Validating an Environmental Education Field Day Observation Tool*

Stephan P. Carlson**

Joe E. Heimlich

Martin Storksdieck

Abstract
Environmental Field Days are held throughout the country and provide a unique opportunity to involve students in real world science. A study to assess the validity of an observation tool for EFD programs was conducted at the Metro Water Festival with fifth grade students. Items from the observation tool were mapped to students' evaluation questions to determine the degree to which observed characteristics of the field day are aligned with student perception. The data support the conclusion that the observation tool not only captures the perspective of a trained observer on the educational potential of a field day, but also the perceived experience of the field day audience (the students): Despite the fact that the observation tool was designed to capture an expert perspective on effective pedagogy and educational practice (rather than student satisfaction), 20 out of 26 items correlated between the observer's and student's assessment tool.

Keywords: Observation tool, environmental field day, validity study, informal science, empirical research

Introduction

Environmental Field Days (EFD) such as Children's Water Festivals, Conservation Days and Agriculture Days provide a unique opportunity to involve students in real world science to build understanding and skill in science, technology, engineering and mathematics (STEM). Field Day programs involve a variety of agencies and organizations like museums, zoos, nature centers arboretums, departments of natural resources, soil and water conservation districts and cooperative extension services. During a Field Day, students usually visit six to eight stations for about 30 minutes

* This research was funded by NSF REESE #0635559, Developing an Evaluation Tool for Environmental Field Days and presented at NAAEE, NARST, and VSA conferences.
** Corresponding author: Stephan Carlson 115 Green Hall, 1530 Cleveland Ave. N. St Paul, 55108-6112, 612 624-8186. e-mail: carls009@umn.edu

ISSN: 2146-0329
www.iejeegreen.com

each, where they engage in hands-on activities, demonstrations and discussions around STEM or environment-related issues (Poudel, Vincent, Anzalone, Huner, Wollard, Clement, DeRamus, & Blakewood, 2005). The stations are often taught by volunteers, many of whom are scientists working for local, state and federal agencies, or nongovernmental organizations (NGO's). Based on an overarching topic (for instance water quality), these professionals provide six to eight independent experiences at their stations in which students simulate various human impacts (for instance erosion on a water table), active models (for instance students become water droplets zig-zagging through the water cycle), and guided exploration (taking water samples from a local stream). The purpose of designing field days around a set of independent, yet related stations or experiences is to allow students a broad introduction to a topic of real-life significance through problem-based learning. Field days can be considered as highly structured and comprehensive field trip experiences for students and teachers.

Informal educators consider well-structured and executed field trips and field days as starting points for young people to gain first-hand knowledge and experience about science as it relates to the environment (Carlson, 2008; Storksdieck, 2006), and as important contributors to positive attitudes towards science and career aspiration in science (Barney, Mintzes & Yen., 2005; DiEnno & Hilton, 2005; Farmer, Knapp, & Benton, 2007; Knapp, & Benton, 2006). However, there is some concern that the field day practice might not always live up to its potential since EFD experiences are generally facilitated by content experts, who tend to be professionals with little or no background in teaching or education. A variety of researchers have addressed "Best Practices" for informal environmental/stewardship education in extended classroom experiences (NAAEE, 1996; Carlson, 2008; DeWitt and Storksdieck, 2008; Meyer & Pardello (Eds.), 2005, Siemer, 2001; McDonnell, 2001; Fortner, 2001; Stevens and Andrews, 2006). However, *few tools exist to measure the effectiveness* or quality of out-of-school learning experiences in ecologically valid ways. Hence, developing an effective observation tool that captures the "best practices" constructs of informal science education is critical to begin measuring the potential educational quality of EFD programs (Carlson, Heimlich, Storksdieck & Meyer, 2009).

