

NAAEE 2007 Research Symposium

Understanding Reliability when Measuring Environmental Knowledge in Elementary Students

Jon E. Berg, Jennifer Bergeron, Jennifer A. Seitz, Martha C. Monroe and Lindsey McConnell

School of Forest Resources and Conservation, University of Florida

ABSTRACT

Project Learning Tree (PLT) and other similar environmental education (EE) programs are becoming more difficult for teachers to use due to current educational reform efforts (Easton and Monroe 2002). These efforts, such as the No Child Left Behind (NCLB) Act of 2002, aim to raise current educational standards and increase teacher accountability with regard to student achievement in reading, writing and mathematics. If environmental education research can be designed to address achievement in these subjects, these results may be useful to educators. This study utilized experienced PLT elementary teachers to explore the benefit of using an EE activity to motivate interest in reading. The teachers who participated greatly appreciated the resources provided, but the results collected were unable to reliably measure student knowledge and intrinsic motivation. This poster summarizes recommendations for designing reliable assessments of elementary student knowledge.

INTRODUCTION

Supplemental environmental education (EE) programs are becoming less feasible for teachers to use due to educational reform efforts (Easton and Monroe 2002). If environmental

education research can be designed to measure reading, writing, and mathematics achievement, these results may help educators use EE materials. Such studies may measure the impact of interactive, engaging EE activities on student intrinsic motivation, interest, science knowledge, and reading comprehension.

In the fall of 2006, six elementary school teachers in Northwest and Central Florida conducted reading and writing lessons about science, with and without Project Learning Tree (PLT) activities as a motivator. It was hypothesized that students participating in a hands-on PLT activity would be more interested in the topic. When paired with a relevant story, this increased motivation should improve reading skills and comprehension. During this study a great deal was learned about reliable assessment measures and that aspect will be the focus of this project.

METHODS

Four PLT activities and relevant reading books were selected for 3rd and 4th grade students. Students were assigned by classroom groups to one of two treatments and all students received reading and writing instruction with all four lessons. One group used PLT activities prior to two lessons; the other group used PLT activities prior to the other two lessons. Student science knowledge was measured before and after the lessons; intrinsic motivation and reading comprehension were measured after the lessons.

Instrument Development

Science knowledge questions were created for the pre and post tests based on the information in each of the reading books as well as teacher recommendations. Since the science

questions were based on the reading books, improvement in science knowledge would be a function of reading comprehension.

Intrinsic motivation (IM) is a well documented, multidimensional concept related to Self-Regulation Theory (Deci and Ryan 1985). The IM instrument assessed participants' perception of the following six constructs: interest/enjoyment, perceived competence, effort, value/usefulness, felt pressure and tension, and perceived choice about an activity.

Recommended items from reliable scales, that were developed and tested, were used (McAuley, Duncan, and Tammen 1989; Plant and Ryan 1985).

Factors Affecting Reliability

In addition to creating instruments that are appropriate for young learners, consideration must be given to the reliability of the assessment instruments. Developing valid and reliable instruments that measure their respective targets is particularly difficult. Reliability, a measure of consistency, provides an indication of the amount of random error that is influencing the scores. Error affecting reliability can be attributed to differences in time, content, scores, memory and guessing (Linn and Miller 2004). Factors within instruments that affect reliability include the following: length (longer tests increase reliability because more items measure similar concepts); objectivity (increase impartiality to increase reliability); time limits (increase reliability by decreasing time between the pre and post tests); spread of scores (a wider spread leads to increased reliability); and test item difficulty (items of middle difficulty increase variability and reliability).

Teacher journals were used to record how teachers used the materials and what they perceived about student interest. To better understand how each lesson was conducted, two of

the six teachers were also interviewed. Information gathered from the teacher journals and interviews provided a basis from which to examine the data. One teacher felt many of her students knew the correct answers on the pre test. This suggests the questions were not difficult enough to increase variability and reliability. Both teachers said the intrinsic motivation (IM) instrument was long and difficult. One admitted her students didn't answer truthfully because they were tired of the test. Both teachers also felt the negative and positive wording on the IM instrument made the test difficult for young students.

