Fine-grained, nomination coding in the support domain: Promising teacher discourse measures in mathematics instruction

Your full name: Sean Kelly

Affiliated authors with institutions: Sean Kelly, University of Pittsburgh; Hadassah Muthoka, Motivate Lab; Erin Vines, University of Florida; Stephanie Wormington, Center for Creative Leadership; Sidney D'Mello, University of Colorado-Boulder

Affiliation: Department of Educational Foundations Organizations and Policy, School of Education, University of Pittsburgh; Motivate Lab, Charlottesville, VA; Department of Family, Youth, and Community Sciences, University of Florida; Center for Creative Leadership, Greensboro, NC; Institute of Cognitive Science & Department of Computer Science, University of Colorado Boulder

Current position of Sean Kelly: Professor, Department of Educational Foundations, Organizations, and Policy, School of Education, University of Pittsburgh

Fine-grained, nomination coding in the support domain: Promising teacher discourse measures in mathematics instruction

Abstract

Contemporary teacher observation protocols offer comprehensive portraits of instruction in mathematics and other subjects and are well designed to provide basic accountability of effective practice. Yet, recent research argues that in order to drive instructional improvement, data on instructional practice are needed that more carefully quantify instructional processes. In this study we report results from a novel coding of the Measures of Effective Teaching Study data that offers evidence on a set of teacher discourse measures in the domain of teacher support, broadly speaking. Several of these measures have promising features, including high lesson- and teacher-level variability, convergent and discriminant validity with existing protocols, and association with achievement growth.

Extended summary

This study focuses on the development of fine-grained measures of instruction instantiated in teacher discourse. We envision these measures as broadly useful in the analysis of classroom instruction, including uses for professional development in a variety of pre- and in-service contexts where analysis of instruction can support teacher learning, not necessarily as narrowly tailored for a particular use. Discourse indicators feature prominently in global protocols as well, but here we measure discourse at the level of the individual utterance. Likewise, we investigate teacher discourse in the support domain

(Shernoff, 2013; Shernoff et al., 2016), broadly speaking, which overlaps considerably with indicators and constructs from global protocols such as TRU or CLASS, but again, with an utterance-level specification. As in Shernoff's framework, our "support" measures pertain principally to socio-emotional supports that address students need for competence (Newmann, Whelage, & Lamborn, 1992) as well as social comparison and equity in heterogenous classrooms (Cohen & Lotan, 1997), although we also include learning strategies. Is supportive discourse a promising candidate for the focus of a fine-grained program of research on instructional practice? Might for example, such measures feature prominently in the rapidly developing field of automated measurement of classroom instruction (see e.g., Jacoby et al., 2018; Jensen et al., 2021; Kelly et al., 2018; Liu & Cohen, 2021; Olshefeski et al., 2020; Ramakrishnan et al., 2021; Watson et al., 2021)?

In this new empirical study, we focus on four paired discourse constructs: **Public Evaluation** of behavior, or valence of evaluation (Praise vs. Admonishment); **Autonomy Support** (vs. controlling language); **Strategy Suggestion** (vs. lack of strategy); and **Learning Mindset** supportive discourse (vs. mindset undermining discourse). These constructs were identified inductively during exploratory coding, guided by both general theories of engagement, discourse, and learning in heterogenous classrooms previously cited, as well as conceptual frameworks in mathematics including the Mathematics Scan in particular (Carpenter & Lehrer, 1999; Langer-Osuna, 2017; Ottmar et al., 2015).

Results are based on a new coding of 156 grade 6–8 math videos from 73 class sections (also 73 teachers; 2.14 video segments per teacher) in the Measures of Effective Teaching (MET) Study (MET Project, 2012, 2014). Data newly coded for this project was merged with existing data including student and teacher assessments, classroom observations

coded with multiple protocols and student perception data from more than 1400 students nested within sample class sections.

We used a nomination-coding process to identify teacher utterances exhibiting a given discourse property, as opposed to an exhaustive coding process where all speech is transcribed, segmented, and coded. After nominating an utterance, the teacher speech was entered into an excel database, along with the code and time-stamp (to facilitate future automated analyses). This approach yields count data, for example, the number of instances of public praise (per 15-minute interval of time), and ratio data within each of the four discourse domains (e.g., the ratio of praise to admonishment, a relative measure of public evaluation). Coding reliability was verified at both the utterance and observation level in an initial sample of 15 double-coded observations (full statistical results will be presented).

To determine whether these constructs showed promise for incorporation into fine-grained systems of teacher observation, we investigated multiple properties including: prevalence, overall lesson-to-lesson variability and teacher-to-teacher variability in particular, internal consistency, convergent and divergent validity with global protocols, and associations with achievement scores. Overall we find that three of these constructs, Evaluation, Autonomy Support, and Learning Mindset Support, appear to hold the most promise for further research. Each of these were prevalent, varied substantially from teacher to teacher, and demonstrated convergent validity with the CLASS protocol. These findings concern central aspects of functionality in discourse measures for instructional research and evaluation, although these features are not necessarily required for certain uses. For example, even measures with relatively low prevalence rates could still be useful in instructional improvement efforts, especially if these measures served as a "tip of the iceberg," being revealing of larger instructional norms, or if when they did occur, their influence was

especially strong. Additionally, in these data Autonomy Support was associated with achievement on a more "authentic" math assessment.

