The development and validation of the Instructional Practices Log in Science: a measure of K-5 science instruction

Elizabeth L. Adams, Sarah J. Carrier, James Minogue, Stephen R. Porter, Andrew McEachin, Temple A. Walkowiak & Rebecca A. Zulli


To link to this article: http://dx.doi.org/10.1080/09500693.2017.1282183

Published online: 06 Mar 2017.
The development and validation of the Instructional Practices Log in Science: a measure of K-5 science instruction

Elizabeth L. Adams, Sarah J. Carrier, James Minogue, Stephen R. Porter, Andrew McEachin, Temple A. Walkowiak and Rebecca A. Zulli

ABSTRACT
The Instructional Practices Log in Science (IPL-S) is a daily teacher log developed for K-5 teachers to self-report their science instruction. The items on the IPL-S are grouped into scales measuring five dimensions of science instruction: Low-level Sense-making, High-level Sense-making, Communication, Integrated Practices, and Basic Practices. As part of the current validation study, 206 elementary teachers completed 4137 daily log entries. The purpose of this paper is to provide evidence of validity for the IPL-S’s scales, including (a) support for the theoretical framework; (b) cognitive interviews with logging teachers; (c) item descriptive statistics; (d) comparisons of 28 pairs of teacher and rater logs; and (e) an examination of the internal structure of the IPL-S. We present evidence to describe the extent to which the items and the scales are completed accurately by teachers and differentiate various types of science instructional strategies employed by teachers. Finally, we point to several practical implications of our work and potential uses for the IPL-S. Overall, results provide neutral to positive support for the validity of the groupings of items or scales.

Recent reform documents continue to outline a vision for science instruction that is focused on engaging students with inquiry-rich science practices. Providing primary level students with varied opportunities to develop science literacy is a shared aspiration of science education reform efforts worldwide (Achieve, 2010; DeBoer, 2011; National Research Council [NRC], 2012; Waddington, Nentwig, & Schanze, 2007). Within this vision, students are encouraged to reason and make sense of science ideas; support their findings with scientific evidence; and participate productively in ways that reflect how science is practised in the real world (NRC, 2012). The Instructional Practices Log in Science (IPL-S) is a daily teacher log designed to signal the type of opportunities teachers provide students to engage in science practices and make sense of science ideas in Grades K-5. The theoretical framework for the IPL-S runs parallel to frameworks used in the development of the major international tests of student performance in science,
Measuring instruction

Measures of instruction include teacher portfolios, teacher or student surveys, daily teacher logs, and observational measures. Evidence exists supporting that observational measures such as the Framework for Teaching (Danielson, 2007) and the Classroom Assessment Scoring System (CLASS; Pianta, Hamre, Haynes, Mintz, & La Paro, 2007) are helpful for understanding aspects of teachers’ general instruction related to emotional, organisational, and instructional aspects of teaching (Bell et al., 2012; Kane & Staiger, 2012). However, scores collected with general measures such as CLASS may not reflect meaningful variation related to content-specific pedagogy. In response to this concern, several content-specific measures of instruction have emerged (e.g. Banilower, 2005; Grossman et al., 2010; Hill et al., 2008). Science-specific observational measures have shown promise in measuring aspects of inquiry-based science instruction (Barton et al., 2012; Marshall, Smart, & Horton, 2010; Piburn & Sawada, 2001; Weiss, Pasley, Smith, Banilower, & Heck, 2003).

Developers of observational measures have made significant strides in supporting the use of those measures for examining general and content-specific instruction. However, the use of observational measures is not always feasible, given that multiple observations of each teacher are often needed. Furthermore, careful rater recruitment, training, and certification are necessary to ensure consistency in scoring (Hill, Charalambous, & Kraft, 2012). Teacher portfolios also require extensive rater training and time for scoring and, therefore, face similar issues in feasibility as observational measures. Student surveys (e.g. Ferguson, 2012) have demonstrated utility for measuring general aspects of classroom climate and rigour, but not content-specific teacher instruction.

Annual instructional surveys are a less resource-intensive tool used for collecting data on instruction (Martone & Sireci, 2009), but are vulnerable to social desirability and memory errors. Most recently, the Science Instructional Practices Survey was designed to measure dimensions of science instruction (Hayes, Lee, DiStefano, O’Connor, & Seitz, 2016). This one-time survey is a powerful contribution to the field of science education. However, annual surveys ask teachers to retrospectively report about their instruction across months of teaching and are, therefore, prone to error in memory and estimation (Mayer, 1999). Annual data self-reported in this way are subject to error from both social desirability of responses and inconsistencies between teachers’ self-appraisals of what they did and the reality of what occurred in the classroom. Previous work has found that individuals tend to respond in socially desirable ways that are less consistent with reality when a question asks about their behaviour over a great length of time as is the case with annual surveys, but respond more accurately when asked about their behaviour on a daily basis (Garry, Sharman, Feldman, Marlatt, & Loftus, 2002). In a comparison of teachers’ daily log and annual survey responses, Camburn, Han, and Sebastian (2015) found that 77% of teachers over-reported aspects of their instructional practice on the annual survey. Therefore, compared to annual surveys, daily logs include less error in teacher responses, in part because they are completed on the same day as instruction.
Why a daily instructional log?

