International Journal of Science Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsed20

Explanation-Construction in Fourth-Grade Classrooms in Germany and the USA: A cross-national comparative video study

Cory Forbes[ae], Kim Lange[b], Kornelia Möller[c], Mandy Biggers[d], Mira Laux[c] & Laura Zangori[e]

[a] School of Natural Resources, Univeristy of Nebraska-Lincoln, Lincoln, USA

[b] Primary Education and Didactics, University of Augsburg, Augsburg, Germany

[c] Institute for Early Science Education, University of Münster, Münster, Germany

[d] College of Education, Pennsylvania State University, College Park, USA

[e] College of Education and Human Sciences, Univeristy of Nebraska-Lincoln, Lincoln, USA

Published online: 09 Jun 2014.

To cite this article: Cory Forbes, Kim Lange, Kornelia Möller, Mandy Biggers, Mira Laux & Laura Zangori (2014): Explanation-Construction in Fourth-Grade Classrooms in Germany and the USA: A cross-national comparative video study, International Journal of Science Education, DOI: 10.1080/09500693.2014.923950

To link to this article: http://dx.doi.org/10.1080/09500693.2014.923950

International Journal of Science Education, 2014
http://dx.doi.org/10.1080/09500693.2014.923950

Explanation-Construction in Fourth-Grade Classrooms in Germany and the USA: A cross-national comparative video study

Cory Forbes[a,e*], Kim Lange[b], Kornelia Möller[c], Mandy Biggers[d], Mira Laux[c] and Laura Zangori[e]

[a]*School of Natural Resources, Univeristy of Nebraska-Lincoln, Lincoln, USA;* [b]*Primary Education and Didactics, University of Augsburg, Augsburg, Germany;* [c]*Institute for Early Science Education, University of Münster, Münster, Germany;* [d]*College of Education, Pennsylvania State University, College Park, USA;* [e]*College of Education and Human Sciences, Univeristy of Nebraska-Lincoln, Lincoln, USA*

To help explain the differences in students' performance on internationally administered science assessments, cross-national, video-based observational studies have been advocated, but none have yet been conducted at the elementary level for science. The USA and Germany are two countries with large formal education systems whose students underperform those from peers on internationally administered standardized science assessments. However, evidence from the 2011 Trends in International Mathematics and Science Exam assessment suggests fourth-grade students (9–10 year-olds) in the USA perform higher than those in Germany, despite more instructional time devoted to elementary science in Germany. The purpose of this study is to comparatively analyze fourth-grade classroom science in both countries to learn more about how teachers and students engage in scientific inquiry, particularly explanation-construction. Videorecordings of US and German science instruction ($n_1 = 42$, $n_2 = 42$) were sampled from existing datasets and analyzed both qualitatively and quantitatively. Despite German science lessons being, on average, twice as long as those in the USA, study findings highlight many similarities between elementary science in terms of scientific practices and features of scientific inquiry. However, they also illustrate crucial differences around the scientific practice of explanation-construction. While students in German classrooms were afforded more substantial opportunities to formulate evidence-based explanations, US classrooms were more strongly characterized by opportunities for students to actively compare and evaluate evidence-based explanations. These factors may begin to help account for observed differences in student achievement and merit further study grounded in international collaboration.

*Corresponding author. School of Natural Resources, Univeristy of Nebraska-Lincoln, Lincoln, USA. Email: cforbes3@unl.edu

Keywords: *Elementary science; Primary science; Inquiry; Video study*

Study Introduction and Rationale

Science is a crucial component of contemporary education reform policy in most countries, including the USA. and Germany (Gesellschaft für Didaktik des Sachunterrichts [GDSU], 2013; Neumann, Fischer, & Kauertz, 2010; National Research Council [NRC], 2013). However, in both countries, elementary students consistently underperform their peers from many other Organisation for Economic Co-operation and Development (OECD)-member countries on the Trends in International Mathematics and Science Exam (TIMSS; Gonzales et al., 2008; Martin, Mullis, Foy, & Stanco, 2012). In 2011, the USA ranked 7th and Germany ranked 17th at the 4th-grade level (9–10-year-old students) out of 57 countries participating in the TIMSS (Martin et al., 2012). The underachievement of students at the earliest stages of the US and German formal education systems is of great concern. By engaging in essential features of inquiry and scientific practices (NRC, 2013, 2000), early learners begin to develop knowledge and skills that lay a foundation for lifelong science learning. Further, research has shown that early learners can develop robust conceptual understanding of natural phenomena. In recognition of its important role in fostering students' lifelong learning, elementary science has reemerged in both countries as a focus of science education reform (Klieme et al., 2003; NRC, 2007).

Education researchers, teachers, administrators, policy-makers, and other stakeholders all share an interest in better understanding how curriculum and instruction shape student outcomes. To address this need, cross-national comparative studies using observational data have been called for (Brückmann et al., 2007). Such studies have thus far been conducted for middle-school science (Roth et al., 2006; Stigler, Gallimore, & Hiebert, 2000), middle-school mathematics (Santagata, 2005), secondary-level physics (Dalehefte et al., 2009; Seidel & Prenzel, 2006), and elementary mathematics (Lan et al., 2009). However, no such studies have been conducted for elementary science and, as a result, knowledge of the day-to-day teaching and learning practices that define elementary science across the globe remains limited. Given the observed differences in student achievement in the USA, Germany, and other countries, as well as the observation that more instructional time is devoted to science in the fourth-grade in Germany than in the USA (Gonzales et al., 2008), it is important to better understand differences in classroom science that may begin to help account for differences in student outcomes. Other research has indicated that elementary science instruction is characterized by a de-emphasis on explanation-construction (Beyer & Davis, 2008; Biggers, Forbes, & Zangori, 2013; Forbes, Biggers, & Zangori, 2013; Metz, 2009; Zangori, Forbes, & Biggers, 2013), a critical dimension of effective, inquiry-based science teaching and learning. While this finding may provide first insight into relationships between instructional practices and student learning outcomes, strong observational evidence is needed to document the extent to which day-to-day elementary classroom science

engages students in scientific practices to foster their science learning. To begin to address this gap in the research, we ask the follow research questions:

(1) To what extent do US and German fourth-grade classrooms exhibit essential features of classroom inquiry, particularly the formulation and evaluation of evidence-based explanations?
(2) How do US and German fourth-grade classrooms afford students opportunities to engage in features of inquiry, particularly the formulation and evaluation of evidence-based explanations?

Background and Theoretical Framework

Video Studies in International Assessment Contexts

Globally, science assessments such as the TIMSS are increasingly used to make comparative judgments about the quality of national systems of formal schooling. At the same time, contemporary research in the field of science education and the learning sciences has shown that early learners can engage in scientific practices such as investigation (Metz, 2011), modeling (Manz, 2012), argument (McNeill, 2011), and explanation-construction (van Aalst & Truong, 2011; Glauert, 2009) in ways that exceed expectations traditionally considered developmentally appropriate. However, to productively do so, they must be provided with substantial support and scaffolding through curriculum and instruction (Hapgood, Magnusson, & Palinscar, 2004; Hardy, Jonen, Möller, & Stern, 2006). Effective science learning environments should therefore be designed around these critical elements to optimally support students' science learning, one important measure of which is the standardized assessment instrument, TIMSS. These are important foundations of contemporary science education reform targeted at the elementary level (Klieme et al., 2003; NRC, 2007).

