Tools for Observing Productive Talk: A Comparison of Two Protocols (RTOP/IQA-SOR)

Authors

Enderle, P., Hagan, C., Morandi, S., Coker, R., Kásper, V., Rhemer, D. M., Schellinger, J., & Southerland, S.A.

Abstract

As part of a larger study focused on supporting high school biology teachers' use of productive science talk, this study compares the use of two different observation protocols, the RTOP and the IQA-SOR. Reviewing a year-long data set of video observations collected from classrooms of teachers participating in the larger professional development study, the two validated instruments produced significantly correlated scores of different scales based on the unique structure of each tool. We posit this demonstrates that both instruments can be useful for analyzing classroom instruction intended to emphasize productive science talk. However, the instruments do possess unique structural and theoretical qualities that warrant this study to understand the insights afforded by each. The similarities and differences emerging from each are explored in the presentation and how they impact the analyses. These considerations can be helpful for scholars who research in-service teacher learning as classroom implementation and impact on student learning activities are general outcomes that most professional development research endeavors to explore. Further, considerations of what a particular observation protocols' foci include will be necessary so that continued research on teacher learning works to make science learning through discourse accessible to all learners.

Subject/Problem

Scientists engage in constructing explanations and arguing from evidence as they develop scientific knowledge and models. Accordingly, the *Framework for K-12 Science Education* (NRC, 2012) emphasizes that in order to learn science, students must experience the process of developing explanations of phenomena, learning that requires a focus on productive epistemic discourse in science classes (Kelly, 2014). "Productive epistemic talk" involves classroom interactions that position students to collaboratively create meaning as they construct explanations, engage in argument from evidence, and evaluate and communicate information—work involving talk, joint attention, and shared activity aimed at the construction and critique of ideas (Ford, 2008). Engagement in productive epistemic talk is an essential aspect of learning science.

Much of the work of science education relies on the notion that teaching is a critical factor in shaping students' learning (Tekkumru-Kisa et al., 2021). Given this, it should come as no surprise that as we strive to shape science instruction to become more effective and equitable for all, the events that occur between teachers and students during the course of instruction become essential. To understand the factors that best support productive epistemic talk in science classrooms, it is imperative to recognize that such insight can only be as good as the descriptions they are based on. Extensive studies involving a wide array of images of teachers, students, and their classroom communities have provided much insight into supportive factors of such classroom discourse (Michaels & O'Connor, 2012). Building on those deep insights, research teams have developed an array of research tools to account for those factors in various ways (Sampson et al., 2011; Marshall et al., 2010; Piburn et al., 2000).

Historically, science educators have used a multitude of methods to understand what is happening in the classroom. As described by Anwar and Menekse (2020), course evaluations, performance assessments, surveys, instructor and student interviews, or classroom observations captured by videos, detailed field notes, or researcher-completed rubrics are some of the tools used to evaluate student and teacher behaviors and classroom dynamics. Classroom observations are the most well recognized approach to closely describing what occurs in classrooms, but they are also labor intensive. There is wide variation even within this particular approach to classroom research, as some observation protocols are open-ended and others provide more structured observations, guided by predetermined sets of criteria. Each of the available observation protocols stem from particular theoretical frames and focus on different aspects of classroom events to generate a systematic description (Anwar & Menekse, 2020). While Anwar and Menekse (2020) speak to the need to carefully partner research questions with a well-paired observation protocol, limited research exists to inform researchers' choice of the appropriate measure to use for a particular study.

This proposal emerged as a problem of practice in ongoing research. Contemplating a large field study of teacher and student learning around the use of productive epistemic talk in science, we were faced with the need to select an observation protocol that would allow for a quantitative comparison of multiple science classrooms across multiple school districts examined through the lens of video and audio recordings. We were also faced with the need to select a measure that

was already established in the literature so that others could understand the descriptions of classroom activities that our work generated. The purpose of this research is to report the results of our comparisons of two observation protocols: the well-established and widely used Reformed Teaching Observation Protocol (RTOP, Piburn et al., 2000) and the more recently developed, Instructional Quality Assessment-Science Observation Rubric (IQA-SOR, Tekkumru-Kisa et al., 2021). We explored their suitability for study of teachers' support of productive science talk.

Research questions include:

1. How do the RTOP and IQA-SOR observation protocols compare to each other?
2. What sorts of descriptions are allowed by each of the protocols?
3. What are the affordances and limitations of each to allow a focus on factors related to productive epistemic discourse?

