# Studying Whole Class Discussions at Scale With Conversation Profile Analysis

Ryan Seth Jones, Middle Tennessee State University, ryan.jones@mtsu.edu
Joshua M. Rosenberg, University of Tennessee, Knoxville, jmrosenberg@utk.edu

**Abstract:** Design research aims to generate theories and empirical evidence about complex, contextually bound mechanisms of learning. This often motivated a methodological toolkit that makes use of small numbers of participants in order to closely analyze talk, gesture, and artifacts related to learning. Yet, design research also aims to inform theory and practice on a large scale. This can create tension because the theory and mechanisms generated from small scale studies can be difficult to implement and study at larger scales. In this paper, we aim to contribute to other efforts to conceptualize design research as an act of scholarship that extends across multiple scales by sharing an approach to characterizing whole class discussions across a large number of diverse students, teachers, and settings using an approach we refer to as Conversation Profile Analysis.

## Introduction

Design research (DR) is motivated by commitments that often lead to tension and challenges with scale. On the one hand, DR is used to develop empirically grounded knowledge about learning that is contextually bounded in a particular setting, and each setting is "a complex, interacting system involving multiple elements of different types and levels" (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003, p. 9). This commitment often requires data and methods that are carried out in a small number of settings, both for practical reasons and for epistemic reasons. After all, if learning mechanisms are contextually bounded in social, cultural, and historical ways, then research methodologies ought to be specified for a particular time, place, and people (Cobb et al., 2003). Yet, DR also aims to do more than just characterize one learning ecology in a particular setting: It also aims to generate theories about relationships between learning and design that can inform a broader set of contexts. As Sandoval (2014) states, "The basic tension in educational design research is the dual commitment to improving educational practices and furthering our understanding of learning processes" (p. 20).

The first commitment of DR - *improving educational practice* - motivates a methodological toolkit that is attuned to generating evidence about conjectured correspondences between elements of a designed learning environment and evidence of student thinking manifested in discourse, action, or inscriptions in classrooms. Sometimes called conjecture mapping (Sandoval, 2014), this research approach typically relies on video and/or audio recordings of classroom activity and interviews, student artifacts, and field notes to create detailed characterizations of the realized learning environments. This methodology allows researchers to test a particular operationalization of their instructional theories, provide evidence about the relationship between researchers' instructional theories and learning theories, and can even produce new theoretical frameworks to model relations between instruction and learning (Cobb et al., 2003; DiSessa & Cobb, 2009; Bakker, 2019). Although these data records can include many units of time (days, weeks, months, or even years) they typically stretch across a much smaller number of participants.

The second commitment - *deepen our understanding of the processes of learning* - creates the need to generate knowledge about how these instructional theories take shape when designs are used across larger scales. This is a challenging endeavor, though, because DR is oriented towards understanding mechanisms for learning that can be difficult to measure or study across large numbers of participants. In addition, DR recognizes that designers are always in a state of prolepsis, imagining a learning environment as if it exists while simultaneously acknowledging that learning environments are rebuilt in each local context. This has motivated the development of design perspectives and research methods for implementing and studying innovations at scale. These perspectives include frameworks for co-design with local practitioners around relevant problems of practice (Penuel, Fishman, Cheng, & Sabelli, 2011), designing infrastructures to coordinate goals and learning across different levels of large institutions (Cobb & Jackson, 2011), generating data about design features that are usable and informative to practitioners on the ground (Penuel, Van Horne, Jacobs, & Turner, 2018), and analytic techniques for studying theoretical constructs related to designed learning environments on a large scale (Sherin, 2013). What these efforts make clear is that design as an act of scholarship extends across multiple scales, and that the simultaneous act of designing and researching brings new questions and challenges on a large scale.

One of the challenges of carrying out DR on a larger scale is to generate evidence about mechanisms of learning across large numbers of participants and diverse contexts. In this paper, we present our efforts to address

this particular challenge by describing work to characterize the structure of dynamic—and often unpredictable—*whole class discussions* in a way that can provide insight about conversational structures across large numbers of participants. We have chosen whole class discussions because of the potential they hold for supporting student learning in mathematics classes, but also because of their dynamic nature which often makes it challenging to study them on a large scale. Conversational structures are often a powerful research finding from DR, though typically at a small scale. For example, the structure of Accountable Argumentation provides insight into role different participants take on during a discussion, and how these roles relate to one another through conversation (Horn, 2008). We aim for a contribution similar in form, to better understand conversational structures across diverse students and settings, and to explore their usefulness in advancing our understanding of whole class discussions. Our goals are twofold, both methodological and empirical. First, we aim to describe an approach to DR on a large scale that explores the structure of whole class discussions across large numbers of participants, and to relate this structure to design conjectures about student learning. Second, we argue that our findings provide new knowledge about how whole class discussions about data, statistics, probability, and inference take shape across large numbers of classes, and how variations in these discussions are related to opportunities for students to discuss key mathematical and statistical ideas.

