Jl. of Computers in Mathematics and Science Teaching (2020) **39**(4), 399-420

Researching Impact: Measuring Technology Enhanced Outcomes from the NASA Space Science Education Consortium

GERALD KNEZEK
Department of Learning Technologies
University of North Texas, USA
gknezek@gmail.com

RHONDA CHRISTENSEN
Institute for the Integration of Technology into Teaching and Learning
University of North Texas, USA
rhonda.christensen@gmail.com

Theoretical foundations for empirical research on the impact of NASA Space Science Education Consortium (NSSEC) technology innovation activities are presented in this paper, then aligned with the NSSEC evaluation framework and historical definitions of comparable psychometric constructs, in order to establish guidelines for research in this field. Measurement considerations and examples of three research designs – a) pre-post assessments, b) treatment versus comparison group studies, and c) retrospective pretest (reflecting on before versus after) – are featured as illustrations of those appropriate for the realm of space science education. Findings from four years of research on the impact of hands-on, technology-infused space science activities in informal learning settings indicate NSSEC technologies combined with rich engagement opportunities are capable of producing large content knowledge gains and an increase in interest in space science, as well as promoting the development of positive dispositions, in particular persistent positive attitudes, toward space science.

Keywords: space science education, empirical research, assessing impact

INTRODUCTION

Measures of cognitive and socio-emotional factors important to STEM interest, intent to pursue higher studies, and persistence toward choosing a STEM career have been refined for two decades (Christensen, Knezek, Tyler-Wood, & Gibson, 2014). These measures can be focused more narrowly on the field of space science. This article describes the process of choosing and adapting several of these measures to conduct research on outcomes of technology-infused learning activities derived through collaborations within the Science, Technology, Engineering, Arts and Mathematics (STEAM) Innovation Laboratory of the NASA Space Science Education Consortium (NSSEC). The University of North Texas (UNT) Institute for the Integration of Technology into Teaching and Learning (IITTL), as a partner of the NASA STEAM Innovation Laboratory, physically located at NASA's Goddard Space Flight Center in Greenbelt, Maryland: 1) selects innovative technologies developed in the STEAM Innovation Laboratory; 2) integrates technologies with NASA content to create activities for middle school aged participants; and 3) measures impact in order to widely disseminate findings, and feed results back to NSSEC to improve space science education best practices. The activities discussed in this paper have been conducted in informal learning settings such as camps. However, planned activities will test the technology-based curricula delivered by teachers to students in school-related environments. Theoretical foundations are presented in the initial portions of this paper, followed by the introduction of three empirical research designs that have been found to be effective in producing measurable evidence of positive impact on learning. Outcomes from four years of research, development, implementation, and assessment of activities, are presented in the latter sections of this paper.

THEORETICAL FOUNDATIONS: SUPPORTING LITERATURE

Theories of Learning

The activities that were used in the studies presented in this paper are based on learning strategies that follow active learning principles. These principles include the use of active, hands-on, relevant, authentic and collaborative activities. Active learning has been shown to improve long-term knowledge retention and deep understanding (Akinoglu & Tandogan, 2007; Bonwell & Eison, 1991; Christensen & Knezek, 2015; Gallagher, 1997).

When using the active learning approach, education becomes more personally meaningful and takes advantage of students' natural curiosity. This approach prepares students for the future by having students communicate, collaborate, and try new approaches in finding solutions to real-world problems.

Active learning principles are rooted in Dewey's "learning by doing and experiencing" principle (Dewey, 1938). Dewey advocated that a child's schoolwork should have meaning and be engaging as well as have connections to other disciplines and life experiences. In an active learning model, the learner takes more responsibility for his/her own learning under the guidance of a teacher. Characteristics that are included in active learning include:

- relevance to real-world applications
- authentic solving of real-world problems
- application of prior knowledge and/or experiences to solve new problems
- collaboration with others
- interdisciplinary integration of subject matters, and
- self-directed learning.

Within the "learning by doing" context, strategies promoting active learning were defined as instructional activities "involving students in doing things and thinking about what they are doing" (Bonwell & Eison, 1991).

Jonassen, Howland, Moore, and Marra (2003) defined meaningful learning as "occurring when students were actively engaged in making meaning" and learning was social, collaborative, intentional, authentic, and active. Jonassen et al. (2003) argued that these five interrelated, interactive, and interdependent attributes provide the most meaningful learning activities when used in combination. Collectively, published literature has established the importance of active, engaged learning in creating learning that is deep and meaningful.

Guiding principles of an active learning paradigm are also based on research findings by Griffin (1998) and others that the primary factors enabling learning to occur in informal settings are: a) having a purpose to learn, b) having a choice over learning, and c) having ownership of the learning process in a social context (Griffin, 1998).

The targeted audience for the NASA activities were middle school aged children. Middle school is an appropriate age to develop an interest in science that may potentially persist through secondary school, into college and beyond into a career. Providing authentic, active learning experiences contributes to the internalization of learning about science.

Importance of STEM

Choosing a STEM career is a dynamic process that requires both *proficiency* and *interest* in STEM content (Bouvier, 2011; Neathery, 1997). The likelihood of student engagement in learning a specific topic increases when students possess an awareness, positive attitude, and interest in the topic (Jolly, Campbell, & Perlman, 2004).

