Improving Understanding of Teaching Practice for Student Learning: A Holistic Measure of Fidelity of Implementation

Eileen McGivney, Emily Gonzalez, Sabrina De Los Santos, Amy Kamarainen, and Tina Grotzer

Presented to the National Association for Research in Science Teaching (NARST)
April 2, 2019

National standards for science education emphasize inquiry-based learning as an effective way to expose students to authentic scientific practice, scientific reasoning, and epistemic knowledge (National Research Council, 2012). Technology-enabled immersive environments can help facilitate authentic inquiry-based learning opportunities (Grotzer et al., 2017), but also require educators to employ various moves, strategies, and social roles that constitute inquiry-based instruction (Minner & DeLisi, 2012; Zhai & Tan, 2015).

The effectiveness of curricula and interventions depends on teachers' implementation and how they utilize resources in their classrooms. Nevertheless, traditional methods of evaluating program effectiveness often fail to account for variation in teacher practice and their fidelity of implementation to explain variation in student outcomes (O'Donnell, 2008; Durlak & DuPre, 2008). A clear and detailed picture of teacher behaviors and practices when implementing inquiry-based curricula would help to more accurately assess its effectiveness, improving causal inference of which curricular components improve student learning (DeSimone & Hill, 2017). It would also lend to a better understanding of the contexts and conditions under which educational technology is most effective (Fishman & Dede, 2016).

Fidelity of implementation is increasingly included in studies of educational interventions. Yet researchers often do not report the criteria under which the measures were developed, analyze outcomes in light of characteristics of implementation, or measure how the quality of delivery varies between teachers (Odom et al., 2010; Mowbray et al., 2003; Swanson et al., 2013; Durlak & DuPre, 2008), and researchers may be lacking tools and procedures for doing so. For example, in a review of fidelity measures in education research, O'Donnell (2008) concluded there are "too few studies to guide researchers on how fidelity of implementation to core curriculum interventions can be measured and related to outcomes" (p 54).

This study contributes to a growing literature to fill this gap by reliably and comprehensively measuring fidelity of implementation and utilizing principles from psychometrics (e.g. Snyder et al., 2013; Southam-Gerow, 2018; Kim et al., 2017). Defining validity as a unitary construct which requires evidence from an observational measure's content and scoring process as well as its coherence, we first outline how our constructs for fidelity were defined (AERA, 2014; Snyder et al., 2013), focusing on those pertaining to the quality of implementation through student-teacher interactions. We report how the content of the measures were developed, the iterative process to develop a reliable scoring procedure, analysis of the measures for their coherence and precision using classical test theory and generalizability theory, and our investigation into the multidimensionality of the scores and possible latent constructs underlying teachers' differing implementation.

The evidence reported here provides a model for assessing the validity of implementation measures based on teachers' interactions with students, as well as an adaptable instrument measuring implementation of programs that rely on inquiry-based teaching practices. The findings have important implications for understanding how teachers vary in their interactions with students in such environments and contribute to better understanding the causal impacts of programs as implemented compared to as intended (Dobson & Cook, 1980).

EcoXPT: Teaching causality through experimentation in virtual environments

This paper describes the fidelity of implementation measures developed for an evaluation of EcoXPT, a problem-based learning curriculum set in a virtual pond to teach ecosystem science concepts.[1] EcoXPT builds off of the EcoMUVE program, an earlier multi-user virtual environment in which students learn about ecosystems through observational inquiry, by incorporating experimental tools that allow students to investigate causal patterns by testing their hypotheses (Dede et al., 2017). The evaluation was conducted in the 2017-2018 school year, and 10 teachers were observed teaching two versions of EcoXPT to 923 students across 40 classes that were randomly assigned to conditions with and without the new experimental tools. The effectiveness of the curriculum was assessed through student pre- and post-tests that measured learning on science content, epistemology, affect, and understanding of causality (Thompson et al., 2016).

Table 1: Study Design for the EcoXPT Evaluation

Curriculum	***Tools***		***No Tools***	
Class number	**1**	**2**	**3**	**4**
Teacher				
A	1	1	1	1
B	1	2	1	1
C	1	1	1	1
D	1	1	1	2
E	1	2	1	2
F	2	1	2	1
G	1	1	1	1
H	1	2	1	1
I	1	1	1	1
J	2	1	1	2

Fidelity data collection and analysis design in which each teacher was recorded in 4 classes. Each cell indicates the number of raters; videos rated by two people are highlighted.

Spanning 13 days, each lesson in the curriculum consists of a combination of teacher-directed activity and student-driven small group work. Lessons may start with a warm-up "do now" activity, teacher-driven introduction to the goals of the day and/or video introducing "thinking move" strategies aligned to Next Generation Science Standards (NGSS) cross-cutting concepts and scientific practices (Achieve, 2013). For the remainder of the lesson, students work in pairs exploring the virtual world by collecting data, conducting experiments, and building a concept map. The role teachers play in helping guide students through their inquiry is central to the

[1] ecolearn.gse.harvard.edu/ecoXPT

program, and as students work in pairs on investigating the ecosystems, teachers are expected to circulate around the classroom providing guidance via small-group interactions. To orient them to the curriculum, teachers either received group professional development or one-on-one guidance that walked them through the program, depending on the school's availability of professional development days.

The primary source of fidelity of implementation data was videotaped observations of each class on the eighth day of the curriculum, resulting in two observed classes for each teacher with the experimental tools treatment, and two classes using the control version without experimental tools. Supplementary sources of data also included students' log file data, and the teachers' and program team's logs of interruptions to the sequence of the curriculum (see Appendix A). A primary rater scored all videos, while a second rater scored a random 20% for reliability analysis. Table 1 provides a visual representation of the study design. The process was carried out by members of the EcoXPT project team from Harvard and the project's outside evaluation team from TERC; the primary rater was an outside evaluator and the secondary rater a member of the project team.

Content: Defining fidelity and a measurement procedure in an inquiry-based curriculum

We define validity as the "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences" based on a score (Messick, 1987, p. 2). In other words, do the scores meaningfully support our interpretations of how teachers differed in their implementation of EcoXPT? By defining validity as a unitary construct in this way, the content of the measure is crucial, and evidence must support that it meaningfully captures the construct of interest (AERA, 2014; Cook & Beckman, 2006).

Fidelity of implementation refers broadly to "how well an intervention is implemented in comparison with the original program design during an efficacy and/or effectiveness study," (O'Donnell, 2008, p. 33). More specifically, fidelity can be conceptualized as how well the implementation adhered to the intended structural components and duration of a program, as well as the process of implementation, including the quality of delivery (Mowbray et al., 2003; Dane & Schneider, 1998; Dusenbury et al., 2003).

For EcoXPT, we determined that fidelity to both the structure, such as whether the materials were utilized and lessons taught in the proper sequence, and process, or the quality of delivery based on teachers' practices and interactions with students, of the curriculum were crucial factors in the program's success, and so the fidelity of implementation measures included aspects of adherence, dosage, and quality. The measures of adherence and dosage are described in Appendix A; the remainder of this paper focuses on the development and validation of the process measures related to the quality of implementation. The focus on process and quality is warranted for several reasons. For one, these measures of student-teacher interactions are more applicable to other curricula, and so have greater implications for further research on inquiry-based programs. Second, compared to adhering to structural components of the intervention (Durlak &DuPre, 2008; Odom t al., 2010), fewer fidelity measures assess the process components of implementation and ask how well delivery met a theoretical ideal. Third, the quality of implementation measures were the most challenging to develop and validate, and we hope describing our challenges and procedures can provide a model for other researchers.

The quality dimension of fidelity of implementation refers to “ratings of provider effectiveness which assess the extent to which a provider approaches a theoretical ideal in terms of delivering program content” (Dusenbury et al., 2003, p.244). In terms of inquiry-based learning environments such as EcoXPT, there is an important ideal for the ways in which teachers interact with students. Teachers must employ various moves, strategies, and social roles that constitute inquiry-based instruction, such as encouraging students to explore their own questions and construct explanations rather than pushing them toward one “right” answer (Minner & DeLisi, 2012; Zhai & Tan, 2015). Further, EcoXPT is designed to foster scientific reasoning skills in line with the NGSS, relying on teachers to encourage students to engage in epistemically authentic work, such as supporting their claims with evidence, reflecting on uncertainties, understanding and adopting accurate scientific language, engaging with authentic tools, and attending to complexity (Berland et al., 2015). In addition to the literature on teaching practices that support inquiry-based learning and scientific reasoning, the program team engaged in an emic, grounded approach by observing teachers during the pilot phase and in previous implementations of EcoMUVE to identify specific behaviors that were expected to help or hinder student learning.