Extant research suggests daily instructional logs capture teachers’ instructional practices in mathematics and reading (Camburn & Barnes, 2004; Rowan & Correnti, 2009; Rowan, Harrison, & Hayes, 2004), but have not yet been validated for use in science. Compared to other types of measures of instruction, instructional logs are a relatively cost-effective way to collect detailed information about enacted classroom practice (Rowan, Jacob, & Correnti, 2009). Rowan and Correnti (2009) state that daily logs are an efficient and accurate way for evaluators to collect data on instruction from a large sample of teachers across many days of instruction, yet we are the first to present the development and validation of a daily science log, the IPL-S. The science education field benefits from the availability of a log focused on science instruction for several reasons, as it provides:

- more detailed information than annual surveys;
- less resource-intensive data collection than observations;
- a tool for evaluating large programmes (e.g. hundreds of teachers); and
- an opportunity for teachers to reflect on their instruction.

The IPL-S is a measure of the frequency of teachers’ science instructional practices, which signals student opportunities to engage in various behaviours during science instruction. The IPL-S captures the frequency and duration (as a proportion of time across lessons) of various instructional strategies during daily and long-term instruction. The IPL-S items include a spectrum of instructional strategies from traditional to reform-minded such that a full range of instruction can be captured with the measure.

Theoretical framework

Our attempt to capture and categorise the types of science learning opportunities provided by K-5 teachers is not unique. Our log’s framework is undergirded by two key domains of science instruction: sense-making and science practices (Figure 1). Like other researchers in the arena (e.g. Campbell, Abd-Hamid, & Chapman, 2010; Hayes et al., 2016; Tekkumru-Kisa, Stein, & Schunn, 2015; Wenning, 2005), our key domains of interest

![Figure 1. Theoretical framework for the IPL-S.](image_url)
emerged from our careful review of reform documents, national and international, both old and new (e.g. Achieve, 2010; American Association for the Advancement of Science, 1994; DeBoer, 2011; NRC, 1996, 2012; OECD, 2016; Waddington et al., 2007). Like other researchers, we make a distinction between ‘levels’ of opportunities. Our classification of high- vs. low-level sense-making and basic vs. integrated process skills is similar in nature to Wenning’s (2005) look at ‘increasing intellectual sophistication,’ Tekkumru-Kisa et al.’s (2015) ‘cognitive demand’ levels, and Hayes et al.’s (2016) levels of ‘cognitive involvement.’ PISA (OECD, 2016) has operationalised cognitive demand using a low (e.g. recall a fact, term, principle, or concept), medium (e.g. organise/display data), or high (e.g. synthesise or evaluate evidence) approach. TIMSS (Jones, Wheeler, & Centurino, 2015) uses knowing (e.g. recall/recognise and describe), applying (e.g. compare/contrast/classify and explain), and reasoning (e.g. design investigations and evaluate). These shared demarcations improve the IPL-S’s sensitivity to a range of teaching practices that signal opportunities for students to develop science proficiency/literacy.

**Sense-making**

Sense-making has been defined as student opportunities that facilitate the process of giving meaning to scientific experiences (Jones et al., 2015; NRC, 2003; OECD, 2016). It is up to teachers to encourage young students to make sense of experiences through student questioning, class discussions, and explanations (Banilower, Cohen, Pasley, & Weiss, 2010). Student sense-making helps build conceptual understanding and helps give meaning to scientific ideas (NRC, 2007). Tekkumru-Kisa et al. (2015) referred to sense-making opportunities as ‘opportunities for thinking’ where students are ‘doing science’ engaged in practices to make sense of the content (p. 5). In our work, we adopt the terms Low-level Sense-making and High-level Sense-making as a way to distinguish the cognitive demand or intellectual sophistication of the particular meaning-making opportunities or tasks.

Teachers who employ High-level Sense-making in their classrooms often ask students to use evidence to explain their understanding of scientific concepts or connect new learning to their prior knowledge, thus facilitating connecting student experiences to science content (Banilower et al., 2010). Sense-making supports students in understanding the nature of science as tentative and grounded in new learning. High-level Sense-making might require students to use cross-cutting concepts (e.g. patterns; cause and effect; systems and system models; stability and change; energy and matter; scale, proportion, and quantity; and structure and function; NGSS Lead States, 2013; NRC, 2012). In contrast, Low-level Sense-making might involve recall of information from previous lessons, learning science vocabulary, or restating scientific facts, what have been called ‘memorisation tasks’ (Tekkumru-Kisa et al., 2015, p. 6). When exposed to mostly low-level sense-making strategies, students miss opportunities to develop true science proficiency (NRC, 2007). Low-level Sense-making (i.e. recalling information from previous lessons, researching science topics, and learning science vocabulary and facts) is qualitatively different from the brand of sense-making we describe as high level (i.e. use of student discourse through argumentation and cross-cutting concepts).

Sense-making opportunities are often cultivated through scientific discourse (NRC, 2007), thus our log is sensitive to a third dimension of sense-making: **Communication**.
Many science educators believe argumentation (a form of verbal communication) contributes to conceptual change in elementary classrooms (Duschl & Osborne, 2002; Hennessey, 2003; Kuhn, 1993; Sampson & Blanchard, 2012; Sampson, Enderle, Grooms, & Witte, 2013) and supports students’ developing abilities to make sense of their world. For example, student argumentation can help students discuss and analyse dialogue in the classroom, thus contributing to their understanding of discourse in the broader science community (Duschl & Osborne, 2002). Argumentation from patterns of evidence requires practice and contributes to students’ metacognitive reflection, which promotes complex connections across scientific understandings (NRC, 2007).

