Teaching with a Fully Digital, Year-long Math Program: Learning Science Futures on the Front Line

Jeremy Roschelle, Digital Promise, jroschelle@digitalpromise.org
Phillip Herman, McREL, pherman@mcrel.org
Karen Bumgardner, McREL, kbumgardner@mcrel.org
Nicole Shechtman, SRI International, nicole.shechtman@sri.com
Mingyu Feng, SRI International, mingyu.feng@sri.com

Abstract: In the context of a large-scale randomized controlled trial, our team investigated "front line" teaching issues as schools implemented a fully digital, blended learning curriculum in mathematics. This paper focuses on observations of instruction within schools that were assigned to use the new digital resources. Compared to a business-as-usual control group, classroom activity and teaching practices changed in the treatment group. Observers, who were blind to student achievement outcomes, found two overall patterns in treatment classrooms across five categories of observations. Later quantitative analysis indeed found the "high" and "low" patterns could account for some of the variance in achievement outcomes within the treatment condition. We explore observed patterns in terms of existing learning science theory and suggest areas where further development of the learning sciences may be needed and how learning sciences can contribute to improvement of digital, blended learning environments.

Introduction

As indicated in this year's conference theme, AI and automation are changing the nature of classrooms as workplaces for teaching and learning. As these changes occur, new complexities arise for learning scientists who study classrooms. As the work of teaching and learning becomes distributed across teacher and technology, learning scientists may need to change or refine understandings of effective teaching and learning processes. The conference theme draws attention to the "imperative to guide commercial development," as well as the need to understand different cultural and educational contexts. Our research investigated a commercially-available digital curriculum and we worked closely with the product team to understand the lessons of the study for improvement. We also worked in schools in West Virginia, a state with a distinctive regional culture. This paper discusses how the learning sciences may need to evolve in order to be responsive to the vision in the conference theme.

We conducted a randomized controlled trial (RCT) aimed at measuring the efficacy of a new digital mathematics curriculum, engaging 46 schools and approximately 2000 students in our research. Overall, we found that the vision and challenges described in the conference program to ring true: AI is changing the classroom workplace. We observed the classrooms in both the treatment condition (TC), which was using the new digital materials and blended learning approach, and in the control condition (CC), which was using "business-as-usual" materials and approaches. As we will later describe, we observed broad differences in the structure of classrooms, such as the predominance of instructor-led mathematics teaching (CC) versus individual student work at computers (TC). For students, the "work of learning" in TC classrooms was also different. For example, TC classrooms emphasized independent learning strategies but CC classrooms did not. For teachers, the change in the role was quite extensive. For example, teachers were no longer the main providers of instruction. Further, teachers in TC classrooms were more likely to use data during class to make instructional decisions; TC teachers spent much time intervening with particular students based on data reports (Singleton et, 2018).

This paper's investigation of this new AI-rich classroom workplace focuses on classroom observations but leverages data from the larger RCT as well. The RCT engaged a large number of schools and our team of observers collected data in all 23 TC schools and with all 38 TC teachers. Thus, we have the opportunity to look at *systematic patterns* of teaching and learning that were emergent across many teachers and schools, whereas many learning science studies only look within a handful of classrooms. Another advantage of the RCT context is that we collected achievement outcome data for all schools, and can control for prior achievement in our analyses. This allows us to ask: controlling for prior knowledge, do the systematic patterns we observed in different TC classrooms predict differential student outcomes in those classrooms? Thus, we can examine whether there is evidence that the observed patterns might be *consequential.*

To achieve the kinds of positive impacts envisioned in the conference program, learning scientists need to know if their theories are a good match to what goes on in new digital, blended learning classroom environments. We argue that the identification of *systematic patterns* that are *consequential* can be a good guide to where the learning sciences, if further developed, could have stronger impacts with regard to new AI-based teaching and learning workplaces, guiding commercial products, and working within specific cultural settings. Within the broad frame of learner-, knowledge-, assessment- and community-centered classrooms drawn from the seminal *How People Learn* (Bransford, Brown & Cocking, 2000), we use our findings to suggest implications for ways in which the learning sciences may be fruitfully developed.

