526-01 2009

Best Practices for Field Days

Validating an Informal Science Education Field Day Observation Tool

Authors: ***Stephan P. Carlson, Martin Storksdieck and Joe E. Heimlich***

A study was conducted at the Metro Children's Water Festival (CWF) in St Paul, Minnesota in the fall of 2008 where 44 schools and more than 1,200 fifth grade students participated in the one day event. The purpose of the study was to assess the validity of an observation tool for informal science education around Field Day programs. Content validity (Modified Delphi) and coder reliability of the observation tool was established the previous years (NSF, #0635559). 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; lack there of, on the other hand, does not indicate that the tool isn't valid.

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 learning stations during the day. Students stayed at each station about 30 minutes and then moved on to the next station. The stations that students visited 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 with a trained observer using the observation tool. Trained observers rated the quality of instructions at each of the learning stations. 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 if they did not want to participate. In addition, all classrooms were given copies of the student survey and 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%.

These two questionnaires had very different purposes and measured very different things. The observers' questionnaire measured the quality of the presentations (for the entire class) at each of 5-7 learning stations. The students evaluated their own experience once, at the end of the day. One would not expect a great deal of overlap among these two questionnaires. Nevertheless, one or more items on both of these questionnaires addressed the constructs in questions *(See Appendix A)*. It would help establish the utility and validity of these questionnaires if we could show some positive relationships between them on these six constructs.

University of Minnesota Extension

Result

A pedagogical framework was created that matched items on observer assessment tool and student survey questions on six constructs: opening, expressing, questioning, physical environment, student engagement, and student satisfaction *(Appendix A)*. The framework had 6 constructs it measured with a total of 12 items from the observer assessment tool and 14 items from the student assessment tool. For the purpose of analysis, we classified 26 items into six basic categories (constructs). Each of these constructs was measured with one or several items. Because of the purpose of the items in the category of expressing, 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.

A t-test was conducted between the sample group (n=16 classes) and the total population (n=44 classes), to see if the sample group differed from the population. None of the classes observed were significantly different from the classes that weren't observed on any of the 7 student variables. In addition, reliability (Cronbach's Alpha) was computed if there were more than two items in each basic category of the observer assessment tool and/or student survey. On the observer assessment tool, this included only student engagement items (alpha = .805). For the student survey, the engagement items had an alpha of .560, and the satisfaction items had an alpha of .789. In all cases, all items contributed positively to the reliability–that is, the reliability was always higher than had the item been deleted.

Limitations

There were some serious limitations with using our data in this fashion.

First, the observers and the students 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. 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 fill out five to seven individual learning station assessment tools. The observer data were specific to each station visited, while the student survey questions were designed to evaluate the overall field trip experience. Each student only filled out one survey at the end of the field trip. The observer data needed to be converted into overall means across all students before it could be correlated with the student data. There might also be a recency effect in that students focus on their latest experiences rather than equally on all of them as is assumed when correlating the average observer scores with the student scores.

Second, the individual observer's field day assessment tool was designed in a three point scale (i.e. not done, partly done, and done), but the student Metro Children's Water Festival survey was designed in a five point 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 observer data, but not the student data. The 3 point scale used by the observers did not show sufficient variation, thus resulting in a ceiling effect with the observation data at each learning station. This effect should be mitigated some when you average the observation scores across 5 observations.

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

Analysis

Observer survey

The means were computed for each construct (Table 1) of all the stations that each observer visited, thus evaluating the average pedagogical experience of that class. If a construct had more than one item, the items were combined to get the means of the construct. The aggregated overall station data was converted into means (basically, averaging the means of all the stations and breaking it down by number of observers).

Student survey

The item means from each construct of the student survey 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 observer classroom scores were correlated with the student classroom scores. A second theory was tested, recency effect. When students evaluated the overall field trip experience, they might have the most vivid memories from the last two stations. Therefore, the last two stations' data were aggregated from each observer and compared to the student data.

Correlation

Assessment items from observer assessment tool and student survey

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

TABLE 1	All Day Learning Station Observation	Last Two Learning Stations Observation
Opening (N=16)	0.118	0.331
Expressing 1	-0.115	0.156
Expressing 2	0.191	0.364
Questioning	-0.097	-0.011
Physical Environment 1	0.562*	0.134
Student Engagement	0.627*	0.17
Student Satisfaction	0.422	0.507*

N=16
** statistical significance, P≤ 0.05 [try 0.1]*

The result showed some interesting phenomena. The assessment items in the basic categories of physical environment (r =0.562, $P \leq 0.05$), and student engagement (r =0.627, $P \leq 0.05$) in the all-day learning station observation were significantly correlated. Also, considering that we had a very small sample size (n=16), the student's satisfaction items from two assessments were also correlated (r =0.422, p< .10), even though the result did not show statistical significance.

On the other hand, in the last two learning station observations, the result showed that student satisfaction items from the two assessment tools were correlated (r =0.507, $P \leq 0.05$). Again, if we considered that we had a very small sample size (n=16), the opening (r =0.331) and expressing 2 (r =0.364) assessment items from observer assessment tool and student survey questions were correlated, even though the result did not show statistical significance. Therefore, a total 20 out of 26 items were correlated between the observer and student assessment tool.

Discussion

These results suggest that there is a positive correlation between the two tools for five of our seven measures and that it validates the observation tool for those measures. This study demonstrates that a correlation was found with an independent student tool that was developed to answer many of the same constructs found in the observation tool. These constructs reflect the Best Practices for Field Days research that comes from and is supported in the informal science education literature.

Recommendation for Further Studies

Even though the observation tool was validated, it is recommended that the observation tool be revised to a 5 point scale with different anchors to prevent a ceiling effect and to better reflect the variance found in each construct. In addition, students should be tested after each learning station along with an overall evaluation of the day when comparing to the observation tool. 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 to increase the number of observers to increase the power of the analyses.

Stephan P. Carlson, Ph.D.
Professor, Extension Educator
Environmental Science Education
University of Minnesota Extension

Martin Storksdieck, Ph.D.
Institute for Learning Innovation
Edgewater, MD

Joe E. Heimlich, Ph.D.
Extension Faculty
The Ohio State University

www.extension.umn.edu/fielddays/

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 learned (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 adjusts 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 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 Satisfaction)	Students 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**