Unfortunately, research has shown that elementary students are too often not afforded substantive opportunities to engage in scientific inquiry and the practices of science (Forbes et al., 2013). Though the field still lacks a formal review of commonly used elementary science curriculum materials (Kesidou & Roseman, 2002), there is evidence that both curricular and instructional components of elementary science learning environments can be improved to better support students' reasoning about natural phenomena through inquiry (Beyer & Davis, 2008; Biggers et al., 2013; Lange, Kleickmann, & Möller, 2009; Metz, 2009; Zangori et al., 2013). Given the increasingly 'high-stakes' use of assessment data in countries around the world, it is critical to learn more about factors that may impact student outcomes. More research is therefore needed in multinational settings to understand how students' participation and engagement in inquiry and scientific practices can be supported in elementary science learning environments, particularly through curriculum and instruction.

To begin to better understand classroom-level dynamics that may help explain student assessment outcomes, attention has turned in recent years to cross-national

comparative video studies (Brückmann et al., 2007). Video studies involve the systematic collection of video-based classroom observation data to gather a representative sample of classroom activity that can be used to compare teaching and learning in multiple settings. Such studies allow for investigation of complex classroom practices that, in the absence of observational data, are very difficult to study. A number of such video surveys have been conducted that are associated with the TIMSS assessment, focusing on international comparisons of eighth-grade mathematics and science instruction in the classrooms of participating countries (Roth et al., 2006; Santagata, 2005; Stigler et al., 2000). However, comparative video studies have also been conducted outside the context of TIMSS (Dalehefte et al., 2009; Lan et al., 2009; Seidel & Prenzel, 2006). Taken as a whole, these studies have highlighted important differences in instructional norms that characterize science and mathematics instruction in different countries using a variety of conceptual frames. Yet no such studies have yet been conducted at the elementary level for science and, as a result, there is little empirical evidence upon which to consider international comparisons of elementary science learning environments.

A Theoretical Framework for Scientific Inquiry and Explanation-Construction

Contemporary perspectives on science teaching and learning in USA and Germany (GDSU, 2013; Möller, 2004; NRC, 2013) are grounded in constructivist views of learning that foreground the role of the learner in actively building new knowledge through cognitive, social, and cultural processes. A core assumption of constructivist views of learning is that students possess ideas about the natural world largely formulated through their own experiences outside of the classroom. Therefore, in effectively designed science learning environments, students' pre-existing ideas should serve as the building blocks of curriculum and instruction through which 'children need to become aware of, build on, and refine their own ideas' (NRC, 2007, p. 312). The intellectual and practical work associated with interrogating and refining ideas over time is grounded in inquiry-oriented scientific practices. The prevailing assumption underlying science education reform worldwide is that students' engagement in classroom inquiry and scientific practices will yield greater learning gains, outcomes that are expected to be observed in standardized science assessment results.

What is inquiry-based classroom science? Students in science classrooms should be involved in a variety of activities that mimic those of scientists, including formulating scientific questions, conducting scientific investigation, collecting and synthesizing information, scientific explanation-construction, scientific modeling, and engaging in scientific argumentation to construct scientifically accurate ideas about the natural world (GDSU, 2013; NRC, 2007, 2013). To operationalize these scientific practices as core elements of effectively designed science learning environment, the NRC (2000) identifies five fundamental features of meaningful classroom inquiry, which include (a) engaging in scientifically oriented *questions*; (b) giving priority to *evidence*; (c) *formulating* explanations from evidence to address scientifically oriented questions; (d) *evaluating* their explanations in light of alternative explanations,

particularly those reflecting scientific understanding; and (e) *communicating and justifying* proposed explanations. Ultimately, explanation is a central element of inquiry-based classroom science, being the focus of three of the five features of inquiry. Evidence-based explanations that students' formulate, which are shared, negotiated, and utilized to answer scientific questions, serve as the 'currency' of the science classroom. The NRC's five-part framework for classroom inquiry, particularly the central role of explanations, serves as both the conceptual and analytical framework for this study.

Taken together, these features of inquiry include constituent processes which define collaborative scientific sense-making about the natural world. However, while students may pose questions, make predictions, and conduct investigations to establish observed relationships (cause and effect), it is crucial that they be afforded opportunities to use that evidence to generate more theoretical propositions for *how* cause brings about effects and *why* natural phenomena occur in the ways they observe. These causal *mechanisms* (Braaten & Windschitl, 2011) are critical for students to go beyond description to posit evidence-based explanations for the natural world. First, students must *formulate* explanations that exhibit particular characteristics, including (a) being supported by evidence; (b) answering a question driving the investigation; (c) being grounded in students' pre-existing ideas; and (d) proposing new understanding about the observed phenomenon (NRC, 2000). Second, students should also be afforded opportunities to *evaluate* their explanations for phenomena by comparing them to their own previous ideas, peers' explanations for the same phenomena, or the scientifically accepted account. In doing so, students should consider (a) whether evidence supports their proposed explanation, (b) whether their proposed explanation answers the investigation question, (c) there are any biases or flaws in reasoning connecting evidence with their proposed explanation, and (d) consider whether alternative explanations can be reasonably derived from the same evidence (NRC, 2000). These two essential features of inquiry—students' *formulation* and *evaluation* of evidence-based explanations—underlie the construction, negotiation, and interrogation of evidence-based claims in elementary science learning environments and comprise the scientific practice of explanation-construction (NRC, 2013).

Study Design and Methods

This empirical study is embedded within two existing research and development projects—one each in the USA and Germany. The first involves a multi-year professional development program designed to support elementary (K-6) teachers in the USA to learn to better engage students in inquiry and scientific practices (Biggers et al., 2013; Forbes et al., 2013; Zangori et al., 2013). The second stems from the initial phase of a longitudinal project investigating the development and interplay of science instruction, classroom climate, and students' science interest in the transition from primary to secondary education in Germany (Lange et al., 2009; Lange, Kleickmann, Tröbst, & Möller, 2012; Möller, Hardy, & Lange, 2012). Here, we draw upon existing, video-based data from both projects to comparatively investigate features of

inquiry and scientific practices in fourth-grade classrooms in the USA and Germany. The objectives and design of this comparative cross-national video survey study are consistent with similar previous studies (Roth et al., 2006; Santagata, 2005; Seidel & Prenzel, 2006; Stigler et al., 2000).

Participants

The US project involved 81 elementary teachers (grades K-6) from 23 elementary schools across 5 school districts in a single Midwestern state. A similar approach was used in German project to identify 60 participant 4th-grade and 54 6th-grade teachers in the single-most populous state in Germany. Though participants were not selected through random sampling, the sampling procedures used in both projects afforded a comparable and representative group of teachers (Möller, 2004; Goldring, Gray, & Bitterman, 2013). For teachers and students in both projects, science was a core component of the primary and elementary school curriculum taught in all primary grades. At the 4th-grade level, teachers in both countries taught 3–4 distinct science topics per year. These topics for science instruction were determined by national, regional, state, and local science curriculum standards (GDSU, 2013; NRC, 2007, 2013) and embodied in curricular resources (lesson plans, student worksheets, investigation materials, etc.) provided to teachers in each country. Both US and German states in which projects were based include larger, urban centers with high population densities and rural, agricultural areas with low population densities. As a result, teachers in both projects were recruited from large, urban school districts, smaller rural districts, and mid-sized suburban districts. There was a wide range of diversity among the campuses. For example, the percentage of US students qualifying for free and reduced lunch, a commonly used measure of socio-economic status, ranged between 7.2% and 89.3% while students in German schools stemmed from the entire range of possible family socio-economic status (in terms of the highest value of the International Socio-Economic Index assigned to father). In both samples, most teachers were female, in their early 40s, and had average class sizes of 22 students. They were also at post-induction stages of their careers, with an average of 14 and 16 years of teaching experience in the USA and Germany, respectively. Teachers in both projects were compensated for their time.