Design/Procedures

Overview, Context & Participants: Data for this study comes from the first year of a four-year NSF funded professional development (PD) project. The PD was designed to foster science teachers' support of student sensemaking through productive science talk. Teachers underwent a 36-hour summer PD and four cycles of collaborative design including *design, teach, analyze* sessions during the school year. Each cycle had its own theme (i.e., the role of anchoring phenomena, using student ideas and reasoning, the role of evidence, and using student ideas towards the end goal) and occurred over approximately 6 hours. Our study focuses on four of these teachers: Monica, Jerry, Kate, and Danny, who were selected because they continued with PD for the cycles of collaborative design. See Table 1 for information about the teachers.

Table 1. Information on Participating Teachers

Teacher	Course Observed	Years Teaching	# of Observations
Monica	Advanced Placement Biology	13	17
Jerry	Middle School Biology	3	17
Kate	Middle School Biology	21	18
Danny	Advanced Placement Chemistry	5	12

Instruments: The **RTOP** was developed by Piburn and colleagues (2000) to measure 'reformed' teaching and practice (Piburn & Sawada, 2000; Sawada et al., 2002), and has been used in a host of mathematics and science classrooms at both the K-12 and postsecondary levels. The RTOP calls for a holistic evaluation of instruction across a lesson by making extensive field notes, then assessing instructor behaviors along 25 items in 5 subscales: (1) lesson design/implementation, (2) propositional knowledge, (3) procedural knowledge, (4) classroom culture, and (5) student teacher relationship (Piburn & Sawada, 2000; Sawada et al., 2002).

The **IQA-SOR** is patterned after the development of the Instructional Quality Assessment in ELA and mathematics, which was designed to provide statistical and descriptive information about the quality of instruction with respect to academic rigor (Boston, 2012; 2014; Matsumura et al., 2008). The IQA-SOR focuses on the extent of students' engagement in rigorous tasks (i.e, lessons) and talk that shapes their science thinking and sensemaking. The IQA-SOR examines

the rigor and talk found in different points of the task including in the potential of the design (R1), in the launch/framing (R2), in the implementation or enactment(s) of the work of the task (R3), and in student discussion at the close of the task (R4, Tekkumru-Kisa et al., 2021).

Data collection: Classroom video recordings of lessons from each of the cycles were collected and analyzed, with the exclusion of Monica's cycle 1 lesson of which all data was lost. One team of four coders used the RTOP to code all the classroom videos. Coders calibrated using data from the instrument developers (Piburn et al., 2000), coded lessons together to reach an interrater reliability of 80%, and then individually coded the remainder of lessons. The coding team developed their own codebook for each of the 25 items that comprise the 5 RTOP subscales describing hallmark features of the lesson with scores for each item ranging from 0 to 4.

Another coding team of three coded the same classroom videos using the IQA-SOR. Coders independently coded each task and came together to reach consensus on each rubric. When consensus could not be reached, scores were discussed with IQA-SOR instrument developers. Once the team reached an interrater reliability of 80% they met only to discuss uncertainties in rubric scores. The IQA-SOR looks across four separately scored rubrics (i.e, R1, R2, R3, and R4) with scores from 5 to NA based on developer-created criteria for each score level. These criteria are built upon a range of lesson characteristics identified from research to be impactful elements necessary for creating rich and rigorous science learning experiences. Score options of NA only occurred in R4 and represent the absence of a summary discussion. For the purposes of the quantitative analysis all NAs were converted to 0. Scores of 0 in this rubric represented that the substance of the discussion was not related to the scientific content or the task. We feel that converting NAs to 0s was appropriate because a discussion void of scientific content or not related to the task sends a similar message to that of not having a discussion. Multiple enactments could occur in the implementation phase (R3) of the task. Each enactment was scored individually and a holistic score for the implementation phase was generated to represent all the enactments of the phase.

Data analysis: Once corollary score sets were generated for the same observation data set, the scores and accompanying researcher notes were compared quantitatively and qualitatively. For quantitative comparison, we compared the overall scores generated from the RTOP protocol for each teacher's lesson to the overall scores generated from the four IQA-SOR rubrics by generating a Pearson's correlation coefficient. To further explore the two instruments, the coding teams met to align specific RTOP items to three of the four IQA-SOR rubrics (R2, R3, R4), which were then compared by calculating Pearson's correlation coefficients for each set. This alignment work focused on looking for similarities between the RTOP item and codebook content and the IQA-SOR predetermined elements for each scoring level. IQA-SOR R1 was excluded from this analysis because it was related to the design and potential of the task, a characteristic not measured in the RTOP. This analysis primarily addresses the first research question. For the qualitative comparisons of the two instruments, the researchers' notes used to determine scores were analyzed along with researcher memos generated by each team documenting particular issues the team experienced while scoring. A basic thematic analysis

involving collaborative, iterative coding among the research team aimed to develop insights addressing the latter two research questions.