We carried out this project in the context of a large-scale efficacy study of an innovative design for supporting middle school students to learn about data and statistics which we call Data Modeling (Lehrer & Kim, 2009; Lehrer, Kim, Ayers, & Wilson, 2014; Lehrer, Kim, & Schauble, 2007). This context was productive for both our methodological goals and empirical goals because the design principles guiding the curricular innovation have a strong base of empirical support from smaller scale DR studies which we aimed to build from, and our findings inform our understanding of whole class discussions in the Data Modeling units. In our work, we were guided by the following research questions:

1. Do stable patterns of talk emerge in middle grades discussions about data, statistics, and probability across large numbers of classrooms?
2. When are the patterns more or less likely during the course of a whole class discussion?
3. How are the patterns related to opportunities for students to discuss mathematical and statistical ideas related to learning goals?

## Whole class discussions in data modeling classes

The Data Modeling instructional sequence engages students with concepts related to data display, statistics, chance, modeling, and inference as tools to answer two (seemingly) simple questions: 1) what is the length of our teacher's arm-span?, and 2) how precise were we as a group of measurers? With these questions as the driving motivation, the Data Modeling materials support teachers to facilitate the development of three epistemic practices in their classrooms: representing variability, measuring variability, and modeling variability. The practices are epistemic because they provide the means by which the original questions (about the teacher's arm span) are answered.

Throughout the instructional design, students have opportunities to *invent* data displays, statistics, and models to help answer their questions, and teachers facilitate *whole class discussions* about the invented approaches to support students to share, compare, critique, and revise their approaches. Each whole class discussion focuses on particular inventions and mathematical ideas, and how these ideas can inform students' practices around representing, measuring, and modeling variability. As students talk about their invented approaches to displaying data and measuring characteristics of the data, the teacher facilitates the conversation to compare different approaches and to discuss how the ideas help them answer the questions that motivate the data.

Teachers are critical in orchestrating classroom discussions that can support students' epistemic learning. Teachers must be able to recognize and sequence student inventions in ways that help students see similarities and differences, and then support students to compare their ideas and approaches, and connect their ways of thinking to conventional mathematical tools (Stein, Engle, Smith, & Hughes, 2008). These conversations are intended to approximate the professional statistical practice of negotiating the value of novel techniques when representing, measuring, and modeling variability. With this in mind, the teacher also has the responsibility to support students in developing goals, values, and discourse norms that are productive in collective activities that resemble disciplinary ways of generating and revising knowledge (Ford & Forman, 2006; Forman & Ford, 2014). As these conversations are carried out interactionally among teachers and students, teachers' facilitation moves and students' contributions create a discourse structure that emerges throughout the conversation.

## Methods

## Research context

This project was conducted during a two-year, large scale efficacy study of the Data Modeling curriculum (Lehrer & Kim, 2009; Lehrer, Kim, Ayers, & Wilson, 2014; Lehrer, Kim, & Schauble, 2007). The 6th grade classrooms in our study were located in 40 schools in a large, Southwestern United States city. These districts had diverse student populations who represented a wide range of economic, racial, ethnic, and linguistic backgrounds.

This project supported teachers to develop new pedagogical practices using focused and sustained professional development and coaching from teachers with experience using Data Modeling. Teachers participated in 12 days of in-person professional development led by the author of the Data Modeling curriculum materials and middle school teachers with experience using Data Modeling. Six of the days were conducted during a one-week summer workshop and six days occurred on Saturdays during the school year. The professional development engaged teachers with opportunities to explore key mathematical concepts and practices related to data and statistics. Teachers also engaged in activities to support the development of knowledge about student thinking in these domains (Ball, Thames, & Phelps, 2008), and had opportunities to develop competencies in core teaching practices necessary to support whole class discussions by rehearsing these practices with colleagues in the professional development sessions (Pfaff, 2017). During the school year, coaches and project staff supported teachers by helping with planning, co-teaching (and debriefing) class sessions, and providing feedback focused on continuous improvement.