Data Collection

During the data collection period, US and German teachers taught topics specified by their normal science curriculum using curricular resources already developed and/or provided to them through local or regional sources. As part of the respective projects, all teachers were involved in documenting their classroom practices, including videorecorded samples of their science instruction. Data used for this study were not related to project-related instructional interventions. The German project was a pure research project and, as such, there was no intervention designed to alter teachers' instructional practices. While the US project was designed around a professional

development program, data used for this study were collected in the first year of the project before any intervention had occurred. The US and German research teams made no effort to modify teachers' instructions or demand implementation of externally developed curricular programs and/or instructional interventions. As such, the video data collected for this study represents a 'snapshot' of typical fourth-grade science teaching and learning in the two countries.

In both projects, a sample of teachers' science instruction was videorecorded. Teachers were asked to capture individual enacted 'complete lessons' for science. Consistent with past video studies, including the TIMSS video studies at the middle-school level (Dalehefte et al., 2009; Lan et al., 2009; Roth et al., 2006; Santagata, 2005; Seidel & Prenzel, 2006; Stigler et al., 2000), the definition of a 'complete lesson' remained country-specific, largely determined by lesson length and structure as written in teachers' normal curricular resources, instructional periods allotted for science in daily school schedules, and institutionalized norms for teaching practice. Whereas science lessons for teachers involved in the US project tended to be daily, stand-alone instructional sequences ranging from 20–60 minutes in length, lessons observed in German classrooms tended to follow a format with longer continuous instructional sequences. As a result, lessons from German fourth-grade classrooms were substantially longer ($\bar{x} = 81$ min) than those from US ($\bar{x} = 42$ min) classrooms, and this difference was statistically significant $t(54) = 5.43$, $p < .001$, $d = 2.85$. As a product of the sampling approach, no effort was made to control for the specific topics or format of observed lessons, similar to past comparative video studies (Stigler et al., 2000). Observed fourth-grade science lessons covered topics across the geosciences, life sciences, and physical sciences, and varied in terms of activity structure (teacher lecture, whole-group discussion, small-group work, etc.) and lesson elements (presentation, use of text, hands-on investigation, worksheets, etc.).

Videorecording of science teaching focused on capturing a 'bird's eye' view of classroom activity, including both teachers and students, to the greatest extent possible. For all videorecorded instruction, the focal point was instruction and whole-class activity around it, not individual students or subgroups of students. The objective was to gather evidence of science teaching and learning at the classroom level, including the teachers' instruction moves, whole-class discourse, and the overall structure of learning experiences for students. In the German project, each lesson was observed live and videorecorded by two project team members. In the US project, 54 lessons were observed live by one of two project team members and videorecorded. The remaining lessons were videorecorded by the teachers themselves. For these teachers, the project provided simple, easy–to-operate video cameras and tripods to record enacted science lessons on SD cards that were submitted to the project team. For US teachers who videorecorded their own lessons, training and detailed instructions were provided to ensure that teachers appropriately set up the equipment and focused the camera on themselves with wide-angle views to capture an inclusive view of events in the classroom (Biggers et al., 2013; Forbes et al., 2013; Zangori et al., 2013). In total, each US teacher provided 4–5

sampled lessons while almost all German teachers were observed once. This resulted in a substantial set of videorecorded science lessons from elementary classrooms in the USA ($n_1 = 367$) and Germany ($n_2 = 114$).

To obtain a comparative sample of videorecorded 4th-grade (students age 9–10-year-olds) science lessons for this study, we sampled data from each project's full dataset. Of the 81 teachers involved in the US project, 42 taught 4th-grade. As such, this characteristic of the US dataset (i.e. 42 individual 4th-grade teachers) determined the functional sample size selected for analysis in this study. For each fourth-grade teacher in the US dataset, a single videorecorded lesson was randomly selected from the full US dataset. A similar number of fourth-grade lessons were then randomly sampled from the German dataset. Using a two-stage sampling approach, we first selected all lessons from 4th-grade teachers and, of those data, randomly selected 42 individual lessons. This sampling approach resulted in a comparable set of videorecorded 4th-grade science lessons from USA ($n_1 = 42$) and German classrooms ($n_2 = 42$). Each videorecorded fourth-grade science lesson used in the analyses for this study was taught by a unique teacher.

Video Scoring and Analysis

The videorecorded science lessons were analyzed with observable activity at the classroom level as the unit of analysis. To score the video data, we used the Practices of Science Observation Protocol (P-SOP; Forbes et al., 2013), a recently developed observation protocol designed for use in elementary science learning environments (see Appendix). The 20-item instrument provides a classroom-level measure of elements of inquiry and scientific practices, as well as sub-measures for each of the five essential features of inquiry, including students' *formulation* and *evaluation* of evidence-based explanations (NRC, 2000, 2013). As such, it is directly grounded in the theoretical perspective on inquiry underlying the study and affords a mechanism through which to make normative judgments about the quality of inquiry occurring in observed classrooms. Each feature sub-measure in the instrument is comprised of four unique instrument items. Each instrument item is scored from 0 ('no evidence') to 3 ('strong evidence'). For example, scoring levels for a sample instrument item are presented in Table 1.

For a given videorecorded sample of science instruction, scores for the five sub-measures are summed to determine a composite P-SOP score (between 0 and 60) which represents an overall evaluation of inquiry occurring in a given classroom. Summed scores for each of the five sub-measures (feature of inquiry) range between 0 and 12 and represent an evaluation of a particular essential feature of inquiry evident in classroom activity. The P-SOP was developed as part of the US team's research and development efforts. In a previous study, it was shown to be valid and reliable in elementary science learning environments (Forbes et al., 2013). More detailed descriptions of the instrument, its development, sub-measures, and its psychometric properties have also been published elsewhere (Forbes et al., 2013; Zangori et al., 2013).

Table 1. Scoring levels and descriptions for P-SOP item 3a

Score	Score description
3	Students formulate explanations for causes of effects or establish relationships about phenomenon of interest that are supported with empirical evidence. Explanations are based on reasoning that connects evidence to claims. Process can be highly scaffolded (by the teacher, curriculum materials, etc.) or open-ended (student-directed)
2	Students formulate explanations for causes of effects or establish relationships about phenomenon of interest that are partially supported by their evidence and exhibit some reasoning connecting evidence to claims. Process can be highly scaffolded (by the teacher, curriculum materials, etc.) or open-ended (student-directed)
1	Students formulate explanations for causes of effects or establish relationships about phenomenon of interest that are weakly supported by their evidence and exhibit limited reasoning connecting evidence to claims. Process can be highly scaffolded (by the teacher, curriculum materials, etc.) or open-ended (student-directed)
0	Students do not formulate evidence-based explanations about the phenomena of interest

Scoring of the video data from the US dataset occurred over two years. In the winter of 2012, a member of the US team (lead author) facilitated a multi-day P-SOP training session for the German team in Germany. The training sessions involved explanations of instrument items, discussion of differences in scoring levels for each item, viewing video examples to illustrate scoring levels, and both collaborative and independent practice scoring videos from both the USA and German dataset. The German team began scoring of the German video data in the summer of 2012. The lead scorer from the German team spent two months at the US team's institution in the summer of 2012, which allowed for informal follow-up with the US team as scoring of the German video data progressed. Video scoring was completed in the summer of 2013.

To formally assess inter-scorer reliability, we used a sample of nine videos, jointly scored by two raters—a member from each research team. Four videos from the German sample were transcribed and subtitled in English. The remaining five videos were from the US sample. These two sets of scores were used to calculate intra-class correlation coefficients (ICCs) for each P-SOP item and feature sub-measure. We used a two-way random model to measure consistency because there were only two raters and all raters scored all videos. We report the mean scores along with results from Tukey's *post hoc* tests (to identify potential interactions between rater and video) and significance of factor rater. Results from video scoring shows Scorer 1's ($\bar{x} = 19.11$) scores were higher than those given by Scorer 2 ($\bar{x} = 18.22$), though this difference was not statistically significant, $t(8) = 0.645$, $p = 0.54$, $d = 0.11$. ICCs for each of the 20 P-SOP items ranged from 0.6 to 1.0, with the exception of one item which scored a 0, for an average item ICC of 0.85. Item scores within each feature were summed to provide feature-specific aggregate scores. Four of the five feature subscores had ICCs > 0.9. The remaining ICC for the feature *engaging students in communicating and justifying proposed explanations* was 0.5, though this was due entirely to scoring of one video from the sample. Based on

these inter-rater reliability analyses, the remaining videorecorded lessons in each dataset were scored independently by the respective project teams.