STEM engagement and identity have been identified in the literature as important for preserving continuity from interest in STEM, to higher studies in STEM disciplines, to matriculation in a STEM career (Aschbacher, Li, & Roth, 2010). *Engagement* is most closely related to activities that promote initial interest, while *Identity* evolves along with longer-term positive dispositions and produces a sense of belonging (Aschbacher, Ing, & Tsai, 2013).

Informal learning experiences have the potential to be transformative experiences for learners (Falk & Dierking, 2000; 2013) with impacts extending months and years post experiences (Anderson, Storksdieck, & Spock, 2007; Falk & Dierking, 1997). Positive impacts can include acquisition of new skills and content knowledge, increased awareness of, and improved attitudes toward STEM. Informal learning opportunities can be particularly meaningful to children from disadvantaged backgrounds because they most likely have fewer opportunities to participate in these types of activities at home (Hooper-Greenhill, Phillips, & Woodham, 2009). Aschbacher, Ing, and Tsai (2013) found that students who persisted in science, engineering or medical aspirations versus those who dropped out of the pipeline were distinguished by having the opportunity to experience compelling, authentic STEM experiences *outside* of school.

Measuring NASA STEAM Innovation Activities

Measuring learning in the 21st century is recognized as a complex, multidimensional issue involving more than one domain. Learning scientists often study human characteristics that do not belong exclusively to one of the traditional domains of psychology: cognitive, affective and psychomotor (Roschelle, Grover, & Kolodner, 2014). The cognitive domain (knowledge, skills, abilities) has traditionally been the focus of outcome measures in STEM disciplines. However in the 21st century, the affective domain has become widely recognized as important as well (Knezek & Christensen, 2018a).

Cognitive domain measures. The academic disciplines blended into the acronym STEM (science, technology, engineering, mathematics) are

foundational areas of content knowledge in space science. More specialized areas such as physics and astronomy are mainstream for the field, and these often culminate in higher specializations such as heliophysics, astrophysics, or astrobiology. For space science both declarative knowledge and procedural knowledge are important, with the latter relying on step-by-step sequences or algorithms for successful completion of a process (Berge & van Hezewijk, 1999). Generally multiple choice, matching or fill-in the blank questions are used to test for declarative knowledge, much like tests for course content in formal learning in schools. Procedural knowledge is often assessed based on successful completion of a sequence of activities required to achieve a goal, complete a mission or develop an artifact. One example of procedural knowledge in the activities discussed in this paper is the completion of the design and printing of a functional 3-D object, such as a pinhole camera to allow safe viewing of a solar eclipse. Successfully functioning procedures are often contributing learning outcomes for space science.

Socio-emotional measures. Human attributes that learning scientists study outside the realm of cognition have evolved to have a designation of their own, called non-cognitive variables. Most of these lie in the affective domain. Attitudes are one human attribute that has been studied extensively related to technology (Knezek & Christensen, 2008), but there are many others that are now recognized as important as well. "Noncognitive is used here to refer to variables relating to adjustment, motivation, and student perceptions, rather than the traditional verbal and quantitative (often called cognitive) areas" (Sedlacek, 2011, p. 191). Non-cognitive skills may include self-concept, leadership abilities, creativity, motivation, accurate self-appraisal, empathy, and persistence (Shechtman, DeBarger, Dornsife, Rosier, & Yarnall, 2013). Non-cognitive variables are becoming more valued in the 21st century because they can function in research designs as important intervening or mediating variables that " … stand between the independent and dependent variables, and […] mediate the effects of the independent variable on the dependent variable (Creswell, 2002, p. 50)."

MEASUREMENT CONSIDERATIONS RELEVANT TO SPACE SCIENCE EDUCATION

Why Measure?

Externally funded projects often require evaluations to be completed to assess: a) process including the number of participants served, number of activities conducted, and b) product evaluations that measure impact in one

or more psychological domains. These empirical data are used to determine how well the project is achieving its goals and objectives. Research designs are often overlaid on top of evaluation criteria in projects focused on education. For NSSEC, the Behavior, Attitude, Skills, Interest, Knowledge (BASIK) model is used to guide evaluation areas to be addressed (Davis, Scalice, Young, Mayo, & Davey, 2018). These categories of possible impact (Table 1) were assembled to develop an evaluation framework for STEM education (Friedman, 2008).

Table 1
Categories of Potential Impact for Informal STEM Education and Outreach Programs

Impact Category	Target	Description
Behavior (related to)	STEM concepts, processes, or careers	Measurable demonstration of assessment of, change in, or exercise of behavior related to a STEM topic.
Attitude (towards)	STEM-related topic or capabilities	Measurable demonstration of assessment of, change in, or exercise of attitude toward a particular scientific topic, concept, phenomena, theory, or careers central to the project or one's capabilities relative to these areas. Although similar to awareness/interest/engagement, attitudes refer to changes in relatively stable, more intractable constructs such as empathy, appreciation for the role of scientists in society or attitudes toward STEM.
Skills (based on)	STEM concepts, processes, or careers	Measurable demonstration of the development and/or reinforcement of skills, either entirely new ones or the reinforcement, even practice, of developing skills. These tend to be procedural aspects of knowing, as opposed to the more declarative aspects of knowledge impacts.
Engagement or Interest (in)	STEM concepts, processes, or careers	Measurable demonstration of assessment of, change in, or exercise of engagement/interest in a particular scientific topic, concept, phenomena, theory, or careers central to the project.
Awareness, Knowledge or Understanding (of)	STEM concepts, processes, or careers	Measureable demonstration of content knowledge gained from informal STEM education/ outreach research or practice activities.