Context: The Learning Sciences as co-evolving with technologies

The learning sciences have always closely connected new understandings of how people learn (HPL) to emerging new technologies for and approaches to learning (Bransford, Brophy & Williams, 2000). For example, earlier advances in technology made new representations of mathematics possible, such as dynamically linked multiple representations. Learning scientists studied how students make sense of mathematics with linked representations (Roschelle, Noss, Blikstein & Jackiw, 2017). Likewise, a long-standing program of artificial intelligence in education made it possible to trace (assess) student knowledge and give targeted feedback, and researchers studied how learner- and assessment-centered AI approaches could improve learning (Luckin, Holmes, Griffiths & Fourcier, 2016). Many earlier learning science studies only examined short curricular unit, because this is what it was feasible to field across many classrooms. Further, earlier studies often examined only a few schools or classrooms, because getting the necessary technology in place was often hard.

Now technological platforms make it feasible to deploy techniques like multiple representations and AI in a curricular resource that spans a full classroom year. Further, the collection and rapid use of student data has become easier, and it is possible to provide teachers with dashboards and reports to guide their work in real time. New approaches such as "blended learning" are becoming popular among educators. In blended learning, it is expected that teachers and technologies will each have a complementary role in the overall instructional program (Means, Toyama, Murphy & Bakia, 2013). Importantly, these infrastructures and approaches have become sufficiently commonplace that they can be studied not just in special research-partner schools, but in a sample of schools recruited from a whole state. Programs that are year-long and could scale state-wide could have big impacts. For learning sciences to play a role in understanding these impacts, it may have to adjust its focus. In this paper, we will look at that evolution in terms of the HPL framework of a learner-, knowledge-, assessment- and community-centered classroom.

The math curriculum impact study

This study, funded by the Institute of Educational Sciences in the US, was intended to investigate the efficacy of a year-long, digital, blended mathematics curriculum with a strong AI component. The main hypothesis was that grade 5 students in schools that implemented the new mathematics curriculum for a full year would have higher mathematics achievement at year end than in schools in a business-as-usual control condition.

Intervention: Reasoning Mind

In the TC, schools were asked to use Reasoning Mind's grade 5 core curriculum (hereafter, "RM") as their main instructional resource. With regard to being *knowledge-centered*, RM's instructional approach (Khachatryan et al, 2014) is closely modeled on an exemplary international approach and seeks to build complementary facets of mathematical ability: fluency with calculations and deep understanding of foundational concepts. It also has a strong problem-solving component, with three levels of progressively harder problems and a "smarter solving" module. With regard to being *assessment-centered,* RM collects copious data as students do mathematical work online and continuously monitors student progress. These data are used to ensure the system is *learner-centered*. The system adapts its instruction, the difficulty of problems, and the pace through the materials based on AI techniques, so that instruction is personalized for each learner. Further, teachers get useful reports that guide their work with specific students (or groups of student). Teachers can assign special assessments to follow up and see if their interventions with students paid off or more support is needed. Another important aspect of the learner-centered approach in RM that aligns with HPL is the focus on metacognition; RM's pedagogical approach seeks to develop *independent learning strategies,* such as students keeping good notebooks and using them when they get stuck, rather than always asking a teacher for help. RM also is notably *community-centered.* The program includes whole class incentives to motivate students. RM builds a strong classroom mathematical culture in part by introducing a "Genie" character to whom students relate and who establishes norms for a mathematics learning. RM envisions and supports a classroom community where students learn individually, but also where they support each other and celebrate successes together, and where teachers have time to care for the needs of individual students. Reasoning Mind was also an attractive intervention to study because it had good prior results (Roschelle, Bhanot, Patton & Gallagher, 2015) and a strong capability to achieve high quality implementation in many schools at once through the role of Implementation Coordinators (Roschelle, Gaudino & Darling, 2016). Previous studies had also found high levels of student engagement in classrooms using RM (Ocumpaugh et al 2013).

