Identifying and developing crosscutting environmental education outcomes for adolescents in the twenty-first century (EE21)

Robert B. Powell, Marc J. Stern, Brandon Troy Frensley & DeWayne Moore

Published online: 20 May 2019.

Submit your article to this journal

Article views: 20

View Crossmark data

Routledge
Taylor & Francis Group

Identifying and developing crosscutting environmental education outcomes for adolescents in the twenty-first century (EE21)

Robert B. Powell[a] (iD), Marc J. Stern[b] (iD), Brandon Troy Frensley[c] and DeWayne Moore[d]

[a]Department of Parks, Recreation and Tourism Management, Department of Forestry and Environmental Conservation, Clemson University, Clemson, SC, USA; [b]Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA; [c]Department of Environmental Sciences, University of North Carolina Wilmington, Wilmington, NC, USA; [d]Department of Psychology, Clemson University, Clemson, SC, USA

ABSTRACT

While multiple valid measures exist for assessing outcomes of environmental education (EE) programs, the field lacks a comprehensive and logistically feasible common instrument that can apply across diverse programs. We describe a participatory effort for identifying and developing crosscutting outcomes for Environmental Education in the twenty-first Century (EE21). Following extensive input and debate from a wide range of EE providers and researchers, we developed, tested and statistically validated crosscutting scales for measuring consensus-based outcomes for individual participants in youth EE programs using confirmatory factor analysis across six unique sites, including two single-day field trip locations, four multiday residential programs and one science museum in the United States. The results suggest that the scales are valid and reliable for measuring outcomes that many EE programs in the United States can aspire to influence in adolescent participants, ages 10–14.

Introduction

Recent reviews of environmental education (EE) programs for youth suggest that they can achieve a wide variety of positive outcomes for participants, including: increased knowledge; more positive attitudes and behavioral intentions toward the environment; enhanced self-confidence and social interactions; and improved academic motivation and performance, among others (Ardoin et al. 2018; Stern, Powell, and Hill 2014; Thomas et al. 2018). Some programs are intended primarily to supplement formal classroom learning in the pursuit of achieving specific knowledge to meet district, state and national curriculum standards. Other programs may focus on building an emotional connection with a site or influencing the attitudes, dispositions or behaviors of participants to become active environmental stewards. Still others might be designed to enhance students' twenty-first century skills, interactions with each other or their teachers or self-confidence. In this article, we ask a rather provocative question: is there a

consistent set of outcomes to which all EE programs for youth could reasonably aspire? And if so, how would we measure those outcomes?

Of course, the first question is largely one of opinion[1], perspective and consensus-building. Recognizing that not all EE programs for youth focus on the same topic, we consider the exercise one of identifying the range of outcomes that EE programs, if done exceptionally well, may aspire to achieve. Well-designed EE programs for youth have the potential to achieve a wide array of desirable outcomes, including not only learning about the environment, human-ecosystem connections, science and other subject matter, but also enhancing environmental and social connection, skills, self-efficacy, motivation, curiosity and inspiration. If we broaden our view beyond the specific factual subject matter of any particular program, we can begin to see the wider potential of what EE programs are actually capable of achieving, as demonstrated through dozens of empirical studies (Ardoin, Biedenweg, and O'Connor 2015; Ardoin et al. 2018; Stern, Powell, and Hill 2014). This wide range of potential programmatic outcomes presents a particular challenge for the field. Because most empirical studies have focused on evaluating single programs with disparate outcomes measurements, current knowledge on the best practices in environmental education is largely anecdotal and based on consensus of the field rather than systematic empirical evidence (National Parks Second Century Commission 2009; NSF 2008; Stern, Powell, and Hill 2014). To identify what works best, a large-scale comparative study is necessary, which requires a psychometrically valid and reliable set of common/shared crosscutting outcomes that are relevant to a wide range of programming, sensitive enough to vary based on program quality, and short enough that students can complete it in a reasonable amount of time (Grack Nelson et al. 2019). Our effort here describes a first step in this direction.

In this article, we describe the development and validation of scales to measure crosscutting outcomes relevant for EE programs for adolescent youth in the twenty-first century. We focus on early adolescents not only because a large proportion of such programs are geared to this age but also because research suggests this developmental period is critical for developing identity, 'twenty-first century skills', environmental literacy and meaningful connections with place and community (Kahn and Kellert 2002; Kohlberg 1971; Kroger 2006; Piaget 1964). We begin with a summary of existing perspectives on appropriate outcomes for early adolescent participants in EE programs. Next, we summarize our extensive participatory efforts to identify, define and develop consensus around a list of shared, crosscutting, aspirational outcomes that are applicable to a range of EE programs for youth ages 10–14. Then, following procedures for scale development outlined by DeVellis (2003) and Presser et al. (2004), we identified and defined outcomes and subsequently developed and collaboratively refined survey items to measure those outcomes. We administered the resulting questionnaire at six different EE programs from across the United States representing a range of program types (day field trips, multiday residential and informal settings) and contexts to examine their practical utility (could students understand the items and complete the survey in a reasonable time period?), construct validity and reliability and psychometric properties. We employed confirmatory factor analysis techniques, including multigroup configural, metric and structural invariance testing (Vandenberg and Lance 2000), to confirm and crossvalidate the hypothesized factor structure and measurement properties. The results suggest the achievement of sensitive, reliable and psychometrically valid scales to measure consensus-based aspirational outcomes for EE for adolescent youth across different contexts in the United States.

Existing perspectives on EE program outcomes

In addition to reviewing guidelines, websites and synthetic documents of key organizations in the field (e.g. Center for Advancement of Informal Science Education, Institute for Museum and Library Services, North American Association for Environmental Education), we examined two

systematic literature reviews that identified the primary outcomes that researchers have measured in peer-reviewed EE research and evaluations over the past fifteen years (Ardoin, Biedenweg, and O'Connor 2015; Stern, Powell, and Hill 2014). While the specific outcomes of individual EE programs vary from program to program, our review of these documents and the broader EE literature revealed four key overarching themes: environmental literacy, positive youth development, the achievement of educational standards, and what many organizations in the United States are calling 'twenty-first century skills'. We summarize each of these perspectives below.

Environmental literacy

Most point to the language of the Tblisi Declaration, which resulted from the world's first intergovernmental conference on EE organized by the United Nations Education, Scientific and Cultural Organization (UNESCO) in 1977, to summarize the general consensus outcomes of EE. These outcomes include (UNESCO 1977, 3):

Awareness—to help social groups and individuals acquire an awareness and sensitivity to the total environment and its allied problems.

Knowledge—to help social groups and individuals gain a variety of experiences in, and acquire a basic understanding of, the environment and its associated problems.

Attitudes—to help social groups and individuals acquire a set of values and feelings of concern for the environment and the motivation for actively participating in environmental improvement and protection.

Skills—to help social groups and individuals acquire the skills for identifying and solving environmental problems.

Participation—to provide social groups and individuals with an opportunity to be actively involved at all levels in working toward resolution of environmental problems.

Today, these same themes are encompassed within the concept of *environmental literacy* and are common across multiple studies and summaries of EE outcomes. Environmental literacy is comprised of the knowledge, attitudes, dispositions and competencies believed necessary for people to effectively analyze and address important environmental problems (Hollweg et al. 2011; Stern, Powell, and Hill 2014).

Positive youth development

Many youth EE programs (e.g. Carr 2004; Delia and Krasny 2018; Stern, Powell, and Ardoin 2010) now focus on elements of positive youth development (PYD), which describes the development of assets essential to human well-being. Recent research has identified that positive character development, which includes emotional intelligence, resiliency, positive self-image or identity, a sense of caring and compassion for others, a sense of right and wrong, self-empowerment, confidence and competence, is important for fostering youth that will excel academically and later in life (e.g. Bowers et al. 2010; Lerner et al. 2005; Seligman et al. 2009). Scholars also commonly consider self-efficacy, prosocial norms and meaningful relationships with peers and adults as components of PYD (Catalano et al. 2004; Delia and Krasny 2018). Eccles and Gootman (2002) classify these factors into four categories of personal well-being associated with PYD: physical (e.g. healthy habits); intellectual (e.g. critical thinking); psychological (e.g. positive self-regard) and social (e.g. connections with others, civic engagement).