Note: Adapted from: The Informal Education and Outreach Framework (Friedman, 2008, p. 11 & 21)

For research on impact presented in this paper, the included constructs are compatible with the BASIK model. Construct assessments are typically administered to participants as pre-post measures of what content knowledge was learned (cognitive) or what dispositions changed (non-cognitive) as a result of participating in the STEAM Innovation activities. The measures developed and demonstrated to be effective by the UNT research team for STEAM Innovation Laboratory impact studies during the first four years of the NASA Space Science Education Consortium have focused on content knowledge acquisition (K = knowledge), enthusiasm (I = interest), and persistent dispositions (A = attitudes). Since almost all assessed activities involve hands-on learning, the research also involves what the BASIK model lists as S = Skills. However these are a byproduct and not directly assessed, but rather indirectly assessed though outcomes such as completion of a product (design and production of a 3D object, for example).

Ways to Assess Learning

Several types of assessment have been used for measuring educational technology activities in the past. Many new types of assessment are emerging for 21st Century learners, especially as a result of the affordances provided by new technologies. These new forms of assessment include a wide range of methods that vary in expense, invasiveness, and difficulty. These include performance-based, observation, rubric, portfolio, self-assessment, and embedded assessments.

Formative and summative assessments are two traditional categories of assessment that are blurring with the availability of technology enhancements. Formative assessment traditionally has a goal of improving the learning and instruction while summative is in place to judge whether good outcomes have been achieved by use of society's space, time, and money resources. Hickey and Itow (2012) have pointed out that new opportunities exist to assess the abilities of new disruptive and/or innovative technologies.

Given that no one form of assessment works for every situation and every learner, a better overview of the depth and breadth of learning can be gained by using more than one type and then combining the information for a more complete evaluation of the learner.

Example Measures Used for Space Science Activities

Multiple choice and short answer questions for content. Well-constructed multiple choice questions have long been regarded as an assessment method that balances the need for efficiency in testing and scoring while still allowing educators to create questions that tap into higher order cognitive skills in students, such as analysis, synthesis and creativity, according to Bloom's Taxonomy of Educational Objectives in the Cognitive Domain (Anderson et al., 2001; Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). The majority of the content items used by the UNT research team to date have been of this type. Content knowledge assessment might include items such as:

1. An eclipse is defined as an astronomical event that occurs when one celestial object moves _____ another, partially or fully obscuring it from view.
   A. **in front of**
   B. next to
   C. in back of
   D. on top of

A short answer question often requires learners to produce knowledge, rather than simply choosing the most correct answer from among those presented. An example of this type of question might be: "Tell how you think the activity in which you participated today increased your knowledge of space science." Short answer questions can often assess richer learning than multiple choice, but short answer are also more difficult to score.

Semantic differentials to measure dispositions. Osgood, Suci and Tanenbaum (1957) are credited with the concept of using adjective pairs as anchors on a 7-point continuum of agreement from 1 to 7. For example, the item might be "To me, Space Science is: Boring __ __ __ __ __ __ __ Interesting." In this type of instrument, the respondent selects a choice closer to boring or interesting, depending on their perception of space science. Semantic differential instruments are time efficient and reliable (providing consistent answers). However, this type of assessment may require more instruction to understand how to complete ratings; subjects often mark one end of the spectrum or the other rather than in the middle space. This type of instrument is commonly used for assessment of socio-emotional variables, where there is no right or wrong answer. One example of a semantic differential used for several UNT studies related to space science is listed in Table 2.

Table 2

Semantic Differential Scale for Assessing Perceptions of Space Science

Instructions: Choose one circle between *each* adjective pair to indicate how you feel about the object.									
To me, space science is									
1.	Fascinating	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	Ordinary
2.	Appealing	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	Unappealing
3.	Exciting	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	Unexciting
4.	Means nothing	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	Means a lot
5.	Boring	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	Interesting

Likert scales for enthusiasm/interest. Likert Scales are among the most common types of items for gathering socio-emotional data, with typical rating choices varying from 1 = strongly disagree to 5 = strongly agree. Some example Likert items related to interest in space science and learning through innovative technologies are provided in Table 3.

