Bringing Assessment *for* Learning Outdoors: developing, using, and reflecting on common measures that drive program improvement and the professional growth of Outdoor Schools' educators' practices.

Evaluation Report.

Prepared for
Outdoor School Program
Multnomah Education Service District
11611 NE Ainsworth Circle
Portland, Or 97220

Prepared by
Emily Saxton
Portland Metro STEM Partnership
18624 NW Walker Road
Beaverton, Or. 97006
esaxton@pdx.edu

July 2015

This project was made possible by the Gray Family Foundation and the Christine and David L. Vernier Fund of the Oregon Community Foundation.

Table of Contents

Executive Summary

Introduction

The Bringing Assessment *for* Learning Outdoors project was designed to strengthen the Outdoor School program's impact on student learning by 1) creating high-quality assessments for the key student outcomes of academic identity, motivational resilience, and application of conceptual knowledge and 2) using assessment data to drive program improvement by providing Outdoor School instructors with reflective professional development that is focused on understanding and strengthening student learning about science concepts that are central to the Outdoor School program. Assessments can be used to enhance student learning, refine educator practices, and drive program improvement, but in order to realize these benefits of assessment *for* learning educators need professional development. This project's joint focus on assessment development and professional development was, therefore, designed to be a first step in bringing the benefits of assessment *for* learning to the Multnomah Education Service District's (MESD) Outdoor School program.

Project Successes.

- In the 2014-15 academic year, the MESD Outdoor School Program served 7,073 sixth grade students.
- Of the Outdoor School instructors who were employed for the full project duration, 88% (*N* = 7) attended 6 or more days of the professional development.
- The project found a positive and statistically significant impact on students' academic identity in science[1] in the fall 2014 session of the Outdoor School program and this finding was replicated with an independent group of students in spring of 2015.
    - In the fall 2014 session of Outdoor School, student survey data demonstrated that only 25.0% of students reported optimal levels[2] of Academic Identity at the time of the pre-survey administration. After the Outdoor School experience, however, 41.6% of the same group of students (*N* = 416) reported optimal levels of Academic Identity on the post-survey. This amounts to a 66.3% change in the number of students reporting optimal levels of Academic Identity.
    - In the spring 2015 session, initial student survey data demonstrated that only 29.1% of students reported optimal levels of Academic Identity on the pre-

---

[1] Academic identity in science is defined as students' deeply help views of themselves and their potential to enjoy and succeed in STEM classes and careers (Saxton et al., 2014). The components of academic identity have been shown to be important predictors of student engagement and achievement in school, and therefore are thought of as an important indicator of college and career readiness.

[2] Please see Appendix C for a definition of optimal levels.

survey; however, after the Outdoor School experience 41.2% of the same group of students (*N* = 704) reported optimal levels of Academic Identity on the post-survey. This is a 41.6% change in the number of students reporting optimal levels of Academic Identity.

- The project successfully recruited 10 science education experts to serve on an expert panel review of the Outdoor School instructor designed application of conceptual knowledge assessment tasks. The results of the expert panel review were largely positive, very constructive, and provided important feedback to help the instructors make improvements on their field study's assessment.
- During the spring professional development, Outdoor School instructors participated in scoring training with the goal of reaching consensus on how to score assessments. Then, together as a team, the instructors scored approximately 1,100 student answers to the revised application of conceptual knowledge assessments.

Summary of Recommendations.

Based on the evaluation findings described in this report, recommendations are made regarding future assessment efforts and the professional development format for Outdoor school staff. The recommendations related to the application of conceptual knowledge assessments include investing further in inter-rater reliability and optimizing assessment conditions. The recommendations related to the Academic Identity and Motivational Resilience Survey detail suggested sampling and sample size techniques and the resources needed for data entry and analysis. Finally, the project's hybrid professional Development format is recommended for future Outdoor School professional development efforts.

Project Rationale, Goals, and Description

Project Rationale

It is often lamented that we, as a community invested in Science, Technology, Engineering, and Mathematics (STEM) education, lose students in middle school as they begin to make determinations about their identities as STEM capable learners. Outdoor School, however, has the potential of being a powerful, positive influence on sixth graders by providing high quality outdoor education experiences that prevent them from 'shutting off' to bright futures in STEM college majors and careers. Assessment of student learning is necessary to ensure the Outdoor School program does in fact reach this potential. The importance of these student outcomes (application of conceptual knowledge, academic identity, and motivational resilience) lies in their relation to college and career readiness in STEM, as well as, students' future abilities to contribute to society as informed citizens and behave as environmental stewards of the land. Success in STEM learning requires student engagement (hard work and follow through) in completing challenging learning activities, as well as, persistence in the face of obstacles and setbacks; when taken together, this engagement and persistence is referred to as *motivational resilience* but this project.

An *academic identity* as a STEM capable learner is also needed if students are to be ready for transitions across school years and to STEM majors in college and STEM careers. The students served by Outdoor School are in a key transition year from elementary to middle school and are in a vulnerable stage of early adolescent development, therefore, the Outdoor School program's role in sixth grade students' evolving identities is of the utmost importance.