Video Coding and Analysis

As part of video scoring using the P-SOP, scorers recorded narrative scoring notes that provide narrative summaries and reference points for video scoring. These scoring notes, as well as item scores for the videos themselves, were used as a starting point for qualitative data analyses. Scoring notes corresponding to each of the essential features of inquiry (NRC, 2000) were used to identify and isolate video segments related to features of inquiry. Scoring notes were compared across multiple scorers on an 8% sample of data in both datasets, which resulted in 81% agreement before and 100% agreement after discussion. Relevant segments of videos from both data sets were further analyzed by each research team to articulate qualitative differences in observed classroom practice for these two features across the two datasets. We then engaged in a collaborative process of pattern-matching (Yin, 2009) to map expected trends related to features of inquiry, largely based upon previous research documenting trends in elementary teachers' science instruction using the instrument (Biggers et al., 2013; Forbes et al., 2013; Zangori et al., 2013), onto observed patterns in this study. The pattern-matching approach afforded the ability to identify trends within and across datasets that helped provide qualitative description for difference observed in the quantitative analysis of P-SOP scores. Through further validation of underlying theoretical propositions about features of inquiry and characteristics of elementary science learning environments, subtle but fundamental differences within individual features of inquiry were identified and isolated to illuminate the complexities of classroom practices, including instruction, in the elementary classrooms under study here.

Results

Comparative Analysis of Features of Inquiry in US and German Classrooms

In research question #1, we asked, 'to what extent do US and German 4th-grade classrooms exhibit essential features of classroom inquiry, particularly the formulation and evaluation of evidence-based explanations?'. Aggregate P-SOP scores ranged from a low score of 0 to a high score of 54 and exhibited metrics of a normal distribution of scores (Skewness = 0.56; Kurtosis = 0.12). Mean aggregate P-SOP scores were 19.67 ($SD = 1.33$) for German 4th-grade classroom videos and 20.83 ($SD = 14.8$) for those from the USA This observed difference in aggregate P-SOP scores between the US and German samples was not statistically significant, $t(68) = 0.58$, $p = .61$, $d = 0.15$. This finding suggests classrooms in both video samples exhibit similar overall levels of scientific inquiry.

Further analysis focused on each of the five P-SOP subscores for the five essential features of inquiry. Group means and standard deviations for P-SOP subscores for each of the five essential features of inquiry are shown in Table 2 and Figure 1.

These P-SOP subscores were all weakly to moderately correlated, as shown in Table 3.

Given the moderate and statistically significant correlations between P-SOP subscores shown in Table 2, a MANOVA was used to compare students' engagement in five features of inquiry in US and German classrooms. All five P-SOP subscores were included as dependent variables. Country was used as the independent variable. The multivariate result was significant for country, $F = 7.6$, $df = (5,78)$, $p < .001$, Wilk's $\Lambda = .671$, partial $\eta^2 = .33$, indicating a main effect for country on at least a subset of the five subscores for essential features of inquiry. Follow-up univariate F-tests showed there was a significant difference between observed features of inquiry in the two datasets for D3, *formulating evidence-based explanations*, $F = 7.1$, $df = (1,78)$, $p < .01$, and D4, *evaluating evidence-based explanations*, $F = 12.2$, $df = (1,78)$, $p < .001$. German fourth-grade classrooms exhibited more evidence of students formulating evidence-based explanations while US classrooms exhibited greater evidence for evaluating explanations in light of alternative explanations.

Because the German lessons were substantially longer than those in the US sample, we also conducted a MANCOVA to compare students' engagement in five features of inquiry in US and German classrooms when controlling for lesson length. The multivariate result was still significant for country, $F = 5.79$, $df = (5,77)$, $p < .001$, Wilk's $\Lambda = .732$, partial $\eta^2 = .268$, indicating an overall main effect for country on at least a subset of the five subscores for essential features of inquiry. Follow-up univariate F-tests showed the statistically significant difference remained between observed features of inquiry in the two datasets for D4, *evaluating evidence-based explanations*, $F = 3.39$, $df = (1,78)$, $p = .004$. However, D3, *formulating evidence-based explanations*, was no longer significant, $F = 7.2$, $df = (1,78)$, $p = .218$. This suggests increased lesson length may have afforded German fourth-grade students more time to engage in the formulation of evidence-based explanation. In contrast, students in US classrooms were afforded particular opportunities to evaluate explanations independent of how long the lessons were.

Table 2. Means (with standard deviations) for features of inquiry observed in fourth-grade classrooms in the USA and Germany ($N = 84$)

Country	Germany	USA
D1. Engaging students in scientifically oriented questions	7.04 (5.15)	7.33 (8.0)
D2. Engaging students in giving priority to evidence in responding to questions	6.38 (2.30)	7.02 (2.71)
D3. Engaging students in formulating explanations from evidence to address scientifically oriented questions	4.45 (3.76)	2.60 (2.56)
D4. Engaging students in evaluating their explanations in light of alternative explanations	0.08 (0.49)	1.45 (2.41)
D5. Engaging students in communicating and justifying their explanations	1.60 (0.99)	2.12 (1.95)

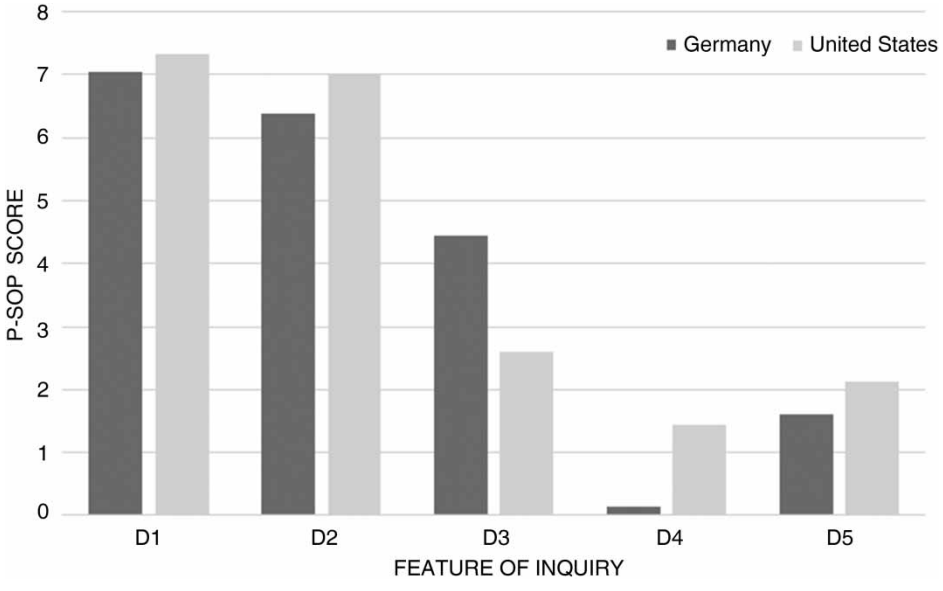

Figure 1. Means for features of inquiry observed in fourth-grade classrooms in the USA and Germany ($N = 84$)

Table 3. Correlation matrix for features of inquiry observed in fourth-grade classrooms in the USA and Germany ($N = 84$)

	D1	D2	D3	D4
D1. Engaging students in scientifically oriented questions	–	–	–	–
D2. Engaging students in giving priority to evidence in responding to questions	.529[b]	–	–	–
D3. Engaging students in formulating explanations from evidence to address scientifically oriented questions	.401[b]	.462[b]	–	–
D4. Engaging students in evaluating their explanations in light of alternative explanations	.311[b]	.370[b]	.249[a]	
D5. Engaging students in communicating and justifying their explanations	.330[b]	.476[b]	.531[b]	.713[b]

[a]Correlation is significant at the 0.05 level (two-tailed).
[b]Correlation is significant at the 0.01 level (two-tailed).