RESULTS AND DISCUSSION

The first challenge to developing reliable instruments was the teachers' requests for short quizzes consisting of only five to ten questions. The aim while developing these was to provide easy and hard questions so students would be able to answer some correctly, but this further undermined reliability. Additionally, the classes within the study were not all performing at the same skill level. This made it difficult to design one knowledge instrument to cover all students.

An item analysis, often used to evaluate and revise instrument items, conducted after the study suggested which questions could be improved to raise reliability. Item analysis calculates mean scores, standard deviations, item correlations, and Cronbach's Alpha to analyze and determine an item's reliability and answers the following questions: did the item function as intended, were the items of appropriate difficulty, were the test items free of irrelevant cues and other defects, and were each of the distracters effective for multiple choice questions?

According to the mean scores of the pre and post test questions, knowledge improved for some lessons. Those questions that did not show an increase in correct answers may have been poorly written or were questions easy enough for students to answer correctly on the pre test.

The following question (Example 1) had a mean of .44 on the pre test and a mean of .37 on the post test. This indicates that more students answered this question correctly before receiving the lesson, rather than after, and creates low variability which leads to low reliability.

- Example 1: What do roots do?
- a.) absorb water from the soil
 - b.) conduct photosynthesis
 - c.) support trees from the wind
 - d.) both absorb nutrients & support trees

According to Cronbach's Alpha, reliability of each test was low due to the small number of items and poor item discrimination, as measured by item-total correlation. In certain instances, items which showed an increase in knowledge showed a decrease in item discrimination. The following question (Example 2) had a mean of .63 on the pre test and a mean of .85 on the post test. This indicates that more students answered this question correctly after the lesson was administered. Although there was an increase in the mean, the item discrimination decreased from .182 on the pre test to .129 on the post test. This may indicate that students who performed poorly on the test guessed the correct answer by chance.

- Example 2: What is the best description of a seed?
- a.) a seed is part of a plant that can grow into another plant
 - b.) a seed grows with sunshine
 - c.) a seed only moves by the wind
 - d.) a seed is a small pebble

The reliability of an instrument can sometimes be increased by removing particularly unreliable items. Cronbach's Alpha if-item-deleted helps identify which items contributed the most to the overall lack of reliability. The pre test that the following question (Example 3) was

taken from had a Cronbach's Alpha of .268, once this question was removed the Cronbach's Alpha for the pretest increased to .394. Questions with low reliability included those closely associated with the reading materials, since they were difficult to pre test, and those with answer choices that were too similar to one another.

Example 3: What does a tree need to make food?
a.) sunlight, water, air
b.) water, nutrients, space
c.) sunlight, water, pollen
d.) soil, water, air

Although an item analysis shows items that can be considered satisfactory from a technical standpoint, the final determination must be made from a logical standpoint (Linn and Miller 2004). Content specific items that contribute to the validity of an instrument should be included.

CONCLUSION

Challenges in the development of assessment instruments for this study included the following: writing knowledge items for young children that are clear, simple, and short; writing pre test knowledge items about books that students have not yet read; creating instruments short enough to be accurately completed by young children and long enough to still be reliable; creating items targeting mid-level difficulty for classrooms with mixed abilities; and using similar assessment instruments repeatedly.

Recommendations for future studies include: work with teachers and students to develop better knowledge questions that measure specific constructs covered by the reading material and activity; develop questions that target the same level of knowledge, rather than creating a scale

for science knowledge; increase the length of each of the assessment instruments by adding more items; and reduce the number of topics covered in the study, so that students complete the intrinsic motivation instrument only once.

REFERENCES

Deci, E. L., and Ryan, R. M. 1985. Intrinsic motivation and self-determination in human behavior. New York, NY: Plenum.

Easton, J. O., and Monroe, M.C. 2002. Project Learning Tree teacher assessment survey. Applied Environmental Education and Communication. 1 (4), 229-234.

Linn, R. L., and Miller, M.D. 2004. Measurement and assessment in teaching (9th ed.). Upper Saddle River, NJ: Merrill (Prentice Hall).

McAuley, E., Duncan, T., and Tammen, V. V. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. Research Quarterly for Exercise and Sport. 60, 48-58.

Plant, R. W., and Ryan, R. M. 1985. Intrinsic motivation and the effects of self-consciousness, self-awareness, and ego-involvement: An investigation of internally-controlling styles. Journal of Personality. 53, 435-449.