In addition, Outdoor School is an ideal setting for students to learn how to apply their conceptual knowledge to real world, environmental issues. The application of conceptual knowledge, in contrast to the memorization of scientific facts or mathematical formulas, is closely tied to the way knowledge is used in STEM careers and is an explicit goal of the new national K-12 science standards, the Next Generation Science Standards. The Outdoor School curriculum is designed to provide multiple, hands-on field study opportunities for student learning of key concepts related to soil, water, plants, and animals, therefore, the Outdoor School program has the potential to promote students' ability to apply their conceptual knowledge. Measuring student learning as a result of their participation in Outdoor School is a powerful way to not only evaluate the program's success, but also drive program improvement and Outdoor School instructor professional growth.

Project Goals

This project's goals were:

1) To develop and iteratively refine common student assessments for the Outdoor School program, including the 1) student survey of academic identity and motivational resilience, and 2) application of conceptual knowledge tasks and scoring rubric.

2) To use high quality assessment data for program improvement and individual instructor professional growth through collection, analysis, and reflection on student assessment data.

This project's joint goals of assessment development and professional development are intricately related and were intentionally woven together in all project activities. All stages of the project's assessment development cycle (described below) were anticipated to have a positive impact on Outdoor School instructor professional learning in assessment practices. The project adopted an emphasis on assessment *for* learning, rather than assessment of learning. Assessment *for* learning shifts the goal of assessment from simply monitoring learning to using assessments to increase learning. In order for educators to successfully incorporate the effective practices of assessment *for* learning, significant professional development is needed. Empirical evidence indicates educator' assessments practices in STEM education are often inconsistent and not representative of best practices (Lingard, Mills & Hayes, 2006; Niemi, Baker, & Sylvester, 2007; Wood, Darling-Hammond, Neill, & Roschewski, 2007). Since assessment of student learning in the non-formal education sector is an area of high needs and low prior investment, assessment focused professional development is key to realizing the benefits of assessment *for* learning which include enhancing student learning, refining educator practices, and driving program improvement.

Professional Development Description

The project's leadership team designed and offered a total of seven one-day professional development workshops (four in-person and three virtual workshops). During the workshops, the Outdoor School instructors learned about Academic Identity, Motivational Resilience, and Application of Conceptual Knowledge, which are key outcomes to College and Career Readiness in STEM (Appendix A). In addition, the instructors learned of research-based instructional techniques that are positively related to each of these important student outcomes. The instructors also developed assessments to measure student's ability to apply their conceptual knowledge at the end of their field study experience. Importantly, the project expanded the assessment development to all four field studies, rather than only focusing on the water and soil field studies as originally proposed.

As planned, the fall session of Outdoor School was used to gather preliminary data including actual student answers to the application of conceptual knowledge assessments, student interviews, and expert panel feedback on the assessments and rubric. These preliminary data were used to guide field study instructor teams in reflecting on their fall assessment work and in revising the application of conceptual knowledge assessments, which were then administered in the spring session.

The professional development concluded with time in person for the field study instructors to practice scoring the application of conceptual knowledge assessments. The discussion of their chosen scores led to rich conversations about student learning, program goals and expectations, and instructional techniques. Finally, instructors were given time to individually score assessments from the students they specifically taught with guided reflection time after the scoring was complete.

Assessment Development Cycle

The project's rigorous assessment development cycle was designed to be iterative in nature and involve multiple stages: expert review of assessments, student interviews, revision of assessments, training for scoring of assessments with a STEM application of conceptual knowledge rubric, and instructor scoring discussions. The assessment development cycle included Outdoor School instructor involvement at multiple stages of the process, and broader involvement of the environmental education community through the Expert Panel Review Committee.

The project's assessment development cycle was designed from research methods that are advocated for by scholars who specifically recommend strategies such as expert review of assessments and student interviews (Stecher et al., 2000; Lissitz & Samuelson, 2007). The expert review of assessments is designed to use expert opinion to vet assessment questions for clarity, developmental appropriateness, and fairness. The consideration of expert opinions also allows for recommendations of additional questions and/or revisions to questions based on the expertise of reviewers. Student interviews (sometimes called cognitive interviews or think aloud interviews in the literature), are a common technique researchers use to investigate the thought processes students undergo while responding to assessment items (Hamilton, Nussbaum, & Snow, 1997; Renkl, 1997). These student interviews provided rich qualitative data for the Outdoor School instructors to deeply examine what their assessments were actually measuring.

Evaluation Findings

Assessment Development Findings

Application of Conceptual Knowledge. The following section of this report describes both the preliminary and revised versions of the application of conceptual knowledge assessments as well as the results of the expert panel evaluation of the assessments and rubric, student interviews, inter-rater reliability, and assessment scores from the spring 2015 session of Outdoor School.

***Preliminary assessment tasks*.** During the Fall 2014 professional development sessions, the MESD Outdoor School instructors began the assessment development process by working in their field study teams (animals, plants, soil, and water) to identify the science concepts that were the best targets for application of conceptual knowledge assessments. The criteria used to arrive at these key concepts included: 1) concepts that were commonly covered in field study instruction across the three Outdoor School sites, and 2) concepts that were already covered thoroughly during instruction or where instruction could be adapted to be covered more thoroughly. The measurement of application of conceptual knowledge for middle school aged students requires the use of novel tasks; therefore, once the key concept(s) for assessment was agreed upon the instructors worked together to develop a task or scenario that was novel to what is covered during instruction. The assessments that each team developed are provided in Table 1; these assessments were administered during the fall 2014 session of Outdoor School.