Setting and sample: West Virginia schools

We conducted this study in West Virginia (WV), a state shaped by its geographical setting amongst the Appalachian Mountains; mountains and rolling hills define the region. Our WV-based McREL team expressed that their region has a love of place, community, and family, and in our initial contact with schools, we felt we could see these attributes reflected in the classroom. WV has low population density and the median household income of $42,000 is considerably lower than the national median of $56,000 (Frohlich, Sauter & Stebbins, 2016). Throughout the US, low family income and lower mathematics achievement are correlated. WV has been a leader in putting strong computing facilities in its schools and connecting schools to the Internet with high bandwidth. Since 2011, access to wired connections has improved from approximately 45% to 91% of West Virginians (Broadbandnow, 2017). The WV State Board of Education has adopted as its goals to "provide a high-quality learning system that (a) encourages a lifelong pursuit of knowledge and skills, (b) promotes a culture of responsibility, personal well-being and community engagement and (c) responds to workforce and economic demands." Just prior to this study, WV also adopted curriculum standards in mathematics that set high expectations for all students, and the teachers we worked with showed strong commitment and effort towards increased mathematics achievement. RM already had an implementation in a few WV schools that was going well, which made recruiting easier. We recruited over 50 schools to participate in the two-year randomized controlled trial from districts spanning the state, and although a few dropped out for various reasons, 46 schools remain in the final data sample we analyzed.

Research design

We planned and conducted a randomized control trial. Schools were matched in pairs that had similar prior math scores and geographic locations, and then a coin was flipped for each pair. Schools assigned to the TC were trained and supported to use RM for two years: a "warm up" year in which the teachers learned the new pedagogical approach and a "measurement year" in which we collected student prior and end-of-year achievement data. Schools assigned to the CC continued with business-as-usual materials and teaching approaches for 5th grade, but as an incentive, they were offered a different RM product for use in grade 2. These students would not reach grade 5 until the study was over. No CC teachers taught both grade 5 and grade 2, to avoid contamination.

Measures

The study collected a very rich array of measures, including teacher interviews and surveys and RM system data. However, in the scope of this paper, we focus on only two measures, a standardized test and observations. We used the required statewide assessment, the WVGSA, for both end-of-year mathematics achievement in grade 5 and as a prior achievement covariate in grade 4. This assessment was designed to be adaptive, to align with the state's curriculum framework, and to measure problem solving and not just procedural fluency.

A team at McREL designed an observational measure. Designing this measure was a challenge, because we wanted a measure that would work in both the TC and CC and the nature of what could be observed in these settings turned out to be quite different. A manuscript under development will describe in more detail how the measure was designed and refined through different phases of pilot testing in schools (Herman & Bumgardner, in preparation). In the course of the refinement process, the McREL team revised the instrument and their training until sufficient interrater reliability (> 80%) was achieved.

We report only on observations from the measurement year, which used an observational instrument that was agreed upon by all the partners in the study. In the instrument, an observer in the classroom typed running record field notes, guided by a framework with five areas in of observation: (a) on task behavior, (b) motivational routines, (c) independent learning strategies, (d) use of data, and the (e) quality of mathematical discussion among teachers and students. After making field notes, each observer rated these five areas on a 1 to 3 scale where "1" (*low* or *not present* or *rarely*) to "3" (*high* or *frequently*), using a rubric that set criteria for the scale levels.

Data

We obtained statewide scores for both grade 4 and grade 5 for students in 46 schools, 23 TC and 23 CC. We conducted a total of 53 observations in 38 TC classrooms and 15 CC classrooms. We included all the TC classrooms because of the greater interest in these, but only a sample of CC classrooms due to limited budget. For each observation, we collected a running record plus ratings for that classroom.

