EXPLORING THE RELATIONSHIP BETWEEN QUALITATIVE LESSON SCORES AND QUANTITATIVE QUALITIES OF INDIVIDUAL CODES

Kathleen Melhuish
Texas State University
melhuish@txstate.edu

Alexander White
Texas State University
whiteale@txstate.edu

Sharon K. Strickland
Texas State University
strickland@txstate.edu

Elizabeth Wrightsman
Texas State University
ewrightsman@txstate.edu

Describing and measuring instructional quality of mathematics lessons is a common goal amongst mathematics education researchers. Such work takes several forms such as classifying and coding instructional moves and student activity or providing high-level rubric-based scores in relation to categories. In this work, we share an innovative mixed methods approach to analyzing lesson data that includes both a time-based classification of instruction and an overall scoring component. Using the Math Habits framework, our project team analyzed a set of 97 fourth-eighth grade mathematics lessons including overall scores. From this qualitative analysis, we developed quantitative models to predict overall scores and better understand the ways that individual codes do or do not contribute to overall lesson score characterizations.

Keywords: Research Methods, Instructional Activities and Practices

In this report, we share recent work aiming to further both our approach to classroom observation tool measures and our understanding of which elements of a classroom are salient in a coding process. This work is situated in a larger validation study focused on the Math Habits Tool (Melhuish, et al., 2020). The Math Habits Tool decomposes a mathematics classroom into four types of codeable activities: teaching routines, catalytic teaching habits, student habits of mind, and student habits of interaction (each of which will be expanded in the next section.) The categories capture teacher and student activity that characterize student-centered, conceptually-oriented classrooms. As in many instruments (e.g., Mathematical Quality of Instruction, Hill, 2014), qualitative coders analyze the lesson at two levels: during the lesson and holistically at the end of the lesson. While coding during the lesson involves identifying time-stamped, individual occurrences, the holistic codes use a rubric-based approached to make a subjective judgement call as to the quality of the teacher and student activity.

We frame our contribution as two-fold. First, we make a methodological contribution -- the development of a quantitative models to estimate overall lesson scores after a qualitative coding process. Notably, we go beyond just using occurrence counts for codes to characterize a class, but also introduce a measure of spread (the degree these occurrences are found at different times in the lesson). We conjectured that although spread was not an explicit portion of the coder's rubrics, it was likely to inform the qualitative evaluations at the lesson level. For example, consider this extreme version. Suppose a classroom has ten rich student contributions, but all occurred within the first five-minute interval. Then the remainder of the class was a lecture. Contrast this situation with a class where student and teachers are interacting, and ten rich student contributions occur throughout the lesson. A frequency-based approach would characterize these two classrooms in the same manner; however, it is unlikely that we would want such classes to be equivalent.

Second, in order to estimate overall lesson scores, we confronted issues of which codes "mattered" and in what ways. Specifically, whether a code occurred spread throughout a lesson sometimes mattered more than how often (and vice versa, along with other combinations). These findings have implications for researchers interested in teaching practices and students' classroom activity.

Background and Framing

Broadly, we take a social cultural approach to the mathematics classroom focusing on social interactions between people in the classroom. Knowledge is co-constructed in these interactions between students and between teacher and students. While we largely assume that individual cognition and social interactions are interrelated (as in Cobb & Yackel, 1996) where individual understanding is developed via social interactions, we focus on the observable social side. Further, we specifically attend to components of classroom interactions that may promote sense-making and mathematical argumentation inclusive of justifying and generalizing. Justifying and generalizing can support the co-construction of mathematical meaning (Brown & Renshaw, 2000; Simon & Blume, 1996) and students' development of conceptual understanding (Staples et al., 2012). Instruction aligned with such goals reflects a standards-based instructional approach (defined in Rubel, 2017 and reflected in standards documents such as the *Common Core*, National Governors Association, 2010)

We use the instructional triangle (Hawkins, 2002) to situate our analytic framing focusing on relationships between teachers, students, and content. We incorporate both Lampert (2001) and Cohen et al.'s (2003) expansion to capture the mediating role a teacher plays in the student-content relationship and the relationships between the students themselves. Figure 1 reflects the components of the BI Framework overlayed on the instructional triangle.