Table 1: Preliminary versions of the application of conceptual knowledge assessments that were piloted during the fall 2014 session of Outdoor School.

Application of Conceptual Knowledge Assessments	
Field Study Topic	**Preliminary Fall Version(s)**
Animals	You're on a journey into a neighboring solar system and you touch down on a new planet. As you step out of your spaceship, you notice the ground is squishy and soggy. It is real hot and you're surrounded by tall trees with branches out of reach. All of a sudden you catch a glimpse of an animal! Draw and describe the creature that you saw, including 4 adaptations that help it get food, water, shelter, and protect itself in this unfamiliar habitat. Explain why you chose those adaptions.
Plants	Design a plant to survive in the crack of a city sidewalk. Explain some challenges the plant may face in order to survive in this location. How is the plant adapted to survive in this location?
Soil	Question 1: The good people of the Earth wish to recognize your greatness by building a statue of you outside. Where should they build it, and what inorganic material should they use to make it last the longest? Explain your thought process. Question 2: You are a soil scientist sent to investigate a new planet that was found to have topsoil up to 15 feet deep. What processes on this planet would you expect to find? What characteristics of that planet affect those processes?
Water	A company developed a goo that allows humans to live without drinking water to survive. The company that invented it claims that our global water crisis is solved! As a water scientist (hydrologist) how would you convince the company that water still matters? Your answer can be a persuasive paragraph, a written speech, or an informative poster. Get creative!

***Expert panel findings*.** The project successfully recruited 10 experts to participate in the Expert Panel Review Committee. The members of this committee represented broad areas of expertise of relevance to this project as summarized in Table 2.

Table 2: The Expert Panel Review Committee's combined areas of expertise as indicated by the panel participants.

Summary of expert panel areas of expertise:
63% life science content
50% non-formal education
50% middle school teaching
50% working with or designing assessment or curriculum for under-represented students
13% with university faculty positions in at least one of these areas: education, STEM, and/or ecology

During their participation on the Expert Panel Review Committee, experts individually evaluated the application of conceptual knowledge rubric and instructor-developed assessment tasks. The application of conceptual knowledge rubric (Appendix B) was rated for clarity and whether the described levels of each rubric category represented a logical, developmentally appropriate progression. For each assessment question, the experts rated the question for its accuracy of the content, word choice, clarity, potential to measure application of conceptual knowledge, language bias, and fairness. Each rating was based on a 4-point scale where a score of a 1 or 2 generally represented the need for major revisions to the assessment and a score of 3 or 4 indicated the need for only minor to no revisions to the assessment. Following the methodology of Rubio and colleagues (2003), Content Validity Indices (CVI) were calculated across all ratings by applying the following formula:

$$\text{CVI} = \frac{\text{\# of experts who rated assessment as a 3 or 4}}{\text{Total \# of experts}}$$

CVI's of 0.8 or above are indicative of evidence of strong face validity.

The application of conceptual knowledge rubric was rated as having high face validity (Table 3) in terms of both clarity and logical, developmentally appropriate progression on both rubric categories. The animals, plants and water field studies' assessment questions were all rated as having adequate to strong face validity across all evaluation criteria (Table 4). Both soil questions, however, were rated as having inadequate face validity, particularly soil question 1 was rated as needing major revisions to meet the evaluation criteria (Table 4). In all cases where an expert rated the rubric or assessment questions in need of major or even minor revisions, the experts were invited to give specific feedback regarding their recommendations for changes. These recommendations were taken under consideration during revisions that were made to the rubric and assessment questions between the fall 2014 and spring 2015 Outdoor School sessions.

Table 3: Content Validity Indices for the STEM application of conceptual knowledge rubric indicated that the experts deemed the rubric to have strong face validity.

STEM application of conceptual knowledge rubric	
Evaluation Criteria	**Content Validity Index**
General level of clarity	0.9
'***Demonstration of conceptual understanding***' rubric category: Does the proposed progression seem logical and appropriate for middle school students?	1.0
'***Application of conceptual understanding***' rubric category: Does the proposed progression seem logical and appropriate for middle school students?	1.0

Table 4: Content Validity Indices for the application of conceptual knowledge assessment questions indicated that the experts deemed the majority of the questions to have strong face validity. The experts' ratings did indicate inadequate face validity for the soil questions, particularly soil question number one (Q#1).

Application of conceptual knowledge assessment					
Evaluation Criteria	**Content Validity Index**				
	Animals	**Plants**	**Soil Q#1**	**Soil Q#2**	**Water**
Accuracy of the content	1.0	.88	.33	.88	1.0
Word choice	1.0	.88	1.0	.78	1.0
Clarity of the question	1.0	1.0	.33	.78	1.0
Potential to measure the application of conceptual knowledge	1.0	1.0	.50	.88	1.0
Language bias	1.0	1.0	.75	1.0	1.0
Fairness of the question	1.0	1.0	.88	1.0	1.0

Student interviews. As part of the project's assessment development cycle, students were interviewed following a cognitive interview (or think aloud) protocol. The protocol began with a practice activity to help students become accustom to thinking out loud in the presence of the interviewer, next student were asked to read and complete the assessment question while verbalizing all their thoughts, and finally the interviewer asked follow up questions after the student had independently completed their assessment answer.