Table 3

Likert Items Related to Interest in Innovative Technologies and Space Science

Instructions: Rate each statement on a scale of 1-5, 1=Strongly Disagree (SD), 5=Strongly Agree (SA)					
	SD	D	U	A	SA
	1	2	3	4	5
1. I want to learn more about the moon.	○	○	○	○	○
2. I want to learn more about Mars.	○	○	○	○	○
3. I want to learn more about the sun.	○	○	○	○	○
4. NASA's Parker Solar Probe mission to the sun will revolutionize our understanding of the sun.	○	○	○	○	○
5. Weather (space weather) that occurs in space can impact my life.	○	○	○	○	○
6. Innovative technologies make learning more engaging.	○	○	○	○	○
7. Innovative technologies help me learn.	○	○	○	○	○
8. I learn better when activities are hands-on.	○	○	○	○	○
9. Using technology to learn gives me more control over my learning.	○	○	○	○	○

With Likert-type items (as well as semantic differentials), individual items are often very useful for research because they are easy to read and understand, but measurement scales are more powerful for contributing to generalizable and replicable findings. For example, in Table 3, a researcher might conjecture that items 1, 2 and 3 have a common underlying core as interest in solar system objects. Likewise, items 6, 7 and 9 appear to be about learning with technology. Formal analysis techniques such as factor analysis enable the researcher to determine whether conjectures about scales were correct, after a sizeable set of data has been gathered using items such as those listed in Table 3. However, the techniques described in the following section allow researchers with smaller data sets such as a single middle school classroom, or participants from two space science summer camps, to confirm whether or not they are justified in combining similar-appearing items into a group to form a scale.

Demographic items. In addition to content and attitude or disposition instruments, it is also useful to gather demographic items like gender, ethnicity, grade level, and so forth. These demographic items can be used for additional analysis that can provide feedback to the implementation team on how to best meet the needs of different sub-populations.

Indicators of Validity and Reliability

Validity and reliability are two important concepts for instruments focused on assessment of psychometric attributes related to space science. Especially in socio-emotional areas, a collection of items that form a scale will generally provide more consistent and accurate assessment than one item alone. Validity is concerned with whether the questions being asked of participants are appropriate for (relevant to) what the researcher wants to determine. Reliability has to do with consistency of measurement, whether the same opinion or answer would be provided, with respect to a set of items, in the same situation at another time. Reliabilities for Likert and semantic differential scales are commonly estimated by calculating internal consistency reliabilities, where the result is reported as an index from 0 (very low) to 1 (very high). As an example, for the five semantic differential attitude/disposition items displayed in Table 2, Cronbach's Alpha for a set of recent middle school students was .90 (Knezek & Christensen, 2018b). For Likert-syle space science interest items (5-point, strongly disagree to strongly agree) such as those focused on the sun, moon and Mars and formed from questions such as "I want to learn more about eclipses." typically produce

internal consistency reliability indices in the range of .75 - .82 (Knezek & Christensen, 2018b). These semantic differential and Likert scale reliabilities could be judged as respectable to very good according to guidelines by DeVellis (1991), as listed in Table 4.

Table 4

DeVellis Guidelines for Interpreting Cronbach's Alpha Internal Consistency Reliability for a Psychometric Scale

Below .60	Unacceptable
Between .60 and .65	Undesirable
Between .65 and .70	Minimally acceptable
Between .70 and .80	Respectable
Between .80 and .90	Very good
Much above .90	Excellent (Consider shortening the scale)

(DeVellis, 1991, p. 85)

Note that calculating Cronbach's Alpha only produces estimates of the consistency of a scale of items if they were combined. It is up to the research team to actually produce a scale score for each person (by averaging responses across relevant items) before proceeding with analysis at the scale level.

Adaptation of Feedback Surveys into Research Instruments

Many education outreach and informal science organizations including several members of the NASA Space Science Education Consortium (NSSEC) distribute feedback surveys at the end of their activities, primarily for formative evaluation purposes. That is, the organizers wish to learn what the participants liked and disliked about the activity just completed, in order to make adjustments before the next group of participants goes through the activity. By making some modifications to a feedback survey and including a research design, it is often possible to expand a feedback survey into a research instrument to create outcome measures that are more robust.

One common modification of surveys is to add additional items similar to those already asked, with consistent rating choices for all, so that the reliability of a group of items as a scale can be assessed. This strengthens the measurement accuracy. Other modifications to the surveys involve altering

administration procedures that result in some basis for comparison of the scores or ratings of the target group after the space science activity has been completed. This can be accomplished by:

a) Having participants complete research instruments before the activity (pre) and after the activity (post); and/or
b) Having a similar group not participating in the activity serve as a comparison against which the anticipated gains in the targeted group can be compared; or
c) Having the instrument administered post test only, but include a retrospective component that asks participants to reflect on their status or state before the activity, and contrast that with ratings after the activity.

Examples of each of these approaches are described in the following section of this paper.

COMBINING METHODOLOGIES & RESEARCH DESIGNS: STUDIES RELATED TO TECHNOLOGY ENHANCED SPACE SCIENCE ACTIVITIES

In order to provide concrete examples of the types of impact studies that as of 2020 are in the design phases and/or underway for space science education activities, three completed research studies are summarized in this section. These education activities are grounded in technologies featured by the STEAM Innovation Laboratory at NASA's Goddard Space Flight Center. Each study illustrates a different type of research design, covering the range of pre-post intervention research, treatment vs. comparison group research, and retrospective pre-post research.

