International Journal of Science Education, Part B
Communication and Public Engagement

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rsed20

The Common Instrument: an assessment to measure and communicate youth science engagement in out-of-school time

Gil G. Noam , Patricia J. Allen , Gerhard Sonnert & Philip M. Sadler

To cite this article: Gil G. Noam , Patricia J. Allen , Gerhard Sonnert & Philip M. Sadler (2020): The Common Instrument: an assessment to measure and communicate youth science engagement in out-of-school time, International Journal of Science Education, Part B, DOI: 10.1080/21548455.2020.1840644

To link to this article: https://doi.org/10.1080/21548455.2020.1840644

Published online: 04 Dec 2020.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, PART B
https://doi.org/10.1080/21548455.2020.1840644

The Common Instrument: an assessment to measure and communicate youth science engagement in out-of-school time

Gil G. Noam[a]*, Patricia J. Allen [a]*, Gerhard Sonnert[b] and Philip M. Sadler[b]

[a]The PEAR Institute, McLean Hospital, and Harvard Medical School, Belmont, MA, USA; [b]Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA

ABSTRACT

There has been a growing need felt by practitioners, researchers, and evaluators to obtain a common measure of science engagement that can be used in different out-of-school time (OST) science learning settings. We report on the development and validation of a novel 10-item self-report instrument designed to measure, communicate, and ultimately help promote science engagement among youth: the Common Instrument (CI). When administered to 7,521 elementary and middle school students participating in OST science programming, CI items were found to have good psychometric properties – showing strong item discrimination and a range of difficulties, little difference in item functioning by grade-level, gender, or race/ethnicity, good unidimensionality, stability over time, and a small standard error of measurement over a large variety of science engagements. Given its properties, the CI is a reasonable way to collect data on science engagement in a wide range of OST science programs. Common measures, like the CI, that are reliable and valid provide a common language that enables programs to describe their strengths and challenges and make decisions about adapting and improving their approaches. Common measures are also essential for collective impact initiatives that need brief, easy-to-administer instruments to assess progress and impact of their change-making processes.

ARTICLE HISTORY
Received 16 January 2020
Accepted 19 October 2020

KEYWORDS
Elementary/primary school; informal education; survey

Introduction

> … [The pursuit of science] is an aspiration shared with all mankind but more with youth and childhood than with adults. In this sense scientists are just children who never grew up, who never lost the nagging urge to ask how, why, and what … (Rabi, 1965, p. 1219)

Rabi's observation made more than 50 years ago portends current studies finding that childhood engagement with science – which encourages science interest, motivation, and further exposure to science – plays an important role in youths' science aspirations and pathway persistence (Maltese & Tai, 2010, 2011; Sinatra et al., 2015; Tai et al., 2006). Science engagement involves cognitive, social, emotional, and behavioral processes (Fredricks et al., 2016, 2018; Osborne et al., 2003; Shernoff, 2013; Sinatra et al., 2015), which are closely linked and develop rapidly from infancy to adolescence (Frazier et al., 2009; Gopnik, 2010; Jipson et al., 2014). For instance, infants experience their world through exploratory play by touching, observing, and experimenting, a type of engagement that predicts longer-term cognitive development (Muentener et al., 2018). Scientific thinking – and potentially science interest, motivation, and identity – grows as children are exposed to more

CONTACT Gil G. Noam Gil_Noam@hms.harvard.edu
*Co-first authors.

science, such as when children engage in highly collaborative interactions using observations to evaluate, challenge, and question each other (Frejd, 2019), or when children talk about science with family and friends (e.g. 'How big is the sun?' 'Why is the sky blue?') to better understand how the natural world works (Jipson et al., 2014).

However, sometime between childhood and adolescence, studies conducted in formal educational settings have found that active engagement in science diminishes for many young people, especially among girls, youth of color, and low-income youth (Fredricks et al., 2018; Gibson & Chase, 2002; Martin et al., 2015; Osborne et al., 2003). This is a concern because science engagement is an important factor in children and adolescents' science-related academic success and pursuit of science-related majors and careers (Maltese & Tai, 2010; Tai et al., 2006; Wang & Degol, 2014). Consequently, apprehension about the American youth's disengagement from science has risen to the top of the national education agenda in the U.S. in the wake of highly publicized reports identifying science, technology, engineering, and mathematics (STEM) as a powerful motor of economic and societal prosperity (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007; Organization for Economic Co-operation and Development [OECD], 2016). Encouraged by findings that early engagement with science can foster a life-long interest – especially in settings outside of formal educational settings – educators now focus on ways to engage children in science and sustain their natural wonder, curiosity, and self-directed inquiry (Jipson et al., 2014; Maltese & Tai, 2010).

Science learning in out-of-school time (OST) – including afterschool programs, summer camps, science centers, museums, and libraries – is viewed as an essential strategy for science education for several reasons (Krishnamurthi et al., 2013). Youth spend more than 80% of their waking hours outside of school each year (Afterschool Alliance, 2014), and OST programs can offer more hands-on and exciting activities than those typically provided in formal school settings to improve young people's understanding, curiosity, interest, and joy in learning science (Bell et al., 2009; Krishnamurthi et al., 2013; Noam & Shah, 2014). Additionally, it is estimated that more than 10 million children participate in afterschool programs each year (Afterschool Alliance, 2014), with many children coming from racial, ethnic, and socioeconomic backgrounds that are underserved and underrepresented in science (Kennedy & Odell, 2014). The OST field seeks to create science curricula that engage youth intellectually, academically, socially, and emotionally. For instance, a synthesis report from the National Research Council (NRC, 2015) highlighted engagement as one of three key criteria for identifying and developing productive STEM OST programming, namely by providing youth with first-hand experiences with science phenomena and materials, engaging youth in sustained STEM practices, and establishing a supportive learning community (p. 2) to build 'a STEM-engaged and STEM-literate society and workforce' (pp. 2–5). An exemplar of these criteria, described in a study by Chittum et al. (2017), engages middle school youth in STEM concepts and practices by including such social-emotional elements as offering meaningful choices (e.g. choices of topics and group members), providing opportunities for decision-making (e.g. lesson pace), and encouraging youth voice (e.g. opinions) – in addition to cognitive elements such as problem-solving activities (i.e. using an inquiry-based approach and interdisciplinary curriculum). While many OST programs are designed to provide engaging experiences, few studies use measures that capture the social, emotional, and cognitive aspects of science engagement from the youth perspective – despite engagement being described as the 'holy grail of learning' (Sinatra et al., 2015).