Sense-making in science involves more that verbal communication. Students engaged in effective communication practices might summarise learning, write about or illustrate the targeted concepts underlying investigations, or communicate information using models or drawings, and these opportunities are captured in the IPL-S. Student-generated drawings, for example, may propel students’ sense-making even further by assisting in the formation of ideas, requiring students to strategically select, organise, and integrate words and images into a coherent mental model of the phenomenon under investigation (Van Meter & Garner, 2005). Student drawings can also be used to elicit student self-explanations, which have been shown to further conceptual understanding (Ainsworth & Loizu, 2003; Edens & Potter, 2003; Gobert & Clement, 1999; Van Meter, 2001). Effective incorporation of student drawings, which are as important as writing, reading, and talking (Ainsworth, Prain, & Tytler, 2011), can enhance student engagement, help students reason about phenomena, and communicate ideas.

Science practices

The second key domain that supported the development of the IPL-S is that of Science Practices, the investigative skills that are associated with scientific inquiry and engagement (NRC, 2012). Research in recent decades has significantly changed our understandings of how learners interact with the natural world (NRC, 2003, 2005). Science educators recognise that children bring ideas and conceptions about the world into classrooms, and these notions influence student learning. In Taking Science to School (NRC, 2007), researchers identified four strands of science proficiency for young children:

(1) know, use, and interpret scientific explanations of the natural world;
(2) generate and evaluate scientific evidence and explanations;
(3) understand the nature and development of scientific knowledge; and
(4) participate productively in scientific practices and discourse (p. 2).

Science practices are investigative skills associated with scientific inquiry that incorporate the transferable abilities that reflect the behaviours of scientists (NRC, 2012). Elementary students engage in a variety of science practices related to the four strands of proficiency from basic to more complex levels of interaction. Not unlike student sense-making, student opportunities to engage in science practices have been classified theoretically as low level or high level. Our dichotomy is similar to the classifications described by Wenning (2005) as rudimentary (e.g. observing and predicting), basic (e.g. constructing a graph), integrated (e.g. identifying a problem to investigate), and advanced (solving...
complex real-world problems). More recently, Hayes et al. (2016) defined Empirical Investigation on two levels (lower and higher cognitive involvement). The former being cookbook-type activities that require prediction, observation, and verification; the latter being activities that require student decision-making based on sense-making or analysis such as identifying testable questions from observations of phenomena. We adopt the terms Basic Practices and Integrated Practices. Low-level practices focus on base-level science activities (e.g., observing, inferring, measuring, classifying, and predicting) and high-level practices are complex and integrated (e.g., controlling variables, formulating hypotheses, interpreting data, and formulating models). While all of these activities are complex (especially for young learners), there is a noticeable progression in practices from basic (foundational skills used in isolation) to integrated (multiple skills used in concert).

Development of the IPL-S

The development and validation of the IPL-S occurred within the context of a longitudinal evaluation study of a STEM-focused elementary teacher preparation programme. The authors of this paper are part of the larger research team. One goal of the larger study is to understand the types of instructional practices that novice teachers, including both graduates of the focal teacher preparation programme and graduates from other programmes, utilise during mathematics and science instruction. The IPL-S was developed to address part of this second goal, to provide insight into novice teachers’ instructional practices in science by examining the types of student opportunities teachers provide. Simultaneously, while not a focus of this paper, the Instructional Practices Log in Mathematics (IPL-M) was developed (Walkowiak, Adams, Porter, Lee, & McEachin, in press).

The IPL-S was developed in several recursive stages, presented in Table 1, where the research team collected data using the log and made revisions to better align the IPL-S among the intended dimensions. During the development stage, three leading science education experts and the project’s external evaluators, who have expertise in science instrument validation, reviewed the log items and provided insight regarding the strength of the items in measuring science practices and sense-making as outlined by recent reform documents (NRC, 2007, 2012).

Stages One through Three covered the early development. We used exploratory factor analysis to support the theoretical scales and inform successive revisions (Greive, Carrier,

### Table 1. Log development stages and participants.

<table>
<thead>
<tr>
<th>Stage</th>
<th># of participants and logs</th>
<th># of logs per teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. First-round cognitive interviews</td>
<td>Five teachers</td>
<td>1 log per teacher during cognitive interviews</td>
</tr>
<tr>
<td>2. First pilot</td>
<td>55 teachers completed 465 logs</td>
<td>4–11 completed logs per teacher (μ = 8.45)</td>
</tr>
<tr>
<td>3. Second pilot</td>
<td>54 teachers completed 571 logs</td>
<td>3–16 completed logs per teacher (μ = 10.57)</td>
</tr>
<tr>
<td>4. Large-scale log implementation, and second-round cognitive interviews</td>
<td>136 second-year teachers completed 3612 logs</td>
<td>1–45 completed logs per teacher (μ = 26.56)</td>
</tr>
<tr>
<td>5. Experienced teacher log pilot</td>
<td>58 teachers with five or more years of teaching experience completed 502 logs</td>
<td>2–14 logs per teacher (μ = 8.65)</td>
</tr>
<tr>
<td>6. Teacher and rater log comparison</td>
<td>11 first-year teachers completed 28 logs</td>
<td>1–2 logs per teacher following observed lessons</td>
</tr>
</tbody>
</table>
Minogue, Walkowiak, & Zulli, 2014). The evidences of validity in this paper focus on the data collected during Stages Four through Six. These stages included a large-scale log implementation, second-round cognitive interviews, experienced teacher log pilot, and completion of logs by trained raters.

**Sample descriptive information**

The large-scale implementation of the IPL-S included 136 second-year teachers, who completed logs during the 2013–2014 or the 2014–2015 school year. All 136 teachers attended one in-person IPL-S training. Trainings were conducted at university facilities across one state in the Southeastern United States. Teachers in this sample graduated from either the teacher preparation programme under evaluation (33%) or another university-system teacher preparation programme in the same state (68%). Almost all teachers were female (97%), and most were White (89%). Most of these teachers were either 23 or 24 years old when they participated in this study (87%). The teachers were relatively evenly distributed across Grades K-2 and 3–5 (56% taught Grades K-2). Five of these teachers also participated in a cognitive interview for the purpose of this paper. Of these five teachers, all teachers were White; four were female, and teachers taught either Grade K, 2, 3, 4, or 5.