Pre-Post Intervention Research

Pre-post assessments have been used in weekend space science camps since 2017. As shown in Table 5, a four-hour Saturday space science camp conducted for sixth grade students in 2019 resulted in significantly ($p < .05$) more positive changes in dispositions toward space science at the time of the post test than prior to camp activities, at the time of the pretest (Christensen et al., 2019). This is based on the semantic differential instrument shown in Table 2. As shown in the last column of Table 5, the magnitude of the gain for this Saturday camp was effect size = .46, which would be considered moderate according to guidelines by Cohen (1988) and well beyond

the effect size = .3 criterion for the point at which the magnitude of the gain becomes educationally meaningful (Bialo & Sivin-Kachala, 1996). For this group of informal learners, long-term attitudes toward space science became more positive, when comparing before versus after the Saturday space science camp.

Table 5

Saturday Space Science Camp Pre-Post Changes in Space Science Dispositions Paired Sample t-tests (2019)

	Mean	N	Std. Dev.	Sig.	Effect Size
Space Science Dispositions Pre	6.41	24	.59		
Space Science Dispositions Post	6.67	24	.55	.035	.46

Pre-post research was also used to assess the impact on knowledge acquisition related to space science from a similar four-hour Saturday camp held two years prior, in 2017. The focus of this camp was solar eclipses, in preparation for the August 2017 Total Solar Eclipse traversing from Oregon to South Carolina across the USA (Christensen et al., 2017). As shown in Table 6, for the paired pre-post 6th grade students, the gains in content were significant (p = .002) and the magnitude of the gains (Cohen's d = 1.12) can be considered educationally meaningful (Bialo & Sivin-Kachala, 1996). An effect size of 1.12 represents a very large gain in knowledge about solar eclipses according to the guidelines provided by Cohen (1988) of small = .2, moderate = .5, and large = .8.

Table 6

Pre-Post ANOVA for Saturday-Camp Participant Eclipse Content Knowledge

	Mean	N	Std. Deviation	Sig.	Effect Size
Eclipse Score Pre	3.80	20	1.673		1.12
Eclipse Score Post	5.40	20	1.142	.002	

Treatment Versus Comparison Group Research Studies

Post test data were also gathered from a comparison group of students who attended the same school and in the same grade level as the participants who attended the 2017 Saturday space science camp (Christensen et al., 2017). As shown in Table 7, pre-post data from the 20 students who at-

tended the Saturday space camp were compared to the group that did not attend the camp. The comparison group included 176 students in the same grade level and school who completed surveys at the time of the post test only. Analysis of variance comparing the post test scores of the treatment group attending the space science camp to the comparison group, indicated the average post test score on content knowledge of solar eclipses for the treatment group was significantly (p = .0005) higher than the average score of the comparison group that did not attend the camp but also completed content knowledge assessments after the camp was completed. The magnitude of the difference between the scores for these groups was Cohen's d = 1.04, very large according to guidelines provided by Cohen (1988).

Table 7

Analysis of Variance for Post Test Content Knowledge Score by Attendance vs. Non-Attendance at a Space Science Camp

	N	Mean	Std. Dev.	Sig.	Effect Size (Cohen's d)
Did not Attend	176	3.77	1.518		
Attended	20	5.40	1.142		
Total	196	3.93	1.563	.0005	1.04

The combined findings from Table 6 and Table 7 are illustrated in Figure 1. As shown in Figure 1, the post test scores of the comparison group were very similar to the pre-test scores of the treatment group, instilling confidence that both groups would have had similar scores if they had both been tested at the pre-test time. The space science camp attendees, however, showed a significant ($p < .05$) increase in their scores, beginning with an average of 3.80 and increasing to an average of 5.40 questions answered correctly. The magnitude of this pre-post gain (effect size = 1.12, see Table 6) would be considered very large based on guidelines by Cohen (1988).

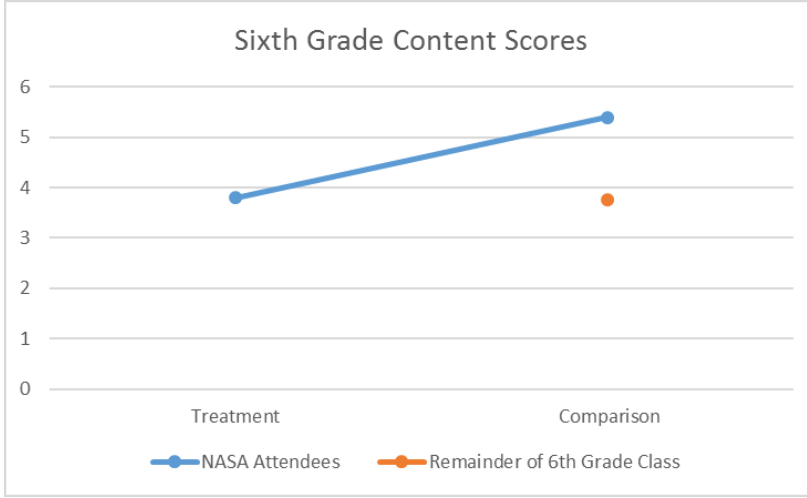

Figure 1. Pre-post data for weekend space science camp attendees vs. comparison group in 2017.