During the fall session, 15 students (4 animals, 3 plants, 3 water, 3 soil question #1, and 2 soil question #2) were interviewed at the end of their field study experience while their class was taking the same field study assessment. Three additional interviews were planned for the last week of the session, but a winter storm caused the session to be canceled, therefore, the additional interviews were not possible. Students who were interviewed were volunteers, however, students were intentionally recruited so that both male and female students would be represented. The audio recordings of these interviews were provided to the Outdoor School Instructors during the next professional development days and were used to inform the revisions of the assessments.

Revision of assessment tasks. During the next two days of professional development, the Outdoor School instructors reviewed three data sources that were collected during the fall 2014 session: student interviews, expert panel feedback, a common, cross-site subset of student answers to each assessment question, and a subset specific to their site and, therefore, their instruction. Each data source was accompanied with a series of reflective questions that guided the review of the data and prompted thinking about revising the assessment question. During the second professional development day (which was in person), more assessments were reviewed as a whole group and the entire team of instructors and the project staff worked together to revise the assessment questions for all field studies. The revised assessment questions were administered during the spring 2015 session (Table 5).

Table 5: Revised versions of the application of conceptual knowledge assessments that were administered during the spring 2015 session of Outdoor School.

	Application of Conceptual Knowledge Assessments
Field Study Topic	**Preliminary Fall Version(s)**
Animals	You're on a journey through space, and you land on a new planet. As you step out of your spaceship, you notice that the ground is <u>squishy</u> and <u>soggy</u>. It is <u>very hot</u> and you see <u>tall trees</u> with branches that you can't reach. Draw and describe a creature that lives in this habitat. Include some adaptations that help it get food, water, shelter, or protect itself. Using clues from the first paragraph explain why the animal needs these adaptations to survive in its habitat.
Plants	Imagine a plant that is growing in the crack in a sidewalk. Keeping in mind all of the things a plant needs in order to live, explain what would make growing here difficult? What parts of the plant help it survive in this area?
Soil	You are a soil scientist studying a new planet. You find that the topsoil is 15 feet deep. What would you expect the duff to be like? Why? What would you expect the subsoil to be like? Why?
Water	Imagine a time in the future in which scientists have it made so that humans no longer have to drink water to survive. How would you convince people that clean and healthy watersheds still matter? Your answer can be a persuasive paragraph, a written speech, or an informative poster. Get creative!

Inter-rater reliability. One full day of scoring training was provided to nine of the Outdoor Instructors at the end of the spring 2015 session. Upon the conclusion of the training, the instructors each independently rated an additional common set of assessments for each field study. Both at the end of the scoring training and after the training, intraclass correlations coefficients were computed for each instructor team to determine their inter-rater reliability (or scoring consistency). Generally, intraclass correlations of 0.70 or above are considered acceptable levels of inter-rater reliability. The results of the four field study teams' inter-rater reliability at the end of the training are displayed in Table 6. These results indicate that at the end of the scoring training: 1) two out of four instructor teams reached acceptable levels of inter-rater reliability on the demonstration of content knowledge rubric category and 2) all teams reached acceptable levels of inter-rater reliability on the application of conceptual knowledge rubric category (Table 6).

Table 6: The inter-rater reliability on the last practice set *during* the scoring training.

		Rubric Category			
		Demonstration of Content Knowledge		**Application of Conceptual Knowledge**	
Field Study Topic	**Number of instructors**	**Single Measure* ICC**	**Average Measure* ICC**	**Single Measure ICC**	**Average Measure ICC**
Animals	3	.75	.90	.81	.93
Plants	2	.62	.77	.81	.90
Soil	2	.56	.72	.85	.92
Water	2	.76	.86	.77	.87

* Single measure ICC is suitable for use when one rater will do the rating of individual assessments in the future; the average measure ICC is suitable if a project plans to always have multiple raters and have scores averaged across those raters.

After the training session, the inter-rater reliability for the field study teams was again tested when each team independently scored a common set of assessments after the training session; these inter-rater reliability results are displayed in Table 7. Across all instructor teams the inter-rater reliability indices decreased for both rubric categories. These findings indicate that the Outdoor School instructors drifted away from the agreement that was established during the training. Further, this drift away from the agreement that was established during the scoring training happened within two weeks of the scoring training. These findings indicate that depending on the field study team in question, the Outdoor School instructors either continued to interpret and score student answers in ways that were not in agreement with their fellow instructors or they lost the agreement they had achieved in the training. While the scoring of assessments served as a valuable professional development purpose for individual instructor reflection on instruction and student learning during this project, the lack of reliable program-wide data about students' ability to apply their conceptual knowledge at the end of Outdoor School limits the uses of these data. Recommendations on two potential ways to address this reliability challenge are proposed below.

Table 7: The inter-rater reliability *after* one day of scoring training. No instructor team reached an acceptable level of inter-rater reliability on both rubric categories. One out of four instructor teams reached acceptable levels of inter-rater reliability on the demonstration of content knowledge rubric category. Two teams reached acceptable levels of inter-rater reliability on the application of conceptual knowledge rubric category.