There is a desire by the OST science field to use measures to assess whether, and to what extent, particular educational practices or strategies are effective at increasing youths' experience of engagement in science, and whether youths' experiences differ based on their background (Allen & Noam, 2016; Grack Nelson et al., 2019; Noam et al., 2017). This desire is even stronger with the formation and expansion of local and national professional learning communities (PLCs) and communities of practice (CoPs) that aim to address the dilemma of science disengagement in the U.S. – such as the STEM Learning Ecosystems Community of Practice (SLECoP), which represents 89 communities across the U.S. and five other countries (Traphagen & Traill, 2014; Wenger

et al., 2002). These communities and initiatives are making significant investments in high-quality STEM-rich learning opportunities, especially in informal/OST settings, such as afterschool programs, summer camps, science centers, and museums (Bell et al., 2009; Dabney et al., 2012; National Research Council, 2015), but require common measures to assess and communicate youth science engagement at the program and community levels (Grack Nelson et al., 2019; Traill & Traphagen, 2015).

This study describes the conceptualization and development of a simple self-report measure of science engagement known as the 'Common Instrument,' which was designed by researchers and practitioners for use across diverse science learning settings. The measure is designed to give OST STEM practitioners and the field a tool to gain robust information about youth experiences of science engagement. By having reliable and valid data on how engaged youth feel before and after OST science programming, programs and the field can measure and communicate what is and is not working for such important endeavors as broadening participation in science for underserved and underrepresented groups. In the following sections, we discuss previous research on engagement in school and OST settings and identify gaps that exist in research and in assessment development. Next, we describe the rationale for developing a common measure of engagement for OST science programs and the steps taken to develop the Common Instrument survey to communicate science engagement for practice and research. We conclude with a discussion of key findings and practical implications of a common measure of science engagement.

Measuring engagement

There have been decades of research on the construct of engagement, with the majority of studies conducted by educational researchers in school contexts (Eccles, 2016; Shernoff, 2013), although there is increasing interest in studying engagement in OST contexts (Fredricks et al., 2014; Shernoff, 2010). Regardless of the learning setting, engagement has been broadly defined as the intensity, breadth, duration, and quality of a student's involvement in academic settings and learning activities through their thoughts, feelings, and actions (Eccles & Wang, 2012). There is general consensus that engagement is a three-dimensional construct that consists of interrelated cognitive, emotional, and behavioral components (Fredricks & McColskey, 2012). However, some researchers have conceptualized student engagement using four-dimensional models, for instance through the addition of agentic engagement (Reeve & Tseng, 2011) or social engagement (Wang et al., 2016).

There are also many measures of engagement – including observation tools and surveys – that have been developed for use in school settings, and most studies of engagement have been conducted in formal classroom settings (Fredricks & McColskey, 2012; Shernoff, 2013). In contrast, there are few measures of engagement designed and validated for use in informal OST settings, and this explains, in part, why there are few peer-reviewed studies of engagement in OST (e.g. Fredricks et al., 2017; Naftzger & Sniegowski, 2018). This is surprising, considering that engagement is a cornerstone of OST practice. One review indicates that engagement is the least researched indicator of OST activity involvement despite potentially being the most important, given that it reflects quality of experiences over quantity (Fredricks et al., 2014).

In terms of engagement with science, in both school and OST research and literature, there has been less focus on rigorously defining and measuring subject-specific engagement relative to general engagement in science learning (Sinatra et al., 2015). Importantly, science engagement in OST settings has been made a priority for both research and practice in a National Research Council (NRC) synthesis report (Bell et al., 2009), and a number of studies reporting science engagement as an outcome have been published over the past several years (Chittum et al., 2017; Fredricks et al., 2018; Fredricks & McColskey, 2012; Sinatra et al., 2015; Wang et al., 2016). Most of these studies have conceptualized science engagement at the behavioral level: the frequency of science course attendance or the amount of time spent participating in science-related coursework,

programming, or activities (e.g. dosage and duration) (Fredricks et al., 2014). Fewer studies have considered science engagement as a psychological construct, leaving much to be learned about its cognitive and emotional components.

One example is a validation study of the Math and Science Engagement Questionnaire (Wang et al., 2016), a multi-dimensional student self-report measure of mathematics and science engagement for classroom settings. The scale measures a global mathematics and science engagement construct, as well as four separate dimensions of mathematics and science engagement (i.e. behavioral, cognitive, social, and emotional). The 33-item student measure was validated based on a racially, ethnically, and socioeconomically diverse sample of middle school and high school students in formal school settings. The measure showed good model fit as well as measurement invariance across gender, race/ethnicity, and socio-economic status. However, like many studies of student engagement, the survey can be viewed as too lengthy to administer in time- and resource-limited OST settings, and the items as too specific to school/classroom settings and goals (e.g. usage of words such as 'school,' 'class,' 'required,' 'work,' 'studying').

There are few peer-reviewed and rigorously tested assessments of engagement for OST science settings (Ben-Eliyahu et al., 2018; Dorph et al., 2016; Fredricks et al., 2014; Shah et al., 2018). One widely-used instrument is the Dimensions of Success (DoS) (Shah et al., 2018), a rubric-based observation tool used to record qualitative and quantitative evidence for 12 indicators of program quality that are organized across four domains: features of the learning environment, activity engagement, STEM knowledge and practices, and youth development in STEM. However, the interpretation of student engagement based on observational data, including DoS data, would be improved by understanding individual student's thoughts or feelings related to science engagement. According to Shah et al. (2018), 'the category of attitudes was not included in the dimensions, as it seemed better measured by student surveys versus observations' (p. 4). Youth self-reported science engagement would help to distinguish the social-emotional aspects of science engagement from the physical/participatory aspects of science engagement (Fuller et al., 2018), especially when data provided by youth are combined with other sources of information (e.g. observation, teacher self-report, attendance, etc.) (Sinatra et al., 2015).

An example of a self-report measure of engagement used in OST science contexts is the Activity Engagement Survey (Ben-Eliyahu et al., 2018), a multi-dimensional student self-report measure of engagement that has been tested in school and OST settings. The scale measures a global engagement construct as well as two separate dimensions of engagement (i.e. affective and cognitive–behavioral). The 17-item measure was validated based on a racially, ethnically, and socioeconomically diverse samples of 6th grade youth in a formal school setting (study 1) and 5th grade youth in an informal museum setting (study 2). The measure showed good model fit in both learning contexts and had a practical length appropriate for OST science programming. However, the survey includes only one item specific to science (i.e. 'I figured out something about science,' with a yes-no response), whereas all other items are more general in nature as they relate to affective engagement (e.g. During today's activity … 'I felt happy or excited.') or cognitive–affective engagement (e.g. During today's activity … 'I paid attention during the activity.'). Although domain-general measures of engagement are valuable – and have the advantage of being used across different learning contexts – evidence suggests that motivation-related constructs like engagement are subject-specific. According to Sinatra et al. (2015), ' … in science, one must be aware of the motivation and emotion factors that interact with how one chooses to engage with science content. Many factors may impact engagement differently in science than in other domains … ' (p. 4). While there exist other measures or frameworks for OST science that capture psychological constructs related to, or inclusive of, elements of engagement – such as science learning activation, which is a construct that includes the dimensions of fascination, values, competency beliefs, and scientific sense-making, and which has been shown to predict engagement levels, using the Activity Engagement Survey above – there does not exist a specific measure of science engagement for OST contexts.