Findings & Analysis

Quantitative Comparisons: Comparing the quantitative scores generated by each instrument demonstrates that both instruments produce significantly correlated scores (See Tables 3). Table 2 provides information regarding the alignment of the three IQA-SOR rubrics to specific RTOP items. Again, these alignments were developed by the coding teams comparing their respective codebooks to determine where significant overlap occurred among items. Resulting from the observation features targeted by the codebook for a specific item, several RTOP items aligned with more than one IQA-SOR rubric. Several RTOP items did not align with any of the IQA-SOR rubrics. Review of the codebooks and researcher memos attributed much of this lack of alignment to the more ambiguous nature of the RTOP stem. For example, RTOP #18, "There was a high proportion of student talk and a significant amount of it occurred between and among students," does not inherently specify the type of talking occurring, whether sensemaking or procedural discussions. Further, the evaluation of this item, as structured, requires accounting for talk across the entire lesson, whereas IQA-SOR R4 focuses on intentional sensemaking talk done by students in a whole class format at the end of the task. Thus, strong alignment with a particular rubric was difficult with some RTOP items due to the more open interpretation of those items.

Table 2. Alignment of IQA-SOR Rubrics and RTOP Items

IQA-SOR Rubric	RTOP Items
R2 - Task Launch/Framing	#1, 3, 4, 5, 10, 17
R3 - Implementation of the Task	#2, 7, 10, 12, 13, 15, 21, 22, 24
R4 - Students' Discussion after Small Group	#2, 5, 11, 12, 15, 16, 19, 22

Table 3 provides the Pearson correlation coefficients for the comparison of the total scores from each instrument and the rubric specific data sets. For most of these comparisons, the results show significantly strong correlation between the instruments, offering support that they can similarly capture important classroom dynamics. However, the R4 comparison shows a relatively weak correlation, which resonates with the interpretation issue alluded to previously, as several '0' scores were given because they did not include a summary whole class discussion.

Table 3. Pearson's Correlation Coefficients for IQA-SOR/RTOP Comparison

Comparison Type	Pearson's *r*	*p* Value
IQA-SOR Total x RTOP Total	0.659	0.005
IQA-SOR R2 x RTOP R2 Items	0.721	0.002
IQA-SOR R3 x RTOP R3 Items	0.805	<0.001
IQA-SOR R4 x RTOP R4 Items	0.290	0.277

Comparing Instrument Descriptions: To demonstrate the different descriptions produced by these tools, we offer an example from Kate's middle school honors biology class with 29 students. The focus of the lesson entailed the decline of salt water fish populations. Kate

introduced the lesson with the phenomenon about the impact of Lionfish, an invasive species to the east coast of the U.S., on saltwater fish populations. Students analyzed data to develop an argument in response to the guiding question: Is our saltwater fish population declining? If so, what policies would be most effective in slowing that decline? (Sampson & Schleigh, 2013). Students engaged in developing analytical procedures for a provided data set, constructed evidence and a justification for their claim, and participated in an argument evaluation session. Kate engaged students in mainly small group discussions using several talk moves like "What evidence do you have?" and "How do you know that?" (Michaels & O'Connor, 2012).

Kate scored 94 out of 100 on RTOP. For the RTOP analysis,we determined evidence concerning the tasks given to students, specifically in lesson design and implementation, propositional knowledge, and procedural knowledge. Lessons that elevated student agency in the task were allotted higher scores, such as the 4's scored on items focused on "student exploration preceding formal presentation" and "students were encouraged to generate conjectures...ways of interpreting evidence" for Kate's lesson. The areas of classroom culture and student teacher relationship were elaborated to include aspects of equity. For instance, lessons that gave space for all students to discuss their ideas in multiple settings scored high. Kate scored 4's on items focused on "intellectual rigor, constructive criticism, and the challenging of ideas were valued" and "communication of their ideas to others using a variety of means and media" as students consistently engaged in various scientific practices (analyzing data, arguing from evidence) working with several representations to construct and critique ideas. Kate's use of various talk moves worked towards a more inclusive classroom, resulting in scores of 4 on items concerning "The teacher's questions triggered divergent modes of thinking." and "Student questions and comments often determine the focus and direction of classroom discourse." If a student hesitantly participated, Kate used talk moves to engage them in the discussion. She also worked to connect students' ideas to each other through those questions throughout the lesson.

Kate scored 11 out of 20 on the IQA-SOR. As designed, the task had some potential to engage students in rigorous thinking (R-1, Score: 3). However, scaffolds in the task and the design of students' work limited some of the opportunities for students to engage in science practices and content together. For instance, some scaffolds in the task implementation constrained the rigor as Kate guided students' thinking towards the ideas embedded in the task. Kate's framing engaged her students in high level thinking by preparing them to address a phenomenon based guiding question (R-2, Score: 4). In addition, she elicited students' ideas and thoughts around human impact and policy on the environment which primed engagement in the activity. Kate continued this support for students' rigorous thinking as they engaged in the science content and practices together (R-3, Score: 4). We coded her use of various instructional moves in small groups to invite students to engage in both the intellectual work of science while also developing a sense of how scientific knowledge is produced. Kate did not facilitate a whole class discussion, so we were unable to assess the extent to which students showed and explained their work to reflect their thinking and sense-making (R-4, Score: 0).