The experienced teacher log pilot included a sample of 58 teachers who had been teaching for at least four years. Experienced teachers logged during 2013–2014 and taught in one of two geographical areas, designated as school districts, in the same state as the beginning teachers (52% taught in one district). More than half the teachers taught Grades K-2 (59%). All teachers attended one of two log trainings which were conducted in person at a school district facility.

The teacher and rater log comparison included 11 first-year teachers who graduated from one teacher preparation programme. Most teachers were White (91%) and female (91%). All teachers were between 22 and 24 years old at the time of the observation. More than half of these teachers taught Grades 3–5 (64%). These teachers attended one of the in-person trainings that were held at university facilities within the state.

**Description of the IPL-S**

The IPL-S is formatted as an online survey which teachers complete shortly after a lesson. Teachers are asked to log only about the behaviours in which at least half of the students in the class engaged during the science lesson. In order to ensure a common understanding of key terms and response options, teachers are provided with a hard copy and an electronic version of the IPL-S User’s Guide, which defines each item and response options, as well as provides examples of each item.

The items on the IPL-S that were included in this validation study are summarised in Table 2. Twelve items capture science content, which provides important contextual information about the science lesson. Items are included on the IPL-S capturing the range of science content that elementary teachers target during science instruction, including:

- Physical Sciences: Motion and Stability, Matter and Its Interactions, Wave Properties, and Energy;
One item captures the number of minutes spent teaching science, providing additional contextual information about the science lesson. The IPL-S includes an item related to the teachers’ use of diagrams during the science lesson, which is scored dichotomously as occurring during the lesson or not. The IPL-S includes four items related to the instructional goals of the science lesson. Response options for these items include a primary goal, a secondary goal, or not a goal today. Thirty-five items on the IPL-S target student activities and behaviours during the science lesson. The response scale for these items measures the proportion of time that the majority of students spent engaged in the item’s specified behaviour. The response options require teachers to select one of the following choices for each of 35 student behaviours:

1. Not today: this behaviour was not done during today’s instruction
2. Little: this behaviour made up a relatively small part of the instruction
3. Moderate: this behaviour made up a large portion, but NOT the majority of the instruction
4. Considerable: this behaviour made up the majority of today’s science instruction

**Evidences of validity**

We describe evidence of validity for the student opportunities measured by the IPL-S using the validation standards outlined by AERA, APA, and NCME (2014). This paper examines three assumptions of validity for the IPL-S: the items and scales (a) are
interpreted as intended by teachers; (b) are completed similarly by teachers compared to raters; and (c) empirically explain instructional variance between teachers. In order to understand the extent to which these assumptions of validity are met, we present sources of validity evidence related to:

- Teachers’ response processes: the fit between the nature of teachers’ responses as they complete the tool and the intended use of the item;
- Relations to other variables: an analysis of the relationship between teachers’ scores and raters’ scores on the tool; and
- Internal structure: the degree to which the relationships between items support the dimensions specified in the theoretical framework.

The evidence of validity presented was informed by the work of Bell et al. (2012) and Kane (2006, 2013) to identify the strengths and weaknesses of the items and scales based on the assumptions above.

**Evidence based on teachers’ response processes**

Evidence based on response processes investigated the fit between the nature of teachers’ responses to an item and the intended use of the item. Two sources of evidence were examined, including (a) five cognitive interviews with teachers as they completed the IPL-S and (b) item descriptive statistics for beginning and experienced teachers.

**Cognitive interviews**

We conducted cognitive interviews to ensure that teachers interpreted the IPL-S items as intended. Three trained researchers conducted a total of five cognitive interviews with second-year teachers. During the interviews, teachers were asked to think aloud as they completed the IPL-S. All five teachers taught in K-5 public schools, were White, and four were female.

The three researchers followed a structured cognitive interview protocol based on Willis (2005). After the lesson observation, researchers asked participants to read items out loud as they completed the IPL-S and to verbalise their thinking when answering each item (Groves et al., 2009; Willis, 2005). Researchers explained that they wanted to hear all of the teachers’ reactions and interpretations, especially when something was unclear, hard to answer, or did not apply. Researchers used probes when teachers paused or seemed unsure. Probes included, ‘Tell me more about [item],’ ‘Why did you say [response]?’ ‘In your own words, what is [item] asking?’ (Willis, 2005). Researchers also asked teachers to describe any items they thought should be added or deleted from the IPL-S and how the IPL-S could be improved overall.

In order to analyse the cognitive interview transcripts, we used a four-component model for survey response called Comprehension–Retrieval–Judgement–Response outlined by Groves et al. (2009) and Tourangeau, Rips, and Rasinski (2000). The model specifies four iterative stages of the survey response process, including (a) comprehension of the item, (b) retrieval of relevant information, (c) use of that information to make required judgements, and (d) selecting and reporting an answer. Results for the IPL-S are described below based on this model.
Comprehension. Comprehension includes the process of attending to the directions and survey items, then understanding what is being asked (Groves et al., 2009). The majority of the items on the IPL-S did not present issues related to item comprehension for teachers. However, the following terms on the IPL-S consistently presented confusion for teachers – objects or specimens, simulations, variable, and inference. In one case, when the teacher referred to the User’s Guide to answer these items, she reported that the terms were clear.