Academic achievement

In the United States, the *No Child Left Behind Act* (2001) and the subsequent *Every Student Succeeds Act* of 2015 require annual standardized testing for grades 3–8 in all publicly supported schools to demonstrate that students are advancing and achieving educational standards. Many EE programs for youth align with state and/or national education standards to assist students in improving academic performance. Standards that are particularly relevant for EE, irrespective of student grade level, focus on *understanding ecological processes*, the *interdependence of organisms*, the *interconnectivity of social and ecological systems*, *how humans may impact the environment* and *how changes in the environment influence ecosystem function* and *human systems* (e.g. Next Generation Science Standards (National Research Council, 2013)). EE can address multiple other standards as well, including those associated with math, history, social studies, economics or others. Moreover, EE has also been shown to influence academic motivation, which contributes meaningfully to multiple forms of achievement (Broussard and Garrison 2004; Stern, Powell, and Ardoin 2010).

Twenty-first century skills

Organizations in the United States, such as the National Park Service, the Institute for Museum and Library Services and the Smithsonian Institute, suggest that informal learning sites such as museums, zoos, aquaria, nature centers and parks, with their nationally and globally significant cultural, environmental and historical resources, provide an opportunity for educational programs to further facilitate the development of 'skills that are critical for addressing twenty-first century challenges', such as climate change, poverty and effective governance (Fenichel and Schweingruber 2010; Institute of Museum and Library Services 2009; National Parks Second Century Commission 2009; National Park Service 2014; National Park System Advisory Board Education Committee 2014; NSF 2008; Smithsonian Institute 2010). Coined 'twenty-first century skills' (e.g. Institute of Museum and Library Services 2009; National Park Service 2014), these include a broad range of knowledge, dispositions, skills and behaviors pertaining not only to the environment, but also to science, culture, health, history and civics. Skills that cut across 'literacies' in each of these topic areas include critical thinking, problem solving, communication, collaboration and social skills, among others (Institute of Museum and Library Services 2009).

Methods

Identifying and defining crosscutting outcomes for EE

With the four broad categories described as a starting point, we began a systematic effort to directly involve EE experts and practitioners in further identifying and defining crosscutting outcomes for EE programs for youth (ages 10–14). First, we coordinated a workshop with the NPS National Advisory Board Education Committee and the executive directors of the Association of Nature Center Administrators (ANCA) and the North American Association for Environmental Education (NAAEE) in December, 2016. The Committee included 20 subject matter experts (SMEs), including academics, practitioners and evaluators and leaders of a wide array of non-profit, government and educational organizations. Through a collaborative process following procedures outlined by Fenichel and Schweingruber (2010) and Powell, Stern, and Ardoin (2006), the SMEs reached preliminary consensus on aspirational crosscutting outcomes for youth EE programs, including clear conceptual definitions for each. Following the workshop, we asked attendees to review the list of outcomes and accompanying conceptual definitions and provide feedback. We then incorporated their feedback to further refine the list of outcomes and definitions.

Next, we engaged an NAAEE Academic Advisory Group (12 leading researchers) to review this list of outcomes and collectively discuss opportunities for improvement. We incorporated

Table 1. Environmental education outcomes for the twenty-first century (EE21).

Outcome	Definition	Items
Enjoyment	Positive evaluation of the experience	1.[a]How would you rate the program on a scale from 0 to 10?
Connection/place attachment	The development of appreciation for and positive personal relationships with the physical location and its story.	How much do you agree with the following statements about . . .? (anchors: not at all, some, totally) 1. It was an amazing place to visit. 2. [b]Knowing this place exists makes me feel good. 3. [b]I want to visit this place again. 4. Even if I never visit this place again, I'm glad it's here. 5. [b]I care about this place.
Learning	Knowledge regarding the interconnectedness and interdependence between human and environmental systems	How much did you learn about each of the following things as a result of . . .? (anchors: nothing at all, a fair amount, a huge amount) 1. [b]How different parts of the environment interact with each other. 2. How what happens in one place impacts another. 3. [b]How people can change the environment. 4. [b]How changes in the environment can impact my life. 5. [b]How my actions affect the environment. 6. How to study nature.
Interest in learning	Enhanced curiosity, increased interest in learning about science and the environment.	Did this . . . make you feel any more interested in any of the following things? (anchors: not at all, more interested much more interested) 1. [b]Science. 2. [b]How to research things I am curious about. 3. [b]Learning about new subjects in school. 4. Learning more about nature.
Twenty-first century skills	Critical thinking, problem solving, communication and collaboration	How much did this . . . help you improve any of these skills? (anchors: not at all, a fair amount, a huge amount) 1. [b]Solving problems. 2. [b]Using science to answer a question. 3. Understanding the difference between facts and opinions. 4. [b]Listening to other people's points of view. 5. Having good conversations with people you disagree with. 6. [b]Knowing how to do research. 7. Working with others. 8. Taking a leadership role.
Meaning/self-identity	A heightened sense of self-awareness, critical reflection and purpose.	Did this . . . do any of the following things for you? (anchors: not at all, a fair amount, a huge amount) 1. [b]Taught me something that will be useful to me in my future. 2. [b]Really made me think. 3. [b]Made me realize something I never imagined before. 4. [b]Made me think differently about the choices I make in my life. 5. Gave me ideas for what I might do in the future. 6. [b]Made me curious about something.
Self-efficacy	Belief in one's own ability to achieve one's goals and influence their environment.	Retrospective pre/post items (anchors: not at all, somewhat agree(d), strongly agree(d)) 1. [b]I believe in myself

Table 1. Continued.

Outcome	Definition	Items
		2. [b]I feel confident I can achieve my goals
		3. [b]I can make a difference in my community.
Environmental attitudes	Sensitivity, concern and positive dispositions towards the environment	Retrospective pre/post items (anchors: not at all, somewhat agree(d), strongly agree(d))
		1. [b]I feel it is important to take good care of the environment
		2. It's important to protect as many different animals and plants as we possibly can.
		3. [b]Humans are a part of nature, not separate from it.
		4. [b]I have the power to protect the environment
Action orientation	Intentions to perform behaviors relevant to the program's content or goals.	1.[a]As a result of the program, do you intend to do anything differently in your life? (yes/no)
Actions: environmental stewardship	Motivations to perform stewardship-related behaviors.	Did this . . . make you any more likely to do any of the following things within the next year? (anchors: no more likely, somewhat more likely, way more likely)
		1. [b]Help to protect the environment.
		2. [b]Spend more time outside.
		3. [b]Make a positive difference in my community.
		4. Talk with my family about ways to protect the environment.
Actions: cooperation/collaboration	Motivation to collaborate more with others	Did this . . . make you any more likely to do any of the following things within the next year? (anchors: no more likely, somewhat more likely, way more likely)
		1. [b]Listen more to other people's points of view.
		2. [b]Cooperate more with my classmates.
		3. Work together with other people to solve problems.
Actions: school	Motivation to work harder in school.	Did this . . . make you any more likely to do any of the following things within the next year? (anchors: no more likely, somewhat more likely, way more likely)
		1. Study science outside of school.
		2. [b]Work harder in school.
		3. [b]Pay more attention in class.

[a]Single items were not included in CFA procedures.
[b]Items in final scale based on results of CFA procedures in Tables 5, 6 and 7.
Ellipses indicate wording that changed from sample to sample (e.g. 'field trip' vs. 'visit' vs. 'experience').

feedback from this group and further refined our list and definitions. We also reviewed and incorporated the results from an unpublished Delphi Study (Clark et al. 2015) that also sought to identify the crosscutting outcomes for EE. In March 2017, we engaged the National Park Foundation Learning Alliance leadership, which included managers from Great Smoky Mountains Institute at Tremont, Teton Science Schools, Yellowstone Forever, North Cascades Institute, Friends of Acadia, Conservancy for Cuyahoga Valley National Park, Golden Gate National Parks Conservancy and NatureBridge. We asked for a critical review of the list of outcomes and conceptual definitions and their applicability to their programs. Table 1 provides the list of crosscutting outcomes that resulted from our efforts, along with broad definitions for each.