In reviewing the literature, we have identified several gaps and limitations pertaining to the measurement of engagement, both generally and specific to science. The literature indicates that assessments of various types of engagement (e.g. student engagement in school, domain-specific engagement) are developed for different purposes, and that the construct of engagement is defined, measured, and reported in different ways (Eccles, 2016; Fredricks & McColskey, 2012; Sinatra et al., 2015). Other limitations identified in the literature on general engagement and science engagement include: many assessments are not rigorously tested (e.g. not peer-reviewed, do not provide validity or reliability values), do not test for measurement invariance to demonstrate whether the survey is similarly interpreted by different populations of students, are too specific to a program or curriculum and therefore cannot be used widely, consist of items that do not apply across learning settings, are too lengthy, focus on narrow age bands, and/or conceptualize engagement too broadly (Fredricks & McColskey, 2012; Sinatra et al., 2015).

Taken together, these issues have made it challenging for the science education field to measure, communicate, and improve science engagement at the program and system levels. It is important to operationally define science engagement and rigorously test it in OST science learning settings to examine whether programs influence levels of engagement, whether there are differences in levels of science engagement between different populations of youth, whether the timing of this engagement matters more at different points in child and adolescent development, and also to determine whether youth who are more engaged in science differ in significant ways from youth who are less engaged (Fredricks et al., 2014).

The Common Instrument project

The dilemma of disengagement from science has led to multiple influential public and private sector initiatives and collaborations in the U.S., designed to support efforts to increase science engagement, knowledge, and awareness with the aim of attracting more young people to science majors and careers (Bell et al., 2009; The White House, 2009). These investments have significantly changed the educational landscape in STEM, leading to the formation and expansion of collective impact initiatives such the statewide afterschool system-building networks (Mott Foundation and STEM Next, 2018) and the STEM learning ecosystems (Traphagen & Traill, 2014). These initiatives promote the collection of data using high-quality measures (i.e. measures that collect reliable and valid data) for improving the implementation of science programming and supporting science learning.

The collective desire for a common vision, common approach, and common language around research, evaluation, and assessment of youth STEM learning experiences has led to a collaborative project to develop a common instrument for science engagement that can be used to support local and national practice, policy, and research (Allen & Noam, 2016; Grack Nelson et al., 2019; Noam et al., 2017; Sneider & Noam, 2019). A common instrument – one that measures the same outcome or construct across a range of OST programs – can begin to answer critical questions of local and national importance, such as why some children, but not others, experience a 'nagging urge' to engage in the pursuit of science (Rabi, 1965), and whether specific programs or strategies can effectively increase science engagement, especially among youth who disproportionately exit from science, including girls, youth of color, and low-income youth (Lyon et al., 2012).

The Common Instrument project began as a collaboration between educational researchers and youth development practitioners implementing OST science programming across the U.S. The primary objective of the Common Instrument project was to develop a measure – a 'common instrument' – capable of assessing the impact of OST STEM programming on youths' perceptions of science engagement, especially youth participating in afterschool programs, summer camps, in clubs, museums, and other venues outside of school. The consensus among a group of program practitioners and researchers involved in the development of the Common Instrument was that

individual science engagement was a common goal of OST programs across a wide spectrum of curricula, activities, and settings, even if programs have unique pathways to achieving this outcome. The group also agreed that a self-report survey would be most appropriate because it would allow practitioners and researchers to determine which individuals in a group (or which groups of individuals) are engaged with science activities, how strongly they are engaged, why they are engaged, or in what ways they are engaged (i.e. cognitively, socially, emotionally engaged). To understand an individual's level of engagement in science, it is essential to directly ask the student themselves. Additionally, self-report instruments have been determined to be the most ecologically valid type of assessment of science engagement in OST settings (Bell et al., 2009), particularly because self-reports are a practical and inexpensive means of assessing large samples of youth (Fredricks et al., 2014).

Practitioners emphasized that it was essential to measure science engagement in a way that is attuned to the informal nature and content of informal science activities – and to do so in the shortest way possible without using an assessment that feels 'test-like.' Researchers emphasized that it was essential to review the literature to build on previous studies and rigorously define the construct of science engagement. Both wanted to collect data from youth participating in science-focused OST settings to test the feasibility of such a survey as well as to provide validity and reliability evidence for the measure.

Study goals and research questions

Our goal was to develop a brief self-report survey that provides practitioners, researchers, and evaluators with a common measure of youth science engagement that can be used across diverse science learning programs and across communities working together to promote science engagement. Our motivating idea was to create a common reporting tool that can be used to communicate findings within and between different OST programs, to increase collaborations between programs (such as through communities of practice or system-building initiatives) that may lead to innovative efforts to measure or promote science engagement, and to create a means to aggregate and analyze trends in youth science engagement – at the local and national levels – to inform changes in educational policies that may impact youth science engagement. The development of this Common Instrument was an iterative process, and practitioner input at all stages (e.g. construct definition, method selection, item development, item revision, interpretation) was essential to ensure the survey was worded so that our definition of science engagement was meaningful and relevant to science activities facilitated in OST settings.

In developing the Common Instrument survey, we aimed to answer the following research questions:

(1) Does the Common Instrument provide valid and reliable data to communicate youth science engagement across a diverse array of OST science programming? We hypothesized that this brief (10-item) survey can comprehensively and accurately measure science engagement when completed by youth attending many different types of OST science programs (that vary in terms of curriculum, setting, duration, etc.).
(2) Is the Common Instrument consistently interpreted across all youth populations, including by age, gender, and race/ethnicity? We hypothesized that the Common Instrument's item characteristics, including discrimination and difficulty, would show no group differences, allowing for cross-group comparisons by age, gender, and race/ethnicity.
(3) Is the Common Instrument interpreted similarly over time, from the start to the end of OST programming? We hypothesized that the interpretation of this measure would be stable over the duration of science learning experiences, making it useful for research or evaluation purposes that aim to measure change in science engagement over time.