## Observational measurement system

To characterize the class discussions, we identified variables (Table 1) related to student contributions and teacher facilitation moves. We developed these variables by drawing on research related to facilitating whole class discussions (e.g., Stein, Engle, Smith, & Hughes, 2008) and from years of design based research conducted in the development of the Data Modeling curriculum (e.g., Lehrer & Kim, 2009). The variables were binary, indicating the presence or absence of the characteristic. We scored these variables in *5-minute, adjacent segments of time* during each whole class discussion. Observers scored a variable if it was observed at least once in a five-minute segment. For example, if one student contributed a comment on the conceptual aspects of a student-invented procedure then the *sInvented* and *sProcedural* variables were scored.

Table 1: Whole class discussion observation indicators

| | Variable Name | Description |
|---|---|---|
| Student Contributions | sInvented | Did students discuss inventions? |
| | sConceptual | Did students make comments or ask questions about the conceptual elements of the inventions? |
| | sProcedural | Did students make comments or ask questions about the procedural or calculational elements of the inventions? |
| Teacher Facilitation | tInitSelect | Did the teacher select student-invented products to be shared? |
| | tCompare | Did the teacher compare different student approaches? |
| | tDiscussQ | Did the teacher use discussion questions? |
| | tConnectOthers | Did the teacher make connections between students' ideas? |
| | tConnectBigIdeas | Did the teacher connect student thinking to important mathematical/statistical learning goals? |
| | tPressExplain | Did the teacher press students to explain their thinking? |

## Data collection

We observed whole class discussions in Data Modeling classes across two years. During Year 1, 21 teachers participated in the Data Modeling professional development and used the materials in their classes. During the

second year, 39 teachers used the Data Modeling materials and PD. During the two summers of the project, a team of observers conducted weeklong training. After training, observers scored whole class discussions in videos of previous Data Modeling classes, and were required to agree with anchor scores at least 80% of the time in order to conduct live classroom observations. We also led ongoing trainings for four different Saturdays across each school year. During the weekend meetings we would discuss issues arising from the ongoing observations and practice the observation variables for upcoming units. During the second year we also conducted random double observations (by two observers) to maintain an ongoing measure of our agreement. The double observations in live classrooms typically produced more agreement than for the videos, so there were no observers that met the original benchmark but failed to agree at or above 80% during live observations.

We observed classes during six different units that focused on data display, statistics, probability, modeling, and inference. In all, we observed 3,249 5-minute intervals across 302 whole class discussions.

### Data analysis

We used a Latent Class Modeling (LCM) approach (Collins & Lanza, 2010) to identify patterns of talk within 5-minute segments of time, which we refer to as Conversation Profile Analysis. In general, LCM is useful because it can inform researchers about groups, or *latent classes*, that describe response patterns with similar characteristics. Moreover, the LCM output is both the estimation of groups as well as the probability that each 5-minute segment response pattern is associated with a particular group. This can reduce the complexity of the response patterns by identifying common categories, but the categories are only informative if they are conceptually meaningful and distinct.

While LCM identifies groups (and probabilities associating each segment with the groups), it does not determine the number of groups. To guide this decision, we made use of a number of criteria, including the Akaike and Bayesian Information Criteria (AIC and BIC) as measures of how satisfactory the model fit the data, the homogeneity and separation of the classes, and concerns of interpretability and parsimony in order to identify a solution that best accounted for the variability in the data. We used the *poLCA* package (Linzer & Lewis, 2011) to carry out the analysis in the R statistical software (R Core Team, 2019). We then qualitatively interpreted the latent class profiles to determine if the categories could be characterized using our understanding of the discussions from previous research on the Data Modeling approach.