From this process, a core list of teacher behaviors was developed to characterize their interactions with students; Table 2 provides a list of the final 25 items on which student-teacher interactions were scored. The items were grouped by “umbrella terms” in order to more easily conduct the rating procedure, grouping the types of interactions that were often observed together, such as discussing patterns and evidence or probing students’ thinking and encouraging them to articulate their ideas. However, these groupings do not imply separate constructs, and we did not hypothesize these groups to be separate domains or different facets of quality implementation.

Alongside defining the content of the measure, the project team and outside evaluators conducted a thorough vetting process of the items and developed the scoring procedure to ensure these behaviors within interactions could be consistently identified. In other words, we wanted to ensure that raters were engaging in the same *response process*, another important area of building validity evidence (AERA, 2014; Messick, 1987). Different scoring procedures were trialed in which videos were divided into distinct student-teacher interactions versus defined time intervals. In the end, the raters scored each teacher on a minute-by-minute basis in which the rubric was used as a checklist, indicating whether the behavior was observed within the minute. Further, as trial scoring proceeded a detailed set of decision rules was created for each item, including examples of interactions in which the code should be scored, as well as examples when it should not, and highlighted items that may frequently be coded together or were similar in nature. During this process, interrater agreement statistics were used to judge the ability for different raters to achieve reliable results, and a number of items were dropped due to ambiguity and difficulty in reliably coding them. For example, the original rubric included an item “overshares key content too early, such that it inhibits student exploration and learning," which raters could not reliably interpret due to the ambiguity of whether telling students content was over-sharing versus helpfully correcting misconceptions.

Table 2: Items for Quality of Implementation

Organizing Groups	***Item Number***	***(+ / -)***	***Description of Teacher Behavior***	***Interrater Percent Agreement***
Inquiry-based practices	1	+	Responds to conceptual uncertainty in an open-ended manner	89.4%
	2	+	Encourages students to feel comfortable with uncertainty/investigation	97.6%
	3	+	Encourages students to give their best explanation or multiple explanations based upon what the evidence suggests.	95.4%
	4	-	Hints that there is a "right answer" to the curricular problem	99.7%
	5	-	Directly tells students answers and content, or dictates their next steps	95.1%
Scientific Reasoning	6	+	Encourages careful observation.	96.6%
	7	+	Encourages attention to evidence.	96.9%
	8	+	Encourages noticing of patterns or relationships.	86.0%
	9	+	Encourages students to question how/why processes/events/patterns are occurring.	80.5%
	10	+	Encourages formation of possible hypotheses.	97.6%
	11	+	Articulates the difference between correlation and causation or asks students to.	99.8%
	12	+	Making explicit that scientists engage in certain practices.	99.1%
	13	+	Encourages making connections between claims, evidence/data, and/or reasoning.	98.6%
	14	+	Assists students in attending to contradictory evidence	100.0%
General thinking supports	15	+	Asks open-ended questions.	81.7%
	16	+	Provides think time after a question is posed.	98.6%
	17	+	Encourages students to articulate their ideas.	87.7%
	18	+	Probes students' thinking	83.1%
	19	+	Stresses the importance of good note-taking for later reference (may include observations/drawings).	93.7%
	20	-	Discredits students' thinking/responses (instead of encouraging them to reassess their idea).	99.3%
Classroom environment	21	+	Circulates around the classroom.	84.6%
	22	+	Offers support when students have a question.	95.7%
	23	+	Shows signs of listening attentively to students.	89.6%
	24	+	Supports students in keeping the goals for the session in mind.	99.1%
	25	+	Offers assistance and/or organizational support with the physical materials or technological aspects of the curriculum.	94.5%

Interrater agreement statistics are presented in Table 2 to confirm that each item reached a sufficiently high agreement, providing evidence of a valid response process in the scoring procedure.[2] Typically, 80% rater agreement is considered sufficient in observational procedures

[2] Interrater agreement is often reported using a Kappa statistic which accounts for agreement by chance, however this calculation is sensitive to skewed distributions. As our measures were largely heavily skewed toward zero

(Hill et al., 2012). This is promising evidence for the validity of our instrument, as consistently identifying practices in student-teacher interactions within authentic classroom environments can be a challenge for observational measures and a potential barrier to measuring the fidelity to the process of an intervention.

Following the coding of all teachers' videos, the data was summarized into class-level indicators of student-teacher interactions. Raw scores were calculated as the class sum of scores on each item, divided by the total minutes in the lesson, to provide a measure of the frequency of each behavior per minute of teaching. Appendix B displays the raw score distributions, illustrating the many items with skewed distributions, as well as highlighting the differing scales each item is on; maximum values ranged from 0.025 to 0.6 occurrences per minute. The negative items (4, 5, and 20) were coded infrequently, and were dropped from further analysis. We log transformed the variables with severely skewed distributions, and standardized all the scores to z-scores, in which the mean for each is zero and the standard deviation is one, making them comparable in standard-deviation units. The transformed item distributions are in displayed in Appendix B.

Coherence: Assessing precision and reliability of observational data

A core goal of developing and validating the fidelity of implementation measures is to assess how precisely characteristics of student-teacher interactions related to inquiry-based teaching can be measured, both for the EcoXPT evaluation and for potential future use on other curricula. While many observational measures rely on interrater agreement as a measure of reliability, those statistics do not provide sufficient evidence that the scores are internally coherent nor do they estimate the level of precision, which has important implications for how scores can be interpreted and generalized (Hill et al., 2012; AERA, 2014). Additionally, we aimed to make the measurement procedure more efficient; understanding sources of error, properties of the items, and whether scores represent one unified construct or multiple dimensions, all provide valuable information for future study designs. We drew on concepts from classical test theory and its extension generalizability theory (Cronbach, 1951; Shavelson & Webb, 1991), as well as conducted exploratory factor analysis (Kline, 2016).

Classical Test Theory and Generalizability Theory

Classical test theory provides a simple and useful framework for understanding the relationships between items and internal consistency of scores. Thus, we begin our analysis by calculating Cronbach's alpha (Cronbach, 1951), which provides an overall estimate of the level of variation due to teachers' differences in their implementation compared to the amount of variation due to measurement error. A reliability coefficient of 1.0 indicates no measurement error: the teacher's "true" score is the only source of variation. Additionally, we used the item-level detail on each item's correlation with the overall score together with consideration of its

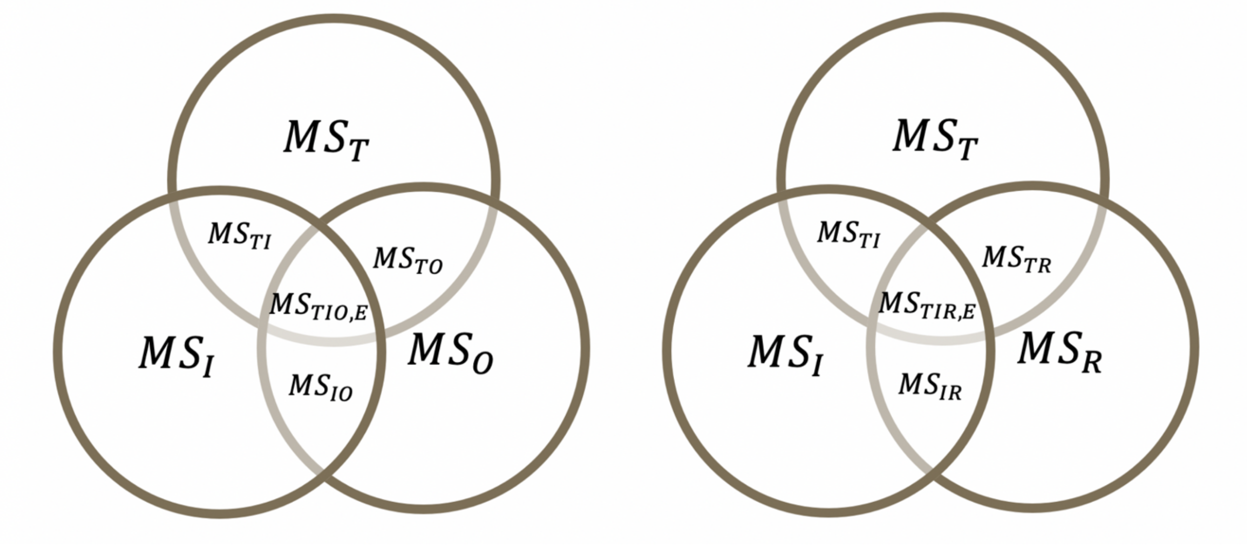
(most items were not observed in most minutes), Kappa does not prove to be a useful statistic. Instead we report the interrater agreement based on the percentage of minutes in which raters marked the same result, counting both present and absent as equally weighted responses. Due to this, items that were rarely observed have high agreement as is the case with item 14, which was only observed in two total minutes in the videos with two raters. We also prefer percent agreement over the Kappa statistic due to its interpretability.