Analysis plan

For the main impact analysis, the SRI research team set up a two-level hierarchical linear model to account for the clustering of students within schools, using the grade 5 student scores as an outcome variable and the grade 4 scores as a co-variate. In later models, ratings from observations were added as potential mediating variables that might account for additional variance. The main impact analysis is the subject of a forthcoming journal submission (Shechtman et al, 2018) and is not reported in detail here. For the analysis of observations, the McREL team looked both at the *contrast* between observations in the CC and TC and also *variation* within only the TC. The McREL team analyzed its observational data for potential variations among classrooms in the TC that might be systematic across a set of teachers. They did so without awareness of which teachers or schools had achieved higher or lower mathematics achievement outcomes. Independently, the McREL team developed its own sense of "low" vs. "high" implementations meeting as a team to review its ratings (through the observation process) of the schools it observed, and then analyzed field notes to see if common themes emerged. For further details on the quantitative rating, see (Herman & Bumgardner, 2018).

Findings

We discuss the main impact findings briefly, for context. Full presentation of findings will be in (Shechtman et al, 2018). We then also consider the *contrast* between TC and CC. We focus thereafter on the findings about *variation* within the TC, and look for systematic ways in which TC classrooms varied and examine evidence as to whether those variations were plausibly related to student achievement outcomes.

Impact findings

In contrast to our hypothesis, the data did not reveal a measurable difference in mathematics achievement between TC and CC schools on the WVGSA test. Although unexpected, this does not mean that the TC was bad for students, indeed the student outcomes did not differ significantly between groups.

We also found considerable variation within the TC. In Figure 1, we illustrate this graphically by drawing a bar for each school where the height represents the average mathematics score at the end of grade 5 for that school. The bars are ranked by score and filled by condition. This shows that some schools in TC had some of the highest end-of-year math scores, but other schools in the TC were among the lowest scoring schools. This distribution led us to wonder about differences between the schools at each end of the distribution, for example, differences in how they did the work of teaching and learning.

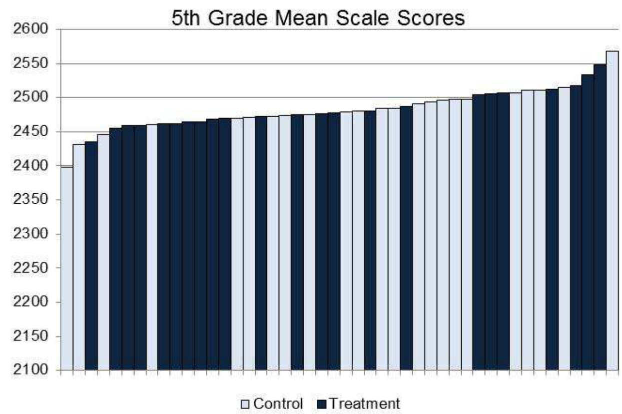

Figure 1. Variation in Achievement by School.

Contrast findings (treatment vs. control)

Our observers found striking differences between TC and CC classrooms. In general, observed compliance to assigned condition was high: TC classrooms were observed to be implementing RM as the central instructional resource and CC classrooms were not. Control classrooms looked reasonably traditional. For example, the teacher often provided instruction from the front of the room and students often worked on math problems individually or in small groups as a teacher (and often additional instructional aides) walked around providing help. Technology was available, but not used frequently. In TC (RM) classrooms, students sat down at computers at the beginning of their mathematics lesson and began working individually; most students this way most of the time. Teachers typically had a station in a corner of the room with a computer that provided reports on student work. The teachers called individual students (or small groups) to their station and worked on targeted issues. Aside from these interventions, there was not as much small group work as in CC classrooms.

With regard to the five scales in the observation instrument, the observers did not find any statistically significant differences between conditions in the degree to which students were on task nor in the observed supports for motivation and engagement. There were two statistically significant difference that favored the TC: as expected, there was more data use (92% vs. 8%) and more emphasis on independent learning strategies (88% vs. 12%). The observers also rated control classroom as higher on the quality of mathematics instruction scale. However, they also noted that it was harder to make relevant observations on this scale in the TC classrooms – for example, more of quality of instruction was mediated by technology and was hard to observe.