After identifying a class solution, we examined how the categories fluctuated across ten-minute time intervals in a conversation through the use of a $\chi 2$analysis for contingency tables (between latent classes and time intervals). We combined the 5-minute intervals into 10-minute intervals because the 5-minute segments proved to be too difficult to analyze due to the number of segments that can occur in one whole class discussion. We also explored the relationship between the 5-minute segment latent classes and mathematical ideas in the curriculum using the same $\chi 2$ approach (for the contingency table between latent classes and the presence of key mathematical ideas). Since the mathematical ideas are different across the units, we created scoring rules to generalize the variables across units through the use of the following three categories: 1) Mathematical or statistical idea not talked about, 2) Mathematical or statistical idea is talked about, but students did not discuss the epistemic role of the idea, 3) Mathematical or statistical idea talked about in ways that are consistent with the epistemic role of the idea. To interpret the results, we first examined the overall $\chi 2$ test statistic to determine whether there were different frequencies across the cells in the contingency table than would be expected by chance overall, and—if differences were found to be present—then interpreting the standardized residuals for each cell of the table. Because the $\chi 2$ test statistic indicated that there were differences for all three steps, we next examined the standardized residuals to explore which cells deviated most from a random, uniform distribution.

## Findings

We determined that a four latent class solution to the LCM provided the best fit to the observation data (Figure 1). These four categories were conceptually and theoretically distinct in the ways they characterized the types of discussion within 5-minute segments. This solution demonstrated good fit in terms of the information criteria relative to three- and five-latent class solutions, adequate class homogeneity and separation, and was satisfactory in terms of concerns of interpretability and parsimony. In addition, the profiles of variables in each latent class were interpretable based on our understanding of the conversations in Data Modeling classes and our theories about how teachers' facilitation moves and students' comments are related to each other. Figure 1 represents the four categories, with each bar corresponding to the estimated probability of an indicator being observed during a 5-minute segment that is classified in the given latent class.

We analyzed the qualities of each latent class profile in order to describe the groups in terms of the aims of the Data Modeling curricular design. After reviewing the data profiles for each latent class, we characterized the four groups as *Low Activity*, *Discussing Ideas*, *Inventing & Discussing*, and *Inventing & Connecting*. There

are some extreme differences in the distribution of probabilities across the latent classes, but we found that close analysis revealed finer grained differences that provided explanatory power in characterizing the nature of dynamic whole class discussions in these settings.