Scale development process

Using these 12 outcomes and their corresponding definitions, we developed and refined survey items to best measure each concept with iterative review by external experts and following

procedures outlined by DeVellis (2003) and Presser et al. (2004). This process also included identifying and reviewing existing scales and items used by other researchers, including those associated with measuring place attachment (e.g. Kyle, Graefe, and Manning 2005), positive youth development (e.g. Bowers et al. 2010), connection to nature (e.g. Cheng and Monroe 2012, Mayer and Frantz 2004, Nisbet, Zelenski, and Murphy 2009), academic motivations (e.g. Powell et al. 2011), environmental knowledge, attitudes, intentions and behaviors (e.g. Bogner 1999; Leeming, Dwyer, and Bracken 1995; Powell et al. 2011; Stern, Powell, and Ardoin 2008) and environmental literacy (Hollweg et al. 2011). We developed primarily retrospective questions (e.g. 'How much did you learn about each of the following things as a result of the program?' 'Did the program help you improve any of these skills?'). For two factors ('Environmental attitudes' and 'Self-efficacy') we developed a retrospective pre-post bank of items in which participants were asked to think back to before the program to indicate their level of agreement with items before participating and then to indicate their current level of agreement after their participation. These retrospective pre-post items were developed to enhance variation and sensitivity measuring changes in these attitudes (Chang and Little 2018; Sibthorp et al. 2007). All items were anchored using 11-point scales, measured as 0 to 10 with three anchors at the low end, the midpoint and the high end of the scale (see Table 1).

We used 11-point scales to counter issues regarding positive skew and lack of variability associated with 'balanced' bipolar Likert-type scales, such as a five-point 'strongly agree' to 'strongly disagree' scale. These scales often curtail the range of variability to one side of the scale (Miller 2018). Issues pertaining to lack of variance and skewness are not unique in scales used to evaluate EE programs (Dawes 2008; Peterson and Wilson 1992; Vezeau et al. 2017). Typically, this problem with measurement reflects a scale's insensitivity, or inability to effectively measure variations in an outcome because of social desirability, poor item wording ('motherhood' items) or a ceiling effect (e.g. high scores in pre-experience surveys limit the ability of scale to measure a change) (Vezeau et al. 2017). This lack of sensitivity ultimately pertains to the design and construction of the scales (Miller 2018; Munshi 2014). According to the literature, there are several ways to improve variation in responses. First, studies have found that by removing unused response options and adding additional options to the skewed end of the Likert-type scale may achieve a greater degree of discrimination with lower mean scores and higher standard deviations (Klockars and Hancock 1993; Klockars and Yamagishi 1988). Although this may appear to limit the possibility of measuring all potential responses to a statement (e.g. from strongly disagree to strongly agree), if the full five-point range of response options are not used, realigning the response options and anchoring one end of the response scale with the neutral response category enhances variability (Streiner 1985). Another technique in cases where there is a lack of variation in responses is to expand the Likert-type scales from five points to seven, nine or eleven points, which according to the literature, does not erode the validity and reliability of a scale (Dawes 2008; Hawthorne et al., 2006; Streiner and Norman 2008). However, if one's sample is children/youth, care must be taken when increasing the number of response options to ensure that they are able to understand the subtle differences between answer choices or validity may be reduced (Clark and Watson 1995). In our case, we employed an 11 point scale, measured from zero to 10, which corresponds to the widely used 'numerical rating pain scale' for youth that is used in health care (Huguet, Stinson, and McGrath 2010; Manworren and Stinson 2016). Our pilot testing revealed that 11-point scales yielded greater variability than more traditional 'balanced' 5-point Likert-type agree/disagree scales. Cognitive testing with early subjects also revealed that youth respondents found them easier to understand.

Sites, samples and data collection

We administered surveys at six different STEM-related EE experiences across the United States. Experiences included two 1-day field trip programs for fifth–eighth grade students, three

multiday residential EE programs for grades 5–7 and one natural science museum, where we encountered youth visitors (ages 10–15) at the end of their visits. These six programs represented a range of geographic contexts from Oregon to Florida in both urban proximate and rural locations that serviced very diverse audiences. We attempted a census of all students that participated in each of the organized programs under study. At the North Carolina Science Museum, we also attempted a census of all youth (ages 10–15) that visited the museum during two weekends in the months of July and August 2018 by positioning researchers to intercept visitors near the exit of the museum. At each research site, response rates were near 100%. Below is a short description of each research site and the study's six samples:

Great Smoky Mountains National Park Appalachian Highlands Science Learning Center (GRSM), NC (https://www.nps.gov/grsm/learn/education/classrooms/fieldtrips.htm) (https://www.nps.gov/grsm/learn/education/north-carolina-ms-programs.htm): Three-hundred and fifty-one sixth–eighth grade students (55% male) from five different middle schools across Western North Carolina completed retrospective surveys after participating in a five-hour field trip program designed to meet state science educational standards. The programs provided place-based, hands-on science education focused on terrestrial invertebrates and soils.

Everglades National Park (EVER), FL (https://www.nps.gov/ever/learn/education/rangerguided.htm): Two-hundred and one fifth grade students completed retrospective surveys after participating in five-hour long ranger-led field trip programs in Everglades National Park designed to meet state educational standards that focused on the Everglades ecosystem and the water cycle. The sample was 57% female and largely Hispanic and African American from Dade and Broward counties.

North Carolina Museum of Natural Sciences (NCMNS), NC (https://naturalsciences.org/index): One hundred and fifty-nine youth visitors between the ages of 10 and 15 completed surveys at the end of their informal visits to the Museum, which contains seven floors of interactive exhibit spaces, research areas and a 3D theatre. Experiences range from passive observation of exhibits, nature-themed art and multimedia presentations to docent programs and opportunities to engage in citizen science, all focusing on natural history, science-related and environmental themes. This sample represents the only sample in the study at the far end of the 'free choice' spectrum, in that visitors were not visiting as part of a school-related program.

NorthBay Adventure Center (NB), MD (www.northbayadventure.org) : Two-hundred and eighty-three sixth grade students (42% male) completed surveys after participating in a five-day residential program, which comingles elements of traditional EE with positive youth development programming. During the day, students participate in traditional EE programs, including investigating wetlands, observing birds and other wildlife, participating in nature walks and exploring the coastal habitat. In the evenings, they participate in multimedia live presentations designed to link the day's lessons with their experiences at home (see Stern, Powell, and Ardoin 2010 for more details). Students were from both urban and rural areas and were highly racially diverse.

Multnomah Education Service District Outdoor School (MESD), OR (https://www.mesdoutdoor-school.org/program-tour1.html): One-hundred and fifty-nine sixth grade students (51% female) completed retrospective surveys after participating in a three-day/two-night residential EE program. Many of the students came from urban settings and constituted a racially diverse sample. The program focuses on natural science through hands-on approaches taught by trained high school volunteers under the supervision and guidance of adult staff members. While the primary focus is on science, the program also focuses on building communities of students by mixing students from various schools and helping foster connections.

Glen Helen Residential Outdoor School (GH), OH (www.glenhelen.org): One-hundred and seventy-six fifth grade students (48% male) completed retrospective surveys following a three-night, 3.5 day program (Tues morning–Friday after lunch). The program included activities such as bird identification, cooperative learning, stream studies, night hikes and evening campfire programs.

Table 2. Data cleaning and final sample sizes.

Sample	Sample size	Missing over 50%	MAH outliers	n Used for analysis
GRSM	351	0	25	326
EVER	201	3	39	160
NCMNS	159	2	8	149
NB	283	0	36	247
MESD	159	2	16	141
GH	176	0	27	149

Data preparation

All data were entered into IBM SPSS statistics for initial screening. First, the data were screened for missing data. Any surveys with data missing more than 50% from any construct were eliminated. Next, we screened for multivariate outliers using Mahalanobis distance (Tabachnick and Fidell 2007). The results of data screening and the final samples are displayed in Table 2.