A variety of observation tools or "learning environments inventories" (LEI) have been developed to measure social and psychological aspects of student outcomes in science education through the use of trained observers (e.g., Dorman, 2003; Fraser, 1998; Fraser & Fisher, 1982; Lawrenz, 1987; Talton & Simpson, 1987; Wahyudi & Treagust, 2004). These may include assessments of the extent to which students are supportive of each other, are actively engaged in learning, or the degree to which curricular or lab materials are appropriate. There is no "perfect" inventory for all informal programs. In fact, most appear to be intended for use in formal science

classrooms. For example, Henderson, Fisher and Fraser (1998) documented the use of a validated Environmental Science Learning Environment inventory for in-class purposes. Brown (1996) utilized a Science Laboratory Environment Inventory, again, for the formal classroom. On the other hand, informal settings have been studied by the High Scope Research Foundation (2006) which developed the Youth Program Quality Assessment tool which was developed as a validated instrument to evaluate the quality of youth programs that are not necessarily specific to science programs. Storksdieck, Kaul and Werner (2006) developed a valid, theory-derived field trip teacher feedback form to assess the quality of overall field trip experiences for museums and other types of informal learning environments that, while comprehensive and based on self-assessment of behavior, was lengthy, potentially burdensome to complete, and required teachers to mail back a questionnaire (a considerable impediment to achieving satisfactory response rates).

Carlson and colleagues at the University of Minnesota have developed an observation tool for trained evaluators to assess EFD (Carlson, Heimlich, Storksdieck, & Meyer, 2009). This tool was based on the curriculum, *Best Practices for Field Days: Program Planning Guidelines for Organizers, Presenters, Teachers and Volunteers* (Meyer & Pardello (Eds.), 2005) and a 2008 study by Carlson that noted that EFD took place in over 75% of Minnesota's counties and annually educated more than 10,000 students. The observation tool was designed to capture a variety of field day characteristics that previous research suggested would provide conditions that are conducive for learning, including (among others) the use of proper introductions to the topics being investigated or discussed, effective techniques for engaging students, student engagment itself, aspects of the social and physical environment, etc. Altogether, six overarching constructs known to positively influence the learning potential of field days were included in the observation tool (see Appendix A). Since the purpose of the tool was to provide feedback on the educational quality of the field day for individual instructors and field day organizors, trained observers would capture the student experiences at each individual leaning station as well as for the field day as a whole. The content validity of the tool was established by linking constructs with each item to "best practice" theory (Carlson, 2008), and then validating them through a modified-Delphi study with a range of experts (Heimlich, Carlson, Tanner & Storksdieck accepted). Coder reliability of the tool was established through rigorous rounds of testing and revisions to reach an acceptable modified Kappa for each item (Storksdieck, Heimlich, Figueriredo & Carlson., 2009). Following the psychometrics for observation research (Hintze, 2005), this validation study was conducted where observation data were compared to student's perceptions.

Method

The validation study was conducted at the Metro Water Festival (CWF) in St Paul Minnesota in the fall of 2008 where 44 schools and over 1,200 fifth grade students participated in a one-day Field Day event. The purpose of the study was to determine whether and in what ways the results obtained from multiple observers on the pedagogical quality of the field day experience, strictly an observation tool, aligns with the perceived experience of students who took part in the observed field day. The study did not aim to show that these two different perspectives necessarily overlap; in fact, one would expect for a variety of theoretical reasons that there could be significant differences between the observation and the student perception. However, in crafting student feedback items that were closely aligned with the observation tool, the study aimed at testing (a) whether the observation tool at least partially captured the audience experience, and (b) develop hypotheses about the connection between observed "best education practice" and student experiences. The results we are presenting represent a form of ecological validity: how does the perspective of experts correspond to the experience of learners. Content validity (Modified Delphi) and coder reliability of the observation tool was established the previous years. Items from the observation tool were mapped to students' evaluation questions to determine the degree to which observed characteristics of the field day are aligned with student perception. It is conceivable that they don't align. Students' assessment of their experience is based on factors that have little to do with what educators care about. Significant correlations support the validity from the perspective of the students experience; lack thereof, on the other hand, does not indicate that the tool isn't valid, at least for capturing the quality of field days based on educational theory and education expert perspectives.

The schools that attended CWF were selected from a large pool of interested schools and all agreed to provide program evaluations. There was no cost to students or schools to attend the event and lunch was also included along with bussing for some of the schools. The Children's Water Festival had 31 different learning stations going on throughout the day; students visited 5 to 7 of the learning stations during the day. Student stayed at each station about 30 minutes and then moved on to the next station. The stations that each student would visit were assigned by CWF crews. Students were greeted at their bus when it arrived and guided through the day by volunteers to each of the learning stations, lunch and back on the bus at the end of the day. Learning stations were taught by volunteers and professionals from state and federal agencies along with non-profit organizations.