conceptual importance in order to identify items that do not cohere well to the overall scores, as well as to eliminate items that do not contribute meaningfully to the scores.

Additionally, an important question when creating a measure based on observational data is how much variation in the measure is due to its sensitivity to the specific occasion that was observed or who rated it. Because we were interested in better understanding the precision of our fidelity of implementation measure not only in our study design, but also how to replicate the measure under other designs and to make the procedure more efficient, we utilized generalizability theory to decompose error due to occasions, raters, and items. Classical test theory accounts for only two facets of measurement, the “true” source of variation and measurement error due to items (Traub & Rowley, 1991), but assumes other facets are fixed. However, generalizability theory (G theory) extends classical test theory to decompose variation among different facets that contribute error in a measurement procedure (Shavelson & Webb, 1991; Brennan, 1992). Decomposing error across multiple facets thus allows for a better estimation of whether the measure can reliably capture the true score under different conditions, such as with different items, raters, or on different occasions. Conducting a generalizability study (G study) provides estimates of error due to the facets, providing a G coefficient similar to classical test theory’s Alpha, and then allows for a decision study (D study) estimate of precision under differing conditions, such as using more less items and raters.

While G studies rely on crossed facets, in which all raters would rate every teacher on every occasion with every item, the EcoXPT evaluation did not allow for such a fully crossed model. Thus, we estimated two G studies: the first in which teachers, items and occasions are crossed, rated only by the primary rater, and the second in which lessons rated by both raters are assessed for variance due to teachers, items, and raters. The G study designs are pictured in figure 1: on the left the design in which variation (means squared) due to teachers (MS_T) is separated from that due to items, occasions, and specific combinations of teachers and items or occasions ($MS_I, MS_O, MS_{TI}, MS_{TO}$), as well as specific teacher-item-occasion combinations, confounded with random error ($MS_{TIO,E}$). On the right, the same is pictured for the subset of lessons rated by both raters to estimate the teacher, item, and rater variance (Shavelson & Webb, 1991; Brennan, 1992; Brennan 2011).

Figure 1: Representation of Generalizability Studies for a Teachers X Items X Occasions Design (left) and a Teachers X Items X Raters Design (Right)



In the teachers-items-occasions design, the value of any observable score at an elemental level of a teacher *t* on item *i* on occasion *o* (X_{tio}) can be represented as:

$$X_{tio} = \mu + v_t + v_i + v_o + v_{ti} + v_{to} + v_{io} + v_{tio,e} \qquad (1)$$

where μ represents the overall mean, and v_n represents that score's deviation from the mean due to facet n or an interaction of multiple facets. The variance of observed scores is:

$$\sigma^2(X_{tio}) = \sigma^2(t) + \sigma^2(i) + \sigma^2(o) + \sigma^2(ti) + \sigma^2(to) + \sigma^2(io) + \sigma^2(tio) \qquad (2)$$

Using these estimates, we can calculate the relative error (σ_δ^2) due to items and occasions, averaging over how many items and occasions are included in the measure (n_i', n_o'):

$$\sigma_\delta^2 = \frac{\sigma_{ti}^2}{n_i'} + \frac{\sigma_{to}^2}{n_o'} + \frac{\sigma_{tio,e}^2}{n_i' n_o'} \qquad (3)$$

Then using the relative error, we can estimate the G coefficient ($E\rho^2$), which represents the amount of variation due to differences in teachers (σ_t^2), out of the total variation due to teacher scores and relative error (σ_δ^2):

$$\mathbf{E}\rho^2 = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_\delta^2} \qquad (4)$$

This coefficient can be interpreted similarly to the Alpha coefficient, and a value of one implies no variation due to items or occasions but that all the score variation is attributable to differences in teachers' "true" scores. We then repeat the calculations in equations 1 through 4 to estimate the teachers-items-raters design as well. Using these variance estimates, we can conduct decision studies (D studies) to estimate the precision of the measure under different designs with more or fewer items, occasions and raters, reporting the G coefficient for different designs to assess the relative precision and efficiency of the measure.

Factor Analysis

The fidelity measures were based on sound theory drawn from the program's pilot and prior literature on important teaching practices for inquiry-based learning. However, in creating a new measure it was not possible to know the scores' underlying structure. How the different items do or do not vary together, and which items capture the greatest amount of variation in actual teaching practice, is an empirical question which can only be answered after data collection. While the items were grouped into categories or "umbrellas" (see table 2), these groupings were not intended to specify different domains but rather to provide a structure to ease the rating procedure.

We utilized exploratory factor analysis (EFA) to assess whether there are multiple dimensions underlying the quality of implementation measure, as well as to assess which items provide the most information regarding teachers' overall scores. Our goals align well with the purpose of EFA, "to arrive at a more parsimonious conceptual understanding of a set of measured variables by determining the number and nature of common factors needed to account for the pattern of correlations among the measured variables" (Fabrigar et al., 1999, p. 275). In other words, knowing that our measure may not be unidimensional, we aim to identify any latent constructs which may underlie the variables we have measured, using both the correlations found in the data as well as our own interpretation of how the variables relate to one another conceptually.

Exploratory factor analysis does not require an *a priori* hypothesis regarding the structure of the measure, making it particularly appropriate in this case where we are creating and validating a new measure. EFA finds common factors that linearly reconstruct the original variables:

$$y_{ij} = z_{i1}b_{1j} + z_{i2}b_{2j} + \cdots + z_{ik}b_{kj} + e_{ij} \tag{5}$$

where y_{ij} is the value of the ith observation on the jth variable, z_{ik} the observed value on the kth common factor, and b_{kj} are the coefficients that represent factor loadings for which the factors reconstruct the original covariance of the variables. A principal factors method was used for factor extraction, and based on our hypothesis that all the variables may be correlated with one another, an oblique rotation was performed to improve interpretability and test whether retained factors are correlated. We did not set a number of factors to be retained at the outset, but rather interpreted the factors based on their loadings and the level of variance explained within the data, aiming to retain factor loadings above 0.5. Interpreting the factor loadings is an inherently subjective procedure that also relies on a conceptual understanding of the variables and how they relate to one another, and so we used our content knowledge about the measures to interpret factor loadings and level of variance explained.

Results:

First, we estimated the scores' internal coherence as the alpha coefficient, decomposing scores' variation due to measurement error from differences in teachers' implementation. Here, we are interested in precision, identifying individual items that are not well correlated with the others, and assessing the substantive contribution of those indicators to determine if items could be excluded from the scores and increase confidence in the scores' reliability.