Determining factor structure using confirmatory factor analysis

We used the EQS v6.1 software (Bentler 2005) to perform the statistical analyses, which progressed in several stages. First, we tested the hypothesized factor structure and psychometric properties of the scales using confirmatory factor analysis (CFA) techniques and developed a final model using the data collected at Great Smoky Mountains National Park Appalachian Highlands Science Learning Center. We then used the data collected from each of the other sites to test the fit of this final model in each context. We also conducted a series of multigroup invariance tests (configural, metric and structural invariance testing) to confirm that the outcomes and items developed to measure these concepts were sensitive, reliable and valid within and across different contexts. This crossvalidation comparison using independent samples determines if the hypothesized model, including the items comprising a scale and the hypothesized factor structure of the scales, is stable across different samples (e.g. Byrne, Shavelson, and Muthen 1989; Powell et al. 2011; Vagias et al. 2012; Vezeau et al. 2017).

CFA explicitly tests a hypothesized measurement model (as opposed to an exploratory approach), accounts for sources of common measurement and method error that is inherent in survey research (such as lack of variance and covariance between items) and provides empirical justification for adjustments to arrive at the most parsimonious model and scale (Brown 2015; Byrne 2006; Kline 2015; Noar 2003). In this article, we report the Satorra–Bentler Scaled Chi-Square (S-B x^2), Robust Comparative Fit Index (CFI), Standardized Root Mean Square Residual (SRMR) and the Robust Root Mean Square Error of Approximation (RMSEA) and its associated 90% confidence interval (Brown 2015; Byrne 2006; Kline 2015). The S-B x^2 should be interpreted like a x^2; for the CFI, values greater than 0.9 indicate an acceptable fit; for SRMR, values less than .09 indicate acceptable fit; and for RMSEA values below .06 indicate acceptable fit (Bentler 1990; Byrne 2006; Hu and Bentler 1999). To identify items and inter-item covariance that degrade the model and that, if removed, would improve the overall fit, we used the Wald and Lagrange Multiplier (LM) Tests (Byrne 2006; Kline 2015). Model adjustment decisions also included theoretical justification (Byrne 2006; Tabachnick and Fidell 2007).

Construct validity and crossvalidation procedures

To further assess the validity and psychometric properties of the final model, we used six independent samples drawn from different program providers in a series of multigroup tests of measurement invariance employing increasingly stringent analyses (suggested by Byrne 2006; Vandenberg and Lance 2000) to examine the stability of the structure and measurement. These crossvalidation procedures provide a rigorous analysis that determines whether the items

comprising a scale, the hypothesized factor structure of the scales and their measurement are stable across different samples (Brown 2015; Byrne 2006; Byrne, Shavelson, and Muthen 1989). Lastly, to test the criterion validity (sensitivity) of EE21 scale, we conducted an analysis of variance to compare the mean composite score of each of the six independent samples. Statistically significant differences in mean composite scores across the six independent programs would provide evidence regarding the sensitivity of the resulting scales across samples.

Results

Using the GRSM data, we explored our conceptually-based hypothesized factor structure as well as alternative models (Breckler 1990; Browne and Cudeck 1993). Our first step in determining the structure was to test the hypothesized 10-factor model and intended structure of the items (Model 1). The resulting fit was good (S-B x^2=1452.77; CFI = 0.940; SRMR=.050; RMSEA = 0.041 (.036, .045)). Model 1 thus served as the base model and starting point for refining the measurement model using procedures outlined by Hatcher (1994). First, we examined the item factor loadings to uncover insignificant paths/low loadings. We also used the Lagrange Multiplier (LM) Test and the Wald Test to identify items that 'crossloaded' on other factors or had high levels of error covariance with another item and thus eroded the model fit and that, if dropped, would increase the overall fit (e.g. Byrne 2006; Kline 2015). To test whether a subsequent model improved fit, we used the change in CFI.

In total, we removed 14 items (items without a "b" in Table 1). Fit indices for Model 2 indicated improvement in fit (S-B x^2=529.90; CFI=.980; SRMR=.040; RMSEA=.029) over Model 1. Factor loadings for each item are provided in Table 3. Next, based on the correlations between factors of this model (Table 4), we tested whether the introduction of a single second-order factor would degrade or improve fit (Model 3). The resulting fit indices (S-B x^2=750.63; CFI=.945; SRMR=.064; RMSEA=.045) were a good fit of the data, but change in CFI indicated a minor erosion of fit (although, less than .05). We then tested the potential of two second-order factors that reflected the question styles used and based on the correlations between factors (Model 4). One second-order factor reflected the Environmental attitudes and self-efficacy first-order factors that were measured using the change scores between self-reported pre and post items. The other second-order factor reflected the other eight first-order factors, which were each measured using retrospective items only. The resulting fit indices of Model 4 (S-B x^2=631.92; CFI=.967; SRMR=.048; RMSEA=.035) were also a good fit of the data. However, the change in CFI indicated minor erosion of fit (Table 5). Thus, we selected Model 2 as the final measurement model, although, Models 3 and 4 also exhibited excellent fit.

Testing validity of the structure

To examine the stability of the structure and measurement of the final model, we conducted increasingly stringent crossvalidation analyses across all six samples. The final measurement model from the GRSM data (Model 2) served as the baseline model and were crossvalidated with the same model constructed in each of the other samples (EVER, NCMNS, NB, MESD and GH).

First, we compared the fit statistics from each dataset independently using the identical model (the same configuration). The results indicated that the fit was acceptable across all samples with most of the differences attributed to the differences in sample size (Table 6). Second, we developed a series of multigroup configural models that tested the invariance of the structure of each sample against the GRSM data sample. The *configural test* of invariance simultaneously compares the 'number of factors and factor-loading pattern' across the groups (Byrne

Table 3. Means, standard deviations and factor loadings of items from Model 2 at GRSM.

Constructs and items	M	SD	Final factor loading
Connection/Place attachment			
It was an amazing place to visit.	8.06	2.32	–
Knowing this place exists makes me feel good.[a]	7.29	2.59	.777
I want to visit this place again.[a]	7.46	2.81	.789
Even if I never visit this place again, I'm glad it's here.	8.08	2.38	–
I care about this place.[a]	7.98	2.45	.833
Learning			
How different parts of the environment interact with each other.[a]	6.99	2.42	.714
How what happens in one place impacts another.	6.92	2.52	–
How people can change the environment.[a]	7.64	2.58	.809
How changes in the environment can impact my life.[a]	7.41	2.61	.826
How my actions affect the environment.[a]	7.36	2.85	.830
How to study nature.	7.50	2.66	–
Interest in Learning			
Science.[a]	6.37	3.08	.782
How to research things I am curious about.[a]	6.26	2.98	.805
Learning about new subjects in school.[a]	5.85	3.25	.766
Learning more about nature.	7.31	2.98	–
Twenty-first century skills			
Solving problems.[a]	5.15	3.21	.771
Using science to answer a question.[a]	6.23	2.99	.764
Understanding the difference between facts and opinions.	5.09	3.48	–
Listening to other people's points of view.[a]	5.73	3.20	.788
Having good conversations with people you disagree with.	5.37	3.51	–
Knowing how to do research[a]	6.37	2.83	.786
Working with others.	6.90	2.92	–
Taking a leadership role.	5.97	3.30	–
Meaning/Self Identity			
Taught me something **that will be useful to me** in my future.[a]	6.10	2.96	.840
Really made me think. [a]	6.20	3.10	.802
Made me realize something I never imagined before.[a]	6.25	3.16	.802
Made me think differently about the choices I make in my life.[a]	5.63	3.39	.809
Gave me ideas for what I might do in the future.	5.14	3.62	–
Made me curious about something.[a]	6.19	3.26	.742
Self-Efficacy (Retrospective pre-post)			
I believe in myself.[a]	0.70	2.40	.499
I feel confident I can achieve my goals[a]	0.68	1.71	.649
I can make a difference in my community.[a]	1.14	1.93	.840
Environmental Attitudes (Retrospective pre-post)			
I feel it is important to take good care of the environment.[a]	1.16	1.80	.605
It's important to protect as many different animals and plants as we possibly can.	1.13	1.91	–
Humans are a part of nature, not separate from it.[a]	1.06	1.96	.695
I have the power to protect the environment.[a]	1.36	2.18	.791
Actions: Environmental Stewardship			
Help to protect the environment.[a]	6.86	2.94	.866
Spend more time outside.[a]	6.60	3.30	.726
Make a positive difference in my community.[a]	6.49	3.06	.899
Talk with my family about ways to protect the environment.	5.18	3.40	–
Actions: Cooperation/Collaboration			
Listen more to other people's points of view.[a]	5.84	3.26	.882
Cooperate more with my classmates.[a]	5.97	3.22	.819
Work together with other people to solve problems.	6.61	3.31	–
Actions: School			
Study science outside of school.	5.46	3.21	–
Work harder in school.[a]	6.41	3.33	.934
Pay more attention in class.[a]	6.46	3.35	.881

[a]Item included in final scale after CFA procedures.