		Rubric Category			
		Demonstration of Content Knowledge		Application of Conceptual Knowledge	
Field Study Topic	**Number of instructors**	**Single Measure* ICC**	**Average Measure* ICC**	**Single Measure ICC**	**Average Measure ICC**
Animals	3	.61	.83	.78	.91
Plants	2	.50	.67	.72	.84
Soil	2	.18	.30	.40	.57
Water	2	.82	.90	.62	.77

* Single measure ICC is suitable for use when one rater will do the rating of individual assessments in the future; the average measure ICC is suitable if a project plans to always have multiple raters and have scores averaged across raters.

Spring assessment scores. After the scoring training, the Outdoor School instructors independently scored student assessments from the spring 2015 session. Together as a program-wide team, the instructors scored approximately 1,100 student answers to the revised application of conceptual knowledge assessments. The percentage of students scoring as proficient[3] on the rubric category of demonstration of conceptual understanding and application of conceptual knowledge are provided in Table 8.

Regarding the relatively low levels of proficient scores on the application of conceptual knowledge assessments, a few interpretations of these preliminary findings are offered. First, logistical issues with administering the assessments may partly explain the relatively low levels of proficient scores on the application of conceptual knowledge assessments. Moving forward, creating assessment conditions that support students in composing their best answers is of the utmost importance for the Outdoor School program to have high quality data to draw conclusions from. Second, it is important to recognize the baseline nature of the assessment of application of conceptual knowledge data. Outdoor School instructors have never had access to this type of formal assessment data on student learning. This project has, therefore,

[3] Scores of 3 or 4 on the PMSP application of conceptual knowledge rubric are currently proposed as proficient scores for middle school students; however, it should be noted that this proficiency cut mark is still somewhat tentative as it has not yet been empirically investigated.

empowered the MESD Outdoor School program with baseline application of conceptual knowledge data. These data will frame future professional development work for the program and allow them to begin to track their progress over time in the future.

Table 8: Percentage of the application of conceptual knowledge assessments that were scored as proficient on the PMSP application of conceptual knowledge rubric. The rubric includes two scoring categories: demonstration of conceptual understanding and application of conceptual knowledge.

Field study topic	Number of assessments scored	% proficient on demonstration of conceptual understanding	% proficient on application of conceptual knowledge
Animal	376	34.1	26.3
Plants	346	35.2	23.4
Soil	261	17.5	7.8
Water	137	26.6	14.4

Academic Identity and Motivational Resilience. The Academic Identity and Motivational Resilience Survey that was used in this project was designed to measure both the overarching outcomes of academic identity and motivational resilience as well as several sub-scales realted to these outcomes (Saxton et al., 2014). The survey's sub-scales are intended to measure student identity, self-system beliefs (including relatedness, autonomy, competence beliefs), and purpose of the discipline, engagement, and constructive coping/persistence. All items are worded as they relate to a specific STEM discipline (in the case of this project, science).

***Survey return rate*.** Despite the surveys being delivered to all teachers by the Outdoor School site supervisors, not all teachers administered the surveys. In addition, it was somewhat common for teachers to administer either the pre-survey or post-survey, but not both surveys. For example, in the fall the project was successful in collecting approximately 1000 pre-surveys, but only 489 post-surveys were returned to the MESD. This attrition in survey completion from pre-survey to post survey makes it challenging to preform analyses that attempt to look at changes in students' Academic Identities and Motivational Resilience. To address this challenge, the MESD Outdoor School program coordinator added an email communication to teachers during the spring to help remind them of the surveys. This resulted in a modest increase in the survey return rate. On a positive note, investigation into which schools tend to return completed and matching surveys did not reveal any concerning trends related to school-wide demographic variables. Outdoor School, however, will need continue to navigate this challenge in the future so that survey data can be collected from a representative sample of the students who attend the program. This finding is discussed further in the recommendations section.

Internal Consistency of the Student Survey. The internal consistency of a survey provides some indication of the coherence of the items the survey is composed of and whether the proposed scales and sub-scales are made of items that measure the same outcome, thus internal consistency indices indicate if survey scales or sub-scales are appropriate to be used in further analyses. This project measured the internal consistency of the Academic Identity and Motivational Resilience Survey by calculating Cronbach's alpha. Levels of 0.70 or above are generally considered indicative of a survey scale having adequate internal consistency. Table 9 shows both the number of items and the internal consistencies of the two main survey scales and its 5 sub-sales. These findings indicate that the two main survey scales of Academic Identity and Motivational Resilience had strong internal consistencies and that four of the five sub-scales met the criteria for adequate internal consistency.

Table 9: Internal Consistencies for the Academic Identity and Motivational Resilience Survey during the pre-Outdoor school administration in spring of 2015. With the exception of the purpose sub-scale, all scales had adequate to strong internal consistencies.

Student Outcome	Survey Scale or Sub-scale	Number of items	Internal consistency
Academic Identity	**Academic Identity (all items)**	**15**	**.86**
	Identity	**5**	**.70**
	Self-system beliefs (relatedness, autonomy, competence)	**6**	**.70**
	Purpose	**4**	**.65**
Motivational resilience	**Motivational resilience (all items)**	**14**	**.83**
	Engagement	**6**	**.73**
	Constructive Coping	**8**	**.74**

Results of the Student Surveys. The project found a positive and statistically significant impact on students' academic identity in science in the fall 2014 session of the Outdoor School program and this finding was replicated with an independent group of students in spring of 2015. In the fall 2014 session, student survey data demonstrated that only 25.0% of students reported optimal levels of Academic Identity (please see Appendix C for a definition of optimal levels) at the time of the pre-survey administration; however, after the Outdoor School experience 41.6% of the same group of students (N = 416) reported optimal levels of Academic Identity on the post-survey. This amounts to a 66.3% change in the number of students reporting optimal levels of Academic Identity. In the spring 2015 session, initial student survey data demonstrated that only 29.1% of students reported optimal levels of Academic Identity on the pre-survey; however, after the Outdoor School experience 41.2% of the same group of students (N = 704) reported optimal levels of Academic Identity on the post-survey. This is a 41.6% change in the number of students reporting optimal levels of Academic Identity.