Retrieval. Retrieval is the process of recalling the information necessary to respond to items from memory (Groves et al., 2009). We did not observe instances where teachers stated that they could not remember if a behaviour happened. In most cases, teachers completed the IPL-S and the cognitive interview immediately after their lesson. Based on the cognitive interviews in which we engaged, when teachers completed the IPL-S immediately following instruction, retrieval did not present issues for teachers.

Estimation. Estimation requires teachers to fill in gaps in what is retrieved in order to answer the item (Groves et al., 2009). The results of the cognitive interviews suggested if a science demonstration occurred, teachers found it difficult to estimate the time that students spent engaged in various practices. In these cases when students mostly watched the teacher carry out an investigation (i.e. demonstrate), teachers struggled to report how much time students spent engaged in behaviours. Based on the cognitive interviews in which we engaged, science demonstration was the only instructional format that presented an issue in estimation for teachers. Small group work, individual work, and class discussion did not present estimation issues for teachers.

Reporting. Reporting is the process of selecting and communicating a response for each item (Groves et al., 2005). Teachers did not express concerns when using the four-point response options, which pertained to the proportion of time students spent engaged in various behaviours (i.e. ‘not today,’ ‘little,’ ‘moderate,’ or ‘considerable’). Teachers felt that it was more difficult to use the three-point response options (i.e. ‘not a goal today,’ ‘secondary goal,’ or ‘primary goal’), and were less clear on how to discriminate between ‘not a goal,’ ‘secondary goal,’ and ‘primary goal.’ This aligns with the work of Burstein et al. (1995), who found that teachers experienced difficulty when reporting their emphasis on instructional goals on a daily log.

Teacher responses to items
The second piece of evidence for the response process looks at item descriptive statistics for 136 beginning teachers, who completed 3612 logs, and 58 experienced teachers, who completed 502 logs about their teaching practices. We included both beginning and experienced teachers to determine how the log items performed across teachers with various levels of experience. Here, we provide evidence that teachers use the response options appropriately and that substantial variation lies between teachers. That is, individual items on the IPL-S allow us to distinguish between teachers based on their self-reported behaviour in the classroom.

In terms of the response options, teachers can report three types of behaviour. First, if every teacher answered ‘not today’ or ‘not a goal’ for an item, the item is clearly not
measuring something that takes place in the classroom. Second, at least some teachers should be choosing both the minimum and maximum response options, indicating that teachers are using the full scale when responding. So while we would not expect to see all teachers choosing ‘not today’ as a response option, we also would not expect to see all teachers avoiding that response option. Third, given the demands on teachers’ time in the classroom, we would expect that the proportion of teachers choosing ‘considerable’ or ‘primary goal’ for an activity to be smaller than the proportion choosing ‘not today.’

**Item descriptive statistics.** Looking across the 35 items that use the ‘not today,’ ‘little,’ ‘moderate,’ and ‘considerable’ response options, the median proportion of lessons for which beginning teachers chose ‘not today’ for an item was .77, with minimum and maximum proportions of .22 and .97. The median proportion of lessons for which teachers chose ‘considerable’ for an item was .04, with minimum and maximum proportions of .00 and .20. These numbers suggest that for a given log item, some, but not all, teachers used the ‘not today’ option, some also chose the ‘considerable’ option, and as expected, the proportions for ‘not today’ were much higher than the proportions for ‘considerable.’ The four items using the three-point options ‘not a goal,’ ‘secondary goal,’ and ‘primary goal’ showed a similar pattern.

In addition, in order to illustrate the strength of the log items in identifying variation between beginning teachers, each log item should show substantial variation (as measured by standard deviations), indicating that teachers are not all choosing the same response category. The median standard deviation is .80 for the four-point scale, with a minimum of .33 and maximum of 1.22, indicating substantial heterogeneity in teacher responses for each log item.

Besides descriptive statistics for each item, another element of our validity argument is that the log distinguishes among teacher classroom practices. If true, we should see substantively significant variation between teachers for each log item. We estimate the intraclass correlation (ICC) using Stata 14 (StataCorp, 2015) to measure the proportion of variation that lies between teachers, versus within teachers. Between-teacher variation refers to behavioural differences among teachers, while within-teacher variation refers to day-to-day differences in a teacher’s instruction. We use .10 as the cut-off to indicate substantive significance, because this is a common threshold used to determine whether enough variation lies between units to merit the use of multilevel modelling (Raudenbush & Bryk, 2002).

The top panel in Figure 2 shows the distribution of the ICCs for all 40 log items for beginning teachers. The ICCs for each item grouped by scale are available in Web Appendix A. For beginning teachers, the minimum ICC was .11, higher than our threshold, and the median ICC was .19, indicating that substantial variation in log item responses lies between teachers. In other words, individual log items appear to be distinguishing between teachers based on their reported behaviours in the classroom.

The data for experienced teachers were similar to the results for beginning teachers in terms of how teachers respond to the log items. The experienced teacher ICCs for 37 of the 40 items were larger than .10, and the median ICC was .29. As seen in Figure 2, one noticeable difference was that the log items did a better job discriminating between experienced teachers than beginning teachers, as evidenced by the higher ICCs for experienced teachers. This might also imply that experienced teachers’ instructional practices are more
varied than that of beginning teachers, who may follow district guidelines for instruction more closely. The beginning teachers also graduated from university-system teacher preparation programmes in one state, which may explain why their instructional practices are more homogeneous compared to that of experienced teachers.

Overall, the cognitive interviews and the item descriptive statistics support that teachers reflect on items and use the response options in ways that align with their intended use. The ICCs for the items suggest that the behaviours on the IPL-S occur during instruction to varying degrees between teachers.