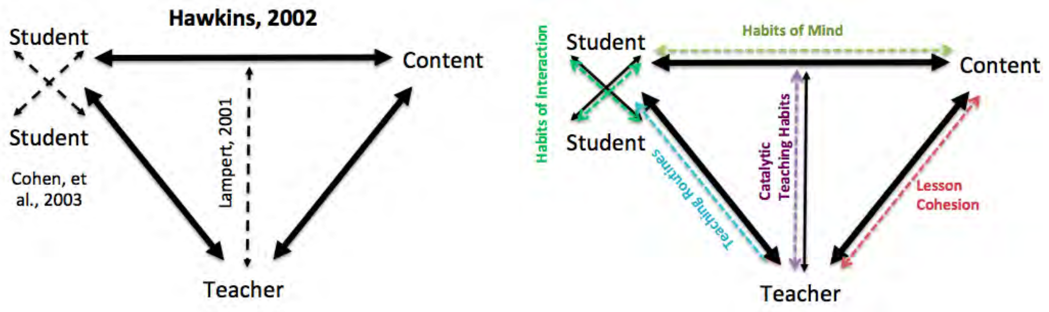

Figure 1: Instructional Triangle and the BI Framework

The blinded framework was developed to operationalize specific instructional routines, moves, and student activity that can be observed within the classroom setting. The coding categories include: Student Habits of Mind which reflect productive ways students engage in mathematics, Habits of Interactions which reflect ways students engage with each other around the mathematics, Catalytic Teaching Habits which capture specific teaching moves that may engender students in engaging with mathematics and each other's mathematical ideas, and Mathematical Productive Teaching Routines. The teaching routines are "recurring, patterned sequences of interaction teachers and students jointly enact to organize opportunities for student learning in classrooms" (DeBarger et al., 2011, p. 244). Unlike the other categories, teaching routines are not identified as instances, but rather over time intervals when they occur. Table 1 includes the categories and subcodes. Each of these categories and subcategories are rooted in the literature on promoting student-centered instruction and mathematical argumentation (e.g., Kazemi, 1998; Staples, 2007; Stein, et al., 2008; Thanheiser & Melhuish, 2022).

Table 1: BI Framework with Variable Names in Parentheticals

Student Habits of Mind	
Mathematical (MathHoM)	Representations; Connections; Mathematical structure
Reflection (ReflectHoM)	Metacognition; Reasoning with mistakes; Making meaning
Capstone (JGHoM)	Justifying; Generalizing
Student Habits of Interaction	
Private Reasoning Time (PRHoI)	Private Reasoning Time
Explaining (ExplainHoI)	Explaining my reasoning; Exploring multiple pathways
Engage with Peer (PeerHoI)	Revoicing; Comparing logic and ideas; Critiquing and Debating
Question (QuestionHoI)	Asking genuine questions
Catalytic Teaching Habits	
Private Reasoning Prompt (ThinkCTH)	Private Reasoning Time Prompt
Sharing Thinking Prompts (ShareCTH)	Prompt to share meaning; Prompt to share thinking; Prompt to share why; Prompt to share representation
Peer Prompts (PeerCTH)	Prompt to analyze strategy; Prompt to analyze mistake; Prompt to compare or connect across strategies; Prompt to revoice or make sense of strategy
Capstone Habit Prompts (JGCTH)	Prompt to justify; Prompt to notice, wonder, or conjecture
Teaching Routines	
Access (AccessTR)	Making meaning of tasks, contexts, and/or language
Public Records (RecordsTR)	Working with selected & sequenced student math ideas Teacher; Working with public records of students' mathematical thinking
Discussion (DiscussionTR)	Orchestrating productive whole class discussions
Groupwork (GroupworkTR)	Structuring mathematically worthwhile student talk; Conferring to understand students' mathematical thinking & reasoning

Methods

Data

This study draws on 96 video-recorded lessons (taken near the end of the school year) from 3 school districts stemming from diverse projects. The samples include 33 lessons from District 1 (Melhuish, et al., 2022), 31 lessons from District 2 (Sorto, et al., 2018), and 33 lessons from District 3 (Kane, et al., 2016). Data on each district can be found in Table 2.