Of the 44 classrooms, a sample of 16 classrooms, (representing 36%) from 5 schools were selected to be followed each by a trained observer who would be using the observation tool to documnt the experience of the particular class being tracked. Trained observers rated the quality of

instruction at each of the learning stations. They scored station presenters on 26 items and students on four items of engagement. The same observers followed the same class throughout the day. Consent forms for participation in the observation study and for the student study were mailed to principals and teachers and sent home to parents to respond with an "opt out" response request. In addition, all classrooms were given copies of a post field day evaluation survey to be completed by students and were asked to return them by the end of the week. Return rate for the 44 classrooms was 90 %. The 16 classrooms in the study had a return rate of 100%. Data from the post field day survey were used in three ways: (1) to provide feedback to the field day organizers; (2) to compare the research sample to the total field day student population for that field day (estimating bias), and (3) to correlate the sample population's feedback data with those obtained through expert observers using the observation tool (the purpose of collecting the student data). The student feedback questionnaire and the observation tool had very different purposes and measured very different things. The observers' questionnaire measured the quality of the educator/student interaction or teaching-learning exchange (for the entire class) at each of 5-7 learning stations while the feedback questionnaire allowed students to evaluated their own experience once, at the end of the day, and for the overall field day rather than individual field day stations.

One would not expect a great deal of overlap among these two aproaches, despite the best attempt to develop items for the student questionnaire that aligned with the observation tool. Nevertheless, one or more items on both of these instruments addressed the six major constructs of the observation tool *opening the field day experience, expressing age appropriate language* and *instruction*, *using a variety of questioning strategies,creating or using a physical environment that did not distract from learning*, *student's engagement*, and *student's satisfaction* (See sample in Table 1.). A positive relationship between the observation and the student feedback would help establish the utility and validity of these six constructs and thereby the observation tool overall for field day organizers, field day educators, teachers, parents and students. In addition, positive correlations between the two assessment methods would strengthen the case that good educational practice in out-of-school experiences are perceived positively by the audience.

The following table show 3 the basic categories (constructs) and criteria of measurement as it is applied to the observation tool and student survey. Appendix A shows all six major constructs and the questions used to measure each construct.

Table 1.:
The framework of observer individual assessment tool and student survey

Basic Categories	Criteria of Measurement	Observer Individual Assessment Tool	Student Survey
Management (Physical Environment 1)	The instructor conveys appropriate voice volume and adjust his or her position to be seen by students when he/she delivers the program	2l. Was seen and heard by all participants nearly all the time	**2h. I could hear and see the presenters at the stations**
Engagement (Student's Engagement)	The instructor and the program attract student's attention all the time	2g. Kept nearly all participants focused on activities most of the time 4a. Listened attentively when expected 4b. Participated fully when expected	**2m. I learned something new at the stations** **2o. I paid attention at the station** **2q. Kids in my class listened when they were supposed to** **2s. Kids in my class really got into the activities at the stations**
Satisfaction(Student's Satisfaction)	Student enjoy the instructor and the learning program during their field trip experience	4c. Showed excitement and enthusiasm	**2g. I enjoyed the presenters** **2t. Kids in my class had fun at the stations** **2p. I found the stations interesting** **3d. I enjoyed being at the Water Festival** **3f. The presenters at the Water Festival were nice to me**

Concurrent validity was tested using the correlation between the observation and student survey items.

Results

A pedagogical framework was created that matched items on the observer assessment tool with student survey questions on the six constructs (see above).. The frameworks six constructs were measured with a total of 12 items from observer's assessment tool and 14 items from the student's feedback survey. For the purpose of analysis, we classified the 26 items into one of the six basic categories (constructs), and each of the six constructs was measured with at least one or mostly several items. Because of the purpose of the items in the category of *expressing age-appropriate language*, we divided this category into two sub-categories, *expressing 1* and *expressing 2*. The questions in "*expressing 1*" examined if the presenter used appropriate language when he or she conveyed his or her message.

The questions in "*expressing 2*" focused on the clarity of instructions during program delivery (see Appendix A or table above.).

A t-test was conducted between the sample group (n=16 classes) and the total population (n=44 classes) to determine sample bias, and none of the classes observed were significantly different from the classes not observed on any of seven student variables used to characterize the field day participants. In addition, reliability (Cronbach's Alpha) was computed if there were more than two items in each basic category of observer's assessment tool and/or student's survey. On the observer's assessment tool, this included only the student's engagement items (α=.81). For the student instrument, the engagement items had an α of .56, and the satisfaction items had an α of .79. In all cases, all items contributed positively to the reliability – that is, when subjected to orthogonal rotation, the reliability α was always higher for the sum than had any item been deleted.