**Evidence based on comparisons of teachers’ and raters’ logs**

A total of 28 science lessons from 11 teachers were observed by one of seven trained IPL-S raters, who also completed the log. These raters attended at least one of the log trainings with the teachers. Four of the authors of this paper were raters. Two other raters were graduate students in education, and one rater was a researcher on the project. The raters completed a hard copy of the log following the lesson observation. We examined the relationship between the pairs of logs for the 28 lessons in two ways. First, we calculated the share of exact matches and comparable matches between the teacher and rater overall and by scale. Exact matches are defined as teacher and rater responses to an item that match perfectly. Comparable matches include exact matches and also include matches that are comparable. We define comparable as teacher and rater responses that are either ‘not today’ or ‘little,’ as well as ‘moderate’ or ‘considerable.’ This is similar to the one-off agreement analysis in Hill (2005). Second, we averaged the items within

---

**Figure 2.** CCs for log items.
each scale to create five simple factor scores for each lesson for both teacher and rater log responses. We graphed the bivariate relationship and computed the Pearson correlation for each of the five scales to establish log item interpretation across raters.

**Item agreement**
In Table 3 we report the agreement analysis between teacher and rater responses to items for the 28 pairs of lessons. The first panel reports the share of exact matches across all items between teacher and rater log responses after observing the lesson in class. On average across the 28 lessons, teachers and raters had exact agreement on 66% of the items, with a minimum of 46% and a maximum of 80%. We also report the exact agreement for each of the five scales. Communication is the only scale with a less than 50% match rate, with exact match rates of 54%, 66%, 43%, 73%, and 82% for Low-level Sense-making, High-level Sense-making, Communication, Basic Practices, and Integrated Practices, respectively.

The bottom half of Table 3 reports results for our comparable agreement analysis. We define comparable agreements as one-off matches for ‘1’ and ‘2’ (not today and little) or ‘3’ and ‘4’ (moderate and considerable). We also include exact matches in the comparable agreements; as such, the results report the share of at least comparable agreements. The overall agreement rate increased slightly to 68% when we included comparable matches with a minimum of 53% and a maximum of 84%. Furthermore, all of the scales had comparable agreements between 55% and 88%. The comparable agreement match rates indicate that the teachers and raters provided qualitatively similar scores for a vast majority of the log items.

**Correlational analysis**
We next present scatter plots between the scale scores for the teacher and rater for each of the five scales, as well as the Pearson correlations. It is important to note that dots on the scatterplots are superimposed because there were cases of lessons sharing the same bivariate relationship. We did not use a confirmatory factor analysis (CFA) or similar data reduction method to create the factor scores for each of the five scales due to the limited sample size ($n = 28$ lessons). The scale scores we used were the arithmetic averages.

| Table 3. Agreement between teachers’ and raters’ log scores. |
|-----------------|-------------------|-------------------|
| **Exact agreement** | Mean (%) | Min (%) | Max (%) |
| All scales | 66 | 46 | 80 |
| Low-level Sense-making | 54 | 0 | 100 |
| Communication | 43 | 0 | 71 |
| High-level Sense-making | 66 | 27 | 100 |
| Basic Practice | 73 | 45 | 100 |
| Integrated Practice | 82 | 38 | 100 |
| **Comparable agreement** | All Scales | 68 | 53 | 84 |
| Low-level Sense-making | 68 | 25 | 100 |
| Communication | 59 | 29 | 86 |
| High-level Sense-making | 55 | 36 | 82 |
| Basic Practice | 84 | 64 | 100 |
| Integrated Practice | 88 | 50 | 100 |

Note: The items with a three-point response scale were not included in the comparable agreement analysis. This explains why the percent comparable agreement is lower than the exact agreement for High-level Sense-making.
of the items in a given scale (i.e. Basic Practices). We report the scatter plots and correlations in Figure 3.

The scatter plots in Figure 3 show the scale scores generated from the teacher logs on the Y-axis and the scale scores generated from the rater logs on the X-axis. Each dot represents the scale score for a given lesson. If the scores from the teachers and raters were identical, all of the dots would fall along the imaginary 45-degree line in the scatter plot. In order to detect a relationship between the two scale scores, we need to observe meaningful variation within the scales and the co-variation between the scales needs to move in the same direction.

The scatter plots show a few interesting patterns. First, in the case of Integrated Practices, the variation in scale scores by one group (i.e. teachers) was much smaller than that of the other group (i.e. raters). The scale scores for the teacher logs ranged from 1 to 1.5.

Figure 3. Bivariate relationships between teachers’ and raters’ log scores.
However, the rater scale scores ranged from approximately 1 to 2.5. While the teacher logs indicated practices do not vary much from teacher to teacher, the rater logs showed teachers do in fact spend different amounts of time providing students opportunities to use Integrated Practices during their observed lessons. The moderate correlation suggests that teacher and rater scores tended to vary in the same direction ($r = .48$). The average difference in scale scores (teacher minus rater) was $-0.18$, indicating teachers tended to report fewer Integrated Practices.

Second, in some instances, it appears that teachers and raters do not agree on the practices that occurred during the 28 lessons. On Low-level Sense-making, the correlation between teachers and raters was effectively zero ($r = -.06$). Similar to other researchers’ findings (Camburn & Barnes, 2004; Desimone, 2006), this may be due to the differing perspectives between teachers and raters or between raters themselves. This suggests that teachers tended to rate less socially desirable practices less frequently compared to raters. On the other hand, teachers tended to agree with raters on scales with more desirable practices (e.g. Communication, High-level Sense-making, and Basic Practices). It is important to note that these results were based on seven raters’ observations of 28 science lessons, and the large number of raters who engaged in this work may have contributed to a low correlation.