Table 3: Cronbach's Alpha for the 22-item Sore

Alpha coefficient		***0.70***		
Item Detail	Average item-test correlation	Item-rest correlation	Interitem correlation	Excluded-item alpha
code1	0.69	0.62	0.08	0.65
code2	0.35	0.23	0.10	0.69
code3	0.51	0.41	0.09	0.67
code6	0.33	0.21	0.10	0.69
code7	0.49	0.39	0.09	0.68
code8	0.49	0.39	0.09	0.68
code9	0.52	0.42	0.09	0.67
code10	0.03	-0.09	0.11	0.72
code11	0.25	0.13	0.10	0.70
code12	0.27	0.15	0.10	0.70
code13	0.39	0.27	0.09	0.69
code14	0.32	0.20	0.10	0.69
code15	0.48	0.38	0.09	0.68
code16	0.19	0.07	0.10	0.70
code17	0.59	0.50	0.09	0.66
code18	0.66	0.58	0.08	0.66
code19	0.52	0.42	0.09	0.67
code21	0.08	-0.05	0.11	0.71
code22	0.18	0.06	0.10	0.70
code23	0.36	0.25	0.09	0.69
code24	0.43	0.33	0.09	0.68
code25	-0.02	-0.14	0.11	0.72

Table 3 shows the results of the item-level alpha estimates using 22 items.[3] Items with lower than a 0.20 correlation are item number 10 (encourages formation of hypotheses), 16 (provides think time after a question), 21 (circulates around the classroom), 22 (supports students when they have a question), and 25 (assistance with technology). Substantively, as well, we could justify the lack of information provided by these items, as in our observations of teachers we found that regardless of other practices, teachers circulated around the classroom, answered student questions, and helped troubleshoot technological problems. Providing think time after a question rarely occurred and was mostly applicable in whole-group instruction. Explicitly encouraging formation of hypotheses was also not frequently observed, particularly compared to encouraging students more generally to articulate their explanations. Based on the poor

[3] Three negatively coded items were removed from analysis, see above section for a description of the summarized item scores.

item-rest correlation as well as these conceptual issues related to quality of student-teacher interactions, we removed these items from subsequent analysis.

In doing so, we use a more parsimonious and internally consistent 17-item measure. Excluding these items from the scores increases the alpha coefficient to .78, indicating that even though we are utilizing fewer measures, those items are more correlated with one another, and capture a larger proportion of teachers' differences in implementation, and less variation due to measurement error. Table 4 displays the item-level detail and indicates there may still be problematic items that are poorly correlated with the others, such as number 11. However, this code refers to discussing the differences between correlation and causation, a central aim of the program, and so we retain it in the scores.

Table 4: Cronbach's Alpha for the 17-item Score

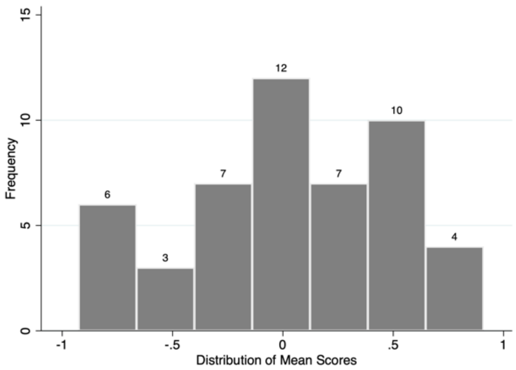
Alpha coefficient		***0.78***		
Item Detail	Average item-test correlation	Item-rest correlation	Interitem correlation	Excluded-item alpha
code1	0.66	0.58	0.16	0.76
code2	0.36	0.24	0.18	0.78
code3	0.56	0.47	0.17	0.77
code6	0.35	0.23	0.18	0.78
code7	0.51	0.41	0.17	0.77
code8	0.52	0.42	0.17	0.77
code9	0.56	0.46	0.17	0.77
code11	0.20	0.08	0.19	0.79
code12	0.37	0.25	0.18	0.78
code13	0.43	0.32	0.18	0.78
code14	0.29	0.18	0.19	0.79
code15	0.58	0.49	0.17	0.76
code17	0.64	0.55	0.16	0.76
code18	0.76	0.70	0.16	0.75
code19	0.41	0.29	0.18	0.78
code23	0.49	0.39	0.17	0.77
code24	0.36	0.24	0.18	0.78

Table 5 displays descriptive statistics for this 17-item mean score. As the variables are standardized to have a mean of zero, it is not surprising that the mean score is approximately zero, and as figure 2 shows, they are relatively normally distributed. While the treatment condition in which the program included experimental tools has a slightly lower overall average score, the difference between the mean scores in the tools and no tools conditions is not statistically significant ($t(38) = .74, p = .45$).

Table 5: Summary Statistics for the 17-item Mean Score

	Total	*Tools*	*No Tools*
Mean	0.00	-0.06	0.06
SD	0.48	0.49	0.47
Variance	0.23	0.24	0.22
Skewness	-0.27	-0.55	0.10
Kurtosis	2.51	2.20	2.47
Percentiles			
10%	-0.82	-0.89	-0.50
25%	-0.25	-0.31	-0.25
50%	0.02	0.05	0.01
75%	0.35	0.28	0.39
90%	0.59	0.52	0.73
N	40	20	20

Figure 2: Histograms of the Frequency of Classes' Mean Scores



G Theory:

Using generalizability theory, we can further estimate the precision of these scores by decomposing error in our scores due to other facets of the measurement procedure, namely the occasions and raters, as well as the items. We estimate two G Studies: teachers-items-occasions (*tXiXo*) with all videos coded by the primary rater, and teachers-items-raters (*tXiXr*) with the videos coded by two raters. Table 6 shows the estimates for each facet's variance at an elemental level: the variance for a score of one teacher on a specific item and specific occasion. In both designs, where teachers are observed on multiple occasions by one rater or on one occasion by two raters, the largest sources of variation come from the teacher-item combination and teacher-item-occasion or teacher-item-rater combined with random error. Error due to specific teacher-occasion variation is also fairly large in the *tXiXo* design.

Table 6: G Study Variance Components Estimates

Teachers X Items X Occasions (tXiXo)						***Teachers X Items X Raters (tXiXr)***					
G Coefficient: .51						***G Coefficient: .69***					
Source	df	MS	$\hat{\sigma}_v^2$	$\hat{\sigma}_v$	*Percent Variance*	*Source*	df	MS	$\hat{\sigma}_v^2$	$\hat{\sigma}_v$	*Percent Variance*
t	9	8.82	0.07	0.26	7%	t	5	5.83	0.12	0.35	9%
o	4	2.69	0.00	0.04	0%	r	16	1.97	0.00	0.00	0%
i	16	0.44	0.00	0.00	0%	i	1	0.05	0.00	0.00	0%
to	26	2.15	0.09	0.31	10%	tr	80	1.86	0.00	0.00	0%
ti	144	1.54	0.20	0.45	22%	ti	5	0.49	0.55	0.74	39%
io	64	1.15	0.06	0.25	7%	ir	16	1.01	0.04	0.21	3%
$tio.e$	416	0.53	0.53	0.73	58%	$tir.e$	182	0.75	0.75	0.87	52%

Note: Three estimates were calculated as negative variance components, an artifact of the ANOVA procedure that can occur due to small sample sizes. They were replaced with zero variances, but total variance calculations retained the negative values to provide unbiased proportion of variance estimates (Shavelson & Webb 1991; Brennan, 2001; Cronbach, 1972). In the tXiXo study, item variance was estimated at -0.03, and in the tXiXr study, item variance was estimated at -0.01, rater variance at -0.01, and teacher-rater variance at -0.02.

However, these estimates refer to the elemental scores, meaning the amount of error observed for each individual combination of items, occasions, and/or raters with each teacher. The reliability of the score overall is of more interest than these individual observations. Thus, the G coefficient is more informative, estimating the combined sources of relative measurement error compared to the amount of variation due to teachers' differences in implementation. In the one-rater, multiple-occasion design, teacher variation accounts for 51% of total score variance, while error due to the facets accounts for 49%. The design in which multiple raters rate the same occasion has less measurement error, and 69% of variation is due to teachers, and only 31% due to error.

Using these estimates, Figures 3 and 4 visualize the D studies, depicting how much the precision would change under different design conditions. In both designs, there is little gained by adding additional items and, conversely, little precision is lost by reducing the number of items. In the *tXiXo* design, 12 items measured on 4 occasions would have a .49 G coefficient, compared to the .51 coefficient for 17 items. On the other hand, a relatively larger amount of precision is lost if the number of occasions is reduced, as only observing 2 occasions with our 17 items yields a .39 coefficient. In this design, however, precision is low overall; even observing teachers on six occasions would result in measures that capture 40% of the variation from error.