A paired samples *t*-test was conducted to evaluate whether the observed increase in student's academic identity after the Outdoor School experience was statistically. The results indicated that the mean academic identity after attending Outdoor School (M = 56.63, SD = 10.247) was significantly greater than the mean academic identity prior to attending Outdoor School (M = 53.87, SD = 9.720), $t(703)$ = -12.916, $p < .001$. For most students, overall academic identity scores increased after the Outdoor School experience. No significant change in motivational resilience was found.

Professional Development Goals. The project's professional development goals were to use high quality assessment data for program improvement and individual instructor professional growth through collection, analysis, and reflection on student assessment data. The following section describes the findings related to this goal including instructor attendance, self-efficacy, reflections and the professional development format.

Instructor attendance. The seasonal nature of Outdoor School's instructor staff created challenges for professional development than occured across the two sessions of the Outdoor School Program. Eight staff members remained in the employment of the MESD Outdoor School Program for both the fall 2014 and spring 2015 sessions. Of the Outdoor School instructors who were employed for the full project duration, 88% (N = 7) attended 6 or more days of the professional development. In total, 14 Outdoor School instructors and three site supervisors took part in some portion of the professional development.

Instructor self-efficacy. In order to quantitatively measure the impact of the professional development on the Outdoor School instructors, a self-efficacy survey was adapted from the STEM Common Measurement System (Saxton et al., 2014). The changes to the survey primarily involved the deletion of items that measured instructional practices that were not anticipated to change as a result of the professional develop and the addition of more assessment-focused questions. The Outdoor instructors were asked to complete the self-efficacy survey at the beginning of the first day of the professional development, at the end of the fall session, and after the conclusion of the last professional development day in spring of 2015. When comparing efficacy beliefs pre- and post- the professional development program, 71% of instructors (N = 7) who attended 6 or more of the professional development days reported some small increase in their assessment-specific efficacy beliefs. However, no significant difference was found in the Outdoor School instructor's assessment for learning efficacy beliefs when comparing either pre-professional development responses and mid-professional development responses (N = 10) or pre-professional development responses and post-professional development responses (N = 7).

Two potential interpretations of these findings relate to the adapted survey itself and the results of the application of conceptual knowledge assessments. First, the adaptation of the self-efficacy survey was done in an effort to create a survey that would be calibrated to the specific goals of the professional development (i.e. the assessment focus); however, it was not possible to test the reliability or validity of the adapted survey prior to the beginning of the professional development. It could, therefore, be that measurement error interfered with any potential opportunity to uncover any positive impact on the instructors' efficacy beliefs. Second, both the logistical issues instructors faced in administering the application of conceptual knowledge assessments and the low percentage of student scoring at proficient levels on the spring assessments could have dampened any positive impact the professional development had on the Outdoor School instructors' assessment-specific efficacy beliefs.

***Instructor reflections*.** Despite the lack of quantitative results supporting the positive impact of the professional development on the Outdoor School instructors, the project does have qualitative evidence in support of the value of the professional development experience. This project's impact is perhaps best conveyed by sharing two excerpts from the Outdoor School instructors' reflections at the end of the project:

> "It was great to have time to truly reflect on the teaching … Seeing what students wrote gave me insight into what was on the top of their brain … It was also highly beneficial to me to have the professional development days at both the start and end to think and talk with coworkers in a productive way about how we work with children: what questions work and which ones do we just ask because we've been asking them? How do we know what students are learning? Can we figure that out better? So, I guess it left me with a lot of questions, but they are the types of questions that I believe are important to keep asking and making time for and I look forward to seeing what this project grows into."

> "… spending this time post session to read all the answers again and think about what concepts stood out to students was really interesting, and beneficial to how I think about the information I present, and also just how we think about assessment of student learning. It was nice to look at these [assessments] and see the connections kids made. For my field study I focused a lot of time/energy into students drawing conclusions on their own, and making connections across not only concepts on my field study but also [all the] field stud[ies], so it was cool to see some students doing that within this question, … I also think it was nice to have the same question across all three sites, and to look at other sites answers during the practice [scoring], but then to only focus on my site during this final grading time. It allowed me to compare student learning, and also to focus on my own students success."

These reflections indicate that, at a minimum for these two instructors, the professional development stimulated deep reflection on their informal and formal assessment practices, students' learning, and instructional practices.

Professional development format. The virtual professional development format was not originally planned, but was adopted to accommodate the seasonal staff. Outdoor School staff often leave the local area immediately following the end of Outdoor School sessions to either begin other jobs or return to their home city. The creation of a virtual professional development platform enabled the project to create greater access to the professional development experiences than would have been possible if all workshops had required in-person attendance in Portland. This hybrid professional development (partially in-person and partially virtual) was a successful technique in this project's professional development efforts.