2006, 233). Third, we tested the *measurement invariance* by constraining the free factor loadings (loadings that are freely estimated and not fixed to 1 for identification and scaling) to be equal. In this step, the pattern between factor loadings of items and factors in the base model is compared against the other samples by constraining them to be equal (Byrne 2006; Kline 2015). Last,

Table 4. Correlations between factors of model 2 at GRSM.

	Place Att.	Learning	Interest	Skills	Meaning	Efficacy	Env. Att.	Act: Env.	Act: Coop.
Place Att.									
Learning	.608								
Interest	.670	.710							
Skills	.576	.770	.855						
Meaning	.635	.782	.810	.910					
Efficacy	.122	.121	.233	.177	.220				
Env. Att.	.172	.185	.210	.171	.229	.824			
Act:Env.	.616	.811	.796	.822	.859	.245	.212		
Act:Coop.	.542	.692	.755	.892	.832	.167	.166	.906	
Act:Schl.	.419	.609	.696	.789	.740	.209	.149	.788	.842

we conducted a series of *structural invariance* tests, which compare the relationships between factors and error covariances by constraining them to be equal across the six samples (Byrne 2006). For each step, we assessed the difference between each of the models using changes in the CFI; changes less than .05 are considered acceptable and indicate invariance between models (Byrne 2006; Little 1997). We used this test because of the complexity (10 factors) of the models and the oversensitivity of other tests (Cheung and Rensvold 2002). Finally to identify sources of variance, such as items and covariances, we used the LM test with a threshold of >.001 (Gould et al. 2011).

The results of the series of multigroup configural tests produced statistics indicative of a good fitting model and suggested that the structure of factors and items (the factor structure and factor loading patterns) are the same across all samples (Table 7). When the factor loadings were constrained to be equal, the results also indicated the models to be invariant. When latent constructs were constrained to be equal, the series of multigroup tests of measurement invariance also showed the models to be invariant. This suggests that the structure and metrics at increasingly stringent levels of analyses were stable across all six sites (Table 7).

Finally, to test the sensitivity of the scales, we conducted analysis of variance (ANOVA) with post hoc comparison of the parceled mean composite scores of the full EE21 scale across the six sites. The results (Table 8) demonstrate significant differences in mean scores between the six sites, indicating that the EE21 scale is sensitive (mean scores vary based on programmatic qualities). This test confirms that the EE21 can be used to discriminate between different programs with different characteristics.

Discussion

We address a challenging and potentially controversial question in this research that we expect will spur debate: Is there a consistent set of outcomes to which all EE programs for youth could reasonably aspire? In an effort to answer this question and develop valid and reliable crosscutting common outcomes for EE programs serving youth, ages 10–14, we undertook an extensive participatory process with practitioners, leaders and academics and subsequent psychometric multivariate statistical crossvalidation procedures across six diverse samples.

This effort established scales exhibiting high degrees of construct validity, composed of content, convergent and criterion validity (Anastasi and Urbina 1998). *Content validity,* often called *face validity,* refers to whether each item in a construct is an accurate representation of the concept of interest (Anastasi and Urbina 1998; DeVellis 2003). Content validity was established by grounding our study in relevant EE literature on outcomes and in the extensive participation of subject matter experts to identify and define the crosscutting outcomes and to iteratively review the item pool developed for measuring each concept (Anastasi and Urbina 1998; DeVellis 2003). *Convergent validity,* or the correlation between factors, is demonstrated in Table 4 and through

Table 5. Results from model building and adjustments using GRSM data.

Model development	S-B x2	df	CFI	SRMR	RMSEA[a]	ΔCFI
Model 1: Hypothesized 10 factor Model	1452.77	944	0.940	0.050	0.041 (.036, .045)	–
Model 2: Best Fitting 10 Factor Model	529.90	419	0.980	0.040	0.029 (.020, .036)	+.040
Model 3: 10 factor model with 1 second-order factor	750.63	453	0.945	0.64	0.045 (.039, .051)	-.035b
Model 4: 10 factor model with 2 second-order factors	631.92	452	0.967	0.048	0.035 (.028, .041)	-.013b

[a]95% confidence interval around the RMSEA.
[b]Compared to Model 2.

Table 6. Comparison of CFA model fit indices for six independent samples.

Confirmatory factor analysis results of final 10 factor EE21 scale

	S-B x2	df	CFI	SRMR		RMSEA[a]	RHO/Cronbach's Alpha
Great Smoky Mountains National Park-Appalachian Highlands Science Learning Center (GRSM)	529.90	419	.980	.040	.029	.020-.036	.974/.950
Everglades National Park (EVER)	479.50	414	.920	.057	.032	.015- .043	.933/.898
NC Museum of Natural Sciences (NCMNS)	581.20	418	.945	.051	.051	.041-.061	.977/.960
NorthBay (NB)	519.48	419	.978	.036	.031	.021-.040	.985/.966
Multnomah Education Service District Outdoor School (MESD)	529.33	418	.941	.050	.044	.031-.054	.975/.959
Glen Helen (GH)	476.28	416	.943	.060	.031	.013-.044	.973/.951

[a]95% confidence interval around the RMSEA.

the extensive CFA procedures across six samples that established high factor loadings and excellent fit across all samples (Kline 2015). *Criterion (predictive) validity* addresses how well a construct either is predictive of another outcome or can differentiate different treatments or conditions in comparative or quasi-experimental designs and is regarded as the 'gold standard' (DeVellis 2003). In the present case, EE21 identified significant differences between different EE programs, which had different programmatic approaches and qualities. Lastly, scale development 'crossvalidation' procedures demonstrated configural, metric and structural invariance of EE21 across six independent samples. These stringent psychometric tests indicate that the items comprising EE21, the factor structure and their measurement were stable across six different samples drawn from six different contexts (Byrne 2006; Byrne, Shavelson, and Muthen 1989).

The results of the participatory process identified 12 outcomes relevant to a wide array of EE programming. The results of our psychometric testing suggest that the resulting EE21 scale is statistically valid and reliable, can be used to assess a range of outcomes across a variety of programs, is not highly skewed, and is sensitive (varies based on program quality). The psychometric properties of the outcomes in combination with the ability to detect mean differences across program sites illustrate the scales' applicability to diverse EE and informal STEM programs, audiences and contexts (Brown 2015; Byrne 2006). The final outcome measures, which comprise what we are calling the 'Environmental Education Outcomes for the twenty-first century', or 'EE21', survey, are summarized in Table 1. Our hope is that this instrument will be useful to the broader field, as it represents the breadth of what EE for youth, and similar informal STEM programs, might aspire to achieve in individual participants.

The EE21 instrument enables the possibility of a large-scale comparative study that could investigate which programmatic characteristics lead to more positive learning outcomes in a wide variety of contexts. As stated, current understanding of what leads to success in EE is limited, because most empirical studies have focused on evaluating single programs (Stern, Powell, and Hill 2014). To identify what works best, a large-scale comparative study is necessary, and without common crosscutting outcomes, such a study would be impossible (Grack Nelson et al. 2019).

Table 7. Multigroup tests of invariance.