Third, the scales Communication, High-level Sense-making, and Basic Practices have meaningful variation within each scale for teacher and rater scores, and the scores co-varied in the same direction. For example, scores on the Basic Practice scale ranged from 1 to 2.5 for both teachers and raters, and lessons with higher scale scores by teachers also had higher rater scores ($r = .82$). Similarly, the average difference scores of approximately zero for the three scales indicate teachers and raters reported qualitatively similar student behaviours relative to these three scales.

**Evidence based on internal structure**

Two-level Cronbach’s alpha coefficients provide evidence for the empirical strength of the internal structure of the IPL-S. Examining the sample of 136 beginning teachers, who logged on average 26.6 science lessons, we calculated two-level Cronbach’s alpha coefficients, which are measures of the extent to which responses to the items are correlated with the other items on that scale. These statistics were estimated for each scale at the within and between levels in Mplus 7 (Muthén & Muthén, 2010) using the two-level covariance–variance matrix as detailed in Geldhof, Preacher, and Zyphur (2014).

The Cronbach’s alpha coefficients in Table 4 were at or above .88 at the between level for all but one scale (i.e. Low-level Sense-making), supporting the strength of the scales. We designed the IPL-S to detect differences between teachers in instruction, not necessarily differences in day-to-day instruction within teachers. The higher between-level alpha coefficients, coupled with the item-level ICCs above .10 presented earlier, support the use of the items and the scales in differentiating types of instruction between teachers.

A two-level CFA was fit to understand the degree to which the relationships between items support the dimensions specified in the theoretical framework. Using teacher codes as the cluster or nesting variable, the CFA was specified at the within-teacher (day) level and between-teacher (teacher) level (see Figure 4 for the theoretical model). Given the positively skewed distributions for the items (e.g. the majority of lessons reflecting ‘not today’), we used the weighted least squares means and variance adjusted estimator
for categorical data (Muthén & Muthén, 2010). Fit criteria were established specifying an overall root mean square error of approximation (RMSEA) below 0.06, as well as standardised root mean square residual (SRMR) values below .10 for the within and between models (Bowen & Guo, 2011; Byrne, 2012).

K-5 beginning teachers (n = 136) logged 3612 science lessons with an average of 26.6 lessons logged per teacher (minimum of 1 and maximum of 45 lessons per teacher). Our primary goal was to provide empirical evidence of the strength of the scales in explaining variability between teachers, rather than within teachers across time.

The CFA results indicated moderate fit between the theoretical model and the data. The RMSEA value indicated adequate fit (.046) with 1476 degrees of freedom. However, the SRMR values are not below the cut-off of .10 for the between or within models (SRMR between = .105 and within = .132). These fit statistics support that the data fit the model better at the between level than the within level.

### Table 4. Two-level Cronbach’s alpha coefficients for factors on the IPL-S.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Between Cronbach’s α</th>
<th>Within Cronbach’s α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-level Sense-making</td>
<td>.756 (.144)</td>
<td></td>
</tr>
<tr>
<td>High-level Sense-making</td>
<td>.913 .600</td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td>.880 .606</td>
<td></td>
</tr>
<tr>
<td>Basic Practices</td>
<td>.896 .747</td>
<td></td>
</tr>
<tr>
<td>Integrated Practices</td>
<td>.925 .622</td>
<td></td>
</tr>
</tbody>
</table>

*Note: The within alpha for low-level sense-making is listed in parentheses because it was not included in the full CFA model.*

Figure 4. Theoretical model for the IPL-S.

*Note: We removed items at the within level with loading below .30, resulting in 40 items at the between level and 26 items at the within level. Table A1 includes a crosswalk between the item numbers and item language.*
Standardised results for the between and within model are summarised in Table 5. The full CFA results are available in Web Appendix B. The significant factor variances for the within and between models indicate that the scales explained a statistically significant amount of variance within and between teachers. For the between model, individual item loadings ranged from .46 to .93, indicating the strength of each item in understanding its specified dimension. Each item loading is statistically significant at \( p < .001 \), suggesting that each item contributes substantially to understanding the dimension.

A correlation matrix at the between level is presented in Table 6. Most scale correlations were below .71, but three scales were highly correlated at the between level. The correlations between High-level Sense-making and Integrated Practices was .86; Basic Practices and Integrated Practices was .94; and High-level Sense-making and Basic Practices was .87. These high correlations may explain why fit statistics did not meet the specified thresholds. Theoretically, these high correlations make sense given the related nature of these dimensions. Teachers who engage in High-level Sense-making often build on both aspects of Basic Practices and Integrated Practices to build students’ conceptual understandings (Padilla, 1990).

Collectively, the Cronbach’s alpha coefficients and CFA results provide empirical evidence supporting the moderate fit between the data and the theoretical scales in describing the type of instruction between teachers.

**Summary and conclusion**

Based on the evidences of validity presented, we claim that the items on the IPL-S:

- Are usually interpreted as intended by teachers,
- Are completed with fair accuracy by teachers when compared to raters, and
- Explain a statistically significant amount of variability in instructional practice between teachers.

Overall, we found the following:

---

**Table 5.** Summary of the two-level CFA results for the between model.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Between factor variance ((\xi))</th>
<th>Between item loading range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-level Sense-making</td>
<td>.42***</td>
<td>.51–.86***</td>
</tr>
<tr>
<td>High-level Sense-making</td>
<td>.12**</td>
<td>.53–.89***</td>
</tr>
<tr>
<td>Communication</td>
<td>.33***</td>
<td>.57–.87***</td>
</tr>
<tr>
<td>Basic Practices</td>
<td>.12**</td>
<td>.50–.78***</td>
</tr>
<tr>
<td>Integrated Practices</td>
<td>.94***</td>
<td>.63–.93***</td>
</tr>
</tbody>
</table>

Significance is indicated with ***\(p < .001\) and **\(p < .01\).