Retrospective Pre-Post Research Studies

Pilot testing of retrospective pre-post analysis (Chang & Little, 2018; Coulter, 2012) as a means of gathering impact data for space science camps for upper elementary and middle school students, began within the UNT research group in 2020. In this assessment method, participants are asked after an event to rate what their perceptions of key concepts or topics were before the event, and then rate their perceptions of the same concepts or topics, at the current time, after the event. Howard and colleagues (Howard, Ralph, Gulanick, Maxwell, & Gerber, 1979) were among the earliest to establish the retrospective pre-post research design as especially appropriate in situations where individuals do not have sufficient information to accurately judge their initial perceptions of a concept or topic of focus, until they participate in the targeted project or event. The UNT research group had observed this limitation to be true for middle school students with regard to space science.

Initial findings from retrospective pretest analysis of socio-emotional values related to space science are encouraging. In February 2020, 21 students attending the 6th grade Saturday space science camp were asked to reply to four retrospective items listed in Table 8. The four items together

were found to form a reliable scale of space science interest with Cronbach's Alpha = .81, which was judged to be very good according to guidelines by DeVellis (1991) as listed in Table 4. The change in space science interest from the retrospective pretest for this group (mean = 3.80, Std. = 1.05) to the post test for this group (mean = 4.08, Std. = 1.01) was found to be significant (t = - 2.96, 20 df, $p < .008$), indicating positive change in space science interest for the group, with a magnitude (ES = .27) approaching the ES = .3 criterion for the point at which an intervention is typically considered educationally meaningful (Bialo & Sivin-Kachala, 1996). As shown in Table 9 and graphically displayed in Figure 2, ratings of space science interest after the event were higher than ratings for prior to the event, for all four items. Item 4, "I am interested in a career in space science," was individually significant ($p = .008$) with an increase in magnitude of ES = .28. Future research is planned to reconfirm these initial findings, and to cross-validate retrospective assessment with pre-post treatment versus comparison group studies.

Table 8

Retrospective Pre-Post Assessment Items for 2020

Thinking of what you thought **BEFORE** you attended this camp, give your impressions that you can recall.					
	SD	D	U	A	SA
	1	2	3	4	5
Pre 1. I am interested in space science.	○	○	○	○	○
Pre 2. I would like to learn more about being a scientist	○	○	○	○	○
Pre 3. I would like to learn more about the moon and space.	○	○	○	○	○
Pre 4. I am interested in a career in space science.	○	○	○	○	○
AFTER participating in the camp,					
Post 1. I am interested in space science.	○	○	○	○	○
Post 2. I would like to learn more about being a scientist.	○	○	○	○	○
Post 3. I would like to learn more about the moon and space.	○	○	○	○	○
Post 4. I am interested in a career in space science.	○	○	○	○	○

Note: 1=Strongly Disagree (SD), 2 = Agree, 3 = Undecided, 4 = Agree, 5=Strongly Agree (SA).

Table 9

T-Test for Retrospective Pre-Post Item Ratings by 6th Graders Attending 2020 Weekend Space Science Camp (Paired Samples Statistics)

Item		Mean	N	SD	Sig.	Effect Size
Pair 1	SS interest1 Before vs.	4.20	20	1.11	.056	.22
	SS interest1 After	4.45	20	1.15		
Pair 2	SS interest2 Before vs.	3.15	20	1.39	.083	.22
	SS interest2 After	3.45	20	1.36		
Pair 3	SS Moon3 Before vs.	4.25	20	1.16	.204	.23
	SS Moon3 After	4.50	20	1.00		
Pair 4	SS Career4 Before vs.	3.35	20	1.50	.008	.28
	SS Career4 After	3.75	20			

Figure 2. Retrospective pre-post ratings for 6th graders attending weekend space science camp in 2020.

DISCUSSION

Multiple measurement strategies and examples from studies were included in this paper to illustrate the breath of alternative designs that are practical and yet sufficiently powerful to produce tangible, quantifiable evidence of positive impact in the realm of space science education. Each method has weaknesses and all have limitations. Some methods have characteristics that are viewed as strengths by some researchers and weaknesses by others. For example, while some research methodologists express concerns about the potential lack of independence of retrospective pretest ratings produced in the same time frame as post test ratings, or the possible inability of young children to remember how they felt before an intervention from the perspective of after the activity (Chang & Little, 2018), other methodologists express concerns about traditional pre-post designs and argue that retrospective pre-post design may reduce response bias (Nimon, Zigarmi, & Allen, 2011). Campbell and Stanley (1963) long ago observed that when conducting research related to what we now call socio-emotional attributes, if bias is created by using a retrospective pre-post design then the bias should be toward more conservative findings rather than inflated results with respect to positive attitude gains. In the words of Campbell and Stanley:

> The reader should be careful to note that the probable direction of memory bias is to distort the past attitudes into agreement with present ones, or into agreement with

> what the tenant has come to believe to be socially desirable attitudes. Thus memory bias seems more likely to disguise rather than masquerade as a significant effect of X [the treatment] in these instances (Campbell & Stanley, 1963, p. 66).

These observations by recognized authorities in the field of social science research, combined with the initial promising results in the realm of attitudes toward and interest in space science reported in this study, provide confidence in the continued pursuit of retrospective pre-post analysis as a promising research method for assessment of the impact of technology-infused, engaged learning activities on socio-emotional attributes related to space science.