There were some serious limitations with using our data in this fashion. First, the observers and the student were not measuring strictly the same thing. Thus one would not expect a large agreement between students and observers. Observers were measuring the teaching efficacy of each learning station while students were measuring their total experience over the course of the day. In our research design, observers evaluated each learning station that the students from their class experienced. Depending on how many learning stations a class visited, one observer might complete five to seven individual learning station assessment tools. The observers' data were specific to each station visited, while the students' assessment tool was designed to evaluate overall field day experience. Each student completed only one student assessment at the end of the field day. The observers' data needed to be converted into overall means across all students and across the various leaning stations they visited before it could be correlated with the student data. In addition, there might be a recency effect at least on some items or constructs, in that students might focus on their latest experiences rather than equally on all of the experiences as is assumed when correlating the average observer scores with the student scores.

Second, the individual observers' field day assessment tool was designed in a three points scale (i.e. not done, partly done, and done), but students' Metro Children's Water Festival assessment tool was designed using a five points scale (i.e. strongly disagree, disagree, not sure, agree and strongly agree). In the process of analysis, a ceiling effect was found to influence the observers' data, but not student data. The 3 point scales used by the observers did not show sufficient variation, thus resulting in a ceiling effect with the observation data at each learning stations. This effect was mitigated some when averaging the observation scores across 5 observations.

Third, this study had only sixteen observers, which greatly reduced the power of the analyses.

Analysis

Observers' tool: The means were computed for each construct of all the stations that each observer visited, thus evaluating the average pedagogical experience for the class observed. If a construct had more than one item, the items were combined to obtain the means of the construct. The aggregated observers' overall station data were converted into means (summed station scores/# stations/# observers).

Student tool: The item mean from each construct of the student tool was computed. These item means and the overall station data from 16 observers using the individual station assessment tool (5-7 observations) were averaged for the class.

Finally, the observers' class scores were correlated (Table 2.) with the students' class scores. A second theoretical threat, the recency effect, was controlled. As students might have the most vivid memories from the last two stations, these stations' data were aggregated from each observer and compared to the student data overall.

Correlation: Assessment items from observer's assessment tool and student's survey

Pearson's correlation was used to compare the relationships among the items from the two assessment tools (individual observers' field day assessment tool and students' Metro Children's Water Festival survey).

Table 2.
Correlations among items

	All Day Learning Station Observation	Last Two Learning Station Observed
Opening	.118	.331**
Expressing1	-.115	.156
Expressing 2	.191	.364**
Questioning	-.097	-.011
Physical Environment	.562*	.134
Student's Engagement	.627*	.170
Student's Satisfaction	.422	.507*

N=16
* $p \leq .05$
** $p \leq 0.10$

The result showed some interesting phenomena. Even with a small N the assessment items in the basic categories of *physical environment* (r = .562, $p \leq .05$), and *student's engagement* (r = .627, $p \leq .05$) in the all day learning station observation were significantly correlated. Also, if considered that we had a very small sample size (n = 16), the student's satisfaction items from two assessment were also correlated (r = .422, $p <$ .10), with significance at the .1 level, which is acceptable for small population studies. On the other hand, in the last two learning station observations, the results showed that student's satisfaction items from the two assessment tools were correlated (r = .507, $p \leq .05$). Again, if we considered that we had a very small sample size (n = 16), the *opening* (r = .331) and *expressing 2* (r = .364) assessment items from observer's assessment tool and student's survey were also correlated.

Results show that 5 of the 7 measure correlated when using $p \leq .10$. Inaddition, a total of 20 out of 26 items that made up the 7 measures were correlated with strength between the observers' and students' feedback instrument. The two measures in pedagogy that did not correlate, Expressing 1 and Questioning, focued on students understanding the presenters questions and asking questions back to the presenter (Apppendix A.). Because of the limitations dicussed earlier of this instrumentation study, it is reasonable to say that the observation tool is validated when correlated with the student self report.

Discussion and Conclusion

People who organize and conduct field days are rarely researchers or evaluators. Indeed, they are often agency personnel with minimal social science or education background. Being able to measure a program against best practices provides field day organizers with an important opportunity to improve practice and to be accountable to participants. Further, the complexity of a field day itself, with multiple sessions, presenters, and sometimes routes for groups to take within a nature or park-like environment increases the value of having a tested instrument that can provide solid evaluative data across sessions, presenters, and the day.