Table 2: Demographic Information on Data Set Districts

District	Race/Ethnicity	Socio-Economic Status	Language

District 1 (grades 4 and 5)	56% White 19% Black/African American, 11% Latino/Hispanic 9% Asian	55% eligible for free and reduced lunch	6% Transitional Bilingual
District 2 (grades 6-8)	99% Latino/Hispanic	95% “economically disadvantaged”	33% Limited English Proficiency
District 3 (grades 4 and 5)	51% Black/African American, 30% White, 13% Latino/Hispanic, 4% Asian	73% eligible for free and reduced lunch	23% Limited English Proficiency

All of these lessons had previously been analyzed with the Mathematical Quality of Instruction (MQI; Hill, 2014) instrument. For the larger databases, we selected a random subset within MQI strata. For District 2, we included all middle school teachers who had opted into recording. We sampled in this manner to assure a range of instructional contexts and practices.

Qualitative Analysis

Each lesson was then coded independently by two researchers according to the BI framework. Any discrepancies were resolved via discussion. After an initial round of coding, the coded lessons were then reviewed by a third member of the research team to identify any coding drift or inconsistences across the coded lessons. Additionally, discrepancies were identified and resolved via discussion. Besides coding using the framework, each coder also assigned an overall rubric-based score for student and teacher activity. Krippendorff’s $\alpha=0.79$ and $\alpha=0.57$ for overall student and teacher, respectively. The levels for overall teaching score are as follows: (1) No evidence of use of Teaching Routines or an attempted teaching routine (but without Catalytic Teaching Habits embedded.) (2) Use of more than one Teaching Routines; some evidence of Catalytic Habits; OR Use of only one Teaching Routine; but many (variety) of Catalytic Habits. (3) Multiple Teaching Routines; Catalytic Habits embedded; (4) Multiple Teaching Routines; Catalytic Habits embedded with pushes towards justifying and/or generalizing. The levels for overall student codes are as follows: (1) Students engaged in at most a Habit of Interaction or two and maybe a Habit of Mind; (2) Students engaged in some Habits of Mind and/or Habits of Interactions (3) Students engaged in multiple Habits of Mind and Habits of Interaction; (4) Students engaged in multiple Habit of Interaction or two and maybe a Habit of Mind with justifying and/or generalizing. As in many rubrics, the overall levels provide some guidance, but also rely on subjective judgements made by the coders. For this reason, consensus was reached through discussion.

Quantitative Analysis

In order to examine how individual codes are associated with the overall Teacher and Student codes, two summary statistics were computed for each lesson and code type: the count and the spread. The count is simply the total number of occurrences of the code during a lesson. To compute the spread, the lessons were partitioned into 10 equal intervals. The spread is the number of intervals in which a code occurred at least once. The two statistics capture the difference between number of times the behavior is observed and the consistency with which it is observed throughout the lesson.

Least Absolute Shrinkage and Solution Operator (lasso) models were used to investigate the relationship between the individual and overall codes. Lasso (James, et al., 2013, pp. 219 - 227) models use l1 regularization to prevent overfitting, reduce the variance of the coefficient estimates of a linear model, and perform variable selection. Unlike stepwise techniques used with standard least squares regression, variable selection in Lasso models does not rely on normality assumptions. The lasso coefficients minimize the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{i=1}^{p}|\beta_j|$$

where the yi is the overall code for the i^{th} lesson, and the xij are the corresponding scaled versions of the count and spread summaries for individual codes. The regularization constant, λ is determined separately for the Teacher and Student models via cross-validation.

Results

In this results section, we share estimates from our models and interpretation; however, the majority of the theorizing and contextualizing of these results can be found in the Discussion section. Recall, our goal is to create a model that can take the human coders' individual timestamped BI codes to predict the human coders' overall student and teacher codes. That is, can we use the detailed BI coding with a model to generate the overall codes? To this end, we needed to examine which (if any) of the individual BI codes were having more or less impact on overall codes. In this section we present what we learned about the role of BI codes in relation to overall codes. We begin with Overall Student codes.