Hypothesis tests	df	x2	S-Bx2	CFI	SRMR	RMSEA [a]		ΔCFI
GRSM-EVER Multigroup configural model	833	1301.31	1009.22	.964	.050	.030	(.022,.036)	–
GRSM-EVER Factor loadings constraints	855	1357.22	1042.48	.961	.065	.030	(.023,.036)	−.003
GRSM-EVER Structural covariance constraints	900	1577.76	1235.30	.931	.413	.039	(.034,.044)	−.03
GRSM-NCMNS Multigroup configural model	837	1335.98	1109.73	.968	.046	.037	(.031,.043)	–
GRSM- NCMNS Factor loadings constraints	859	1385.41	1149.58	.966	.057	.038	(.032,.043)	−.002
GRSM- NCMNS Structural covariance constraints	904	1472.94	1221.32	.963	.076	.039	(.033,.044)	−.005
GRSM-NB Multigroup configural model	838	1376.24	1048.76	.979	.038	.030	(.023,.035)	–
GRSM-NB Factor loadings constraints	860	1425.53	1092.23	.977	.044	.031	(.025,.036)	−.002
GRSM-NB Structural covariance constraints	905	1526.61	1172.50	.973	.070	.032	(.027,.037)	−.006
GRSM-MESD Multigroup configural model	838	1343.25	1070.21	.966	.046	.035	(.028,.040)	–
GRSM-MESD Factor loadings constraints	860	1397.02	1108.12	.964	.057	.035	(.029,.041)	−.002
GRSM-MESD Structural covariance constraints	905	1462.58	1167.95	.962	.114	.035	(.029,.041)	−.004
GRSM-GH Multigroup configural model	836	1379.28	1007.56	.964	.051	.029	(.022,.036)	-
GRSM-GH Factor loading constraints	858	1413.96	1039.03	.962	.054	.030	(.023,.036)	−.002
GRSM-GH Structural covariance constraints	903	1517.57	1122.80	.953	.194	.032	(.025,.038)	−.011

[a]95% confidence interval around the RMSEA.

Table 8. ANOVA comparing EE21 outcomes scale scores across the six independent samples.

	(1)		(2)		(3)		(4)		(5)		(6)		ANOVA F (df) p	Post hoc
Construct	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD		
EE21	5.47	1.86	7.49	1.01	5.43	1.89	6.03	2.03	6.70	1.62	6.88	1.88	42.51 (5,1166) p<.001	1,3 < 2,5,6*** 1,3 < 4** 2 > 1,3,4,5*** 2 > 6* 4 > 3* 4 < 5** 4 < 6***

*<.05,**<.01, ***<.001.

Note: Because the research did not collect a representative sample from each program, we do not identify them here to avoid misinterpretation of the findings as a comparative evaluation.

The EE21 instrument can be adapted and applied in multiple situations for research or evaluation purposes. For short duration programs, we recommend a post-experience retrospective design, such as the one we have employed here. Our analysis revealed the ability of the instrument to discriminate between programs in terms of differential learning outcomes. Moreover, retrospective designs address concerns regarding response shift bias (i.e. when respondents' personal understanding of a construct changes over time) (Chang and Little 2018; Sibthorp et al. 2007), which has been alternately described as a situation in which respondents overstate their knowledge, attitudes or behaviors because they 'don't know what [they] don't know' (Browne 2018, 2). Participants may have limited knowledge or experience to accurately assess their attitudes or behaviors prior to an experience. A retrospective design, including retrospective pre/post items, ensures a more consistent understanding of items. Moreover, in short duration

programs in particular, the retrospective design limits testing bias, in which pre-experience questionnaires influence (typically inflate) post-experience scores (e.g. Dimitrov and Rumrill 2003). For longer duration programs, such as multiday residential experiences or repeated experiences, the EE21 questionnaire can be adapted into a pre/post design. This would require adapting item stems/prompts. Alternatively, retrospective post-experience only surveys are also valid for longer programs.

In either case, the use of consistently measured outcomes provides the basis for comparative studies across organizations and programs, enabling researchers or organizations to determine which approaches tend to most consistently achieve desired outcomes. EE21 also enables the opportunity for collective evaluation of multiple programs (where the outcomes can be summarized and quantified) and the development of learning networks between program providers. For example, if results from using EE21 reveal that a certain program excels in reaching one subset of outcomes and another program excels at a different set, the opportunities for sharing techniques and approaches become clearly visible. Oregon Outdoor Schools (made up of multiple program providers) are already embarking on efforts for programs to learn from each other based on EE21 survey results (Braun 2019). The breadth of outcomes in the EE21 survey reflect the potential of EE (see Ardoin et al. 2018) as well as the roots of the field and future directions and enable organizations to gauge outcomes and reflect on how their programs are designed. Our ultimate hope is that through the use of the EE21, an expanding 'learning' network of EE organizations and program providers will share evidence-based best practices for achieving those outcomes.

The instrument also addresses a persistent measurement challenge facing EE research regarding positive skew and lack of variability associated with 'balanced' bipolar Likert-type scales, such as a five-point 'strongly agree' to 'strongly disagree' scale (Miller 2018; Vezeau et al. 2017). We used an 11 point scale anchored by 'not at all'(0) and 'totally', 'a huge amount' or 'strongly agree' (10), which corresponds to the widely used 'numerical rating pain scale' for youth that is used in health care (Huguet, Stinson, and McGrath 2010; Manworren and Stinson 2016). Results suggested that these unbalanced 11-point scales yielded greater variability than more traditional 'balanced' 5 or 7 point Likert-type agree/disagree scales.

That is not to say that EE21 does not have limitations. Similar to any survey questionnaire, EE21 is prone to social desirability bias, or respondents' tendencies to respond in a way they expect the survey administrator desires, and other typical survey biases (see Podsakoff et al. 2003, for example). Moreover, some elements of the survey proved challenging for the lower end of the age range in lower achieving groups, in particular the retrospective pre/post items measuring environmental attitudes and self-efficacy. Our statistical analyses revealed that these items could be dropped from EE21 if problems arise with younger or lower achieving audiences without eroding the statistical validity of the other subscales. Future researchers could also consider testing alternative question formats for these items such as the retrospective style questions used in the other EE21 scales to reduce the cognitive burden for respondents. Perhaps most importantly, our effort to build crosscutting outcomes that can apply to a diverse range of programming necessitated the development of general, rather than specific, measures of knowledge, attitudes and behaviors. In other words, EE21 does not measure students' dispositions regarding specific pressing environmental issues such as climate change, biodiversity loss, ocean acidification, conservation economics or other topics that are program-specific. We invite researchers and practitioners to add important specific content-related items at the end of the survey as needed. Placing these items at the end of the survey will avoid the erosion of the psychometric performance of the scales. Moreover, because EE21 is written to assess the influence of a program on the individual, we invite other scholars to develop additional outcomes and metrics that measure 'on-the ground' conservation outcomes, group-level civic engagement, policy implications or other collective actions.

Finally, the EE21 scales were specifically developed to be parsimonious, meaning that the minimum number of items necessary to measure a concept was based on the results of the psychometric testing; any items that did not vary sufficiently, did not measure the intended concept, or were too highly related with another were removed. From a face validity standpoint, certain items that are not present in the final scales may appear conceptually important or appropriate. However, because the goal of this research was to develop an efficient measure of these outcomes, redundant items (those that explain roughly the same variance) have been dropped. We encourage researchers to use and to continue to refine our scales to address the inevitable shortcomings inherent in measuring such complex topics.

Note

1. See Biesta (2010) and Grack Nelson et al. (2019) for a discussion regarding concerns regarding the use of shared common measures.

Acknowledgements

Disclosure statement

Funding

Notes on contributors

Robert B. Powell is the George B. Hartzog, Jr. Endowed Professor in the Department of Parks, Recreation, and Tourism Management at Clemson University. He is also the Director of the Institute for Parks, which is an interdisciplinary institute focused on providing research, training, and outreach to support park and protected area management. His research and outreach program focuses on environmental education and interpretation, ecotourism, and protected area management.

Marc J. Stern is a professor in the Department of Forest Resources and Environmental Conservation where he teaches courses in environmental education and interpretation, social science research methods, and the human dimensions of natural resource management. His research focuses on human behavior within the contexts of natural resource planning and management, protected areas, and environmental education and interpretation.

B. Troy Frensley is an assistant professor in the Department of Environmental Sciences where he teaches courses in environmental education and interpretation, global environmental issues, and nonprofit management and leadership. His research focuses on environmental education; environmental interpretation; program evaluation; motivation and engagement; and citizen/community science.

DeWayne Moore is professor emeritus in the Psychology Department at Clemson University. His interests and teaching focus on quantitative methods. Moore taught four quantitative methods courses for PhD students from several different departments. He has served on over 100 dissertation committees and as a quantitative consultant for the CDC and on grants from NSF and NIH. Moore has published over 100 articles in peer-reviewed journals.