Methods

This section describes the definition used to operationalize science engagement as well as study participants, procedures, and statistical analyses used to examine the psychometric properties of the Common Instrument survey, including dimensionality, IRT characteristics, and differential item functioning.

Definition and theory

We aimed to develop items that encompass behavioral, emotional, and cognitive elements of engagement that are consistent with popular and empirical conceptualizations of engagement (Fredricks & McColskey, 2012), but with an application of these elements to OST science learning (Noam & Shah, 2014; Sinatra et al., 2015). The survey has a cognitive element: engaging to understand, observe, or figure out science phenomena. It also has an emotional component: engaging out of a sense of excitement for science learning. And it lastly has a behavioral element: engaging for the physical, hands-on experiences of science activities or projects. Because thoughts, feelings, and actions appear to combine when it comes to science engagement, we planned for our psychometric testing to examine the dimensionality of the survey (i.e. whether the survey items would load as three separate dimensions or as one global scale) (Eccles, 2016; Sinatra et al., 2015; Wang et al., 2016).

Our discussions with OST science practitioners helped to shape our conceptualization of science engagement; they indicated that science engagement connects to feelings of excitement, attraction, and motivation to embrace science-related questions and phenomena (especially those that are meaningful and relevant to everyday life), which are 'sparked' or sustained by fun, hands-on activities – such as learning the physics behind trebuchet catapults and then building and using them to launch potatoes. With this contextual feedback in mind, survey items were developed or revised to prompt young people to reflect on how they engage (cognitively, behaviorally, or emotionally) with science in their everyday lives, like protecting the environment and figuring out how electric cars work, which is relevant to the activities facilitated in OST as well as twenty-first-century life and work.

Our conceptualization of science engagement is grounded in the expectancy-value perspective, which bridges the concept of engagement with other highly interrelated concepts including interest and motivation (Wigfield & Eccles, 2000). The expectancy-value theory has previously been used in the development of constructs related to motivational beliefs and science learning (e.g. Chittum et al., 2017; Fredricks et al., 2018). According to the expectancy-value theory, there are two motivational factors that influence engagement – expectancy beliefs and subjective task values (Wigfield & Eccles, 2000). Expectancy beliefs refer to how well an individual believes they will do on an academic-related activity or task, whereas task value refers to how much an individual values an academic-related activity or task. Task value is further subdivided into four components: (1) intrinsic and interest value, or the enjoyment that an individual feels when engaging in an activity, (2) attainment value, or the importance of an activity for validating aspects of the individual's self-concept or identity, (3) utility value, or how useful the activity is for future goals, and (4) cost, or the risks or negative implications of engaging in an activity. Evidence suggests that an individual's expectancy and task values most directly influence engagement in terms of the individual's level of performance, effort, and persistence in the near or distant future (Wigfield & Eccles, 2000). Applying this theory to science learning, an individual is more likely to experience positive science engagement – cognitively, emotionally, and behaviorally – if they perceive a good chance of succeeding in the science activity (i.e. expectancy) and if they view what they are learning in science as interesting, purposeful, and important (i.e. task value). Expectancy-value theory has been shown to be useful in the study of motivational beliefs related to science as a function of gender, culture, and social contexts (Eccles, 2016; Fredricks et al., 2018; Guo et al., 2017; Perez et al., 2019).

An advantage of anchoring the Common Instrument in the expectancy-value theory is that this theory provides a framework that is frequently used to study, measure, and explain related concepts that are very relevant to science education. This is important because it is difficult to describe engagement without other constructs such as interest, enjoyment, and motivation. For example, according to Renninger and Hidi (2016): ' … meaningful engagement includes intensity as well as positive enjoyment' (p. 77). Additionally, according to Shernoff (2013): ' … enjoyment and concentration likely represent different dimensions of engagement, especially the emotional and cognitive dimension, respectively' (p. 3). Renninger and Hidi (2016) discuss how it is not uncommon to consider other constructs in measurement, as interest, enjoyment, motivation, and other constructs are characteristic of engagement:

> Engagement, like motivation, is beneficial and productive when it is accompanied by interest. A person whose interest is developing is a person who is meaningfully engaged … We suggest that articulating the synergy between interest and engagement could be particularly useful for researchers and practitioners alike.' (p. 15).

Thus, in developing the Common Instrument, the researcher and practitioner team were careful to acknowledge the relationships between these related concepts while respecting the concerns that the term engagement can easily be misused and overgeneralized in education, learning, instructional, and psychological sciences (Azevedo, 2015). The expectancy-value theory will allow for future research to articulate the synergy between youth science engagement, interest, and other concepts relevant to improving educational practice.

Participants and Procedure

The Common Instrument underwent three phases of testing: survey development (field test and item revisions), pretest-posttest study (stability over time), and validation (internal structure, dimensionality, and construct validity) (see Figure 1). We describe the substantial work undertaken during the development, study, and validation phases of the Common Instrument project, including details about programs, participants, and procedures, and then focus our results reporting on the study and validation phases of the final, refined Common Instrument.

Phase 1: development

The pilot version of the Common Instrument consisted of 20 items developed by researchers and practitioners from OST science programs (Phase 1). The self-report survey was designed in traditional pretest-posttest format using a 4-point Likert scale (i.e. Strongly Disagree, Disagree, Agree, Strongly Agree), and it was initially field-tested with youth from 50 OST science programs located across eight U.S. states. Programs were recruited from national initiatives designed to elevate STEM in OST by providing OST educators with high-quality STEM trainings and resources, and participation in the project by programs and youth was voluntary. Most programs were affiliated with national youth-serving organizations, and many of these program sites were concentrated in urban communities with many low-income youth and youth of color. For this reason, the project oversampled underprivileged and underserved youth.

Programs participating in the study focused on many different topics – such as astronomy, environmental science, robotics, and zoology – and were convened in a variety of settings, including afterschool programs, summer camps, and museums and science centers, for a minimum of one hour per week. The pilot sample consisted of 1,200 elementary and middle school-age students ($n = 465$ males, 38.8%; and $n = 735$ females, 61.2%) representing different demographic backgrounds: 24.2% African American/Black, 10.7% Asian/Asian American, 41.3% Latino/a or Hispanic, 12.8% White/Caucasian, and 11.0% Other. Participants' average age was 12.3 years (S.D. = 2.2). At the time of the survey, 48.4% of study participants had attended the program for at least four weeks, and 84.6% of students reported spending at least one hour doing informal science-related activities in the program.