The design with multiple raters, however, shows much more promising results. In our design of 17 items and 2 raters, the G coefficient is already high. Reducing the number of raters to only one, however, decreases the precision significantly, while adding items and raters could marginally increase precision. 5 raters with 20 items, for example, yields a G coefficient of .79, while adding one additional rater with the 17-item scale increases the coefficient to .73.

Figure 3: D Study: Teachers, Items, and Occasions Designs

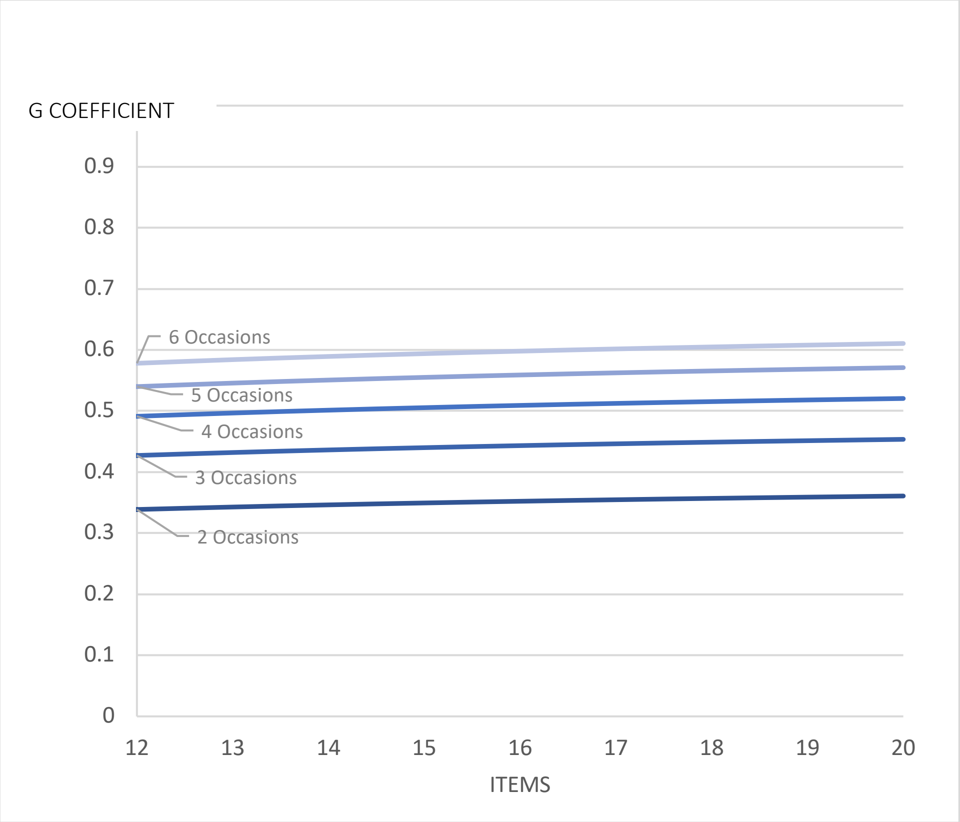
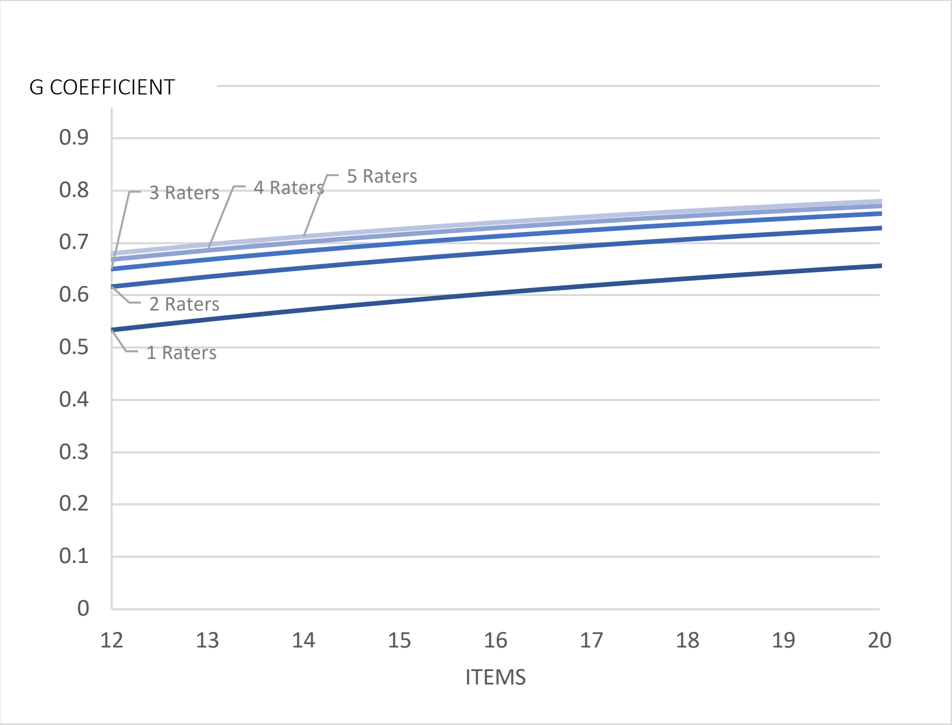


Figure 4: D Study: Teachers, Items, and Raters Design



In summary, these results help answer the question not only of how precise our measure is as it stands, but also what future designs might be the most efficient. These results suggest the most efficient option would be to observe teachers for fewer occasions, but with multiple raters. Reducing or adding items has a smaller impact on reliability, in which case it may be best to reduce the number to ease the burden on raters.

Exploratory Factor Analysis:

To understand more about the ways in which teachers varied in their implementation, we utilized exploratory factor analysis (EFA) to describe the underlying dimensional structure of the measure, and how the individual items do or do not covary. The results thus provide information about the latent constructs being measured, and whether distinct sub-domains should be measured by specific items.

Figure 5 depicts the scree plot of eigenvalues and factors, and table 7 reports the results for four factors. The scree plot indicates that indeed there is multidimensionality in our measure, but the plot and eigenvalues do not supply a clear-cut solution to how many factors should be retained. There are four factors with Eigenvalues above 1, but there are two factors present before the scree plot depicts an "elbow." We consider 0.5 to be a sufficiently high loading for an item to be considered part of a factor, and in table 7 have displayed all loadings above 0.4, suggesting there are in fact only two factors. For one, factors three and four only have one and zero items that load onto them at the 0.5 cutoff, respectively. Further, factors one and two each capture a much larger amount of variation than three and four. Together, factors one and two account for 60% of all the variation in the data. Item number nine loads onto both factors one and two, but is both more statistically and substantively aligned with the second factor. An oblique factor rotation helped confirm the structure of the two factors, as well as determined they are weakly correlated at 0.21.

Figure 5: Exploratory Factor Analysis: Scree plot of Eigenvalues

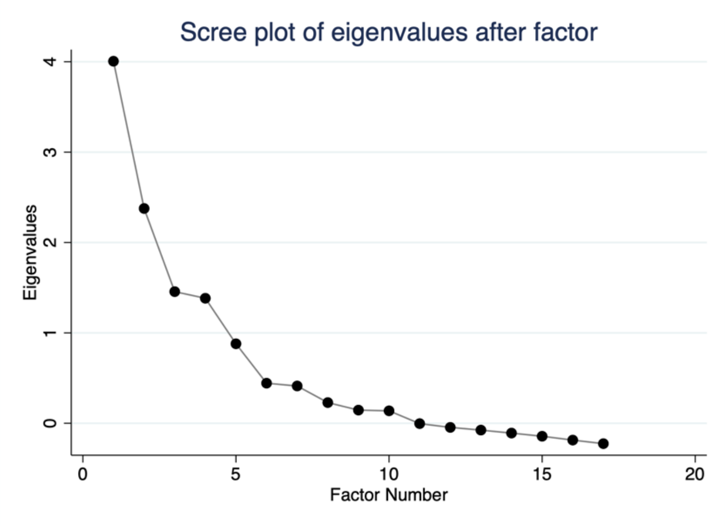


Table 7: Exploratory Factor Analysis Results (4 Factors)

Variable	*Factor1*	*Factor2*	*Factor3*	*Factor4*	*Uniqueness*
code1	0.60				0.42
code2				0.42	0.60
code3	0.56				0.49
code6					0.90
code7	0.45		0.57		0.38
code8	0.48	0.70			0.22
code9	0.54	0.65			0.21
code11					0.73
code12					0.63
code13	0.41		0.49		0.43
code14					0.68
code15	0.68				0.20
code17	0.66				0.30
code18	0.87				0.15
code19		0.61			0.38
code23	0.60				0.29
code24					0.78
Explained variance	38%	22%	14%	13%	
Eigenvalue	4.01	2.38	1.46	1.38	

Note: Only factor loadings above 0.40 are reported here; our threshold for retaining factors was 0.5.