Developing an observation tool for measuring program elements of a field day based on norm-referenced criteria ("best practices") creates a complex set of challenges. Best practices must be deconstructed and then critically considered in terms of what elements are observable and evaluative. These observations, however, must somehow be related to outcomes of those for whom the field day is offered, in this case, fifth grade students, or else the findings of the observational evaluation may be inherently flawed. In short, educational practices based on "best practices" need to be tied to the audience. To address this concern, the tested observation tool was used by trained observers and compared with, among other self-report measures, *student satisfaction*, considered a low-level

outcome measure that yet captures some of psychological conditions known to support science learing (National Research Council, 2007; 2009).

One finding in this study shed light on the researchers' concern about recency effects that may bias an observation tool towards an artifical objectivity across the entire field day experience. Observers note somewhat objectively over the course of the day the nature of the teaching-learning exchange and the environmental and social conditions under which this exchange occurs; yet, one might reasonably argue that students may put a stronger weight on experiences they have during the end of the field day, i.e., their overall assessment of the day might be biased by the last field day stations they visited. The concern of thepotential recency effect was that it might create a systematic bias in the observation tool or, conversely, in student feedback surveys, which would limit the usefulness of the observation to gauge student impact.

We found only a mild recency effect on two variables (*Opening* and *Expressing 2*), where the all-day observation data did not correlate with student self-report, while the observation data that were averaged across only the last two visited field day station visits showed a weak correlation. More importantly, the results support an interpretation that states just the opposite: the observation tool may more faithfully reflect student self-report when observation data are averaged across the entire day than when they are averaged only across the last two visted field day stations: *Student Engagement* correlated significantly and relatively strongly with all-day observation data (r = .627) while it did not correlate significantly with observation data averaged across the last two visited stations. Similar with the *Physical Environment*: The correlation between all-day observation data and student feedback data was relatively strong and signicant (r=0.56) while those between the observation of the last two visited field day stations and student feedback was not (r = .13). The results for *Student Satisfaction* seem to suggest the opposite, but the correlation coefficients are very close (r = .42 vs r = .51), and if anything, suggest that there is no significant recency effect even in a measure that might reasonably be seen as most sensitive to recency. Overall, these findings suggest that the observers were able across the day to measure the *environment, engagement* and even *satisfaction* in ways that are congruent with student experiences. Moreover, the results indicate that student self-report could be baised toward the novelty in the earlier part of the day, and fatigue toward the end of the day for at least some measures of the field day experience.

The results from this study suggest there is a positive correlation between the two tools for five of our seven measures and that this study validates the observation tool for those measures in terms of concurrent validity and in terms of being able to transfer claims from observation to student engagement. While not designed to do so, the results show that the observation tool can capture some of the felt experience of students. Further, these findings would support the belief that these are, indeed, best

practices that come from and are supported by the informal learning literature (Carlson, 2008, Heimlich, Carlson, Tanner & Storksdieck accepted).

With these findings, it is possible for field day coordinators to use the observational tool in combination with observer training as a valid resource for examining field days against observable best practices. Additionally, if these elements are satisfied, there is a positive relationship to student engagement and satisfaction with the overall field day experience (Wang & Carlson, 2011).

Recommendation for Further Studies

Although the observation tool was validated, data from this and other studies led the researchers to recommend that the observation tool be revised to a 5 point scale with different anchors to prevent ceiling effect and to better reflect the variance found in each construct. In addition, for more rigorous testing of the observation tool students should be tested after each learning station along with an overall evaluation of the day, using items that are closely aligned with the observed elements of the experience and with a cognitive outcome measure. This would allow us to directly compare apples to apples and would create an analysis with less noise in the data. Last but not least, it is recommended that the number of observations (observers and across stations) be increased to strengthen the power of the analyses in comparison studies. For the purpose of documenting field days with the observation tool, however, a limited number of observers may suffice.