Variation within treatment findings

High rated classrooms

Nine of the thirty-eight treatment classrooms (24%) received overall ratings of 3 across all subscales. The instruction strategies, techniques, and procedures observed across each of these nine classrooms were quite similar, with recurring themes emerging. High scoring treatment classrooms had teachers who demonstrated comfort and control in managing student behaviors in their classrooms, structuring their classes and lessons in a manner conducive to student engagement and learning. Students in such classrooms regularly demonstrated familiarity with behavioral expectations and performance objectives, as well as standard classroom procedures. In many treatment classes observed, students entered the classrooms at the beginning of the period, obtained laptop computers, and signed in to RM to begin their coursework without prompting from instructors.

Almost all highly-rated treatment classrooms concluded the lessons with a review of students' performance for the day, with teachers highlighting mathematical achievements at both the whole-class and individual level. Students in highly-rated treatment classrooms demonstrated apparent investment in their mathematical performance—for example, in their commitment to engaging with the RM system. All highly-rated treatment teachers were observed employing adaptive learning strategies and techniques that aligned with RM. For example, they focused on independent learning strategies, by asking students to use available resources to resolve mathematical difficulties before calling a teacher over. Teachers in these classrooms also frequently used formative performance data in real-time, using it both to motivate students and to select groups of students for one-on-one interventions. With regard to motivation, teachers in the high rated classrooms, tended to exhibit community-centeredness by featuring the whole classes daily performance statistics before discussing any individual students. In the one-on-one interventions, the mathematics talk in these classrooms more often involved students in doing significant amounts of mathematical work; the teacher didn't do the math for students. However, overall, it was hard to observe mathematical knowledge building, in part because teachers often directed students to do work on the computers; further, the knowledge-building work that was available in discourse was like an "intervention" than a longer-term process of developing understanding.

Low rated classrooms

McREL observers rated six of thirty-eight (16%) of classrooms as lower quality across all categories of observation. A wider variety of instructional strategies, policies, and procedures were observed across lower-rated compared to higher-rated treatment classrooms, though several patterns were consistent. Perhaps the most overt trend observed across lower-rated treatment classrooms was the extent to which many of the teachers appeared to struggle in implementing effective classroom management. Some students spent extended periods of time disengaged from mathematical material in these classrooms. These teachers demonstrated an ability to recognize off-task behavior, but demonstrated difficulty in sufficiently addressing it.

Lower-rated treatment classrooms also differed from one another as far as the extent to which teachers established RM-related routines and expectations as well as the extent to which objectives were evident and communicated to students. Like higher-rated treatment classrooms, several teachers in lower-rated classrooms clearly communicated and referenced established procedures and ensured that students understood what was expected of them. In other scenarios, teachers were not observed making significant effort to motivate students to use RM as intended. Teachers from three classrooms were not observed incentivizing or encouraging engagement with mathematical material to a significant degree, and, thus, students in these classrooms appeared predominantly ambivalent as to the extent to which they accomplished RM objectives. Additionally, none of the teachers in lower-rated treatment classrooms were observed implementing strategies or practices to facilitate student autonomy or independent learning strategies. Across each of the lower-rated treatment classrooms, students were reliant on course instructors to make significant progress through the RM curriculum, with several students exhibiting an inability to engage in any mathematical work independently. Further, in some classrooms, teachers actively inhibited students from independently completing assignments—emphasizing more teacher control in the mathematical work of the classroom. In lower-rated treatment classrooms where teachers allowed students to collaborate with classmates, students nevertheless appeared to have trouble engaging with mathematical content without assistance from instructors. None of the teachers in these classrooms were observed making references to resources that students could use for math learning without involving the teacher. On occasion, a student or small group of students appeared to consult the hints provided in RM or refer to the RM library, but these behaviors were rare. For the most part, teachers of low-rated treatment classrooms were not observed making frequent use of data to inform instruction. Those teachers who used data did so in a more supplemental manner.