Table 8 summarizes these two latent constructs more substantively, and includes the alpha coefficients, indicating each latent construct has a high level of coherence despite the relatively few items included. The first factor includes items that we had originally grouped among different “umbrellas” of interaction characteristics, yet they all pertain to asking questions and listening to students. A teacher who is exhibiting high scores on these items would be frequently asking students what they are thinking about, encouraging them to discuss their ideas and explanations, as well as listening to them and helping their exploration without trying to directly guide them. These interactions would be characterized by more talk from the student than the teacher.

The second factor only includes three items, two of which are clearly linked and logically happen together: encouraging students to notice patterns and to question why they are occurring. The third item relates to these in that teachers who were encouraging students to notice and explain patterns should also be likely to encourage students to document these relationships in their notebooks, in order to keep track of their thinking and build their explanations with evidence.

Table 8: Latent Factors

Latent Factor	***Alpha***	***Items***	***Item loadings***
Inquiry Dialogue	0.82		
		Responds to conceptual uncertainty in an open-ended manner	0.60
		Encourages students to give their best explanation or multiple explanations based upon what the evidence suggests.	0.56
		Asks open-ended questions.	0.68
		Encourages students to articulate their ideas.	0.66
		Probes students' thinking	0.87
		Shows signs of listening attentively to students.	0.60
Pattern seeking and documenting	0.79		
		Encourages noticing of patterns or relationships	0.70
		Encourages students to question how/why processes/events/patterns are occurring	0.65
		Stresses the importance of good note-taking for later reference	0.61

These latent variables can explain much of the variation in our data, indicating that teachers were most likely to differ in their interactions related to these items than on the other items broadly. Table 9 reports teachers' mean scores on each of these factors, along with their overall mean score on the 17-item scale.

Table 9: Mean Scores by Teacher: Latent Factors and Total Scores

Teacher	*Questioning and Listening*	*Pattern Seeking*	*Overall Mean Score*
A	1.31	-0.18	0.34
B	-0.08	-0.30	-0.08
C	-0.01	-0.18	0.21
D	0.61	-0.17	0.39
E	0.13	0.64	0.40
F	0.06	0.74	0.07
G	-0.06	0.04	-0.07
H	-0.73	-0.02	-0.35
I	-0.62	-0.91	-0.75
J	-0.61	0.34	-0.03

Limitations

Our conclusions for this study are limited by a number of factors. For one, our sample is admittedly small, with just 10 teachers and 40 classes. With low power, it is difficult to find statistically significant associations or differences between sub groups, such as the 20 classes with tools and 20 without, or between individual teachers.

Furthermore, while we have taken a holistic view of validating our measures of fidelity, they are limited by the data collected. For one, the observations took place in the context of a controlled study, in which researchers were present in the classes and broadly supporting the setup and implementation of the curricula. Teachers were not randomly selected, but rather they opted in to the study and were compensated for their participation. Teachers were also fully aware of being videotaped during the lesson in which we assessed fidelity, which may have changed their behavior. These characteristics of the study may raise concerns regarding the generalizability of our findings, and whether the data represents what teachers' practices would have been with less involvement from the research team. The measures would benefit from being trialed and tested with teachers implementing inquiry-based curricula in a less controlled environment, independent of a research study. Trialing the measures in a large and diverse sample would provide more evidence for their generalizability and whether greater variation is observed among different populations.

Discussion and implications

This paper describes a thorough approach to measuring the fidelity both to the structure and to the process of implementation, drawing on principles from psychometrics to provide evidence for the measures' validity. We described the process through which fidelity was defined as adhering both to structural components of a curriculum, such as fully utilizing materials and having functioning technology, as well as to the process that defines the quality of its delivery through teachers' interactions with students. Focusing on our measure of quality, we discussed how we defined the construct as a core set of practices teachers employ in their interactions with students, developed a sound observational scoring procedure, and assessed the properties of the scores in terms of their coherence, precision, and multidimensionality.

Our findings have implications for future efforts to measure fidelity of implementation, as well as understanding teaching practice within inquiry-based curricula more broadly. In terms of measuring fidelity of implementation, we provide a model for developing and validating measures that emphasizes their content and procedures. Looking to teachers' interactions with their students provides a rich description of meaningful differences in how teachers implemented the curriculum. By devoting time to the response processes used by raters, this procedure can identify distinct differences in how teachers implement a curriculum, and thus can support inferences about teaching and implementation within the context of an evaluation. We also find that, despite capturing teaching practices in authentic classroom environments, these measures can be reasonably coherent.

From our experience, researchers adapting our measures or developing their own should consider the balance between capturing the complexity of learning environments with the need for efficient measures. While our approach aimed to capture teachers' practices at a granular level of detail, with a large number of items at a minute-by-minute level, a fewer number of items that target specific practices, such as inquiry dialogue, may be more useful to capture variation efficiently. We also find that where resources are scarce, prioritizing more raters over more occasions can also maximize precision efficiently. For those who are adapting our measures, particular attention should be paid to the content of the measures, whether the items are relevant to the program's theory of change, and should assess the evidence for validity in different samples. There may be items that better represent the variation in practices in different settings or under different curricula. One particularly important dimension is likely whether the teachers are participating in a controlled study or a larger-scale implementation of a curriculum.

Regardless of these questions around generalizability, implementation measures such as ours show promise for better capturing the ways teachers interact with students, an important consideration in determining the effectiveness of a curriculum or educational technology program. Our findings are also promising for better measuring how teachers guide students in inquiry-based programs and the types of guidance teachers provide students. We found that the greatest area of variation in teacher practice was characterized by their level of questioning and listening. These are core practices for inquiry-based learning, in which students benefit from guidance that pushes them to construct explanations and elaborate on their thinking, as well as characterize interactions in which students should be talking more than teachers. The fact that teachers did not vary as much across behaviors related to classroom structure or fostering scientific reasoning indicates inquiry-based curricula should focus on supporting teachers to engage in interactions around inquiry dialogue.

Our future work will extend these findings by validating the measures based on their correlation to student learning outcomes. Future research questions include how quality of implementation scores predict student learning outcomes, and whether the sub-domains are as predictive of student learning as the full scores. We also plan to describe how the different implementation measures are related to one another to understand whether teachers with high quality of delivery also exhibited stricter adherence to structural components, and which of these dimensions of fidelity are most important for student learning.

References:

AERA. (2014). *Standards for Educational & Psychological Testing (2014 Edition)* (2014 Edition). Washington, D.C.: American Educational Research Association. Retrieved from https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2015). Epistemologies in practice: Making scientific practices meaningful for students. Journal of Research in Science Teaching, 53(7), 1082-1112.

Brennan, R. L. (1992). Generalizability Theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brennan, R. L. (2010). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Chiesa, M., & Hobbs, S. (2008). Making sense of social research: how useful is the Hawthorne Effect? *European Journal of Social Psychology*, *38*(1), 67–74. https://doi.org/10.1002/ejsp.401

Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, *119*(2), 166.e7-166.e16. https://doi.org/10.1016/j.amjmed.2005.10.036

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dane, A. V., & Schneider, B. H. (1998). PROGRAM INTEGRITY IN PRIMARY AND EARLY SECONDARY PREVENTION: ARE IMPLEMENTATION EFFECTS OUT OF CONTROL? *Clinical Psychology Review*, *18*(1), 23–45. https://doi.org/10.1016/S0272-7358(97)00043-3

Dede, C., Grotzer, T. A., Kamarainen, A., & Metcalf, S. (2017). EcoXPT: Designing for Deeper Learning through Experimentation in an Immersive Virtual Ecosystem. *Journal of Educational Technology & Society*, *20*(4), 166–178.