Biographical statements

Stephan P. Carlson, Ph.D is a Professor/Extension Educator with the University of Minnesota Extension and the College of Food Agriculture and Natural Resource Sciences. He teaches classes in environmental education/interpretation and researches STEM learning in informal science settings such as parks, nature centers, zoos, arboretums, and museums. **Phone #** 612 624 8186, **E-mail:** carls009@umn.edu

Joe E. Heimlich, Ph.D is a Professor in Extension, the School of Environment and Natural Resources, and the Environmental Science Graduate Program at The Ohio State University. He is also a Director and Senior Research Associate with the Institute for Learning Innovation in Edgewater, Maryland. His research focuses on learning in informal environmental settings. **Phone** # 614.288.2674x2425,
E-mail: heimlich.1@osu.edu

Martin Storksdieck, Ph.D. is director of the Board on Science Education at the National Academy of Sciences/National Research Council where he oversees a wide range of studies related to science education (ranging from climate change education to new national science education standards). His specific research interests lie in free-choice and informal learning in the areas of science, environment and sustainability. **Phone** # 202 334 3987,
E-mail: mstorksdieck@nas.edu.

References

Barney, E., Mintzes, J., & Yen, C. (2005) Assessing knowledge, attitudes, and behavior toward charismatic megafauna: The case of dolphins. *The Journal of Environmental Education*, *36*(2), 41-55.

Brown, F. S., (1996). *The Effect fo an Inquiry-oriented Environmental Science Course On Preservice Elementary Teachers' Attitudes about Science.* Presentation to Annual Meeting of the National Association for Research in Science Teaching.

Carlson, S. P. (2008) Environmental field days: Recommendations for success. *Applied Environmental Education and Communication 4*, 94-105.

Carlson, S. P., Heimlich, J. E., Storksdieck, M. & Meyer, N. (2009). Best Practices for Field Days: Assessment Tool and Observation Protocol. U on MN Extension #08653.

Dorman, J. P. (2003). Cross-national validation of the what is happening in this class? (WIHIC) questionnaire using confirmatory factor analysis. *Learning Environments Research, 6*(3), 1573-1855.

DeWitt, J. & Storksdieck, M. (2008). A short review on school field trips: Key findings from the past and implications for the future. *Visitor Studies 11*(2): 181-197.

DiEnno, C., & Hilton, S. (2005) High school students' knowledge, attitudes, and levels of enjoyment of an environmental education unit on nonnative plants. *The Journal of Environmental Education, 37*(1), 13-23.

Farmer, J., Knapp, D., & Benton, G. (2007) An elementary school environmental education field trip: Long-term effects on ecological and environmental knowledge and attitude development. *The Journal of Environmental Education*, *38*(3), 33-42.

Fortner, R. W. (2001). The right tools for the job: How can aquatic resource education Succeed in the classroom? The Ohio State University.

Fraser, B. J. (1998). Classroom environment instruments: Development, validity and applications. *Learning Environment Research, 1*, 7-33.

Fraser, B. J. & Fisher, D.L. (1982). Predicting students' outcomes from their perceptionof classroom psychosocial environment. *American Education Research Journal, 19,468-518.*

Heimlich, J. E., Carlson, S. P. Tanner, D. & Storksdieck, M. (accepted 2010). Building face, construct and content validity through use of a modified Delphi: Adapting grounded theory to build an observation tool. *Environmental Education Research.*

Henderson, D. G., Fisher, D. L., & Fraser, B. J. (1998). Learning Environment, Student

Attitudes and Effects of Students' Sex and Other Science Study in Environmental Science Classes. Presentation to Annual Meeting of the American Educational research Association, April 13-17.

High Scope Educational Research Foundation (2006) Youth Program Quality Assessment. High Scope Press, Ypsilanti, MI.

Hintze, J. M. (2005). Psychometrics of Direct Observation. *School of Psychology Review, 34*(4), 507-519.

Knapp, D. H., & Benton, G. (2006) Episodic and semantic memories of a residential environmental education program, *Environmental Education Research*, *12,* 165-177.

Lawrenz, F. (1987) Gender effects for student perception of the classroom psychosocial environment. *Journal of Research in Science Teaching, 24*, 689-697.

McDonnell, J.D. (2001). Best practices in marine and coastal science education: Lessons learned from a National Estuarine Research Reserve. In: Fedler, A.J. (Ed.), *Defining Best Practices in Boating, Fishing, and Stewardship Education*, 173-182.

Meyer, N.J. & Pardello, R. (Eds.). (2005). Best practices for field days: A program planning guidebook for organizers, presenters, teachers and volunteers. St Paul: Environmental Science Education Work Group,University of MN Extension.