ORCID

Robert B. Powell 🔟 http://orcid.org/0000-0003-2775-2571
Marc J. Stern 🔟 http://orcid.org/0000-0002-0294-8941

References

Anastasi, A., and S. Urbina. 1998. *Psychological Testing*. 7th ed. Upper Saddle River, NJ: Prentice-Hall.

Ardoin, N. M., K. Biedenweg, and K. O'Connor. 2015. "Evaluation in Residential Environmental Education: An Applied Literature Review of Intermediary Outcomes." *Applied Environmental Education and Communication* 14 (1): 43–56. doi:10.1080/1533015X.2015.1013225.

Ardoin, N. M., A. W. Bowers, N. W. Roth, and N. Holthuis. 2018. "Environmental Education and K-12 Student Outcomes: A Review and Analysis of Research." *The Journal of Environmental Education* 49 (1): 1–17. doi:10.1080/00958964.2017.1366155.

Bentler, P. M. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107 (2): 238–246.

Bentler, P. M. 2005. *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software.

Biesta, G. J. 2010. "Why 'What Works' Still Won't Work: From Evidence-Based Education to Value-Based Education." *Studies in Philosophy and Education* 29 (5): 491–503. doi:10.1007/s11217-010-9191-x.

Bogner, F. 1999. "Towards Measuring Adolescent Environmental Perception." *European Psychologist* 4 (3): 138–151.

Bowers, E. P., Y. Li, M. K. Kiely, A. Brittian, J. V. Lerner, and R. M. Lerner. 2010. "The Five Cs Model of Positive Youth Development: A Longitudinal Analysis of Confirmatory Factor Structure and Measurement Invariance." *Journal of Youth and Adolescence* 39 (7): 720–735. doi:10.1007/s10964-010-9530-9.

Braun, S. M. 2019. *Outdoor School for All: Diverse Programming and Outcomes in Oregon: 2018 Pilot Study Evaluation*. Portland, OR: The Gray Family Foundation

Breckler, S. J. 1990. "Applications of Covariance Structure Modeling in Psychology: Cause for Concern." *Psychological Bulletin* 107 (2): 260–273. doi:10.1037/0033-2909.107.2.260.

Broussard, S. C., and M. B. Garrison. 2004. "The Relationship between Classroom Motivation and Academic Achievement in Elementary-School-Aged Children." *Family and Consumer Sciences Research Journal* 33 (2): 106–120. doi:10.1177/1077727X04269573.

Brown, T. A. 2015. *Confirmatory Factor Analysis for Applied Research*. 2nd ed. New York, NY: Guilford Press.

Browne, L. (2018). Let's get retrospective. American Camp Association. Retrieved from https://www.acacamps.org/news-publications/blogs/research-360/lets-get-retrospective.

Browne, M. W., and R. Cudeck. 1993. "Alternative ways of assessing model fit." In *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long, 445–455. Newbury Park, CA: Sage.

Byrne, B. M. 2006. *Structural Equation Modeling with EQS: Basic Concepts, Applications and Programming*. 2nd ed. Mahwah, NJ: Erlbaum.

Byrne, B. M., R. J. Shavelson, and B. Muthen. 1989. "Testing for Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105 (3): 456–466. doi:10.1037/0033-2909.105.3.456.

Carr, D. 2004. "Moral Values and the Arts in Environmental Education: Towards an Ethics of Aesthetic Appreciation." *Journal of Philosophy of Education* 38 (2): 221–239. doi:10.1111/j.0309-8249.2004.00377.x.

Catalano, R. F., M. L. Berglund, J. A. M. Ryan, H. S. Lonczak, and J. D. Hawkins. 2004. "Positive Youth Development in the United States: Research Findings on Evaluations of Positive Youth Development Programs." *The Annals of the American Academy of Political and Social Science* 591 (1): 98–124. doi:10.1177/0002716203260102.

Chang, R., and T. D. Little. 2018. "Innovations for Evaluation Research: Multiform Protocols, Visual Analog Scaling, and the Retrospective Pretest–Posttest Design." *Evaluation & The Health Professions* 41 (2): 246–269. doi:10.1177/0163278718759396.

Cheng, J. C. H., and M. C. Monroe. 2012. "Connection to Nature: Children's Affective Attitude toward Nature." *Environment and Behavior* 44 (1): 31–49. doi:10.1177/0013916510385082.

Cheung, G. W., and R. B. Rensvold. 2002. "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling* 9 (2): 233–255. doi:10.1207/S15328007SEM0902_5.

Clark, C., J. Heimlich, N. M. Ardoin, and J. Braus. 2015. "*Describing the Landscape of the Field of Environmental Education*." North American Association for Environmental Education Annual Conference, San Diego, CA.

Clark, L. A., and D. Watson. 1995. "Constructing Validity: Basic Issues in Objective Scale Development." *Psychological Assessment* 7 (3): 309–319. doi:10.1037/1040-3590.7.3.309.

Dawes, J. G. 2008. "Do Data Characteristics Change according to the Number of Scale Points Used? An Experiment Using 5 Point, 7 Point and 10 Point Scales." *International Journal of Market Research* 50 (1): 61–77. doi:10.1177/147078530805000106.

Delia, J., and M. E. Krasny. 2018. "Cultivating Positive Youth Development, Critical Consciousness, and Authentic Care in Urban Environmental Education." *Frontiers in Psychology* 8: 2340.

DeVellis, R. F. 2003. *Scale Development: Theory and Applications*. 2nd ed. Thousand Oaks, CA: Sage Publishing.

Dimitrov, D. M., and P. D. Rumrill. Jr. 2003. "Pretest-Posttest Designs and Measurement of Change." *Work* (Reading, Mass.) 20 (2): 159–165.

Eccles, J. S., and J. A. Gootman. 2002. "Features of Positive Developmental Settings." *Community Programs to Promote Youth Development* 86–118. In Community programs to promote youth development. Eccles, J.S. and J. A. Gootman, eds. National Research Council and Institute of Medicine. Washington, DC: National Academy Press.

Fenichel, M., and H. A. Schweingruber. 2010. *Surrounded by Science: Learning Science in Informal Environments*. Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Gould, J., D. Moore, N. J. Karlin, D. B. Gaede, J. Walker, and A. R. Dotterweich. 2011. "Measuring Serious Leisure in Chess: Model Confirmation and Method Bias." *Leisure Sciences* 33 (4): 332–340. doi:10.1080/01490400.2011.583165.

Grack Nelson, A., M. Goeke, R. Auster, K. Peterman, and A. Lussenhop. 2019. "Shared Measures for Evaluating Common Outcomes of Informal STEM Education Experiences." *New Directions for Evaluation* 2019 (161): 59–86. doi:10.1002/ev.20353.

Hatcher, L. 1994. *A Step-by-Step Approach to Using SAS for Factor Analysis and Structural Equation Modeling*. Cary, NC: SAS Institute.

Hawthorne, G. J. Mouthaan, D. Forbes, and R. W. Novaco. 2006. "Response Categories and Anger Measurement: do Fewer Categories Result in Poorer Measurement?." *Social Psychiatry and Psychiatric Epidemiology* 41 (2):164–172. doi:10.1007/s00127-005-0986-y

Hollweg, K. S., J. R. Taylor, R. W. Bybee, T. J. Marcinkowski, W. C. McBeth, and P. Zoido. 2011. *Developing a Framework for Assessing Environmental Literacy*. Washington, DC: North American Association for Environmental Education.

Hu, L.-T., and P. M. Bentler. 1999. "Cutoff Criteria for Fit Indices in Covariance Structure Analysis: Guidelines, Issues, and Alternatives." *Structural Equation Modeling* 6 (1): 1–55. doi:10.1080/10705519909540118.

Huguet, A., J. N. Stinson, and P. J. McGrath. 2010. "Measurement of self-reported pain intensity in children and adolescents." *Journal of Psychosomatic Research* 68 (4): 329–336. doi:10.1016/j.jpsychores.2009.06.003.

Institute of Museum and Library Services. 2009. *Museums, Libraries, and 21st Century Skills*. Washington, DC: Library of Congress. doi:10.1037/e483242006-005.

Kahn, P. H., and S. R. Kellert. 2002. *Children and Nature: Psychological, Sociocultural, and Evolutionary Investigations*. Cambridge, MA: MIT Press.