We conclude with a call for further development of fine-grained discourse-based measures of instruction, and by distinguishing between implications for use in research and practice. Concerning further development, we emphasize several design features which might be altered in future study. First, a critical feature of this study was the nomination coding approach. This feature streamlined coding, allowing us to efficiently code more discourse constructs, but may be less precise than more intensive coding. Second, while we are relatively confident that these measures vary substantially from teacher-to-teacher, a study featuring more intensive study (more observations) for each participant, would sharpen this understanding. Finally, because MET did not visually roster/map classrooms with student ids, we could not match teacher utterances to individual participants. Considering the many research design possibilities in fine-grained research on instruction, together with the inherent complexity of conceptual models of learning, the continued collection of high-quality instructional data like MET, that can then be flexibly analyzed, will be critical to advancing the educational sciences.

Finally, we must distinguish between implications for research and use in practice. For example, Goldring et al. (2015) find that even as global protocol scores are often unreliable and clustered in the middle of the scale, principals report that the observational frameworks themselves both focus teacher attention and reflection on important domains of instruction and enhance professionalization through a shared technical language. Similarly, we must assess the various uses and use-value of fine-grained observational systems. For example, we know such systems have been put to productive use in teacher education and professional development settings (Caughlan et al., 2013; Lehesvouri et al., 2017; Reznitskaya & Wilkinson, 2021; Sherry, Messier-Jones, & Morales, 2018). Yet, less is

known about how less structured use of such tools, for example, how teachers might use fine-grained, automated systems. As fine-grained measures are evaluated across a range of uses, their possibilities and limitations will be brought into sharper focus.

References

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. *Mathematics classrooms that promote understanding*, 19–32.

Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S., & Fine, J. G. (2013). English teacher candidates developing dialogically organized instructional practices. *Research in the Teaching of English*, *47*, 212–246.

Cohen, E. G., & Lotan, R. A. (1997). *Working for equity in heterogeneous classrooms: Sociological theory in practice*. New York: Teachers College Press.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value-added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*, 96–104.

Jacoby, A. R., Pattichis, M. S., Celedon-Pattichis, S., & LopezLeiva, C. (2018). Context-sensitive human activity classification in collaborative learning environments. *IEEE Southwest Symposium on Image Analysis and Interpretation* (*SSIAI*), pp. 1–4).

Jensen, E., Pugh, S. L., & D'Mello, S. K. (2021). A deep transfer learning approach to modeling teacher discourse in the classroom. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 302–312.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47, 451–464.

Langer-Osuna, J. M. (2017). Authority, Identity, and Collaborative Mathematics, *Journal for Research in Mathematics Education*, *48*, 237–247.

Lehesvuori, S., Hahkioniemi, M., Jokiranta, K., Nieminen, P., Hiltunen, J., & Viiri, J. (2017). Enhancing Dialogic Argumentation in Mathematics and Science. Studia Pedagogica, 22, 55-76.

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*. Online first.

MET Project (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains*. Bill and Melinda Gates Foundation.

MET Project (2014). *Measures of Effective Teaching: 1 – Study Information, User Guide*. Bill and Melinda Gates Foundation.

Newmann, F.M., Wehlage, G. G., & Lamborn, S. D. (1992). The significance and sources of student engagement. Pp. 11-39 in F. M. Newmann (Ed.) *Student engagement and achievement in American secondary schools*. Teachers College Press.

Olshefski, C., Lugini, L., Singh, R., Litman, D., & Godley, A. (2020). The discussion tracker corpus of collaborative argumentation. *Proceedings of the 12th Conference on Language Resources and Evaluation* (*LREC 2020*), 1033–1043.

Ottmar, E. R., Rimm-Kaufman, S. E., Larsen, R. A., & Berry, R. Q. (2015). Mathematical knowledge for teaching, standards-based mathematics teaching practices, and student achievement in the context of the responsive classroom approach. *American Educational Research Journal*, *52*, 787–821.

Ramakrishnan, A., Zylich, B., Ottmar, E., LoCasale-Crouch, J., & Whitehill, J. (2021). Toward automated classroom observation: Multimodal machine learning to estimate class positive climate and negative climate. *IEEE Transactions on Affective Computing*.

Reznitskaya, A., & Wilkinson, I. A. (2021). The Argumentation Rating Tool: Assessing and supporting teacher facilitation and student argumentation during text-based discussions. *Teaching and Teacher Education*, *106*, 103464.

Sherry, M. B., Messier-Jones, L. M., & Morales, J. (2018). Positioning in prospective secondary English teachers' annotations of teaching videos. *English Teaching: Practice & Critique*.

Shernoff, D. J. (2013). *Optimal learning environments to promote student engagement*. New York: Springer.

Shernoff, D. J., Kelly, S., Tonks, S. M., Anderson, B., Cavangah, R. F., Sinha, S., & Abdi, B. (2016). Student engagement as a function of environmental complexity in high school classrooms. *Learning and Instruction*, *43*, 52–60.

Watson, G., Youngs, P., van Aswegen, R., & Singh, S. (2021). Automated classification of elementary instructional objects and activities: Analyzing consistency of manual annotations. Paper presented at the 2021 annual meeting of the *American Educational Research Association* (April, online pandemic accommodation).