Exploratory model

To explore whether the high and low patterns identified by McREL's observation team might relate to student outcomes, we conducted an additional analysis. For this analysis, we included only students of TC teachers in either the high (n=10) or low (n=6) pattern classrooms. We conducted a 2-way ANOVA in which the outcome variable was assessment score, and the factors were (1) Year (Grade 4, Grade 5), (2) McREL Group (Low, High), and (3) the interaction term (Year x McREL Group). Means are shown in Figure 2. We found that both main factors were significant. As would be expected, WVGSA scores were higher in Grade 5 than in Grade 4, $F(1,748)=12.95$, $p <.001$ (while reported on the same scale, the tests were different and aligned with respective grades). There was also a main effect of McREL Group, such that students of teachers in High classrooms had higher assessment scores than did students of teachers in Low classrooms, $F(1,748)=8.40$, $p <.01$. The interaction term was not significant, $F(1,748)=.43$, $p = .51$, n.s. In everyday terms, this means there was no closing nor expanding of the achievement gap.

Discussion

Overall, the work of teaching and learning was quite different with the digital, blended learning approach – treatment classrooms were quite different from control classrooms. Further, based on observations, we were able identify systematic "high" and "low" patterns within the treatment schools. We explored the importance of these patterns in a quantitative model that included student prior achievement and student outcomes. The observed patterns appear to be linked to classroom *mean prior achievement,* which raises equity issues. Overall, only 16% (6 of 38) of the observed classrooms fit the low pattern. While examining these classrooms is useful in looking for improvements, these classrooms are not representative. We emphasize *exploratory* investigation of these classrooms, and do not use this small sample to reach generalizable conclusions. We frame our discussion in terms of "uptake" – the uptake of unique RM features and possibilities was different across the two classroom groups.

Learner-centered

In high functioning TC classrooms, there was more uptake of learner-centered opportunities. For example, high functioning classrooms were observed to place an emphasis on independent learning strategies, but low functioning classrooms did not. Likewise, on the quality of mathematics instruction scale, we found that math talk in the higher functioning classroom gave the responsibility for doing mathematical work to the students, whereas in lower functioning classrooms, teachers did more mathematical work. The suggests that not all classrooms take advantage of the opportunities for learner-centered instruction equally. However, one caution is that the observed low group also had lower prior mathematics achievement scores; it could be that some students (who are about 10-11 years old in grade 5) are not ready for independent learning strategies in mathematics and may benefit from a more teacher-centered approach. Likewise, teachers may have a stronger repertoire for engaging students with higher mathematics achievement in doing mathematical work, but may default to doing more of the mathematical work for their lower-performing students. In a traditional classroom, this may be less evident, because there may be enough mathematical knowledge in the classroom overall for the teacher to sustain a high quality of mathematical discourse. The learning sciences may need to elaborate how teachers could enact learner-centered instruction when they are working one-on-one with a large group of students who are coming in with low existing knowledge in mathematics and really struggling.

Assessment-centered

Overall TC classrooms used real-time data reports to make instructional decisions, whereas CC classrooms did not. Moreover, within TC classrooms that were observed to be lower functioning, there was much less use of real-time data reports. Again, we urge some care in interpretation. It could be that implementation coordinators should help teachers to make better use of the data reports. But it also could be that certain factors have to be in place before teachers can sensibly use data reports in real time. For example, if students do not stay on task during individual work at computers, it may not make sense for teachers to be looking at data reports during classroom time. Likewise, if students are uniformly struggling with the mathematics (the observed-low group had weaker prior math achievement), it may make less sense to teachers to work with individual students. The learning sciences could help us understand better how to leverage a more assessment-centered teaching structure with classrooms that are more or less ready to engage in grade-level mathematics.