Kline, R. B. 2015. *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford Press.

Klockars, A. J., and G. R. Hancock. 1993. "Manipulations of Evaluative Ratings." *Psychological Reports* 73 (3_suppl): 1059–1066. doi:10.2466/pr0.1993.73.3f.1059.

Klockars, A. J., and M. Yamagishi. 1988. "The Influence of Labels and Positions in Rating Scales." *Journal of Educational Measurement* 25 (2): 85–96. doi:10.1111/j.1745-3984.1988.tb00294.x.

Kohlberg, L. 1971. "Stages of Moral Development." *Moral Education* 1 (51): 23–92.

Kroger, J. 2006. *Identity Development: Adolescence through Adulthood*. Thousand Oaks, CA: Sage Publications.

Kyle, G., A. Graefe, and R. Manning. 2005. "Testing the Dimensionality of Place Attachment in Recreational Settings." *Environment and Behavior* 37 (2): 153–177. doi:10.1177/0013916504269654.

Leeming, F. C., W. O. Dwyer, and B. A. Bracken. 1995. "Children's Environmental Attitude and Knowledge Scale: Construction and Validation." *Journal of Environmental Education* 26 (3): 22–33. doi:10.1080/00958964.1995.9941442.

Lerner, R. M., J. V. Lerner, J. B. Almerigi, C. Theokas, E. Phelps, S. Gestsdottir, S. Naudeau, et al. 2005. "Positive Youth Development, Participation in Community Youth Development Programs, and Community Contributions of Fifth-Grade Adolescents: Findings from the First Wave of the 4-H Study of Positive Youth Development." *Journal of Early Adolescence* 25 (1): 17–71. no.doi:10.1177/0272431604272461.

Little, T. D. 1997. "Mean and Covariance Structures Analyses of Cross-Cultural Data: Practical and Theoretical Issues." *Multivariate Behavioral Research* 32 (1): 53–76. doi:10.1207/s15327906mbr3201_3.

Manworren, R. C. B., and J. N. Stinson. 2016. "Pediatric Pain Measurement, Assessment, and Evaluation." *Seminars in Pediatric Neurology* 23 (3): 189–200. doi:10.1016/j.spen.2016.10.001.

Mayer, F. S., and C. M. Frantz. 2004. "The Connectedness to Nature Scale: A Measure of Individuals' Feeling in Community with Nature." *Journal of Environmental Psychology* 24 (4): 503–515. doi:10.1016/j.jenvp.2004.10.001.

Miller, Z. D. 2018. "Finding the Unicorn: Evidence-Based Best Practices for Improving Quantitative Measures." *Journal of Park and Recreation Administration* 36 (4): 149–155. doi:10.18666/JPRA-2018-V36-I4-8889.

Munshi, J. 2014. "A Method for Constructing Likert Scales." *Social Science Research Network*. doi:10.2139/ssrn.2419366.

National Research Council. 2013. Next generation science standards: For states, by states. Washington, DC: National Academy Press.

National Park Service. 2014. *Achieving Relevance in Our Second Century*. Washington, DC: National Park Service.

National Park System Advisory Board Education Committee. 2014. *Vision Paper: 21st Century National Park Service Interpretive Skills*. Washington DC: National Park Service.

National Parks Second Century Commission. 2009. *Education and Learning Committee Report*, 1–11. Washington, DC: National Park Service.

National Science Foundation (NSF). 2008. Framework for evaluating impacts of informal science education projects. http://www.informalscience.org/documents/Eval_Framework.pdf.

Nisbet, E. K., J. M. Zelenski, and S. A. Murphy. 2009. "The Nature Relatedness Scale: Linking Individuals' Connection with Nature to Environmental Concern and Behavior." *Environment and Behavior* 41 (5): 715–740. doi:10.1177/0013916508318748.

Noar, S. M. 2003. "The Role of Structural Equation Modeling in Scale Development." *Structural Equation Modeling* 10 (4): 622–647. doi:10.1207/S15328007SEM1004_8.

Peterson, R. A., and W. R. Wilson. 1992. "Measuring Customer Satisfaction: Fact and Artifact." *Journal of the Academy of Marketing Science* 20 (1): 61–71. doi:10.1007/BF02723476.

Piaget, J. 1964. "Cognitive Development in Children." *Journal of Research in Science Teaching* 2 (3): 176–186. doi:10.1002/tea.3660020306.

Podsakoff, P. M., S. B. MacKenzie, J. Y. Lee, and N. P. Podsakoff. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies." *Journal of Applied Psychology* 88 (5): 879–903. doi:10.1037/0021-9010.88.5.879.

Powell, R., M. Stern, and N. Ardoin. 2006. "A Sustainable Evaluation Program Framework and Its Application." *Applied Environmental Education and Communication* 5 (4): 231–241. doi:10.1080/15330150601059290.

Powell, R. B., M. J. Stern, B. Krohn, and N. M. Ardoin. 2011. "Development and Validation of Scales to Measure Environmental Responsibility, Attitudes toward School, and Character Development." *Environmental Education Research* 17 (1): 91–111. doi:10.1080/13504621003692891.

Presser, S., M. P. Couper, J. Lessler, E. Martin, J. Martin, J. M. Rothgeb, and E. Singer. 2004. "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly* 68 (1): 109–130. doi:10.1093/poq/nfh008.

Seligman, M. E., R. M. Ernst, J. Gillham, K. Reivich, and M. Linkins. 2009. "Positive Education: Positive Psychology and Classroom Interventions." *Oxford Review of Education* 35 (3): 293–311. doi:10.1080/03054980902934563.

Sibthorp, J., K. Paisley, J. Gookin, and P. Ward. 2007. "Addressing Response-Shift Bias: Retrospective Pretests in Recreation Research and Evaluation." *Journal of Leisure Research* 39 (2): 295–315. doi:10.1080/00222216.2007.11950109.

Smithsonian Institute. 2010. *Strategic Plan: Inspiring Generations through Knowledge and Discovery: Fiscal Years 2010-2015*. Washington, DC: Author.

Stern, M. J., R. B. Powell, and N. M. Ardoin. 2010. "Evaluating a Constructivist and Culturally Responsive Approach to Environmental Education for Diverse Audiences." *The Journal of Environmental Education* 42 (2): 109–122. doi:10.1080/00958961003796849.

Stern, M. J., R. B. Powell, and D. Hill. 2014. "Environmental Education Program Evaluation in the New Millennium: What Do We Measure and What Have we Learned?" *Environmental Education Research* 20 (5): 581–611. doi:10.1080/13504622.2013.838749.

Streiner, D. L. 1985. *Diagnosing Tests: Using and Misusing Diagnostic and Screening Tests for Educational and Psychological Testing*. 3rd ed. Washington, DC: National Academy Press.

Streiner, D. L., and G. R. Norman. 2008. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th ed. Oxford, UK: Oxford University Press.

Tabachnick, B. G., and L. S. Fidell. 2007. *Using Multivariate Statistics*. 5th ed. Needham Heights, MA: Allyn and Bacon.

Thomas, R. E., T. Teel, B. Bruyere, and S. Laurence. 2018. "Metrics and Outcomes of Conservation Education: A Quarter Century of Lessons Learned." *Environmental Education Research* 1–21. doi:10.1080/13504622.2018.1450849.

UNESCO 1977. Final Report Intergovernmental Conference on Environmental Education, Tbilisi, USSR, 14–26 October 1977. Paris: UNESCO.

Vagias, W., R. B. Powell, D. Moore, and B. A. Wright. 2012. "Development, Psychometric Qualities, and Cross-Validation of the Leave No Trace Attitudinal Inventory and Measure (LNT AIM)." *Journal of Leisure Research* 44 (2): 234–256. doi:10.1080/00222216.2012.11950263.

Vandenberg, R. J., and C. E. Lance. 2000. "A Review of and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3 (1): 4–69. doi:10.1177/109442810031002.

Vezeau, S., R. B. Powell, M. J. Stern, D. Moore, and B. A. Wright. 2017. "Development and Validation of a Scale for Measuring Stewardship Behaviors and Elaboration in Children." *Environmental Education Research* 23 (2): 192–213. doi:10.1080/13504622.2015.1121377.