Table 3: Coefficients from the LASSO model for Student Codes

	Model Coefficients	
Code	**Count**	**Spread**
MathHoM	-0.208	0.443
ReflectHoM	0.143	0
JGHoM	0	0.273
PRHoI	0	0.045
ExplainHoI	0.089	0.249
PeerHoI	0	0.060
QuestionHoI	0.120	0

The coefficients for the lasso model for Overall Student code are shown in Table 3. Recall that these habits reflect observable ways that students engaged with the mathematics and with each other mathematically. The zero coefficients for the count variables for JGHoM (capstone habits of justifying and generalizing) and PRHoI (private reasoning time), and the spread variable for QuestionHoI (asking genuine questions) indicate that lasso dropped those respective count or spread variables from the model. We use *z*-scores to interpret the coefficients of the remaining predictors. For example, for a lesson with one standard deviation more ReflectHoM (reflection habits of mind) than the average lesson, the predicted overall code increases by 0.143. We can unpack the slightly more complex case of the mathematical habits of mind (MathHoM) coefficients where we see a negative relationship. Consider two lessons where the count for MathHoM differs by one. If the additional code occurs in an "empty" interval, the count and the spread both increase and the predicted overall code increases. On the other hand, if the additional code occurs in an interval where MathHoM has already been observed, the predicted overall code decreases. That is, observing the code throughout the lesson is more beneficial than simply counting a total. On the other hand, some codes only mattered in terms of count. For example, students asking a genuine question (QuestionHoI) is associated with an increase in the overall code, no matter where it occurs. That is all to say, that some codes matter where they occur in a lesson (indicated by spread) and some only matter how often they occur (indicated by count).

Table 4: Coefficients from the LASSO model for Teacher Codes

	Model Coefficients	
Code	**Count**	**Spread**
GroupworkTR	0.280	0.096
RecordsTR	0	0.241
DiscussionTR	0	0
AccessTR	0.253	0
ThinkCTH	0	0.058
ShareCTH	0	0.104
JGCTH	0.144	0
ReflectCTH	0.139	0
PeerCTH	0	0
RevoiceCTH	0.056	0.097

Table 4 provides the coefficients for the lasso model for the Overall Teacher codes. Again, we can notice that different types of activities are differently related to the overall codes. The teaching routines related to groupwork (GroupworkTR) matter both in terms of how frequently they occur (counts) and how they spread throughout the lesson (spread). In contrast, the teaching routines related to public records (RecordsTR) were only significant in terms of how spread they were throughout the lesson and the meaning making teaching routine (AccessTR) was significant only in terms of frequency, not spread. If we turn to individual teaching moves (the catalytic teaching habits), we can see spread is significant for prompts related to private think time and to share thinking (ThinkCTH; ShareCTH), but overall frequency, but not spread, is significant for reflection prompts (ReflectCTH). For teacher revoicing (RevoiceCTH), both frequency and spread were small, but positive predictors of overall score. Finally, we note that the orchestrating discussion routine (DiscussionTR) and the prompts to engage with peers' reasoning (PeerCTH) did not contribute to predicting overall teacher scores.

We also briefly share results about how closely these models fit our overall scores. Table 4 and Table 5 present a crosstabulation comparing the true and predicted codes. In grey, we have emphasized the lessons where the coder overall code matched the predicted code. For the students, the model correctly predicts 71.1% of our lessons. For the overall teacher score, the model correctly predicts 76.77% of our lessons. Further, only one lesson in each case is predicted more than one level off. We can calculate Krippendorf's α, as we would when comparing coders. For the overall teacher scores, α=0.879 and for the overall student scores α = 0.813. Both numbers are over 0.80 indicating substantial agreement. That is, our models are doing a relatively good job predicting the overall scores arrived at by coders.