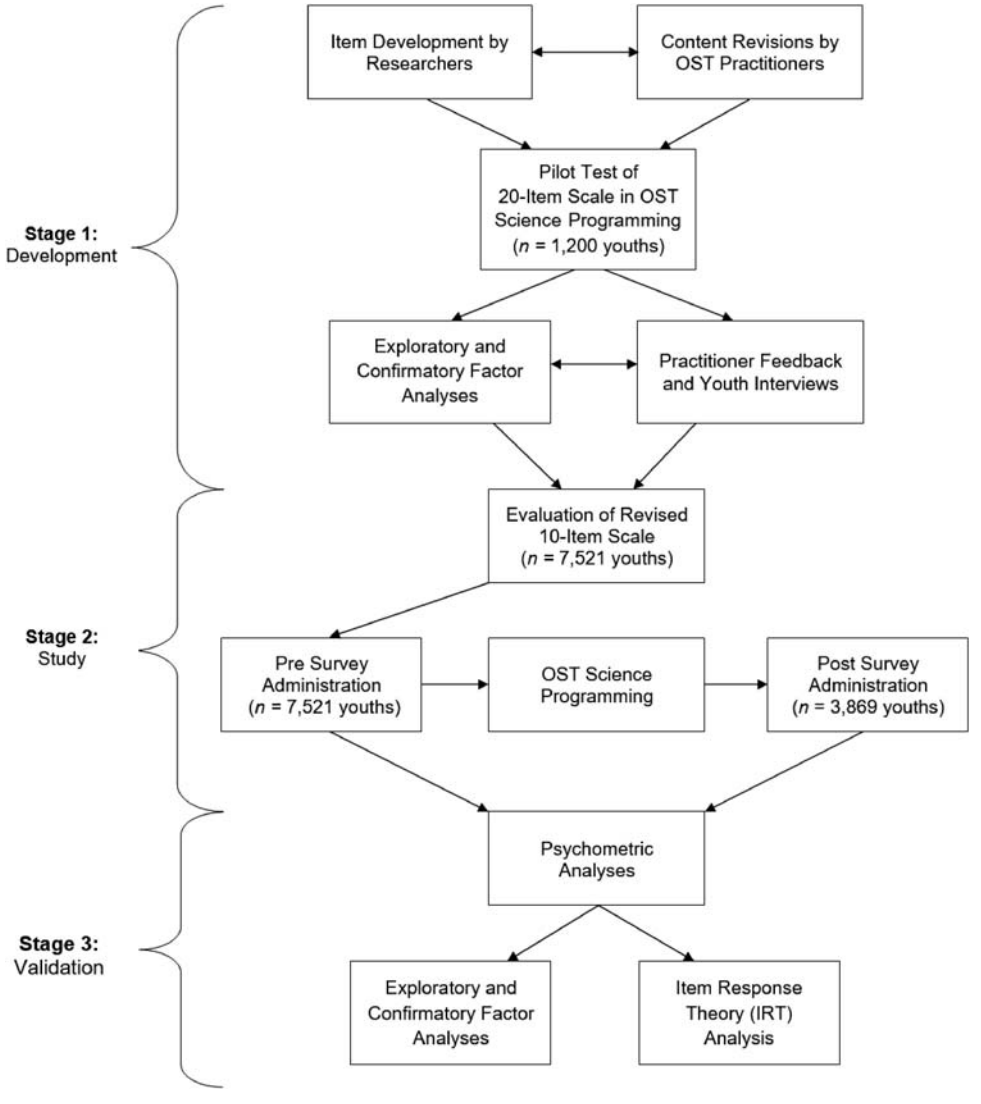

Figure 1. Visualization of Common Instrument development and validation process

Group administration procedures were used to survey participants. Students were gathered in a common area and provided information about the confidentiality of their answers. They were informed about the purpose of the survey and that they were not required to participate. Students were told that they could stop filling out the survey at any time, and that accepting or refusing to take the survey would not in any way jeopardize their participation in the program. The instructions for completing the surveys and the individual survey items were read aloud by a trained administrator. Additional research staff was available to answer questions, and a subset of youth, from a subset of programs, was interviewed upon completion of the survey (youth and programs were chosen at random). Because the survey was relatively short, we had high participation rates and low levels of item non-response. All study procedures were reviewed and approved by the Institutional Review Board at our research institution.

Through this extensive pilot testing phase, the 20 items initially under consideration were pared down to a 10-item survey based on partner feedback, exploratory and confirmatory factor analyses,

distributional analyses, and interviews with youth participants. The team endeavored to balance research considerations (e.g. ensuring that there were items representative of the behavioral, emotional, and cognitive elements of engagement, that items had adequate statistical properties) with practice considerations (e.g. keeping survey brief, ensuring items align strongly with OST program goals based on facilitator feedback, removing items that were hard for youth to relate to or understand based on follow-up interviews with youth). Feedback from educators and cognitive interviews with youth greatly informed the researcher-practitioner team's decision to remove or keep specific items for the next study phase. Items and phrasings that were identified as redundant, ambiguous, unreliable, unrelatable, and/or lacking variation or explanatory power were removed or changed. For example, several items related to excitement were reported as too similar by both educators and youth (e.g. the following item was removed: 'I get excited to find out that I will be doing a science activity.'). There were also items that were not viewed as well-aligned with activities run by OST programs (e.g. 'I like to watch programs on TV about nature and discoveries.'). In this last example, the word 'program' was hard for youth to relate to because the word is not commonly used when referring to TV shows. There were also items that practitioners felt strongly about keeping – such as 'I like online games or computer programs that teach me about science' – as these activities are exciting for youth and OST programs increasingly engage youth in computer games or programs that encourage experimentation, problem-solving, and creativity. Notably, the final set of 10 items chosen for the study and validation phases of this project did not require revision and tested well with both program practitioners and youth.

Phase 2 and 3: study and validation

The revised 10-item form of the Common Instrument, consisting of the best-performing items (those that had the best statistical properties and worked best in practice), was administered to a new cohort of students following the same procedures used in the item development phase. Programs that volunteered to participate in the validation study were also involved in the same national initiatives designed to elevate STEM in OST by providing OST educators with high-quality STEM trainings and resources (see Phase 1, above). These programs, like those that participated in the development phase, focused on a diverse array of science topics and were convened in a variety of learning settings (e.g. afterschool programs, summer camps, science centers or museums) for a minimum of one hour per week.