Recommendations and Next Steps.

Application of Conceptual Knowledge Assessments.

Investing further in inter-rater reliability. The inter-rater reliability levels reached in this project did not consistently meet adequate levels. These findings indicate that at the end of the project, the Outdoor School instructors were still interpreting and scoring student answers in ways that are not in agreement with their fellow instructors. This lack of inter-rater reliability is a threat to the quality of the application of conceptual knowledge data particularly for any analyses that aggregate the data across sites (and therefore across raters). The lack of inter-rater reliability could be addressed with one of two solutions. First, the instructors made clear progress in the consistency of their scoring (or their inter-rater reliability) over the course of the one-day training, therefore, longer trainings may be required in the future to reach acceptable levels of inter-rater reliability. Second, when considering the average measure intraclass correlations (Table 6 and 7) rather than the single measure intraclass correlations, the instructor teams met acceptable levels 100% of the time at the end of the training and 63% of the time after the scoring training. This indicates that Outdoor School could adopt a scoring protocol that calls for 1) a sub-sample of assessments to be rated by more than one instructor and 2) the use the average of those raters scores to draw interpretation of the program's impact of students' ability to apply their conceptual knowledge. To make the implementation of either solution feasible, each of these recommendations requires that Outdoor School have an appropriate budget to make both scoring training and staff scoring time a possibility moving forward.

Optimizing assessment conditions. It is undeniable that implementing formal assessment tasks in a non-formal setting is a challenging change for both educator and students. The conditions under which students are assessed will have an impact on their performance on the assessment. Determining the best time in the day, a dry, warm place conducive to concentration, and the introduction to the assessment task that sets expectations for the quality of student answers are all examples of the strategies the MESD Outdoor School team implemented to try to address this challenge. While many instructors reported progress and valuable learnings in the spring session, it remains undeniable that this will be a continual challenge in the future at Outdoor School. Moving forward, the MESD Outdoor School team should strive towards creating assessment conditions that support students in composing their best answers. These efforts are of the utmost importance for the Outdoor School program to have high quality data to draw conclusions from. Assessment times that are after dinner, in competition with field day or other enticing activities, or in locations that are cold, wet or otherwise likely to hinder student concentration should be avoided in the future.

Academic Identity and Motivational Resilience.

Strategic sampling and sample size. This project attempted to collect data from all participating students, but found significant barriers related to survey return rates. In the future and in order to ensure a representative sample of students is surveyed, a better strategy may be to strategically sample from participating schools. This recommended entails several steps. First, demographic data should be collected from the Oregon Department of Education's annual School report cards for all participating schools in order to gather school-level data about the socioeconomic status (% economically disadvantaged), numbers of languages spoken, and ethnicity of the student population at each participating school. Second, a ranking system could be devised to classify schools into two or three demographic categories (i.e. > 50% economically disadvantaged and more than < 50% economically disadvantaged). Finally, schools from each demographic category should be randomly selected for inclusion in the survey data collection. If the MESD Outdoor school program continues to serve approximately 3500 students per session, it is further recommended that at least 784 student be surveyed per session. This minimum sample size and the stratified sampling method described above should ensure that the sample is representative of the population MESD Outdoor School serves and the subsequent analyses are adequately powered to detect significant findings.

Resources needed for data entry and analysis. In order for the MESD Outdoor School program to be able to continue to collect and reflect on student survey data, it is important to consider the resource that are needed to effectively collect, enter, and analyze the data. While a strength of this project was the offer of a choice of paper or online surveys to participating teachers, this choice was made possible by the additional staff resources and printing budget created by the project's external funding. The use of paper surveys is associated with the potential benefit of preventing teachers, who work in schools without adequate technology or with limited computer access[4], from administer the surveys. On the other hand, the paper surveys are associated with a significant cost of staff time[5] to enter the survey data. Moving forward, the MESD Outdoor School program either needs proper staff resources to

[4] Computer access was particularly challenging during the spring session because in many schools computers were otherwise committed for the use of standardized testing for much of spring 2015.
[5] It is estimated that each survey takes a trained staff member approximately 1.5 minutes to enter. If hypothetically the minimum sample recommendations are followed and all students complete pre and post surveys on paper, data entry time could be as high as 40 hours per session.

ensure paper surveys can be entered or the program should move to exclusively online surveys. Finally, beyond entering survey data the MESD Outdoor School program either needs qualified staff or a supporting partner to aid in the analysis of the survey data once it is collected.

Professional Development Format. As previously discussed, the seasonal nature of Outdoor School staff presented challenges for implementing a professional development plan over the course of 2 sessions. It is recommended that the hybrid professional development (partially in-person and partially virtual) is considered a valuable technique for future Outdoor School professional development efforts.

Appendix A: Key Terms

STUDENT OUTCOMES: Effective STEM learning environments

Academic identity
Students' deeply held views of themselves and their potential to enjoy and succeed in STEM classes and careers.

Components: identity, belonging/relatedness, competence/efficacy, autonomy/ownership, and purpose.

Rationale: This is the fundamental student transformation that needs to be accomplished if we are going to see the effort and determination students need to achieve in STEM.

Motivational resilience
Characterized by students' enthusiastic hard work and persistence in the face of challenging STEM coursework.