Table 5: Crosstabulation of Overall Student Codes: True vs. Predicted

		Predicted Overall Code				
		1	2	3	4	total
Coder Overall Code	1	16	10	0	0	26
	2	2	24	6	0	32
	3	0	6	13	0	19
	4	0	1	4	15	20
Total		18	41	23	15	97

Table 6: Crosstabulation of Overall Teacher Codes: True vs. Predicted

		Predicted Overall Code				
		1	2	3	4	total
Coder Overall Code	1	24	7	0	0	31
	2	1	26	3	0	30
	3	0	4	12	1	17
	4	0	1	6	12	19
Total		25	38	21	13	97

Discussion

When we code data, we often make (explicit or implicit) inferences based on quantitative characteristics such as frequency of a code. However, when we consider a complex setting like a mathematics classroom, we often rely on subjective judgement calls about other features such as the degree it feels like a particular activity characterizes a lesson. In fact, this sort of expert judgement is why qualitative coding can be powerful. Yet, we have found that it can be quite challenging to come to agreement on lesson level scores because, by nature, coders are not noticing or perhaps not weighting elements of instruction in the same way. Our goal with this study was to develop quantitative means to estimate overall scores. We focused on the measure of spread in addition to count to avoid collapsing some of the dimensionality in a classroom.

If we consider our results, we can see that in some cases spread was more important, some cases count, and in yet others, they played out in more complex ways. We return to a couple of examples to conjecture what might account for these differences. If we turn to the overall teacher scores, we can see that both orchestrating discussion and prompts to engage with peers' ideas did not contribute to predicting overall scores. First, these are types of codes that are theoretically related. In order to orchestrate discussion, we required that multiple students are engaged with each other's ideas in some way. This typically occurs when teachers make related prompts for engagement. From a simplistic view, the rubric would reward both types of activities with higher overall scores. However, if we examine the data, we can note two features that may account for this result. First, the DiscussionTR and the PeerCTHs were only meaningfully different for teachers with an overall high score. For example, the mean spread for the DiscussionTR for overall high teachers was 3.1 (meaning, on average, discussion happened in 3 of 10 intervals) and mean count was 6.9 (meaning on average discussion happened 7 times per lesson). In contrast, spread was less than one for all other levels of overall score and less than two for mean count. Rather than gradual increases, this routine served to discriminate between high level lessons from other levels. This leads to the second point, the high-level lessons all had higher spreads and counts in other categories. Theoretically this makes sense. It is likely that in a class where a teacher orchestrates discussion, there is overlap with other teaching routines like using public records of students thinking – in many cases the discussion is about such a record. Thus, these codes are not contributing new information about the overall teaching in the lesson.

Now let's contrast two teaching routines that both were significant but in different ways. Working with public records of student thinking and selecting and sequencing (RecordTR) mattered in terms of spread whereas Making meaning of task and terms (AccessTR) mattered in terms of counts. In this case, we conjecture the way these routines operate in the classroom may account for the difference. Engaging students in making meaning around tasks and terms is a routine that comes up when students encounter an idea, task, or piece of language in which they may be unfamiliar. This is likely to occur at specific points in the lesson such as when a task is launched. In contrast, working with records of students' ideas may be threaded throughout. Thus, the number of occurrences of meaning making may be more salient than the spread of meaning making; while the spread of record use is likely a more salient feature of overall teaching.

If we turn to the student codes, we see the relationship between observed math habits of mind (including capstone habits) and overall scores are more linked to spread. These habits include things like reasoning with representations, structure, connections, and justifying and generalizing. Frequencies alone may paint a misleading picture because a single student contribution may embed many of these habits (e.g., justifying a result by use of a pattern within a table). A short span of time with high counts is not as meaningful as occurrences spaced throughout a lesson (reflected in spread), thus theoretically spread is likely to be more salient. A second explanation may be that at a certain threshold, frequency might not contribute new information. That is, a class with 20 instances and a class with 30 instances of habits of mind are both likely at a high level and the difference of 10 instances does not contribute something new. In contrast, the spread is capped at the interval number and any difference has the potential to provide meaningful information that characterizes a lesson across time.