Desimone, L. M., & Hill, K. L. (2017). Inside the Black Box: Examining Mediators and Moderators of a Middle School Science Intervention. *Educational Evaluation and Policy Analysis*, *39*(3), 511–536. https://doi.org/10.3102/0162373717697842

Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning*, *3*(4), 269–276. https://doi.org/10.1016/0149-7189(80)90042-7

Durlak, J. A., & DuPre, E. P. (2008). Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Dusenbury, L. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Fabrigar, L. R., Wegener, D., MacCallum, R., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.

Fishman, B., & Dede, C. (2016). Teaching and Technology: New Tools for New Times. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching* (Fifth, pp. 1269–1334).

American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6_21

Grotzer, T. A., Kamarainen, A. M., Tutwiler, M. S., Metcalf, S., & Dede, C. (2013). Learning to Reason about Ecosystems Dynamics over Time: The Challenges of an Event-Based Causal Focus. *BioScience*, *63*(4), 288–296. https://doi.org/10.1525/bio.2013.63.4.9

Gwet, K., & Road, B. S. G. (n.d.). Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters, 5.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Kim, J. S., Burkhauser, M. A., Quinn, D. M., Guryan, J., Kingston, H. C., & Aleman, K. (2017). Effectiveness of Structured Teacher Adaptations to an Evidence-Based Summer Literacy Program. *Reading Research Quarterly*, *52*(4), 443–467. https://doi.org/10.1002/rrq.178

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling, Fourth Edition* (Fourth edition). New York: The Guilford Press.

Mendive, S., Weiland, C., Yoshikawa, H., & Snow, C. (2016). Opening the black box: Intervention fidelity in a randomized trial of a preschool teacher professional development program. *Journal of Educational Psychology*, *108*(1), 130–145. https://doi.org/10.1037/edu0000047

Messick, S. (1987). Validity. *ETS Research Report Series*, *1987*(2), i–208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Minner, D., & DeLisi, J. (2012). Inquiring into science instruction observation protocol (ISIOP): Data collection instrument. Education Development Center: Waltham, MA.

Mowbray, C. T., Holter, M. C., & Bybee, D. (2003). Fidelity Criteria: Development, Measurement, and Validation. *American Journal of Evaluation*, *24*(3), 315–340.

Odom, S. L., Fleming, K., Diamond, K., Lieber, J., Hanson, M., Butera, G., … Marquis, J. (2010). Examining different forms of implementation and in early childhood curriculum research. *Early Childhood Research Quarterly*, *25*(3), 314–328. https://doi.org/10.1016/j.ecresq.2010.03.001

O’Donnell, C. L. (2008). Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K–12 Curriculum Intervention Research. *Review of Educational Research*, *78*(1), 33–84. https://doi.org/10.3102/0034654307313793

Pianta, R. C., Hamre, B. K., & Paro, K. M. L. (2008). *Classroom Assessment Scoring System*. Paul H. Brookes Publishing Company. Retrieved from https://books.google.com/books?id=gkLwGQAACAAJ

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. SAGE.

Snyder, P. A., Fox, L., Hemmeter, M. L., Crowe Bishop, C., & Miller, M. D. (2013). *Developing and Gathering Psychometric Evidence for a Fidelity Instrument* [Data set]. *Journal of Early Intervention*. https://doi.org/10.1037/t33369-000

Southam-Gerow, M. A., Bonifay, W., McLeod, B. D., Cox, J. R., Violante, S., Kendall, P. C., & Weisz, J. R. (2018). Generalizability and Decision Studies of a Treatment Adherence Instrument. *Assessment*, 107319111876536. https://doi.org/10.1177/1073191118765365

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention Fidelity in Special and General Education Research Journals. *The Journal of Special Education*, *47*(1), 3–13. https://doi.org/10.1177/0022466911419516

Thompson, M., Tutwiler, M. S., Kamarainen, A., Metcalf, S., Grotzer, T., & Dede, C. (2016). A Blended assessment strategy for EcoXPT: An Experimentation-driven ecosystems science-based multi-user virtual environment. Washington D.C.: American Educational Research Association.

Traub, R. E., & Rowley, G. L. (1991). Understanding Reliability. *Educational Measurement: Issues and Practice*, *10*(1), 37–45. https://doi.org/10.1111/j.1745-3992.1991.tb00183.x

Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen’s kappa. *Journal of Consulting and Clinical Psychology*, *82*(6), 1219–1227. https://doi.org/10.1037/a0037489

Zhai, J., & Tan, A. L. (2015). Roles of teachers in orchestrating learning in elementary science classrooms. Research in science education, 45(6), 907-926.

Appendix A: Other fidelity measures

Adherence measures

Adherence was defined as how well the program was implemented related to structural components, such as whether teachers discussed specific features of the curriculum with students, utilized the PowerPoint slides provided, if the technology was working properly, and whether the class suffered interruptions throughout the 13 days of teaching EcoXPT.

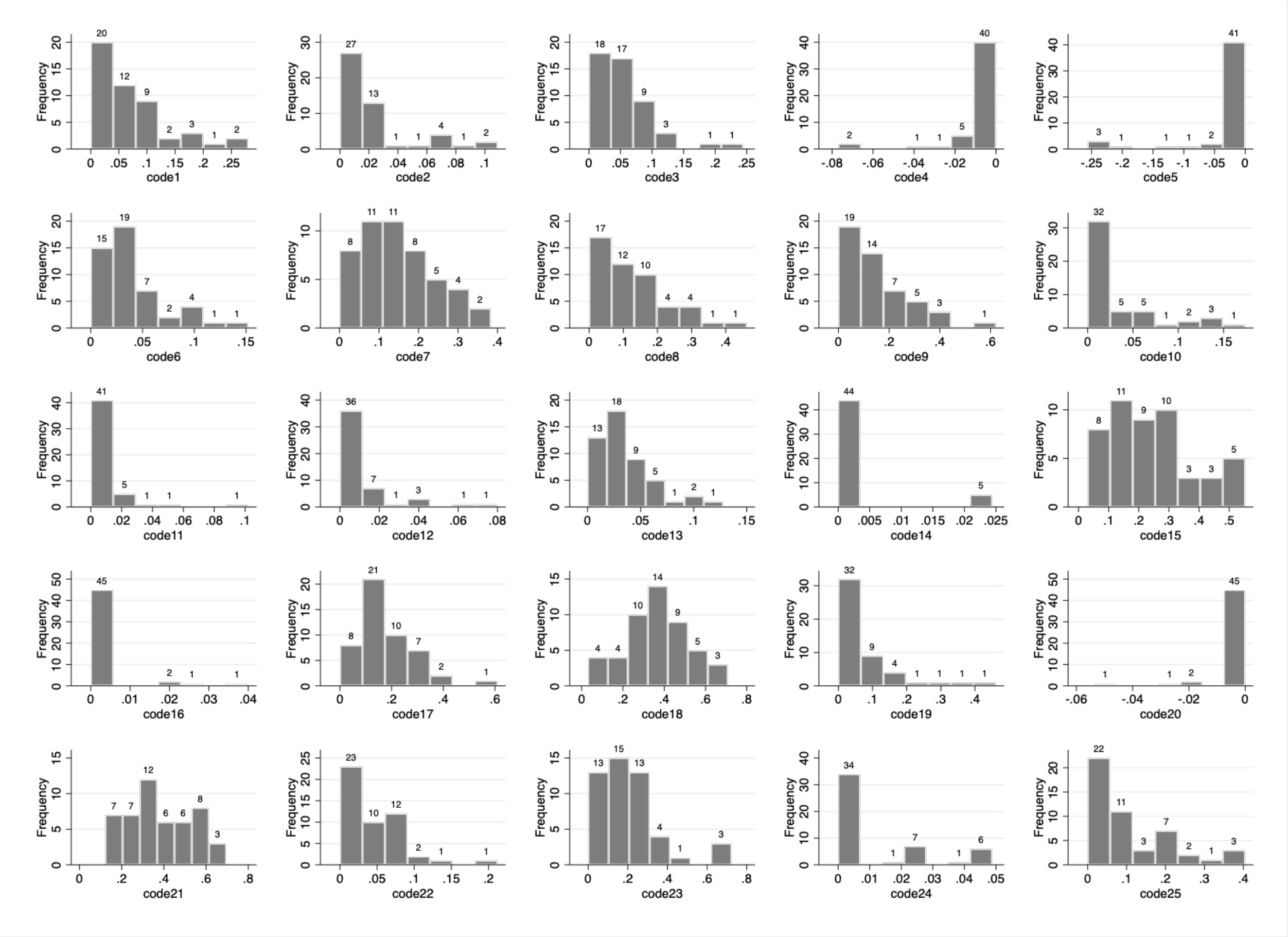
Student-teacher interactions		**Item**	**Rater Percent Agreement**
		References/draws student attention to concept map.	*94.1%*
		Draws student attention to data (tables and graphs).	*91.2%*
		Discusses or encourages use of relevant tools (those that are in both conditions and those specific to Experimental Tools Condition)	*88.0%*
		Discusses findings/results from tools.	*83.2%*
		Encourages use of Field Guide.	*98.4%*
		Reminds students of information learned in Powerpoint, from Jade, on the Posters or as Thinking Moves.	*99.2%*
		Discusses interventions/the kinds of experimentation that ecosystem scientists engage in.	*98.6%*
Implementation Checklist		*Completes the Do Now*	
		Teacher covers material in the PowerPoint for the day: • *First few slides, half of slides, whole presentation* • *Read it* • *Read it and explained some parts* • *Read it, explained parts and checked for understanding*	
		Teacher outlines tasks and expectations for the day	
		Hands out the thinking move posters: Deep Seeing, Evidence Seeking, Pattern Seeking	
		Every pair of students has a working computer	
		EcoXPT is working on every computer	
		Wrap-up discussion	
Interruptions Score (0-3)	3	*Major interruptions due to lack of access to internet and computers in which students completed offline XPT lessons, learned other content, and split implementation of XPT by 4 weeks*	
	2	*Interruptions due to lack of access to computers, teacher absences in which substitute teachers taught XPT, days of other content and lessons unrelated to XPT taught throughout the curriculum, some major delays including two-week winter break occurring in the middle of implementation.*	
	1	*Minor interruptions including snow days, field trips, and school events during which EcoXPT was not taught, no major technological issues*	
	0	*EcoXPT implemented with no interruptions apart from weekends, each day of the curriculum taught sequentially, no technological delays, no other lessons taught, substitutes or snow days*	

