational organizations as loosely coupled systems, *Administrative Science Quarterly* 21(1), 1-19.

Wenglinsky, H. (2005). *Using technology wisely: The keys to success in schools*. New York: Teachers College Press.

Wolf, M. A. (2010). *Innovate to Educate: System [Re]design for Personalized Learning: A Report from the 2010 Symposium*. Software & Information Industry Association.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69.

Wood, L., Kiperman, S., Esch, R. C., Leroux, A. J., & Truscott, S. D. (2017). Predicting dropout using student-and school-level factors: An ecological perspective. *School Psychology Quarterly*, *32*(1), 35.

Woodworth, J. L., Raymond, M. E., Chirbas, K., Gonzales, M., Negassi, Y., Snow, W., & Van Dongle, C. (2015). *Online charter school study*. Stanford, CA: Center for Research on Education Outcomes. Retrieved August 30, 2016 from https://credo.stanford.edu/pdfs/OnlineCharterStudyFinal2015.pdf

Yazdani, M. (1987, July). Intelligent tutoring systems: an overview. In *Artificial intelligence and education* (Vol. 1, pp. 183-201). Ablex Norwood, NJ.

Yeung, K. Y., & Ruzzo, W. L. (2001). An empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, *17*(9), 763-774.

Appendix

Chart 4: Percentage of Lessons with Completed Exit Slips by Student

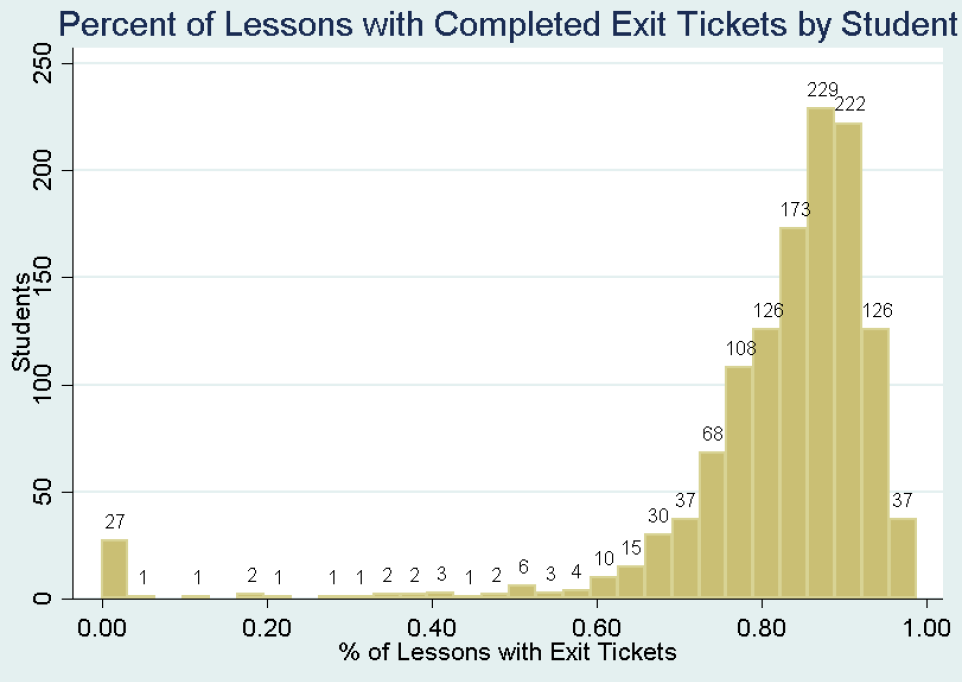


Chart 5: Distribution of Standardized Exit Slip Scores

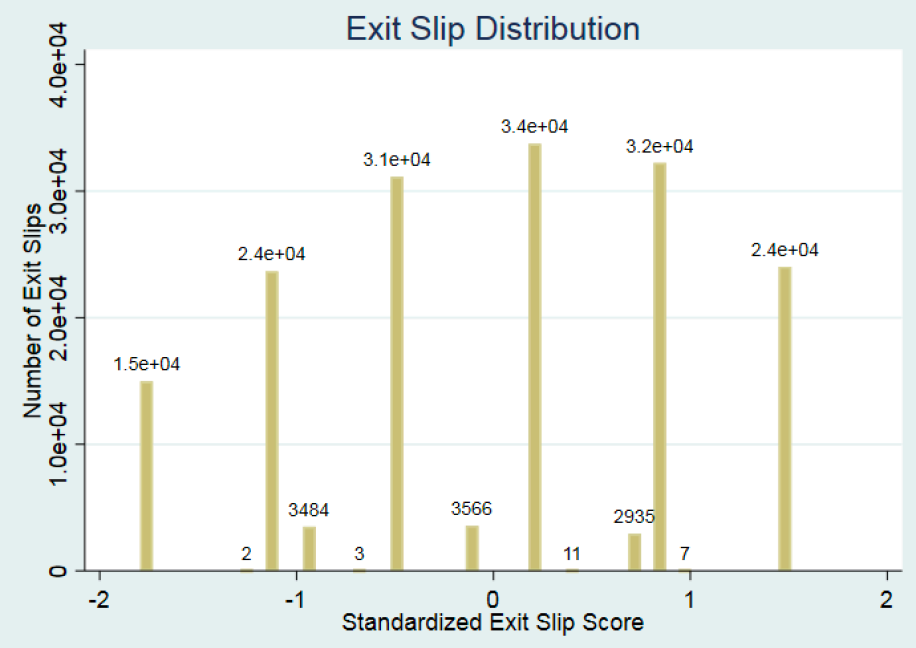


Chart 6: Distribution of Standardized Mean Group MAP Scores

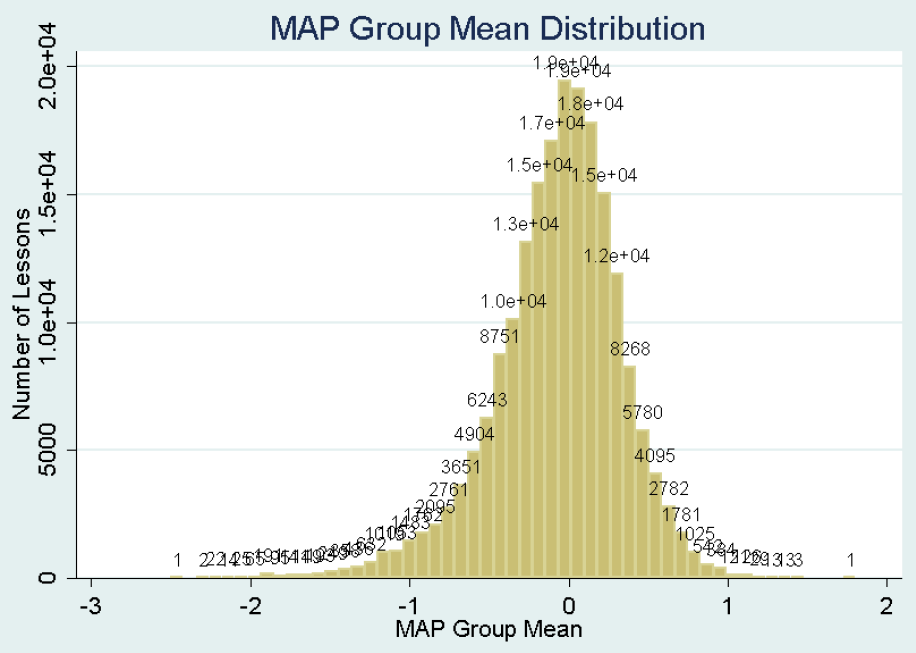


Chart 7: Distribution of Content Gap

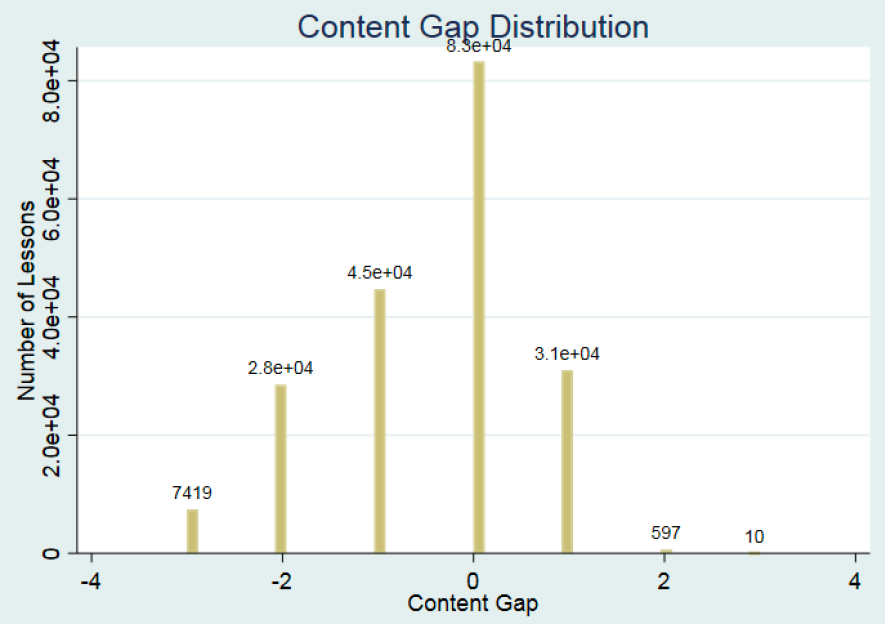


Chart 8: Distribution of Standardized Fall MAP Score

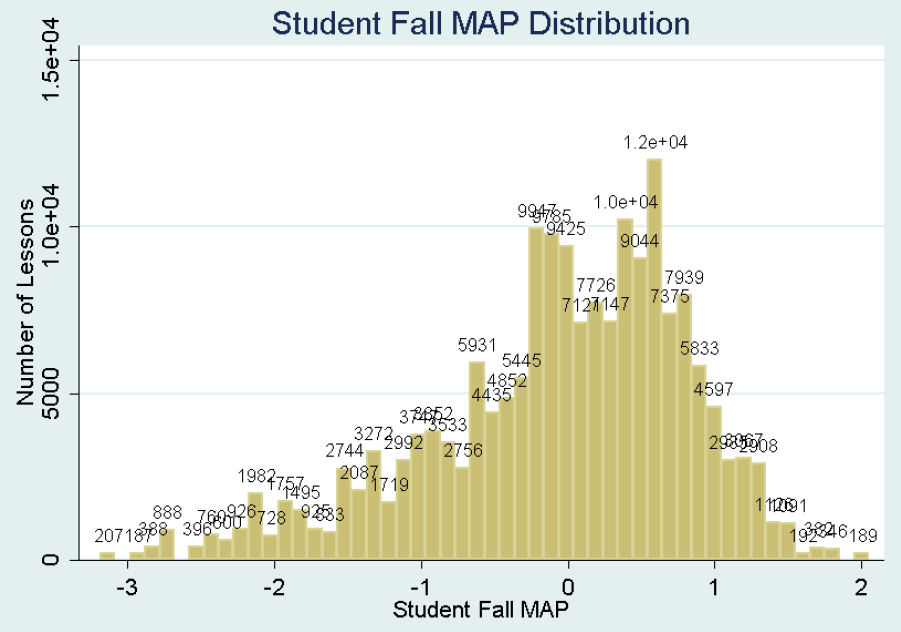


Chart 9: Distribution of Centered Group Size