Community-centered

As learning communities, the observed-low classrooms were more chaotic, with many behavioral problems. It is unclear why the 16% of classrooms in the "low" group had lower engagement. We are reluctant to attribute it to the students alone, because control classrooms with low achieving students did not have the same level of behavioral problems. It could be that the experience of digital and blended learning was less satisfying to some groups of students, and the behavioral problems emerged from their frustration and confusion. However, this was not observed in prior studies of student engagement with RM (e.g. Ocumpaugh et al, 2013). Our team of observers sometimes wondered whether the mix of instructional activities in the TC classrooms had too much of an emphasis on individual time at a computer, and whether the classroom community might fare better with social activities like small group work and full classroom discussions. We particularly wondered whether the rather quiet and individualistic classrooms in the blended learning condition may not have fully utilized the community-centeredness otherwise evident in the WV classrooms that we visited. Yet, in the observed-high group, we were able to see many positive community-centered aspects of classrooms, such as peers helping fellow students and new norms being established through the character of the Genie. We noted that teachers spent lots of time working one-on-one with students, which can be good for strengthening relationships.

Knowledge-centered

It was hard to observe how knowledge building worked in the new, digital learning blended environment. This may be a weakness of observational methods. With regard to the method, classic learning science "knowledge building" environments make more use of collaborative, social and full-classroom learning – the TC classrooms were less collaborative and social and thus knowledge building was less public. Methodologically, we were able to observe strengths of the RM instructional materials as we watched individual students use them, for example, the recommendation to read "Genie solutions" to understand problem solving processes, which related to the well-known self-explanation effect (Van Lehn, Jones & Chi, 1992). The RM materials also give strong conceptual presentations to students, and these may be stronger that many teachers would typically achieve in their own presentations of concepts. Yet, although there is more time and space in RM classrooms for teachers to work with individual students on their conceptual understanding of mathematics, we observed variability in the quality of the mathematical discourse in TC classrooms – many of the conversations we observed were more procedural than knowledge-building oriented. Overall, a challenge for the learning sciences is to come to a better understanding of how to measure knowledge-building in environments like those we observed in TC classrooms, which less available public knowledge-building discourse.

Equity

Overall our analysis revealed a potential equity issue. We observed less uptake of the capabilities of the digital, blended curriculum and more behavioral management problems in a cluster of classrooms, and then later found the cluster had lower mean prior achievement. Conversely, when our observers noted a cluster with uniformly high quality of implementation, we found that the prior achievement scores in those classrooms were higher. Causality cannot be determined from this analysis. Either (a) while the digital, blended approach is appropriate for classrooms with lower mean prior achievement, specific additional support for implementation is needed or (b) the adaptive, blended learning approach may have been less appropriate for classrooms with lower prior mean achievement, running into classroom problems despite good implementation support. Overall, we remind readers that there is a distribution of low-to-high achieving students in every classroom, so it *cannot* be inferred from this analysis whether the approach has differential benefits for individual students who have higher or lower achievement (e.g., there were some students with lower achievement in the classrooms that had higher mean achievement). The only relationship we explored was between *classrooms* that were observed to be making lower use of the HPL-related features of RM and *classrooms* that as a whole had lower mean prior achievement.

Conclusion

The ICLS 2018 conference program envisions a future in which the learning sciences helps schools to make sense of new teaching and learning approaches, including those with strong AI components, and guides improvement in the quality of products. The MCIS project provided an opportunity to investigate a future-oriented learning environment, with a full-year digital curriculum that incorporated AI features and a blended learning instructional approach. We found that schools were able to implement this curriculum throughout a state. Relative to classrooms in the control condition, there was a strong contrast in how the new classrooms functioned as workplaces for teaching and learning. We also looked systematically across implementing classrooms for systematic patterns that might explain variation in classroom outcomes. We found a pattern in which a group of classrooms with lower uptake of the HPL-related features of the new technological approach also were classrooms with lower prior mathematics achievement. This points us to one way in which Learning Scientists could help product developers – by examining the differential uptake of research-aligned features and examining relationships to equity factors, like low prior achievement. As our data does not reveal causality, we would encourage future researchers to explore why uptake of HPL-aligned features was lower in some classrooms.