For space limitations, we will not unpack all of the differences in how the codes are operating but will spend a brief amount of time comparing the difference in the ExplainHoI (students explaining their thinking) and the MathHoMs. Explaining mattered in terms of both count and spread, although with a relatively small coefficient for count. The threshold to explain one's thinking is much lower than the threshold for that thinking to include math habits of mind (which reflect higher level reasoning). Explaining was by far the most frequently observed student activity at all levels of overall student code. The mean number of occurrences of explain was 6.4 for the Overall Student =1 classes and 39.8 for the Overall Student=4 classes. However, for lessons that received the lowest overall score, the mean spread was less than 3 intervals whereas in highest lessons, it was nearly 9 intervals. Both frequency and spread appear to provide important information to characterize a lesson.

This leaves several questions open for future research. First, are these relationships a consequence of the coders and rubrics or a consequence of how these activities unfold in the classroom? This is work that could be further addressed with additional qualitative analysis of the classrooms as well as attention to classrooms where overall codes as assigned by the model diverged from overall code as determined by the qualitative coders. Second, how might we develop more accurate models and measures when considering coding at this grainsize? Spread and count provided a start, but other measures such as interrelated (different types of codes within a timespan), existence (binary), or alternative models (such as non-linear models) could lead to additional insights.

Acknowledgments

This research was supported by the National Science Foundation grant DRL-1814114.

References

Brown, R., & Renshaw, P. (2000). Collective argumentation: A sociocultural approach to reframing classroom teaching and learning. In H., Teoksessa Cowie & G. van der Aalswoort (Eds.), *Social interaction in learning and instruction. The meaning of discourse for the construction of knowledge*. Amsterdam: Bergamon Press.

Cobb, P., & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational psychologist*, 31(3-4), 175–190.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). *Resources, instruction, and research. Educational evaluation and policy analysis,* 25(2), 119–142.

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Schank, P. (2011). Teaching routines to enhance collaboration using classroom network technology. In *Techniques for fostering collaboration in online learning communities: Theoretical and practical perspectives* (pp. 224-244). IGI Global.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

Hawkins, D. (2002). I, thou, and it. In The informed vision: Essays on learning and human nature (pp. 51–64). Algora Publishing.

Hill, H. (2010). Mathematical Quality of Instruction (MQI). Coding tool. Lansing, MI: University of Michigan.

Kazemi, E. (1998). Research into Practice: Discourse That Promotes Conceptual Understanding. Teaching Children Mathematics, 4(7), 410-414.

Lampert, M. (2001). *Teaching problems and the problems of teaching.* Yale University.

Landis, J.R.; Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics. 33 (1): 159-174.

Melhuish, K., Thanheiser, E., Heaton, R., Sorto, A. Strickland, S., & Sugimoto, A. (2020). The
Math Habits Tool - Research Version [Measurement instrument] Retrieved from http://mathhabits.wp.txstate.edu

Melhuish, K., Thanheiser, E., White, A., Rosencrans, B., Foreman, L., Shaughnessy, J. M., Riffel, A. & Guyot, L. (2022). The Efficacy of a "Mathematics for All" Professional Development. *Journal for Research in Mathematics Education*.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common core state standards for mathematics. Washington, DC: Authors.

Rubel, L. H. (2017). Equity-directed instructional practices: Beyond the dominant perspective. *Journal of Urban Mathematics Education,* 10(2).

Simon, M. A., & Blume, G. W. (1996). Justification in the mathematics classroom: A study of prospective elementary teachers. The Journal of Mathematical Behavior, 15(1), 3–31.

Sorto, M. A., Wilson, A. T., & White, A. (2018). Teacher knowledge and teaching practices in linguistically diverse classrooms. In *Language and communication in mathematics education* (pp. 219-231). Springer, Cham

Staples, M. (2007). Supporting whole-class collaborative inquiry in a secondary mathematics classroom. *Cognition and instruction*, 25(2-3), 161-217.

Staples, M. E., Bartlo, J., & Thanheiser, E. (2012). Justification as a teaching and learning practice: Its (potential) multifacted role in middle grades mathematics classrooms. *The Journal of Mathematical Behavior,* 31(4), 447-462.

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.

Thanheiser, E., & Melhuish, K. (2022). Teaching routines and student-centered instruction: a case of divergent lessons. [Manuscript submitted for publication].