There are several limitations to the impact studies reported in this paper. Among these are small sample sizes ($n < 32$) that typically increase the difficulty of concluding that changes in interest, attitudes, or knowledge regarding space science are rare by chance ($p < .05$). Plans are in place for a five-year continuation of the NASA Space Science Education Consortium (NSSEC) to include studies with larger treatment group samples.

Future research might also include additional aspects of the five-component BASIK model presented in Table 1, since not all are being assessed in the research designs presented. However, each presents certain difficulties. Specifically, Behavior (measurable demonstration of assessment of, change in, or exercise of behavior related to a STEM topic) as defined in Table 1, and as operationalized in enrollment in college courses leading to careers in space science or other STEM fields, is beyond the scope of assessment for the time frame spanning middle school. Association with collaborating partners who have such longitudinal tracking capabilities is being considered for the future. Also the category of Skills (measurable demonstration of the development and/or reinforcement of skills, either entirely new ones or the reinforcement, even practice, of developing skills) has not been the focus of impact research by the UNT research team during the first four years of the NASA Space Science Education Consortium (NSSEC). Skills are best confirmed through the production of an artifact or completion of a project more comprehensive than is practical in one day or one week, which has been the limitation of time span for findings reported in this paper. Studies of activities continuing over a longer time frame would make verification of these types of skills practical.

Nevertheless, even with the many limitations acknowledged for the situated context as well as the impact designs and findings reported in this paper, evidence is beginning to emerge that measurable, desirable, impacts

result from the infusion of NASA innovative technologies into active learning contexts, to produce meaningful positive shifts in young learners in the areas of knowledge related to space science, interest in learning more about space science, and in the development of long-term, persisting attitudes (dispositions) toward space science.

CONCLUSION

A conceptual framework for empirical study research regarding the impact of NASA Space Science Education Consortium (NSSEC) STEAM Innovation Laboratory activities has been introduced in this paper. Findings from fours years of research on the impact of hands-on, technology infused space science activities have provided evidence for large content knowledge gains and an increase in interest in space science, as well as the development of positive dispositions (persistent positive attitudes) toward space science as an area of further study and future careers. Studies featured in this paper have provided evidence that: a) pre-post assessments, b) treatment versus comparison group studies, and c) retrospective (before versus after) analyses upon completion of an engaging event, can all yield useful empirical data for quantitative analysis of impact. Nevertheless, regarding measureable impact of STEAM Innovation Laboratory supported activities of the NASA Space Science Education Consortium, rigorous studies have only begun. These types of research activities are anticipated to increase in breath and depth over the coming years.

Acknowledgment

This research and development was supported in part by the U.S. National Aeronautics and Space Administration (NASA) Grant # NNX16AL63A.

References

Akinoglu, O., & Tandogan, O. R. (2007). The effects of problem-based active learning in science education on students' academic achievement, attitude and concept learning. *Eurasia Journal of Mathematics, Science & Technology Education, 3*, 71-81.

Anderson, D., Storksdieck, M., & Spock, M., (2007). Understanding the long term impacts of museum experiences. In J. Falk, L. Dierking, & S. Foutz (Eds.), *In principle, In practice*. (pp. 197-215).

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Pearson, Allyn & Bacon.

Aschbacher, P.R., Ing, M., & Tsai, S.M. (2013). Boosting student interest in science. *Kappan Magazine, 95*(2), 47-51.

Ashbacher, P.R., Li, E., & Roth, E.J. (2010). Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. *Journal of Research in Science Teaching, 47*(5), 564-582.

Berge, T. & van Hezewijk, R. (1999). Procedural and declarative knowledge: an evolutionary perspective. *Theory and Psychology. 9*, 605-624. 10.1177/0959354399095002.

Bialo, E.R. & Sivin-Kachala, J. (1996). The effectiveness of technology in schools: A summary of recent research. *School Library Media Quarterly, 25* (1), 51-57.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain.* New York: David McKay Co Inc.

Bonwell, C., & Eison, J. (1991). *Active learning: Creating excitement in the classroom.* AEHE-ERIC Higher Education Report No. 1. Washington, D.C.: Jossey-Bass.

Bouvier, S. (2011). *Increasing student interest in science, technology, engineering, and math (STEM): Massachusetts STEM pipeline fund programs using promising practices.* University of Massachusetts Donahue Institute: Hadley, MA.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi- experimental designs for research*. Boston: Houghton Mifflin.

Chang, R., & Little, T.D. (2018). Innovations for evaluation research: Multiform protocols, visual analog scaling, and the retrospective pretest–posttest design. *Evaluation & the Health Professions*, *41*(2), 246–269. https://doi.org/10.1177/0163278718759396

Christensen, R., & Knezek, G**.** (2015). Active learning approaches to integrating technology into middle school science classrooms: Reconceptualizing a middle school science curriculum based on 21st century skills. In X.Ge, D. Ifenthaler, & J.M. Spector (Eds.). *Full Steam Ahead: Emerging Technologies for STEAM*. New York: Springer Academic.

Christensen, R., Knezek, G., Darby, D., Den Lepcha, S., Jiang, B., Kuo, A. & Wu, A. (2017). Outcomes from Technology Enhanced Informal Learning Activities: Total Solar Eclipse. In J. Johnston (Ed*.), Proceedings of EdMedia 2017* (pp. 1064-1071). Washington, DC: Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/p/178499/.