Components: academic engagement and constructive coping/persistence.

Rationale: Whole-hearted engagement and tenacity in demanding STEM classwork is essential to student learning and achievement.

- Furrer, & Skinner, 2003; Skinner, Kindermann, & Furrer, 2009.

Application of Conceptual Knowledge
Students' understanding of and thinking about ideas, theories and perspectives considered critical or essential within an academic or professional discipline or in STEM interdisciplinary fields recognized in authoritative scholarship. "....References to isolated factual claims, definitions, or algorithms are not indicators of significant disciplinary content unless the task requires students to apply powerful disciplinary ideas which organize and interpret information."

- Definition adapted from Lingard, Mills, & Hayes, 2006

Rationale: The focus on deep understanding and application of conceptual knowledge is key to student success in STEM because it more accurately reflects the way concepts are applied in the real world by scientists, engineers, and other STEM professionals. This outcome stands in stark contrast to rote memorization of isolated facts, definitions, formulas, or algorithms because application of conceptual knowledge results in longer lasting understanding of STEM content.

Appendix B: PMSP Application of Conceptual Knowledge Rubric

Score	Demonstration of conceptual understanding	**Application of conceptual knowledge:** Apply powerful disciplinary ideas, which organize and interpret information and evaluate them in *new* contexts.	Score
4	Demonstrates a firm, correct understanding of relevant concept(s) by clearly and accurately utilizing all appropriate vocabulary. No misconceptions.	Provides strong evidence of an ability to apply the target concept(s) to a novel context through the correct application of relevant science concepts to draw conclusions, extend key concepts to make predictions, or explain observations.	4
3	Demonstrates a correct understanding of relevant concept(s) by utilizing some vocabulary, however, there may be slightly inaccurate use of vocabulary or minor misconceptions.	Provides satisfactory evidence of an ability to apply the target concept(s) to a novel context through the mostly correct application of relevant science concepts to draw conclusions, extend key concepts to make predictions, or explain observations.	3
2	Demonstrates a partially correct understanding of relevant concept(s), but uses imprecise vocabulary and some misconceptions are revealed.	Provides incomplete evidence of an ability to apply the target concept(s) to a novel context through a partially correct application of relevant science concepts to draw conclusions, extend key concepts to make predictions, or explain observations.	2
1	Demonstrates a completely incorrect understanding of relevant concept(s), with inaccurate vocabulary and several misconceptions revealed.	Provides little to no evidence of an ability to apply the target concept(s) to a novel context OR Provides conclusions, predictions, or explanations that are incorrect, limiting evidence of application of relevant science concepts.	1
0	Evidence either missing or too insufficient to score.	Evidence either missing or too insufficient to score.	0

Appendix C: Percent optimal

Explanation of percent optimal analysis:
Calculating the percentage of students who fall within the optimal threshold on the PMSP academic identity and motivational resilience is an additional useful way to analyze and interpret these data. At this point in time, the percent optimal is theoretically based rather than being empirically based. The theoretical rationale for this analysis is that on the 5-point likert scale (1 = Not true at all; 5 = totally true), scores of 4's (mostly true) and 5's (totally true) are at the most positive end of the scale*; therefore, the optimal answer pattern for a student would logically be answers of only 4's and 5's on all survey items. Following this logic, percent optimal is calculated by multiplying the number of questions by 4 to create an 'optimal' cut mark. Then the number of students that tabulate at or above the cut mark are counted and that number is divided by the total number of students who took the survey.

*Please note that negatively worded items on the survey are reverse coded so that a higher score always indicates a more positive answer.

Appendix D: References

Hamilton, L.A., Nussbaum, M. & Snow, R.E. (1997). Interview Procedures for Validating Science Assessments. *Applied Measurement in Education*, 10(2), 181-200.

Lingard, B., Mills, M., & Hayes, D. (2006). Enabling and aligning assessment for learning: some research and policy lessons from Queensland. *International Studies in Sociology of Education*, 16(2), 83-103.

Lissitz, R.W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Education Researcher*, 36(8), 437-448.

Niemi, D., Baker, E.L. & Sylvester, R.M. (2007). Scaling Up, Scaling Down: seven years of performance assessment development in the nation's second largest school district. *Educational Assessment*, 12(3), 195-214.

Renkl, A. (1997). Learning from Worked-out examples: A study on Individual Differences. *Cognitive Science*, 21(1), 1-29.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., & Rauch, S. (2003). Objectifying content validity: conducting a content validity study in social work research. *Social Work Research*, 27(2), 94-104.

Saxton, E., Burns, R., Holveck, S., Kelley, S., Prince, D., Rigelman, N., & Skinner, E. (2014). A Common Measurement System for K-12 STEM Education: adopting an educational evaluation methodology that elevates theoretical foundations and systems thinking. *Studies in Educational Evaluation*, 40, 18-35.

Stecher, B.M., Klein, S.P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R.J. & Haertel, E. (2000). The effects of content, format, and inquiry level on science performance assessment scores. *Applied Measurement in Education*, 13(2), 139-160.

Wood, G.H., Darling-Hammond, L., Neill, M. & Roschewski, P. (2007). Refocusing accountability: using local performance assessments to enhance teaching and learning for higher order skills. Washington, D.C.: Briefing paper prepared for members of the congress of the United States.