Qualitative Analysis for Formulating and Evaluating Evidence-Based Explanations

In research question #2, we asked, 'how do US and German 4th-grade classrooms afford students opportunities to engage in features of inquiry, particularly the formulation and evaluation of evidence-based explanations?'. As shown in the findings from quantitative analyses for research question #1, German fourth-grade classrooms afforded students more effective opportunities to *formulate* evidence-based explanations about natural phenomena while students in US classrooms were afforded

more effective opportunities to *evaluate* evidence-based explanations. Qualitative analysis of the video datasets yields crucial insight into specific elements of classroom practice that help explain these differences. In the sections below, we present qualitative examples from classrooms in both countries that are illustrative of broader trends observed in the study.

Formulating Evidence-based Explanations

To better support students to formulate evidence-based explanations, teachers in German classrooms employed multifaceted forms of curricular and instructional scaffolds during investigations and whole class discussions. While there were highly effective examples of students articulating scientific claims in both US and German classrooms, on average, German classrooms afforded students more productive opportunities to formulate evidence-based explanations. A critical difference that emerged was the extent to which teachers foregrounded and supported students' use of data and evidence to ground their claims. German fourth-grade teachers more actively facilitated students' use of direct, observational evidence to formulate mechanism-based claims for how and why natural phenomena occurred. In all cases teachers heavily scaffolded the process of linking evidence that students gathered during classroom investigations to claims that they were making. As shown in the results from quantitative analyses, the amount of instructional time available to German teachers as compared to those in the USA contributed to the opportunities students were afforded to formulate evidence-based explanations.

There were many illustrative examples of this trend in videorecorded samples of instruction. In one German classroom, for example, the teacher chose the question 'what helps water evaporate faster?' as a driving question for a lesson on evaporation. After a short introduction where she discussed what it means to construct an explanation, students worked on nine different parts of an investigation in which they observed the influence of particular variables (temperature, wind, surface area, etc.) on the rate of evaporation. For each part of the investigation, students were asked to record observations, first making a prediction, then writing down an observation and constructing an explanation. Students quickly identified 'heat' as one of the factors that helps accelerating the rate of evaporation. When writing down 'heat' as an answer to the research question, she probed students' responses, asking 'In which activity could you see that?' In the subsequent discussion, students identified activities and the teacher recorded their observations on the blackboard (V_007; 1:28:00–1:30:15). The teacher repeated this procedure when students claimed that 'wind' and 'surface area' also influence the rate of evaporation (V_007; 1: 01:32:15–01:32:45). Using this prompt, the teacher tied her students' explanations back to their observations, scaffolding their use of evidence.

In a second classroom, another fourth-grade teacher also encouraged her students to ground their explanations in evidence by establishing links between the activities that students conducted during the lesson. In contrast to the teacher in the previous

example, this teacher did not wait to provide this support until the end of the lesson, but challenged her students while the students carried out the investigations. The teacher afforded her students six different activities in which changes of the state of water were observed. For example, two of these activities illustrated the phenomena of condensation. The teacher asked her students to observe a mirror that has just been taken out of a refrigerator as well as a jar that has been filled with cold water and ice cubes. Another set of activities centered around evaporation and condensation. In one of the activities, which the teacher called 'Let it rain', students were asked to heat ice cubes in a pan that is covered by a light. Students were to open the light every two minutes and to observe what happened. The students were not only asked to document their predictions and all their observations in detail, but also to construct explanations for each of the six phenomena. While the students were completing their investigation activities and constructing their explanations in groups, the teacher engaged small groups of students in discussions like the following concerning the 'Let it rain' investigation (V_056_00:32:15-00:34:30):

Teacher: [Stud]ent #18, are you working on an explanation?
Student #18: When water is heated, it rises.
Teacher: What rises? The water?
Student #10: Water vapor.
Teacher: Exactly. The water vapor rises.
Student #22: Then it starts raining because the water droplets fall down.
Teacher: How do the droplets get to the lid?
Student #22: It is water vapor.
Teacher: What happened to the water vapor?
Student #18: It could not get out.
Teacher: How would you describe the lid?
Student #22: Wet.
Teacher: I meant before you put it on the pan.
Student #22, 17, 10 and 18 together: Cold!
Teacher: There you go. So how can you explain [the forming of water droplets] now?
Student #22: The water rose, just a little bit.
Teacher: The water vapor rose. Where to?
Student #18: To the lid.
Teacher: To the cool lid. And then, [Student #22]? What happened to the water vapor when it reached the cold lid?
Student #22: It changes into water again.
Student #18: Yes!
Teacher: Right. And this is what you see on the lid. There [pointing at the lid] you can still see it. It is still there. Droplets of water—it is raining. This is how you can frame your explanation.

As shown in this excerpt, as the students interacted with the material they were asked to construct an explanation. This process was triggered by scaffolds that are integrated in the student worksheets. While they were trying to construct explanations with the investigation material at hand, the teacher helped them to connect evidence to the claims they were making by repeatedly asking them to refer back to what they observed during the activity. In doing so, she ensured that the students used this

information to construct an explanation and helped them to ground their explanation in evidence by referring back to the experiment (the cold lid). Both of these examples illustrate the ways in which German teachers better supported students to ground their explanations in evidence, a critical component of explanation-construction.

In contrast to these examples from German classrooms, US teachers less frequently and effectively asked students to explicitly articulate evidence-based explanations for the natural phenomena they investigated. Often, teachers would emphasize identifying trends in data but provide limited opportunities for students to postulate mechanism-based explanations for those trends. One US teacher's lesson on electrical circuits is illustrative of this trend. During this lesson, students worked in pairs to use a series circuit to light a light-bulb. In the initial stages of the lesson, the teacher asked a pair of students to report out to the whole class how they might try to light the bulb using a series circuit. She then drew their ideas on the overhead projector and instructed all of the students in the class to copy these drawings in their own science notebooks and predict whether or not each arrangement might work. Students then conducted tests of each circuit with their physical materials and reported their observations to the class. As they worked, the teacher consistently scaffolded students' thinking. For example, in one exchange, she worked to clarify trends in students' observations:

Teacher: OK, some of these worked and some did not. Group 1, show me how you have your battery
Student #5: We have the two right together
Teacher: So I want everyone to draw this. [she models what the students say on the overhead] A bunch of you did something just like this but it did *not* work. So [Student #11], how were your batteries?
Student #11: Um, we did it like they did but our lights were not bright . . . they were kind of dim.
Teacher: OK, did anybody do a battery facing a different way? [Student #19], what did your group do?
Student #19: flat side to flat side
Teacher: So the flat side to the flat side, or another way to say that is negative-to-negative, right? [She proceeds to draw the example again with the batteries in the incorrect order] So I want you to look at their circuit . . . is this going to light? Why or why not? (06-02: 12:45)

In this instance, the teacher asked students to consider evidence from a variety of sources as she supported them in identifying trends and formulating an explanation for those trends. The discussion ended with her asking why the different circuit configurations may or may not have resulted in the bulb lighting. However, she did not provide an opportunity for students to respond to the question either in writing in their notebooks or vocally in a class discussion. Therefore, students did not have an opportunity to use their evidence to formulate an explanation for how and why the circuit configuration worked to light the bulb. This was consistent within the observed US lessons—explanation-oriented questions were frequently posed to students without opportunities for them to draw upon their evidence to formulate effective explanations.