National Research Council (2007). *Taking science to school: Learning and teaching science in grades K-8*. Committee on Science Learning, Kindergarten Through Eighth Grade. R.A. Duschl, H.A. Schweingruber, and A.W. Shouse (Eds.). Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (2009). *Learning Science in Informal Environments: People, Places, and Pursuits*. Committee on Learning Science in Informal Environments. P. Bell, B. Lewenstein, A.W. Shouse, and M.A. Feder (Eds.). Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

North American Association for Environmental Education (NAAEE). (1996). Environmental education materials: Guidelines for excellence. Troy, OH: author.

Poudel, D.D., Vincent, L.M., Anzalone, C., Huner, J., Wollard, D., Clement, T., DeRamus, A. &Blakewood, G. (Summer, 2005). Hands-on activities and challenge tests in agricultural and environmental education. *The Journal of Environmental Education*. *36*(4), 10-22.

Siemer, W.F. (2001). Best practices for curriculum, teaching, and evaluation: Components of aquatic stewardship education. Curriculum, Teaching and Evaluation Components. Cornell University.

Stevens, M. & Andrews, E. (Eds.). (February, 2006). Outreach that makes a difference: Target audiences for water education – A research meta-analysis. A study conducted for the National Extension Water Outreach Project. University of Wisconsin.

Storksdieck, M., Heimlich, J.E., Figueriredo, C., Carlson, S. P. Environmental Field Days Assessment Tool: Reliability Study U of MN Extension, #523-04. (2009).

Storksdieck, M., Kaul, V. & Werner, M. (2006). Tackling it together: Local partnerships to improve field trip experiences. NARST 2006 Annual Meeting, San Francisco, CA, April 3-6, 2006.

Storksdieck, M. (2006). *Field trips in environmental education*. Berlin, Germany: Berliner Wissenschafts-Verlag.

Talton, E.L. & Simpson, R.D. (1987). Relationships of attitude toward classroom environment with attitudes toward and achievement in science among tenth grade biology students. *Journal of Research in Science Teaching, 24*, 507-525.

Wahyudi, & Treagust, D.F. (2004). The status of science classroom learning environments in Indonesian lower secondary schools. *Learning Environments Research, 7*(1), 1387-1579.

Wang, H. H. & Carlson, S. P. (2011). Factors that Influence Student's Learning in an Environmental Field Day*. International Electronic Journal of Environmental Education*. *1* (2), 129-139.

APPENDIX A

The framework of observer individual assessment tool and student survey (match up)

Basic Categories	Criteria of Measurement	Observer Individual Assessment Tool	Student Survey
Pedagogy (Opening)	The instructor sets up stage to attract students' attention to the learning program	2b. Introduced self- clearly	**2b. Presenters told us who they were**
Pedagogy (Expressing 1)	The instructor conveys age appropriate language when he/she delivers the program.	2h. Used appropriate language (clearly defining new terms when necessary) 2i. Presented content information appropriate for participants' knowledge and ability	**2c. Presenters asked us questions that I could understand even though I didn't know the answer**
Pedagogy (Expressing 2)	The instructor gives clear instruction when he/she delivers the program.	2j. Provided clear instructions 2c. Stated upcoming activities clearly	**2a. At the learning station, I knew what would happen**
Pedagogy (Questioning)	The instructor applies variety of questioning skills when he/she delivers the program	2m. Used questions that allowed participants to voice what they already knew or just learn (i.e. recall questions) 2n. Used questions that challenged participants to apply knowledge to new situations and/or made them think critically about an issue	**2d. I had a chance to ask my questions**
Management (Physical Environment 1)	The instructor conveys appropriate voice volume and adjust his or her position to be seen by students when he/she delivers the program	2l. Was seen and heard by all participants nearly all the time	**2h. I could hear and see the presenters at the stations**
Engagement (Student's Engagement)	The instructor and the program attract student's attention all the time	2g. Kept nearly all participants focused on activities most of the time 4a. Listened attentively when expected 4b. Participated fully when expected	**2m. I learned something new at the stations** **2o. I paid attention at the station** **2q. Kids in my class listened when they were supposed to** **2s. Kids in my class really got into the activities at the stations**
Satisfaction(Student's Satisfaction)	Student enjoy the instructor and the learning program during their field trip	4c. Showed excitement and enthusiasm	**2g. I enjoyed the presenters** **2t. Kids in my class had fun at the stations**

	experience		**2p. I found the stations interesting** **3d. I enjoyed being at the Water Festival** **3f. The presenters at the Water Festival were nice to me**