Dosage measure:

Dosage was defined as the amount of time students spent working in the program across the 11 days of the curriculum in which exploring the program is part of the lesson. Classes varied in the amount of time due to class length, teacher decisions regarding warm-up activities and teacher-directed instruction, technological issues, and other factors within school days. Class-averages were calculated using students' log file data, and ranged from 4 hours, 10 minutes to 6 hours, 45 minutes.

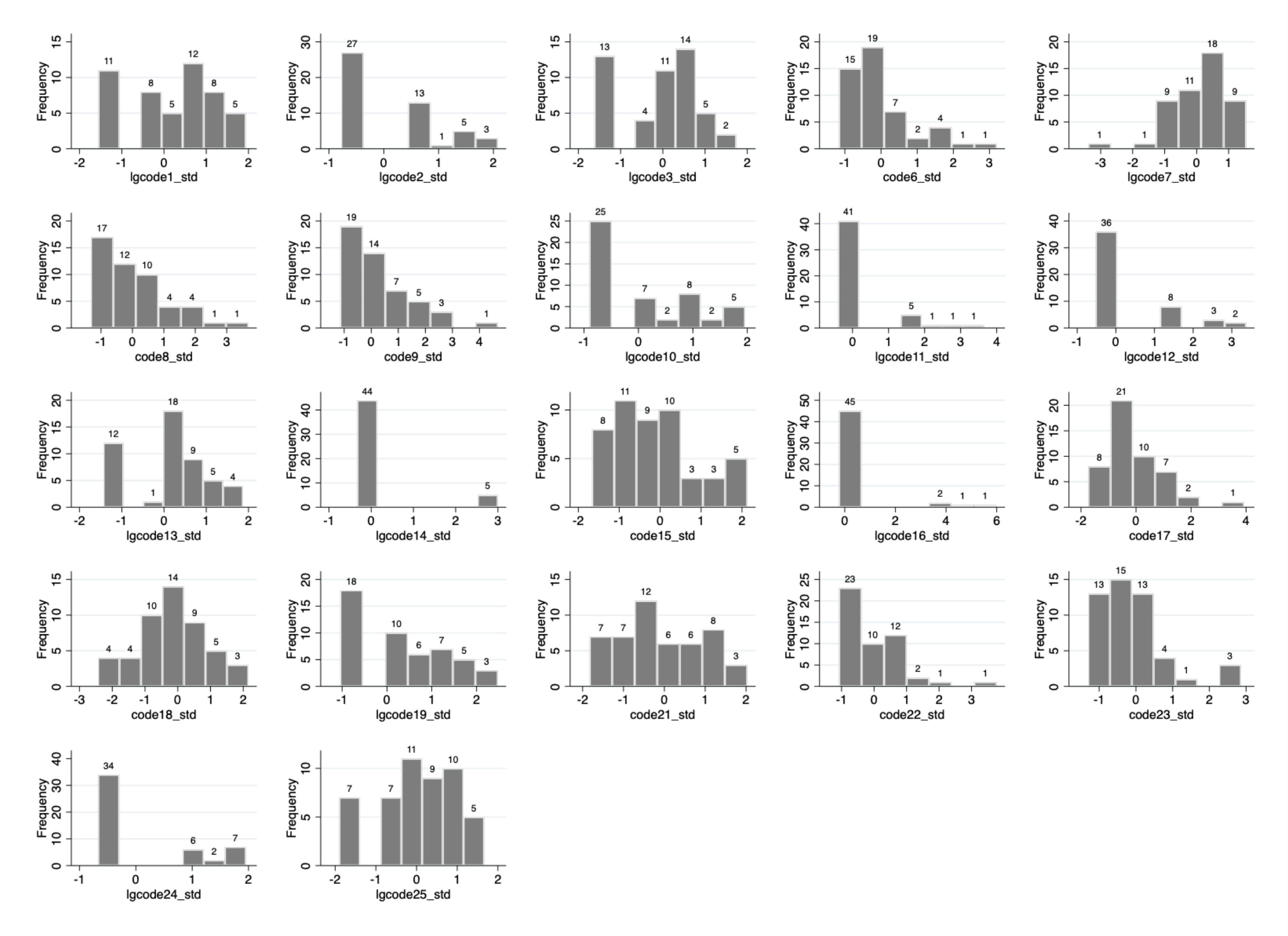
Appendix B: Item histograms

Raw data



Transformed and standardized:

Histograms for standardized, z-scores of items. Reverse-coded items have been removed (4, 5, and 20), and skewed variables have been log-transformed (1, 2, 7, 10, 11, 12, 13, 14, 16, 19, 24, and 25). All codes have been standardized to have a mean of zero and standard deviation of one.


Frequency
11 8 5 12 8 5
lgcode1_std
27 13 1 5 3
lgcode2_std
13 4 11 14 5 2
lgcode3_std
15 19 7 2 4 1 1
code6_std
1 1 9 11 18 9
lgcode7_std
17 12 10 4 4 1 1
code8_std
19 14 7 5 3 1
code9_std
25 7 2 8 2 5
lgcode10_std
41 5 1 1 1
lgcode11_std
36 8 3 2
lgcode12_std
12 1 18 9 5 4
lgcode13_std
44 5
lgcode14_std
8 11 9 10 3 3 5
code15_std
45 2 1 1
lgcode16_std
8 21 10 7 2 1
code17_std
4 4 10 14 9 5 3
code18_std
18 10 6 7 5 3
lgcode19_std
7 7 12 6 6 8 3
code21_std
23 10 12 2 1 1
code22_std
13 15 13 4 1 3
code23_std
34 6 2 7
lgcode24_std
7 7 11 9 10 5
lgcode25_std