Affordances and Limitations: These protocols were selected for comparison because they represent a well established tool and a newly validated tool that both assess the nature and

quality of classroom activities including productive science talk. This alignment is supported by the strong correlations between the scores. Considering the structural characteristics of each instrument, the RTOP offers a stronger focus on instructor behaviors, across multiple classroom activities in a lesson, with talk being central to many items while instructional activities are considered more generically. However, the need for elaborating and calibrating protocol items can also surface issues with distinguishing types of talk and activity that may lessen the precision of the scores. The IQA-SOR shifts focus towards the observable rigor in students' thinking, primarily through the tasks that teachers and students are engaged in and the productive science talk among them. Yet, researchers may run into scoring issues if some of the targeted rubric elements are ignored or do not manifest in specific ways, as happened in our data set with all of the teachers' first lessons that did not include whole class debrief discussions. Separated by almost 20 years of scholarship, the two instruments are grounded in related but different visions of effective science classrooms. The emphasis on science classroom discourse that RTOP represents has been further elaborated through extensive research (Kelly, 2014) and the IQA-SOR reflects those advancements, as well as other instructional emphases, such as phenomena based lesson launches and justifying knowledge claims. We strongly concur with Anwar and Menekse (2021) that structural and functional elements of research tools should be chosen to reflect the focus of their questions.

Contribution to the Teaching and Learning of Science and NARST Members

In this proposal, we provide a systematic account of a comparison of two very different classroom observation protocols for their relative suitability for a research project focused on teachers' support of productive epistemic talk. Such accounts can be useful to other researchers and NARST members as we clarify some considerations needed to explore the suitability of a different research tool. As the field moves toward more large-scale examinations of classroom activities, such accounts can assist others in "thinking through" factors that need to be considered in matching their research questions to appropriate tools.

This material is based upon work supported by the National Science Foundation under DRL #1720587. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Anwar, S., & Menekse, M. (2021). A systematic review of observation protocols used in postsecondary STEM classrooms. *Review of Education, 9(1), 81-120.* https://bera-journals.onlinelibrary.wiley.com/doi/epdf/10.1002/rev3.3235

Boston, M. (2012). Assessing instructional quality in mathematics. *Elementary School Journal*, 113(1), 76–104.

Boston, M. D. (2014). Assessing instructional quality in mathematics classrooms through collections of students' work. In Y. Li, E. A. Silver, & S. Li (Eds.), *Transforming mathematics instruction* (pp. 501–523). Switzerland: Springer International Publishing.

Ford, M. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education, 92* (3), 404-423.

Kelly, G.J. (2014). Discourse practices in science learning and teaching. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education,* volume 2, (pp. 321-336). Mahwah, NJ: Lawrence Erlbaum Associates.

Marshall, J. C., Smart, J., & Horton, R. M. (2010). The design and validation of EQUIP: An instrument to assess inquiry-based instruction. *International Journal of Science and Mathematics Education*, 8(2), 299-321.

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, 13(4), 267–300

Michaels, S., & O'Connor, C. (2012). *Talk science primer*. Cambridge, MA: TERC.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts and core ideas*.Washington, DC: The National Academies Press.

Piburn, M., & Sawada, D. (2000). *Reformed Teaching Observation Protocol (RTOP) reference manual*. Technical Report.

Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed teaching observation protocol (RTOP): Reference manual (ACEPT Technical Report No. IN00–3)*. Tempe , AZ : Arizona Collaborative for Excellence in the Preparation of Teachers. (Eric Document Reproduction Service, ED 447 205.).

Sampson, V., & Schleigh, S. (2013). *Scientific Argumentation in Biology: 30 Classroom Activities*. Arlington, VA: NSTA Press.

Sampson, V, Enderle, P, & Walker, J. (2011). The development and validation of the Assessment of Scientific Argumentation in the Classroom (ASAC) observation protocol: A tool for evaluating how students participate in scientific argumentation. In M. Khine (Ed.) *Perspectives in Scientific Argumentation: Theory, Practice and Research* (pp. 235-264). Springer.

Sawada, D., Piburn, M., Judson, E., Turley, J., Falconer,K. K., Benford,R.. & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science & Mathematics, 102* (6), 245.

Tekkumru-Kisa, M., Preston, C., Kisa, Z., Oz, E., & Morgan J. (2021). Assessing instructional quality in science in the era of ambitious reforms: A pilot study. *Journal of Research in Science Teaching*, 58, 170–194. *https://doi.org/10.1002/tea.21651*