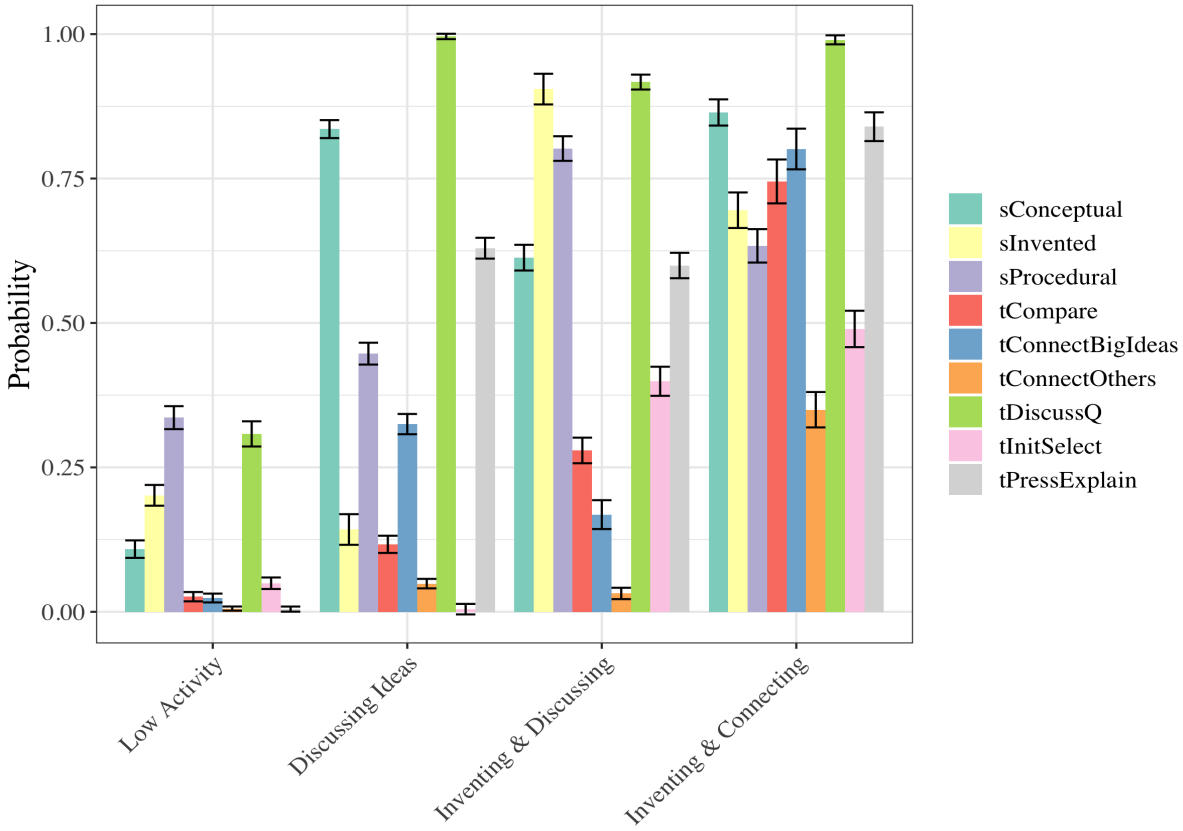


Figure 1. Latent class profiles.

## Characteristics of latent classes

The *Low Activity* category represents segments of discussion in which all of the observation variables have a low probability of being observed within the 5-minute segment of time. Our label for this category is not meant to suggest that teachers and students in the class were not engaging in any activity, or any activity that was meaningful to them, but rather that they were not engaged in the types of conversation that the Data modeling materials were designed to support. In Low Activity segments, student procedural talk and the presence of teachers' discussion questions were more likely to be observed than any other variable. However, we note that the discussion question variable is scored generously since it is difficult to determine the quality of a discussion question in real time. In Low Activity segments, there is a 40% probability of observing students talking procedurally and a 25% chance of observing a teacher using discussion questions. Additionally, there is a 13% chance of seeing students talking about their invented methods. All other variables have less than 10% probability of being observed. The co-occurrence of teachers' questions, students talking procedurally, and students talking about their inventions suggests that the majority of questions about the invented methods were about procedural elements of the methods, which elicited primarily procedural answers.

The *Discussing Ideas* category is characterized by a high probability that teachers are asking discussion questions, students are talking both procedurally and conceptually, and the teacher is pressing students to elaborate their thinking and attempting to connect student thinking to key mathematical learning goals. There is a very high—99% chance—the teacher will ask discussion questions, 61% chance that he or she will press students to further describe their thinking, a 34% chance they will make connections between students' ideas and key mathematical concepts, and a 12% chance that they will work to compare different students' ideas during these segments. For students, there is an 81% chance of talking conceptually, a 45% chance of talking procedurally, and a 13% percent chance of talking about their invented methods. This category highlights why it is not sufficient

to simply determine if students are talking about invented methods when characterizing the conversations, because the probability is identical between Low Activity and Discussing Ideas. However, teacher facilitation moves and student contributions are very different, and in ways that suggest that when teachers are asking questions they are eliciting conversations about both conceptual and procedural elements, and connecting these comments to relevant mathematical learning goals.

The *Inventing & Discussing* category describes 5-minute segments of discussion with the highest probability of students talking about their invented approaches both conceptually and procedurally, and a high probability that the teacher is using multiple facilitation moves in order to support the whole class discussion. There is a 91% chance that students are discussing their inventions, a 79% chance they are talking about procedural aspects, and a 60% chance they are talking conceptually in these segments. Teachers in these segments have a 90% chance of using discussion questions, a 59% chance of pressing students to elaborate their thinking, a 40% chance of intentionally selecting which invention to discuss, and a 27% chance of comparing different students approaches to each other. There is only a 14% chance that teachers connect student ideas to mathematical learning goal, and a 5% chance teachers connect students' ideas to each other. This suggests that these segments of discussion are characterized by students talking about their inventions, and both the procedures that define them and the concepts that the procedures embody. In addition, teachers are supporting this conversation with multiple facilitation moves, but are not working as often to connect students' ideas to each other or to mathematical learning goals.

Finally, the *Inventing & Connecting* category was similar to *Inventing & Discussing*, but with increased probability of teachers working to connect students' ideas to each other and to mathematical learning goals. Students still have high probabilities of talking conceptually and procedurally about their inventions, with a 71% chance they are talking about inventions, a 64% chance of talking procedurally, and an 87% chance of talking conceptually during the 5-minute segment. Teachers also use multiple facilitation moves, with a 99% chance of asking discussion questions, an 85% chance of pressing students to explain their thinking, and a 49% chance of intentionally selecting which invented methods to feature. In addition to these, though, the teacher also has an 81% chance of connecting students' ideas to mathematical learning goals, a 74% chance of comparing different student approaches to each other, and a 34% chance of connecting different student ideas to each other.