Evaluating Evidence-based Explanations

While teachers in German classrooms more actively supported students' use of evidence to ground claims, students in US classrooms were afforded more robust opportunities to evaluate evidence-based explanations through comparison. It is important to note that the evaluation of evidence-based explanations was the least frequently observed feature of inquiry observed in enacted science instruction in both countries (see again Table 2 and Figure 1). US fourth-grade students were rarely afforded opportunities to evaluate their explanations in light of others' ideas at all and, when they did, opportunities afforded them were relatively weak. There was virtually no evidence of this practice in German classrooms. As a result, we do not present contrasting examples from both countries here, but rather provide multiple examples from US lessons in which teachers provided at least some active support for students to evaluate the quality of explanations by comparing (a) the evidentiary basis of their own ideas over time or (b) others' explanations for the same phenomena. Both approaches were consistently observed in US teachers' efforts to support students' evaluation of explanations.

One way students were supported to evaluate evidence-based explanations was by being afforded opportunities to compare their own explanations over time. In these cases, students may make a prediction or an initial claim and later (after some form of science instruction or investigation) make a new, revised claims which they can compare to their earlier ideas and use as a basis for evaluation. Teachers in US classrooms more often asked students to consider how evidence from classroom observations and investigations influenced their evolving ideas. For example, one teacher enacted a magnets lesson in which she began the lesson by passing out slips of paper to the students which included the sentence starter, 'A magnet is . . . '. She instructed students to 'glue this strip under the top line of your notebook and finish the sentence' (03-3c: 2:26). During the lesson, students rotated through six stations that engaged them in various activities with magnets. Each station presented a different guiding question, i.e. 'What objects stick to a magnet and what objects don't stick to a magnet?' Students recorded observations and data at each station. After the students rotated through all six stations, they came back together as a whole class for discussion, which the teacher referred to as a 'science conference'. To begin the discussion, the teacher asked a few students to read what they wrote about a magnet I, saying:

> I want you to think about what you wrote, what you just read, and what you heard, and I want to challenge you to change what you wrote when we're done with science conference. I want you to be comfortable with not copying the same thing again . . . thinking about, hmmm . . . do I want to change that just a little to show that I learned something new? (03–3c: 4:45)

The discussion continued with the teacher guiding the students through what they learned at each station. She scaffolded the students' thinking about how to change what they initially wrote about the nature of magnets by asking probing questions such as, 'how might what you learned at this discovery box change your thinking?',

and even voicing examples of specific ways to change what they have written based on their observations. After they discussed students' recorded observations from all six stations, she asked students to consider the evidence to compare their new explanations to their previous ideas, saying:

> Open your science notebooks. I'm going to ask you to talk to your elbow partner and this is what it's going to sound like. I think I want to change my definition by adding . . . then my partner is going to tell me how she's going to change her definition. I'm going to bring you another slip of paper just like the first one and I want you to glue it right underneath the first one. What I'm looking for today is 'Is this definition different from this definition?' I don't care if you only change one word but I want to challenge you to at least change one thing. (03–3c:19:20)

Through this type of reflection, discussion, and writing, the teacher had the students use their observations of magnets during the lesson to compare what they had learned to what they initially wrote about the nature of magnets.

Another way students can evaluate evidence-based explanations is by comparing their own explanation to those of peers or classmates. In these instances, students learn to look for bias in their own reasoning by seeing how other students explain the same scientific phenomenon being studied. One teacher from the USA taught a lesson in which students dissected owl pellets and through which he consistently scaffolded students to compare their explanations for the owls' dietary habits to that of their peers. Early in the lesson, as students made claims about how owls ingest their food, he instructed students to pay close attention to evidence, saying:

> I would like to see in your science notes something about what your peers (that's your other classmates) . . . What are they saying? Connect that with what you think. I want to see you write down your thoughts about that. You should think about what's been said, what you know, write down what you think right now about this. (05-3b: 17:45)

As students began dissecting their own owl pellets, they made claims about what kind of bones they were observing. The teacher circulated from student to student asking them to make a claim about what they were finding and what evidence they have to back up their claims. In the following conversation (representative of conversations with several students in the class), he encouraged one student to seek out peers' explanations to help support the student's claim that his owl pellet contains more than one animal's remains:

Student:	I found three of these
Teacher:	So what do you think?
Student:	I think it eats a certain amount and then it coughs it up
Teacher:	So you think it eats a certain amount and then coughs it back up? See if you can find more evidence to support your claim. If you make a claim like that, that's fine, but what you want is a lot of evidence to back it up . . . what else would be more evidence?
Student:	If I found, like, 5 of the same bone because that has to mean something because no animal has 5 legs
Teacher:	Sure about that? So would other students' results help your argument? If you look at what [other students] had would their results help you with your claim?

Student: Maybe
Teacher: So I want to see you checking that out (05-3c: 4:18)

He continued to encourage students to compare their explanations with classmates' explanations to reflect upon the evidence they used to ground their claims about their owl pellets. These illustrative examples highlight opportunities afforded students in US classrooms to compare and evaluate evidence-based explanations, a critical element of explanation-construction.

Discussion and Implications

Elementary science remains a crucial focus of international science education reform, including in the USA and Germany (GDSU, 2013; Klieme et al., 2003; NRC, 2007). To support students to become lifelong science learners, elementary science learning environments must afford them opportunities to engage in scientific practices and features of inquiry (GDSU, 2013; Möller, 2004; NRC, 2007, 2013) through which to develop robust conceptual understanding of the natural world. Cross-national comparative video studies can yield important insights into classroom practices that support students' learning, particularly those that might help explain national trends in student achievement in science (Brückmann et al., 2007; Gonzales et al., 2008; Martin et al., 2012; Neumann et al., 2010). Though such studies have been conducted around middle-school mathematics and science using TIMSS data (Roth et al., 2006; Stigler et al., 2000), at the secondary level for physics (Dalehefte et al., 2009; Seidel, & Prenzel, 2006), and at the elementary level with a focus on mathematics (Lan et al., 2009; Santagata, 2005), no such research has been carried out for elementary science. Findings from cross-national comparative video studies, such as those presented here, provide a window into the day-to-day activities occurring in science classrooms in various cultural and institutional settings and, as such, help promote broader understanding of effective science teaching and learning amongst researchers, teacher educators, and practitioners. Results from this study specifically begin to shed light on elementary science instruction and science learning environments in the USA and Germany, and should be of interest to science teacher educators, science curriculum developers, and science education researchers.

First, despite evidence that early learners are capable of successfully engaging in scientific inquiry and scientific practices to refine their ideas and construct knowledge of natural phenomena (Glauert, 2009; Manz, 2012; McNeill, 2011; Metz, 2011; van Aalst & Truong, 2011), elementary science instruction often deprioritizes scientific explanation in lieu of more active, hands-on elements of classroom science (Forbes et al., 2013). Evidence from this study reinforces this trend, with questioning and investigation practices far more evident in US and German elementary classrooms than those focused on the formulation, evaluation, and communication of scientific explanations. In many cases of instruction observed in this study, relatively small instructional adjustments could have dramatically impacted opportunities afforded

student to engage in features of inquiry. Such changes include introducing an explicit investigation question, structuring students' observation and data collection, asking students to provide evidence for claims they make about natural phenomena, facilitating small-group social interactions, or integrating opportunities for them to compare their ideas over time and with peers. Each of these examples represents a critical dimension of inquiry and scientific practices (GDSU, 2013; Lange et al., 2012; NRC, 2007, 2013).