More than half of the 82 OST science programs participating in the Common Instrument study and validation operated during the school year (58.5%); the remainder took place during the summer months (41.5%). Programs that participated in the validation of the Common Instrument represented 26 states plus Washington, D.C. and all four U.S. regions defined by the U.S. Census Bureau: Northeast (36.0%), including Maine, Massachusetts, New Jersey, New York, Pennsylvania, Rhode Island; Midwest (25.2%), including Illinois, Indiana, Kansas, Michigan, Minnesota, Missouri, Nebraska; South (23.2%), including Alabama, Florida, Georgia, Maryland, North Carolina, Oklahoma, Texas, Tennessee, Virginia, and Washington, D.C.; West (15.2%), including Alaska, California, Idaho, and Oregon. These programs were largely located in cities (67.6%), with less representation from suburbs (22.9%), towns (3.0%), and rural areas (2.7%), if the location was known (3.7% unknown). Program size, estimated based on youth survey responses, varied widely from 1 to 85 participants.

The validation sample consisted of 7,521 elementary and middle school-age students ($n = 2{,}610$ males and $n = 4{,}911$ females). The sample was 38.2% African American/Black, 5.1% Asian American, 21.4% Latino/a or Hispanic, 17.3% White/Caucasian, and 18.0% Other. Participants' average age was 11.93 years (S.D. = 2.4). At the time of the survey, 48.6% of study participants had attended the program for at least four weeks, and 63.1% of students had spent one or more hours doing informal science-related activities in the program each week. The results describe the validation of the instrument based on this sample of students.

Statistical analysis

The final 10-item Common Instrument (Phases 2 and 3) was subjected to a battery of psychometric diagnostics to ascertain its suitable psychometric characteristics.

First, we performed an exploratory factor analysis (EFA) to re-examine the factor structure of the shortened instrument and to ensure there were no remaining problem items in the validation sample. We then used confirmatory factor analysis (CFA) to test the robustness of the model fit (Hooper et al., 2008).

Second, we conducted an IRT analysis to ascertain the major psychometric characteristics of the Common Instrument. Xcalibre 4.1 software was used for all IRT analyses. Because the item responses were 4-point rating scales (i.e. Strongly Disagree, Disagree, Agree, Strongly Agree), Samejima's Graded Response Model was applied. Discrimination of items was assessed using Baker's (2001) taxonomy.

Third, we explored potential Differential Item Functioning (DIF) by gender, race/ethnicity, and grade/age to find out whether the items and scale functioned similarly for different groups of students (Price, 2011). We assessed DIF in terms of item discrimination and difficulty. The term difficulty stems from IRT analyses of tests. Difficult items are those that are answered correctly only by subjects with relatively high theta; easy items are those that are answered correctly also by subjects with relatively low theta. In our case, the items are rating scales of science engagement, so that high 'difficulty' means that an item is rated highly (in the direction of strong agreement) only by those with a strong science engagement, whereas low difficulty means that an item is rated highly also by those with a relatively weak science engagement. Because the rating scale consists of four response categories, there are three transition points (or 'cut points') between them, which represent the average theta at which the majority of subjects shifts from giving the lower response to giving the higher response.

Fourth and last, we explored DIF by time to find out whether item response patterns are stable when the survey is repeated (also in terms of item discrimination and difficulty). We analyzed pretest-posttest responses from a subset of youth who completed the survey at both the beginning and end of their program. We did not conduct a classic test-retest reliability study (in which no relevant characteristic of the participant is supposed to change between administrations) because, in this case, an intervention occurred between the two tests that was intended to boost the students' science engagement.

Results

Dimensionality

EFA showed a strong first factor (eigenvalue 4.60), which explained 46% of the variance. All other factors had eigenvalues considerably below 1. All ten items loaded fairly evenly on the first factor (Table 1), with Item 7 having the strongest loading (0.76) and Item 15, the weakest loading (0.59). Various fit parameters obtained from CFA (Goodness of Fit Index = 0.970; Adjusted Goodness of

Table 1. Factor Pattern of the Final 10-Item Common Instrument

Item ID	Survey Statement	Factor Loading
CI1	Science is something I get excited about.	0.74
CI3	I like to participate in science projects.	0.67
CI5	I like to see how things are made (for example, ice-cream, a TV, an iPhone, energy, etc).	0.59
CI7	I am curious to learn more about science, computers, or technology.	0.76
CI10	I would like to have a science or computer job in the future.	0.63
CI11	I want to understand science (for example, to know how computers work, how rain forms, or how airplanes fly).	0.75
CI13	I get excited about learning about new discoveries or inventions.	0.71
CI15	I pay attention when people talk about recycling to protect our environment.	0.59
CI16	I am curious to learn more about cars that run on electricity.	0.66
CI20	I like online games or computer programs that teach me about science.	0.65

Note: The factor loadings for all items are based on a sample size of 7,521 elementary and middle school students.

Fit Index = 0.952; Root Mean Square Residual = 0.025; Standardized Root Mean Square Residual = 0.032; Root Mean Square Error of Approximation = 0.065; Normed Fit Index = 0.955) indicated an acceptable model fit (Hooper et al., 2008). These results strongly supported the unidimensionality of the 10-item test, with all items measuring the same construct. In addition, all 10 items contributed similarly to that measurement. This justifies the use of a straightforward scoring method for the Common Instrument – simply adding, or averaging, the 10 item scores.

IRT characteristics

Figure 2 shows the Conditional Standard Error of Measurement (CSEM) function of the Common Instrument. The CSEM is an inverted function of the test Information Function; it estimates the amount of error in theta estimation for each level of theta. The Common Instrument was able to estimate theta at close to 0.3 in a range between about −2.5 and 1 theta. Beyond this range, floor and ceiling effects set in.

Inspection of the category response function plots for each of the 10 items revealed appropriate item functioning. The discrimination of the items ranged between 0.76 and 1.44. According to Baker's (2001) taxonomy, most items would thus be considered to fall into the 'moderate' range, with the highest item (Item 7) being in the 'high' range. The item cut points show an ample spread, which is reflected in the spread of the item means (from 2.58–3.35). Table 2 presents the discrimination (a) and the three cut points (b1, b2, b3) from the IRT analysis, in addition to the item means.

Differential item functioning

We explored potential DIF by gender, race/ethnicity, and grade/age. In addition, we examined the stability of the item pattern over time.

Gender

Separate IRT analyses were carried out for males (n = 2,610) and for females (n = 4,911), and Figure 3A plots the discrimination coefficients of the 10 items by gender. For each item, indicated by a blue dot, the x-value is its discrimination for males, and the y-value is its discrimination for females. If an item had exactly the same discrimination for males and females, it would be located on the black diagonal line of equality, shown in the figure. The results indicated that the discrimination coefficients generally lined up in close proximity to the diagonal line, which indicated the absence of major gender differences in discrimination.