## Latent classes and design conjectures

Whole class discussions are not static, but rather unfold over time and develop in ways that are sometimes unexpected. Because of this, we hypothesized that the four categories would not be equally likely across the arc of a whole class discussion, but that some would be more likely earlier and others more likely later. Also, the instructional theories that guide the Data Modeling curricular materials suggest that the four categories should be differentially related to the mathematical ideas students talk about. For example, we expected that students would talk about mathematical ideas in more sophisticated ways during the moments of discussion that were assigned to the *Inventing & Comparing* and *Inventing & Discussing* than those assigned to the *Low Activity* category. In this section we report on our analyses to test these conjectures, and the implications for the meaningfulness of the categories to characterize the dynamic, complex conversations.

Table 2: Frequencies of segments in each category across time. Green highlights reference frequencies that are statistically significant in the positive direction and red in the negative

| | Class Discussion Timeline | | | | | |
|---|---|---|---|---|---|---|
| | 0-10 Minutes | 10-20 Minutes | 20-30 Minutes | 30-40 Minutes | +40 Minutes | Total |
| Low Activity | 206 | 102 | 95 | 85 | 266 | 754 |
| Inventing & Connecting | 32 | 73 | 92 | 93 | 147 | 437 |
| Inventing & Discussing | 130 | 157 | 164 | 149 | 306 | 906 |
| Discussing Ideas | 156 | 191 | 169 | 179 | 457 | 1152 |
| Total | 524 | 523 | 520 | 506 | 460 | 3249 |

We found that these profiles were not randomly distributed across conversational arcs in Data Modeling classes, but were more or less likely at different time points in the discussions. Table 2 reports on the frequencies

of segments classified in each latent class by 10-minute segment, with red cells indicating those where a category is significantly less likely to be observed and green those that are significantly more likely to be observed as judged by a $\chi 2$test with $p < .05$. *Low Activity* segments were more likely within the first ten minutes of a conversation, but less likely to be observed between the 10-40 minute marks. *Discussing Ideas* was more likely to occur after 40 minutes of conversation. *Inventing & Discussing* was more likely to occur between ten and 40 minutes of the conversations. Last, *Inventing & Connecting* was less likely to occur within the first ten minutes, but more likely between the 20-40 minute marks.

This table also shows that these categories were highly dynamic, as the majority of cells have frequencies that are not more or less than what we would expect if the categories were randomly distributed across time. This suggests that for much of the time, the categories are highly unpredictable, which is not surprising in a whole class discussion that is responsive to students' ideas. Yet the aggregate does suggest a structure that is sensible given our conceptualization of the groups. For example, it makes sense that in the first ten minutes of a class, discussion would be more likely to be characterized as *Low Activity*, as teachers and students might be engaging in informal discussion about past or upcoming assignments or reminders about previous activities and discussions. It is also sensible that discussing students' inventions in ways that attend to conceptual and procedural elements, and that make connections among students' ideas and mathematical learning goals, are more likely between 20-40 minutes of conversation. Although we did not previously have an empirical basis for identifying these segments, it is sensible that it would take considerable time to build the conversation to this point because it takes time for students to discuss their inventions, and for the teacher to have enough inventions in the conversation to compare across them. This analysis also suggests that whole class discussions less than 20 minutes in length are not as likely to spend time discussing inventions in this way.

Finally, we found that some classifications supported discussion about key mathematical ideas more than others. Table 3 shows that 5-minute segments classified as *Inventing & Discussing* or *Inventing & Connecting* were significantly more likely to also have students talking about mathematical ideas in ways that address their epistemic role in making claims with data. *Low Activity* segments were less likely for students to be discussing any of the mathematical ideas. During moments in the *Discussing Ideas* category, students were more likely to be discussing mathematical ideas, but not in ways that addressed the epistemic role.