This finding has important implications for the design of science learning environments. Past research has highlighted aspects of elementary teachers' knowledge, orientations, and commitments to science teaching and learning that influence their instructional decision-making and may therefore partially help explain these trends (Beyer & Davis, 2008; Biggers et al., 2013; Metz, 2009; Zangori et al., 2013). It is therefore critical to support teachers' learning to identify and implement concrete instructional strategies and lesson elements that can better foster students' engagement in inquiry and scientific practices. However, observed instruction is also shaped by the curricular resources that teachers use for science, particularly at the elementary level. Though no comprehensive review of commonly used elementary science curriculum materials has yet been conducted (Kesidou & Roseman, 2002), limited past research suggests that elementary teachers tend to enact science curriculum with high fidelity (Biggers et al., 2013). To craft elementary science learning environments that engage students in explanation-based scientific sense-making, elementary teachers also need access to effective, well-developed science curriculum materials. If such materials are designed to afford students concrete opportunities to generate scientific questions, engage with data and evidence, and formulate, compare, and communicate explanations, evidence of such curricular elements is likely to be observed in classroom practices.

Second, a notable difference that emerged between US and German fourth-grade classrooms studied here was the extent to which science instruction engaged students in *formulating* and *evaluating* their evidence-based explanations, both critical components of classroom inquiry (NRC, 2000) that lie at the heart of the scientific practice of explanation-construction (NRC, 2013). The key difference revolved around the sequencing of the process of using evidence to ground claims. German teachers more actively supported students to ground their explanations in evidence at the formulation stage through prompts and supports that reinforced the need for empirical evidence from classroom investigations. In contrast, teachers in US classrooms tended to first allow students to formulate claims that were often not entirely based upon evidence. They then often facilitated students' comparison of explanations as a means to collaboratively highlight students' non-scientific reasoning and have them consider the evidentiary basis of their claims after initially formulating them. Both approaches emphasized the documentation of cause and effect as evidence for mechanism-based claims (Braaten & Windschitl, 2011; NRC, 2013) necessary to promote early learners' explanation-construction.

This finding also has important implications for teacher education, professional development, and the design of elementary science curriculum materials. It highlights

that reform-based science instruction is not formulaic and stepwise. Instead, the teacher plays a critical role in mobilizing and implementing various scaffolds, both curricular and instructional, that support unique groups of students to engage with and make sense of natural phenomena (Hapgood et al., 2004; Hardy et al., 2006). However, it extends past research by illustrating two approaches to supporting students' explanations defined by points along a continuum from more open-ended to structured inquiry (NRC, 2000). Examples such as these can provide guidance for teachers to support students' *formulation* and *evaluation* of explanations. An emergent question related to this finding, however, revolves around whether students should be supported to engage in iterative experiences around the formulation and evaluation of explanations rather than foregrounding the formulation of evidence-based explanations as a precursor to evaluating them. To answer this question, more research is needed to explore subtleties of instructional approaches that support students' engagement in the *formulation* and *evaluation* of evidence-based explanations, as well as their respective impact on students' learning.

Third, study findings suggest that the formulation of evidence-based explanations is tied to the amount of instructional time available for science, while the comparison and evaluation of explanations are not. It is typically assumed that increased instructional time is broadly advantageous for students' science learning (NRC, 2007). However, TIMSS data show that top-performing countries actually devote fewer hours to science instruction than many lower-performing countries, including the USA and Germany (Gonzales et al., 2008), suggesting that the amount of instructional time may not alone predict student achievement. Rather, we hypothesize that the targeted use of available instructional time involves engaging students in specific, crucial, value-added scientific sense-making practices. Results from this study show that opportunities for students to *evaluate* evidence-based explanations through comparison occurred independently of how long teachers' enacted science lessons were. On the other hand, students' *formulation* of explanations was time-dependent. Study findings may therefore suggest that while both practices are important components of scientific inquiry and, in particular, explanation-construction, comparing, and evaluating explanations could be a fundamental element of classroom science that has a disproportionate impact on observed classroom characteristics and, possibly, student outcomes. Given that students' evaluation of evidence-based explanations through comparison was the least emphasized feature of inquiry in both 'German and USA classrooms, a potential implication of this finding is that models of science instruction must foreground particular practices, such as the evaluation of explanations. Specific elements of science learning experiences that help students engage in this practice include structured opportunities to reflect upon changes to their explanations over time (i.e. metacognition) and to critically evaluate alternative explanations from other sources, such as peers or authoritative sources. More empirical work is needed to identify which features of inquiry and scientific practices contribute most significantly to observed classroom practice and student outcomes.

Limitations and Future Research

Findings from this study, though preliminary, begin to shed light on similarities and differences in elementary science learning environments in two OECD-member countries—the USA and Germany—that lag their peers on internationally administered assessments of science learning. Though illuminating, they are limited by the relatively small sample size and sampling approaches afforded by the datasets. However, they do provide guidance and direction for future research predicated on the hypothesis that observed practices in elementary science learning environments are related to student outcomes. First, additional studies must be conducted based on data from classrooms illustrating widely variant features of inquiry and scientific practices. Subsequent cross-national comparative studies must draw from observational data from more countries, particularly those that fall at the lower and upper ends of score distributions for assessments such as the TIMSS. Second, future studies must leverage affordances of TIMSS sampling procedures to draw upon a larger, more representative, randomly sampled set of observational data. Such expanded participant sampling will allow for stronger and more valid comparisons of elementary science learning environments from different countries. Finally, third, efforts should be made to link observational, classroom-level evidence to measures of student learning, such as assessment data. Such work could include descriptive, single-subject studies designed to provide evidence of relationships as well as longitudinal research investigating relationships between classroom practice and student outcomes over time. Eventually, such work must be intervention-based and studied using randomized controlled trials to establish causal relationships between observed patterns of classroom practices and relevant student outcomes for science. Each step of this research agenda brings additional complexity which can only be addressed adequately through collaboration between international research teams and practitioners. Results from these research and development efforts, carried out through future partnership-driven multinational observational studies, may ultimately help explain comparative differences between TIMSS scores of fourth-grade students in Germany, the USA, and other OECD countries.

Acknowledgements

This research is funded by the German Research Foundation, Roy J. Carver Charitable Trust, University of Iowa Measurement Research Foundation, and University of Iowa International Programs. However, any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors. Forbes and Lange share primary authorship and appear in alphabetical order. Möller is second author. Biggers, Laux, and Zangori are third authors and appear in alphabetical order. An earlier version of this paper was presented at the 2013 meeting of the National Association for Research in Science Teaching in Rio Grande, Puerto Rico. We appreciate the interest and cooperation of the teachers who made this research possible. We also thank Madison Fontana, Cornelia Sunder, and anonymous

journal reviewers for their help in thinking about these issues and their thoughtful comments on earlier versions of this paper.

References

van Aalst, J., & Truong, M. S. (2011). Promoting knowledge creation discourse in an Asian primary five classroom: Results from an inquiry into life cycles. *International Journal of Science Education*, *33*(4), 487–515.

Beyer, C. J., & Davis, E. A. (2008). Fostering second graders' scientific explanations: A beginning elementary teacher's knowledge, beliefs, and practice. *Journal of the Learning Sciences*, *17*(3), 381–414.

Biggers, M., Forbes, C. T., & Zangori, L. (2013). Elementary teachers'curriculum design and pedagogical reasoning for supporting students' comparison and evaluation of evidence-based explanations. *The Elementary School Journal*, *114*(1), 48–72.

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, *95*(4), 639–669.