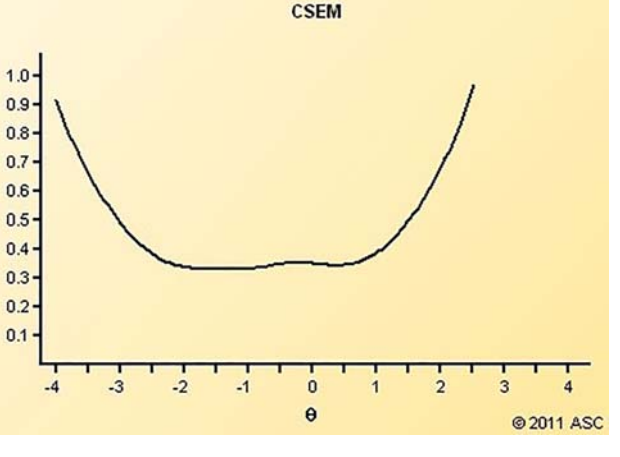

Figure 2. The Conditional Standard Error of Measurement (CSEM) function of the Common Instrument

Table 2. IRT Characteristics of the Final 10-Item Common Instrument.

Item ID	*N*	*M*	A	b1	b2	b3
CI1	7856	2.96	1.20	−1.91	−0.88	0.77
CI3	7794	3.19	1.00	−2.47	−1.43	0.41
CI5	7841	3.35	0.82	−2.97	−1.89	0.03
CI7	7787	3.09	1.44	−1.98	−0.95	0.43
CI10	6295	2.58	0.79	−1.54	−0.11	1.17
CI11	7809	3.05	1.33	−1.91	−0.91	0.51
CI13	7782	3.12	1.21	−2.16	−1.09	0.44
CI15	7762	2.95	0.76	−2.53	−1.02	0.99
CI16	7768	2.88	0.91	−2.08	−0.66	0.81
CI20	6588	2.85	0.85	−2.02	−0.58	0.95

Note: N = sample size, M = mean, a = discrimination, b1 = cutpoint 1–2, b2 = cutpoint 2–3, b3 = cutpoint 3–4.

We also made a rough estimation of the extent to which the items functioned similarly or differently between the genders in terms of their 'difficulty.' Figure 3B presents the cut points for the 10 items as a function of theta from the separate male and female IRT analyses. The closer the cut points for males and females lie on a straight line, the less of a gender difference in the relative difficulty of the items is observable. Again, our results showed an absence of marked gender differences. The cut points line up fairly well along the dotted regression line, which in turn is close to the diagonal of equality. However, in the higher regions of theta, the cut points are overall somewhat higher for females than for males. To some extent, this is an expected effect caused by the lower average science engagement of the females, as compared with the males.

Grade (Age)

For the DIF analysis by grade, we split the sample into two groups: 5th grade and lower (n = 2,147); and 6th grade and higher (n = 4,881). In effect, this split separated elementary school students from middle school students. The analysis followed the pattern described for gender above. The discrimination coefficients lined up well for most of the items (Figure 4A). The graph of the cut points (Figure 4B) shows that the relative 'difficulty' pattern of the Common Instrument is similar for younger and older students. A slight upward displacement is to be expected because of the lower average science engagement of the older students, as compared with the younger students.

Race/ethnicity

The participants reported their race/ethnicity by choosing from 10 different descriptors. For this analysis, we used the survey categories 'African-American/Black' (n = 2,571), 'Asian/Asian-American' (n = 341), 'Latino/a or Hispanic' (n = 1,443), and 'White, Caucasian (non-Hispanic)' (n = 1,161) and collapsed the remaining six categories into one category we labeled 'Other' (n = 1,214). (Those six categories were 'American Indian/Native American or Alaskan Native,' 'Caribbean Islander,' 'Middle Eastern or Arab,' ' Native Hawaiian or Other Pacific Islander,' 'more than 1,' and 'Other.') For our DIF analyses, we compared each of the non-white categories to the white category as the baseline. The discrimination of the items is ordered similarly for the white group and all the non-white groups (Figure 5A). Moreover, the regression lines are close together for all racial/ethnic groups and fairly parallel. For each of the non-white race/ethnicity groups, the 'difficulty' pattern of the items is similar to that of the white group, as indicated by the tight alignment of the cut points to the regression lines (Figure 5B).

Stability over time

We examined whether the item patterns of discrimination and difficulty were conserved when the Common Instrument was administered at the beginning and end of programming. We used data for 3,869 participants for whom we had two tests. The tests were, on average, about 63.5 days or