Table 3: Relationship between discussion category and students' discussion of big ideas. Green highlights reference frequencies that are statistically significant in the positive direction and red in the negative

| | How students Talked About Big Mathematical Ideas | | | |
|---|---|---|---|---|
| | No Big Ideas | Big Ideas Only | Big Ideas & Epistemic Issues | Total |
| Low Activity | 534 | 151 | 69 | 754 |
| Inventing & Connecting | 28 | 54 | 355 | 437 |
| Inventing & Discussing | 184 | 244 | 478 | 906 |
| Discussing Ideas | 282 | 428 | 442 | 1152 |
| Total | 1028 | 877 | 1344 | 3249 |

These findings suggest that although the conversations are dynamic, and constantly in flux, the relationship between the categories and students' discussion of mathematical ideas is much more predictable than the relationship with time. The relationship with time suggested a general structure, but there was much more randomness compared to students' talk about mathematical ideas. For example, only 5% of the segments of time where students are talking about mathematical ideas and epistemic issues related to the ideas were classified as *Low Activity*. Moments of discussion where students were talking about their inventions, and teachers were using multiple facilitation strategies, were much more likely to also exhibit student talk about epistemic roles of mathematical ideas. This structure aligns with our conjectures about the relationships, providing additional evidence that the categories and our conceptualization of them are useful in characterizing the discussions.

## Discussion

The ICLS conference strand on scale attunes researchers to the complexity of implementing and studying designed learning environments across large numbers of participants and contexts. This means *methodological* innovation is needed to drive *conceptual* innovation. The conversation profile analysis we report on here is an example of how these types of innovation can inform each other when studying designed learning environments on a large scale.

Methodologically, indexing discrete characteristics of the dynamic discussions provided a data structure where patterns informed more than the discrete variables could on their own. The Conversation Profile Analysis approach provided a probability model that allowed for the complexity and unpredictable nature of the discussions, but also supported us to find patterns that informed our understanding of the conversations. This is how the *methodological* innovation drove *conceptual* innovation in this project. For example, the students' talk about their invented methods differed in terms of the extent to which they compared different approaches. This distinction was related to mathematical and statistical concepts in epistemically meaningful ways, which is important knowledge that has the potential to inform future teacher support and research. It is possible that these categories may be stable across diverse settings, although more work needs to be done to generate evidence about this question. In addition, the aggregate structure of how the categories unfold across the timeline of a conversation has the potential to inform our understanding of time constraints on teachers and students. For example, the *Inventing & Connecting* category was most likely to occur between 20 and 40 minutes into a whole class discussion, and this latent class was more likely to support students to talk about mathematical and statistical ideas in epistemically congruent ways, this suggests that teachers need to be supported to allot more than 20 minutes of time to carrying out these discussions. This work also suggests many new questions and areas for further work. Are there better indicators to look for during whole class discussions? Can this approach inform the study of other participant structures across large scale implementation? How might we leverage findings such as these to inform future large-scale implementation efforts? Although there are many questions left unanswered, we believe that conversation profile analysis provides a contribution to the challenge of studying designed learning environments on a large scale, and has the potential to inform similar work within the Learning Sciences.

## References


Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching. *Journal of Teacher Education, 59*(5), 389-407.

Bergman, L. R., & El-Khouri, B. M. (1999). Studying individual patterns of development using I-states as objects analysis (ISOA). *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 41*(6), 753-770.

Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Cobb, P., & Jackson, K. (2011). Towards an empirically grounded theory of action for improving the quality of mathematics teaching at scale. *Mathematics Teacher Education and Development, 13*(1), 6-33.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *Journal of the Learning Sciences, 13*(1), 77-103.

Ford, M. J., & Forman, E. A. (2006). Redefining disciplinary learning in classroom contexts. *Review of Research in Education, 30*, 1-32.

Forman, E. A., & Ford, M. J. (2014). Authority and accountability in light of disciplinary practices in science. *International Journal of Educational Research, 64*, 199-210.

Lehrer, R., & Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal, 21*(2), 116-133.

Lehrer, R., Kim, M. J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning, 12*(3), 195-216.

Lehrer, R., Kim, M.-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In J. Confrey & A. Maloney (Eds.), *Learning Over Time: Learning Trajectories in Mathematics Education* (pp. 31-60). Information Age Publishers.

Linzer, D. A., & Lewis, J. B. (2011). *poLCA: An R Package for polytomous variable latent class analysis. Journal of Statistical Software, 42*(10), 1-29. URL http://www.jstatsoft.org/v42/i10/.

Penuel, W. R., Van Horne, K., Jacobs, J. J., & Turner, M. (2018). *Developing a validity argument for practical measures of student experience in project-based science classrooms.* In Annual Meeting of the American Educational Research Association, New York, NY.

Pfaff, E. (2017). *The Role of Teacher Rehearsal in Classroom Mathematics Discourse* (Doctoral dissertation, Vanderbilt University).

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, 10(*4), 313-340.