Brückmann, M., Duit, R., Tesch, M., Fischer, H., Kauertz, A., Reyer, T., ... LaBudde, P. (2007). The potential of video studies in research on teaching and learning science. In R. Pintó and D. Couso (Eds.), *Contributions from science education research* (pp. 77–89). Dordrecht: Springer.

Dalehefte, I., Rimmele, R., Prenzel, M., Seidel, T., Labudde, P., & Herweg, C. (2009). Observing instruction "next-door": A video study about science teaching and learning in Germany and Switzerland. In T. Janík and T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 83–102). Münster: Waxmann Verlag.

Forbes, C. T., Biggers, M., & Zangori, L. (2013). Investigating essential characteristics of scientific practices in elementary science learning environments: The Practices of Science Observation Protocol (P-SOP). *School Science and Mathematics*, *113*(4), 180–190.

Gesellschaft für Didaktik des Sachunterrichts (Society for Teaching Natural and Social Sciences and Technology in Elementary School). (2013). *Perspectives framework for general studies in primary education*. Bad Heilbrunn: Klinkhardt.

Glauert, E. B. (2009). How young children understand electric circuits: Predection, explanation, and exploration. *International Journal of Science Education*, *31*(8), 1025–1047.

Goldring, R., Gray, L., & Bitterman, A. (2013). *Characteristics of public and private elementary and secondary school teachers in the United States: Results from the 2011–12 schools and staffing survey (NCES 2013–314)*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hapgood, S., Magnusson, S., & Palinscar, A. (2004). Teacher, text, and experience: A case of young children's scientific inquiry. *Journal of the Learning Sciences*, *13*(4), 455–505.

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology*, *98*(2), 307–325.

Kesidou, S., & Roseman, J. E. (2002). Do middle school science programs measure up? Findings from project 2061's curriculum review. *Journal of Research in Science Teaching*, *39*(6), 522–549.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., ... Vollmer, H. (2003). *Zur Entwicklung nationaler Bildungsstandards* [Regarding the development of National Education Standards]. Berlin: Bundesministerium für Bildung und Forschung.

Lan, X., Ponitz, C., Miller, K., Li, S., Cortina, K., Perry, M., & Fang, G. (2009). Keeping their attention: Classroom practices associated with behavioral engagement in first grade mathematics classes in China and the United States. *Early Childhood Research Quarterly*, *24*(2), 198–211.

Lange, K., Kleickmann, T., & Möller, K. (2009). Zusammenhänge zwischen PCK von Grundschullehrkräften und dem Verständnis naturwissenschaftlicher Konzepte bei Grundschülern [Teachers' pedagogical content knowledge and students' learning gains: A multilevel analysis of elementary science classrooms]. In D. Höttecke (ed.), *Gesellschaft für Didaktik der Chemie und Physik: Chemie- und Physikdidaktik für die Lehramtsausbildung* [Society for the teaching of chemistry and physics: Chemistry and physics education for teacher training] (pp. 404–406). Berlin: Lit.

Lange, K., Kleickmann, T., Tröbst, S., & Möller, K. (2012). Fachdidaktisches Wissen von Lehrkräften und multiple Ziele im Sachunterricht [Teachers pedagogical content knowledge and multiple outcome gains in elementary science classrooms]. *Zeitschrift für Erziehungswissenschaft* [Review of Education], *15*, 55–75.

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. *Science Education*, *96*(6), 1071–1105.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McNeill, K. L. (2011). Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. *Journal of Research in Science Teaching*, *48*(7), 771–792.

Metz, K. (2009). Elementary school teachers as "targets and agents of change": Teachers' learning in interaction with reform science curriculum. *Science Education*, *93*(5), 915–954.

Metz, K. (2011). Disentangling robuse developmental constraints from the instructionally mutable: Young children's epistemic reasoning about a study of their own design. *Journal of the Learning Sciences*, *20*(1), 50–110.

Möller, K. (2004). Naturwissenschaftliches Lernen in der Grundschule - Welche Kompetenzen brauchen Grundschullehrkräfte? [Elementary science teaching – what competencies do elementary teachers need?]. In H. Merkens (ed.), *Lehrerbildung: IGLU und die Folgen* [Teacher education: PIRLS and the consequences] (pp. 65–84). Opladen: Leske + Budrich.

Möller, K., Hardy, I., & Lange, K. (2012). Moving beyond standards: How can we improve elementary science learning? A German perspective. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible – learning outcomes in science education* (pp. 31–54). Münster: Waxmann.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academies Press.

National Research Council. (2007). *Taking science to school: Leaning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

National Research Council. (2013). *Next generation science standards: By states, for states*. Washington, DC: National Academies Press.

Neumann, K., Fischer, H., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*(3), 545–563.

Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., … Gallimore, R. (2006). *Highlights from the TIMSS 1999 video study of eighth-grade science teaching (NCES 2006–17)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics, U.S. Government Printing Office.

Santagata, R. (2005). Practices and beliefs in mistake-handling activities: A video study of Italian and US mathematics lessons. *Teaching and Teacher Education*, *21*, 491–508.

Seidel, T., & Prenzel, M. (2006). Stability of teaching patterns in physics instruction: Findings from a video study. *Learning and Instruction*, *16*(3), 228–240.

Stigler, J., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist*, *35*(2), 87–100.

Yin, R. K. (2009). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

Zangori, L., Forbes, C. T., & Biggers, M. (2013). Fostering student sense-making in elementary science learning environments: Elementary teachers' use of science curriculum materials to promote explanation-construction. *Journal of Research in Science Teaching*, *50*(8), 887–1017.

Appendix – Practices of Science Observation Protocol (P-SOP; Forbes, Biggers, & Zangori, 2013)

Engaging students in scientifically oriented questions				
1a. Students engage with an investigation question that is contextualized, motivating, and meaningful for students	0	1	2	3
1b. Students engage with an investigation question that focuses on standards-based content/phenomena	0	1	2	3
1c. Students engage with an investigation question that is answerable through scientific inquiry	0	1	2	3
1d. Students engage with an investigation question that is feasible and answerable in the context of the classroom	0	1	2	3
Engaging students in giving priority to evidence in responding to questions				
2a. Students engage with phenomenon of interest	0	1	2	3
2b. Students work with data related to phenomena of interest	0	1	2	3
2c. Students generate evidence by organizing and analyzing data	0	1	2	3
2d. Students reflect upon and verify the data collection process, accuracy of data, and transformation of evidence from data	0	1	2	3
Engaging students in formulating explanations from evidence to address scientifically oriented questions				
3a. Students formulate explanations about phenomenon of interest that are based on evidence	0	1	2	3
3b. Students formulate explanations about phenomenon of interest that answer investigation question	0	1	2	3
3c. Students formulate explanations about phenomenon of interest that propose new understanding	0	1	2	3
3d. Students formulate explanations about phenomenon of interest that build on their existing knowledge	0	1	2	3
Engaging students in evaluating their explanations in light of alternative explanations				
4a. Students evaluate their explanations by comparing to alternative explanations to consider whether evidence supports their proposed explanation	0	1	2	3
4b. Students evaluate their explanations by comparing to alternative explanations to consider whether their proposed explanation answers the investigation question	0	1	2	3
4c. Students evaluate their explanations by comparing to alternative explanations to consider any biases or flaws in reasoning connecting evidence with their proposed explanation	0	1	2	3
4d. Students evaluate their explanations by comparing to alternative explanations to consider whether alternative explanations can be reasonably derived from the same evidence	0	1	2	3
Engaging students in communicating and justifying their explanations.				
5a. Students clearly share and justify their investigation question	0	1	2	3
5b. Students clearly share and justify their procedures, data, and evidence	0	1	2	3
5c. Students clearly share and justify their proposed explanation and supporting evidence	0	1	2	3
5d. Students clearly share and justify their review of alternative explanations	0	1	2	3