We also believe that the Learning Sciences itself needs to change in order to become more relevant. We need better methods for making sense of knowledge building activities when distributed and so thoroughly mediated by technology that there is more individual and less "community" knowledge building time. The learning sciences could say more about how to make sense of the kinds of observations we made of lower functioning classrooms: were students ready for an emphasis on independent learning? How can teachers have good mathematical conversations in one-on-one settings, where they may less variety of students' ideas to draw on than in full class discussions? How do teachers decide when to make use of student progress reports, and are there circumstances in which using such reports is more or less useful? Overall, we suspect the learning sciences could make stronger contributions as envisioned in the conference program with greater attention to systematically describing variability at scale and helping to uncover aspects of variability (and equity) which are consequential for learning outcomes. In this way, we foresee more relevance for the learning sciences by aligning with improvement sciences and through a focus on measuring and addressing undesirable variability across settings.

References

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school.* Washington: National Academy of Sciences.

Bransford, J., Brophy, S., & Williams, S. (2000). When computer technologies meet the Learning Sciences. *Journal of Applied Developmental Psychology, 21(1),* 59-84.

Broadbandnow. (2017, May 16). Internet access in West Virginia. Retrieved from https://broadbandnow.com/West-Virginia

Frohlich, T.C., Sauter, M.B., & Stebbins, S. (2016, September 15). America's Richest (and Poorest) States. Retrieved fromhttp://247wallst.com/special-report/2016/09/15/americas-richest-and-poorest-states-4/2/

Herman, P. & Bumgardner, K. (2018). Developing a Classroom Observation Protocol When Classrooms Don't Look Alike. In preparation.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. London: Pearson.

Khachatryan, G. A., Romashov, A. V., Khachatryan, A. R., Gaudino, S. J., Khachatryan, J. M., Guarian, K. R., & Yufa, N. V. (2014). Reasoning Mind Genie 2: An intelligent tutoring system as a vehicle for international transfer of instructional methods in mathematics. *International Journal of Artificial Intelligence in Education, 24(3),* 333-382.

Ocumpaugh, J., d Baker, R. S., Gaudino, S., Labrum, M. J., & Dezendorf, T. (2013, July). Field observations of engagement in Reasoning Mind. In Int Conf on AI in Education (pp. 624-627). Berlin: Springer.

Means, B., Toyama, Y., Murphy, R., & Bakia, M. (2013). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. *Teachers College Record, 115(3),* 1-47.

Roschelle, J., Bhanot, R. T., Patton, C., & Gallagher, L. (2015). Strength of Research for Reasoning Mind. Menlo Park, CA: SRI International.

Roschelle, J., Gaudino, S. & Darling (2016). Designing for consistent implementation of a 5th grade digital math curriculum. *International Journal of Designs for Learning, 7(2),* 41-56.

Roschelle, J., Noss, R., Blikstein, P., & Jackiw, N. (2017). Technology for learning mathematics. In J. Cai (Ed.), *Compendium for Research in Mathematics Education*. Reston, VA: National Council of Teachers of Mathematics. 273-296.

Shechtman, N., Roschelle, J., Feng, M. & Singleton, C. (2018). *An Efficacy Study of a Digital Core Curriculum for Grade 5 Mathematics.*

Singleton, C., Roschelle, J., Shechtman, N., & Feng, M. (2018). *A framework to evaluate the design and implementation of blended learning instructional environments*. Manuscript in preparation.

VanLehn, K., Jones, R. M., & Chi, M. T. (1992). A model of the self-explanation effect. *Journal of the Learning Sciences, 2(1),* 1-59.

Acknowledgements

This material is based upon work supported by the Institute of Educational Sciences (IES) of U.S. Department of Education under Grant Number R305A130400. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of IES. We thank the West Virginia Department of Education for providing access to the data. We also thank the students, teachers, and schools who participated in the study as well as the project staff from Digital Promise, SRI International, McREL, and Reasoning Mind.