**Table 6.** Correlation matrix for the two-level CFA at the between level.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Low-level Sense-making</th>
<th>High-level Sense-making</th>
<th>Communication</th>
<th>Basic Practices</th>
<th>Integrated Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-level Sense-making</td>
<td>–</td>
<td>.71</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>High-level Sense-making</td>
<td>.72</td>
<td>.87</td>
<td>.64</td>
<td>–</td>
<td>.94</td>
</tr>
<tr>
<td>Communication</td>
<td>.48</td>
<td>.86</td>
<td>.60</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Basic Practices</td>
<td>.47</td>
<td>.86</td>
<td>.60</td>
<td>.94</td>
<td>–</td>
</tr>
<tr>
<td>Integrated Practices</td>
<td>.47</td>
<td>.86</td>
<td>.60</td>
<td>.94</td>
<td>–</td>
</tr>
</tbody>
</table>

Note: Bold correlations are at or above .86. Correlations are all significant at \(p < .001\).
Teachers self-report their daily enacted practice fairly accurately

Evidence of this included teachers’ verbalised thinking as they completed the log and the exact agreement between teachers’ and raters’ log scores, which ranged from 43% to 83% of the time. However, teachers did not report their practice as accurately when the behaviour or concept was not directly observable (e.g. instructional goals). We found that teachers and raters were more likely to disagree on items that may be considered less desirable by teachers (e.g. Low-level Sense-making). However, this may also have been attributable to the varied perspectives of the seven raters, who included faculty content experts, graduate students, and researchers (Desimone, 2006). We did not observe systematic over- or under-reporting by teachers on the scales, with the exception of Integrated Practices. Teachers tended to under-report Integrated Practices compared to raters. Camburn and Barnes (2004) found that in a comparison of teacher and rater reports of the same lesson, teachers tended to not report behaviours that happened frequently in their classrooms. Therefore, teachers who use these practices frequently may do so without realising the extent to which they use the practices, given that they are common in some teachers’ classrooms.

Daily logs work well for discriminating between teachers in their instructional practice

To our knowledge, a multilevel CFA has not been applied to data collected with a science log. Based on our CFA, we found that the items and the scales on the IPL-S identified substantial variability between teachers in their science instruction based on the item ICCs and the CFA. This finding aligns with the work of Rowan, Camburn, and Correnti (2004), who found that even with large variance within teachers in their instructional choices on any particular day, daily logs can be used to discriminate between teachers in their patterns of enacted teaching. This work supports that daily logs differentiate between teachers in their practice in statistically meaningful ways.

Limitations

The research team acknowledges several limitations to this work. First, reliability of the log scores, as in how many logs are needed for a teacher to obtain a stable estimate of her teaching, was not included in this paper. Second, grade-level differences were not examined. Additionally, this validation work primarily relies on logging responses for beginning teachers. Finally, the IPL-S was validated with elementary teachers and the applicability of the IPL-S beyond elementary has not been examined.

Implications

The IPL-S is a daily teacher log designed to signal the type of opportunities teachers provide students to engage in science practices and make sense of science ideas in Grades K-5. The practical implications of our work lie in the potential usefulness of the IPL-S at various levels of science education reform efforts, both here and abroad (Abd-El-Khalick et al., 2004; Achieve, 2010; DeBoer, 2011; Waddington et al., 2007). The use
of the IPL-S can support teachers, administrators, teacher educators, and researchers by providing a resource-efficient instrument that yields important data. The IPL-S is not a measure of teacher effectiveness. When used appropriately, it can help (a) facilitate teachers’ reflective practice and self-assessment (Appleton, 2013; Zeichner & Liston, 2013); (b) inform science specialists in prioritising professional development efforts; (c) aid administrators in identifying effective science strategies in their classrooms (Grisom, Loeb, & Master, 2013; Millar & Osborne, 1998; National Center for Educational Research & Development, 1997; Neumerski, 2013); and (d) inform teacher educators and researchers regarding the status of science teaching in their local context. IPL-S data can be used to assess the impact of a programme or curricular reform efforts on the ground through measuring teachers’ enacted instruction. The IPL-S shines a light on instructional practices that lie at the core of primary school science reform. We believe the diagnostic information generated by the IPL-S can ultimately translate into more and better reform-minded elementary science teaching across the globe.

Note

1. The IPL-S User’s Guide is available starting on page 31 at http://go.ncsu.edu/ipls_users_guide.

Acknowledgements

We greatly appreciate the support of Teresa Frasca and feedback from staff at Horizon Research, Inc. during the development of this paper. We also would like to acknowledge the contributions of Dr Ellen McIntryre and Dr Daniell DiFrancesca during the development of the log. We would like to thank those who provided feedback on this work at the 2013 AEA Conference in Washington, D.C.; 2014 AERA Conference in Philadelphia, PA; and 2015 NARST and NCME Conferences in Chicago, IL. We would also like to express our sincere gratitude to the teachers who participated in this study. 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.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is funded by the National Science Foundation [award #1118894].

References


Sampson, V., Enderle, P., Grooms, J., & Witte, S. (2013). Writing to learn by learning to write during the school science laboratory: Helping middle and high school students develop argumentative writing skills as they learn core ideas. *Science Education, 97*(5), 643–670.