Christensen, R., Knezek, G., Hobbs, F., Kelley, J., Den Lepcha, S., Dong, D., Liu, S., Wang, K., Yotchoum Nzia, H.A. & Kelley, D. (2019). Creating technology enriched activities to enhance middle school students' interest in STEM. In J. Theo Bastiaens (Ed.), *Proceedings of EdMedia + Innovate Learning* (pp. 1326-1334). Amsterdam, Netherlands: Association for the Advancement of Computing in Education (AACE).

Christensen, R., Knezek, G., Tyler-Wood, T., & Gibson, D. (2014). Longitudinal analysis of cognitive constructs fostered by STEM activities in middle school students. *Knowledge Management and ELearning, 6*(2), 103-122.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Coulter, S.E. (2012). Using the retrospective pretest to get usable, indirect evidence of student learning. *Assessment & Evaluation in Higher Education, 37*(3), 321-334, DOI: 10.1080/02602938.2010.534761.

Creswell, J.W. (2002). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Los Angeles, CA: Sage.

Davis, H.D., Scalice, D., Young, A., Mayo, L., & Davey, B. (2018). *Measuring effects across NASA space science education consortium activities using NSF impact categories*. DC: American Geophysical Union.

DeVellis, R.F. (1991). *Scale development*. Newbury Park, NJ: Sage Publications.

Dewey, J. (1938). *Experience and education. A touchstone book*. Kappa Delta Pi, New York.

Falk, J.H., & Dierking, L.D. (1997). School field trips: assessing their long-term impact. *Curator 40*(3), 211-218.

Falk, J.H., & Dierking, L.D. (2000). Learning from museums: Visitor experiences and the making of meaning. *Altamira Press.*

Falk, J.H., & Dierking, L.D. (2013). The museum experience revisited. *Left Coast Press.*

Friedman, A. (Ed.). (2008). *Framework for evaluating impacts of informal science education projects.* http://insci.org/resources/Eval_Framework.pdf.

Gallagher, S. (1997). Problem-based learning: Where did it come from, what does it do and where is it going? *Journal for Education of the Gifted, 29*(4), 332-362.

Griffin, J. (1998). *School-museum integrated learning experiences in science: A learning journey*. Unpublished PhD thesis, University of Technology, Sydney.

Hickey, D.T., & Itow, R.C. (2012). *Re-mediating assessment.* Retrieved from http://remediatingassessment.blogspot.fr/2012/03/some-things-about-assessment-that-badge.html

Hooper-Greenhill, E., Phillips, M., & Woodham, A. (2009). Museums, schools and geographies of cultural value. *Cultural Trends, 18*(2), 149 – 183.

Howard, G.S, Ralph, K.M., Gulanick, N.A., Maxwell, S.E., & Gerber, S.K. (1979). Internal validity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*, 1-23.

Jolly, E., Campbell, P., & Perlman, L. (2004). *Engagement, capacity, and continuity: A trilogy for student success.* Minneapolis, MN: GE Foundation.

Jonassen, D.H., Howland, J.L., Moore, J.L., & Marra, R.M. (2003). *Learning to solve problems with technology: A constructivist perspective.* Merrill Prentice Hall, Upper Saddle River, New Jersey.

Knezek, G., & Christensen, R. (2008). The importance of information technology attitudes and competencies in primary and secondary education. In J. Voogt and G. Knezek (Eds.). *International Handbook of Information Technology in Primary and Secondary Information*. (pp. 321-331).

Knezek, G., & Christensen, R. (2018a). The evolving role of attitudes and competencies in information and communication technology in education. In J. Voogt, G. Knezek, R. Christensen, & K-W. Lai. (Eds.) *Springer: Second Handbook of Information Technology in Primary and Secondary Education*, (pp. 239-253).

Knezek, G., & Christensen, R. (2018b). Capitalizing on informal learning opportunities to enhance formal learning in classroom environments. In E. Langran & J. Borup (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 1920-1926). Washington, D.C., United States: Association for the Advancement of Computing in Education (AACE). Retrieved March 19, 2020 from https://www.learntechlib.org/primary/p/182791/.

Neathery, M.F. (1997). Elementary and secondary students' perceptions toward science and the correlation with gender, ethnicity, ability, grade, and science achievement. *Electronic Journal of Science Education* 2(1).

Nimon, K., Zigarmi, D., Allen, J. (2011). Measures of program effectiveness based on retrospective pretest data: Are all created equal? *American Journal of Evaluation 32(*1), 8-28.

Osgood, C.E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Roschelle, J., Grover, S., & Kolodner, J. (2014). CIRCL primer: Learning sciences. In *CIRCL Primer Series*. Retrieved from http://circlcenter.org/learning-sciences/

Sedlacek, W.E. (2011). Using noncognitive variables in assessing readiness for higher education. *Readings on Equal Education, 25*, 187-205.

Shechtman, N., DeBarger, A.H., Dornsife, C., Rosier, S., & Yarnall, L. (2013). *Promoting grit, tenacity, and perseverance: Critical factors for success in the 21st century.* Washington, D.C.: U.S. Department of Education Office of Educational Technology.