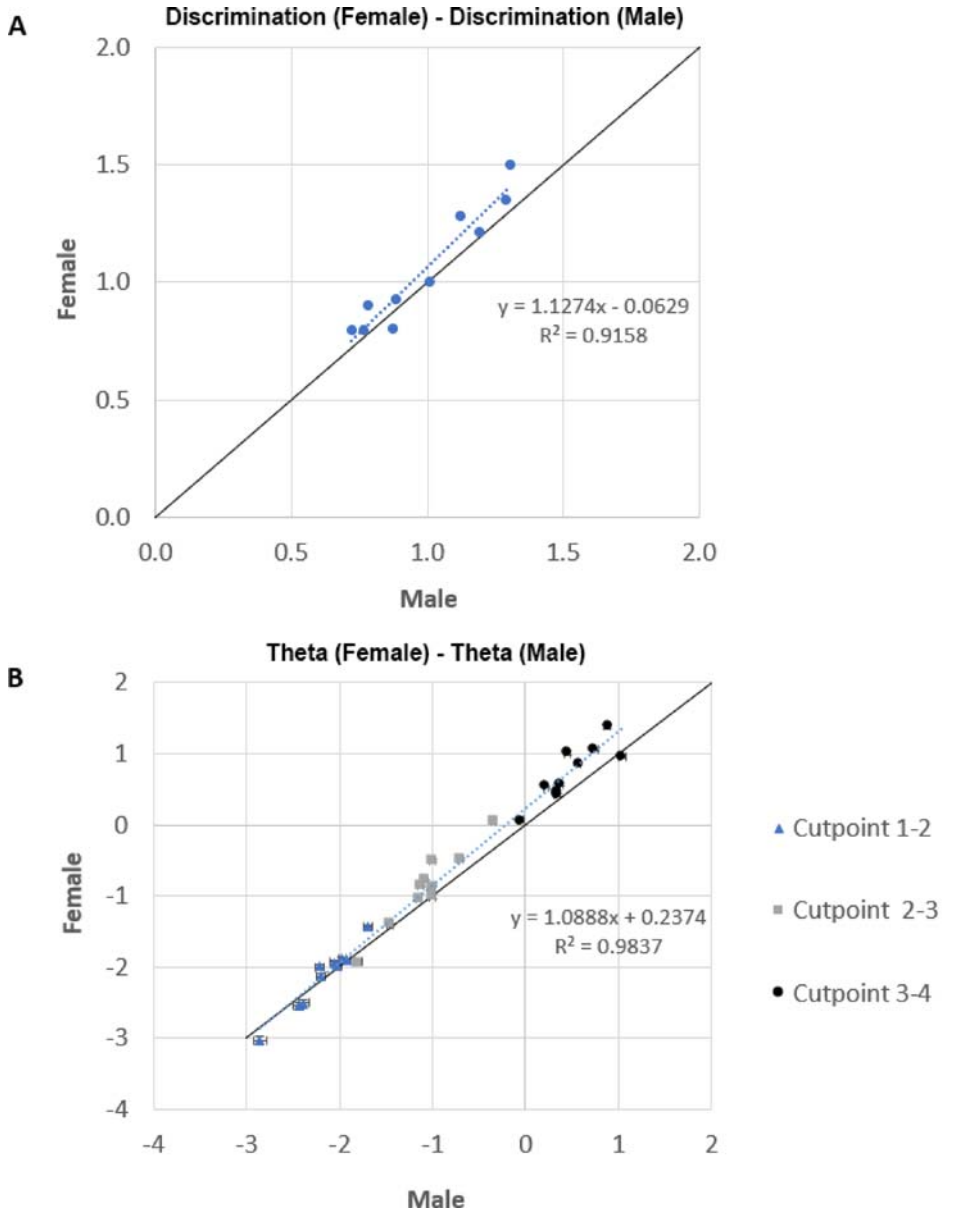

Figure 3. Measurement invariance of the Common Instrument by gender

approximately nine weeks apart. Both item discrimination (Figure 6A) and item difficulty (Figure 6B) were found to be remarkably stable over time.

Discussion

The informal science field is moving toward a common vision around research and evaluation (Allen et al., in press), which requires common measures to aggregate data across a diverse array of OST science programs (Grack Nelson et al., 2019). Researchers and practitioners participating in the Common Instrument project identified a specific need for a measure of science engagement. Although the programs using the Common Instrument were very different in their implementation

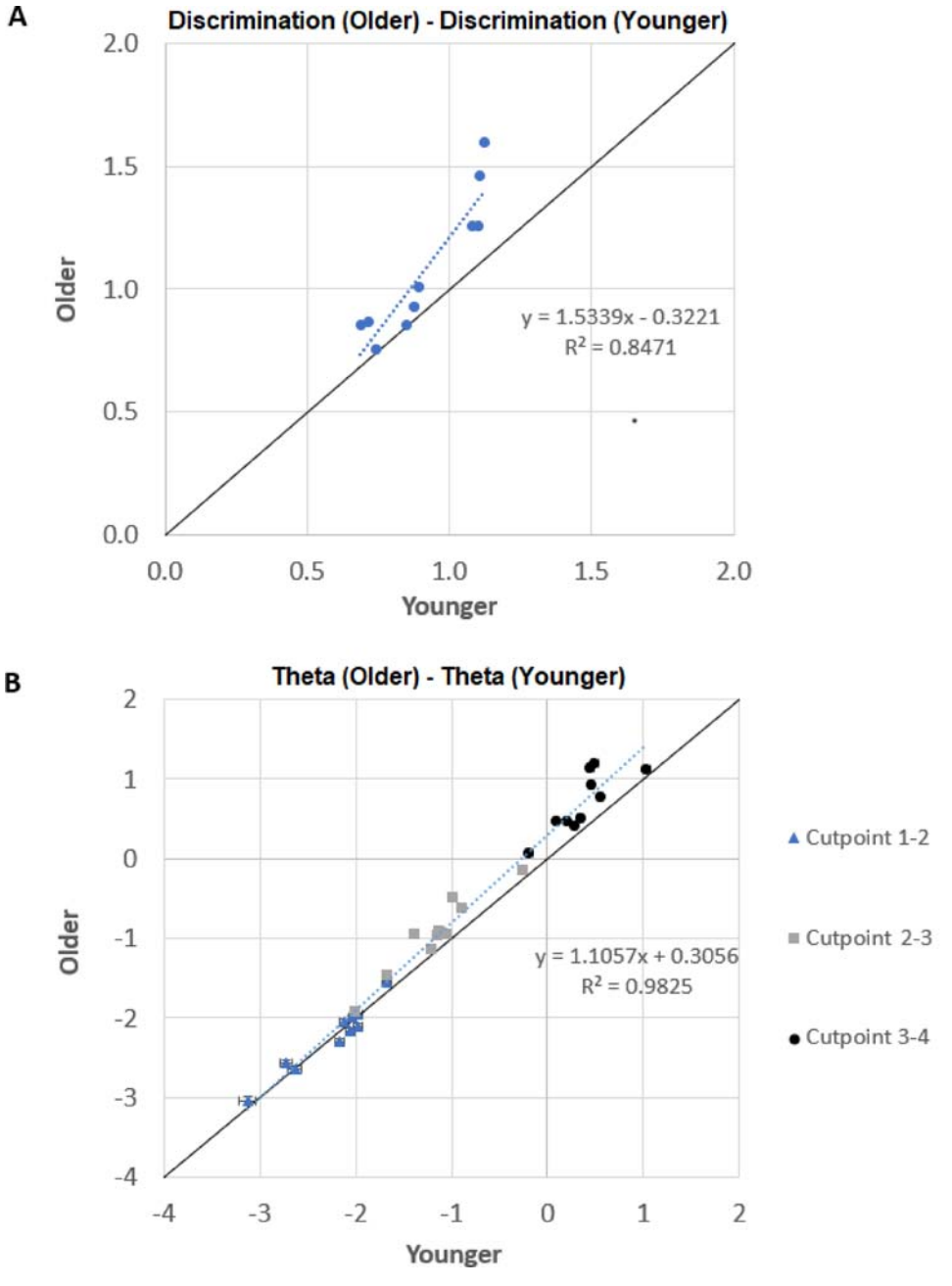

Figure 4. Measurement invariance of the Common Instrument by grade/age

– in terms of curricula, activities, duration, locale, and other variables – all program providers were interested in learning about the youths' experience in their program through the lens of science engagement. This project led to the development of a brief youth self-report survey of science engagement for elementary, middle, and high school youth representing different demographic backgrounds. This study provides evidence that the Common Instrument collects valid and reliable data so that it can be used to measure, communicate, and promote youth science engagement in OST practice and research. A better understanding of factors and processes that influence youth

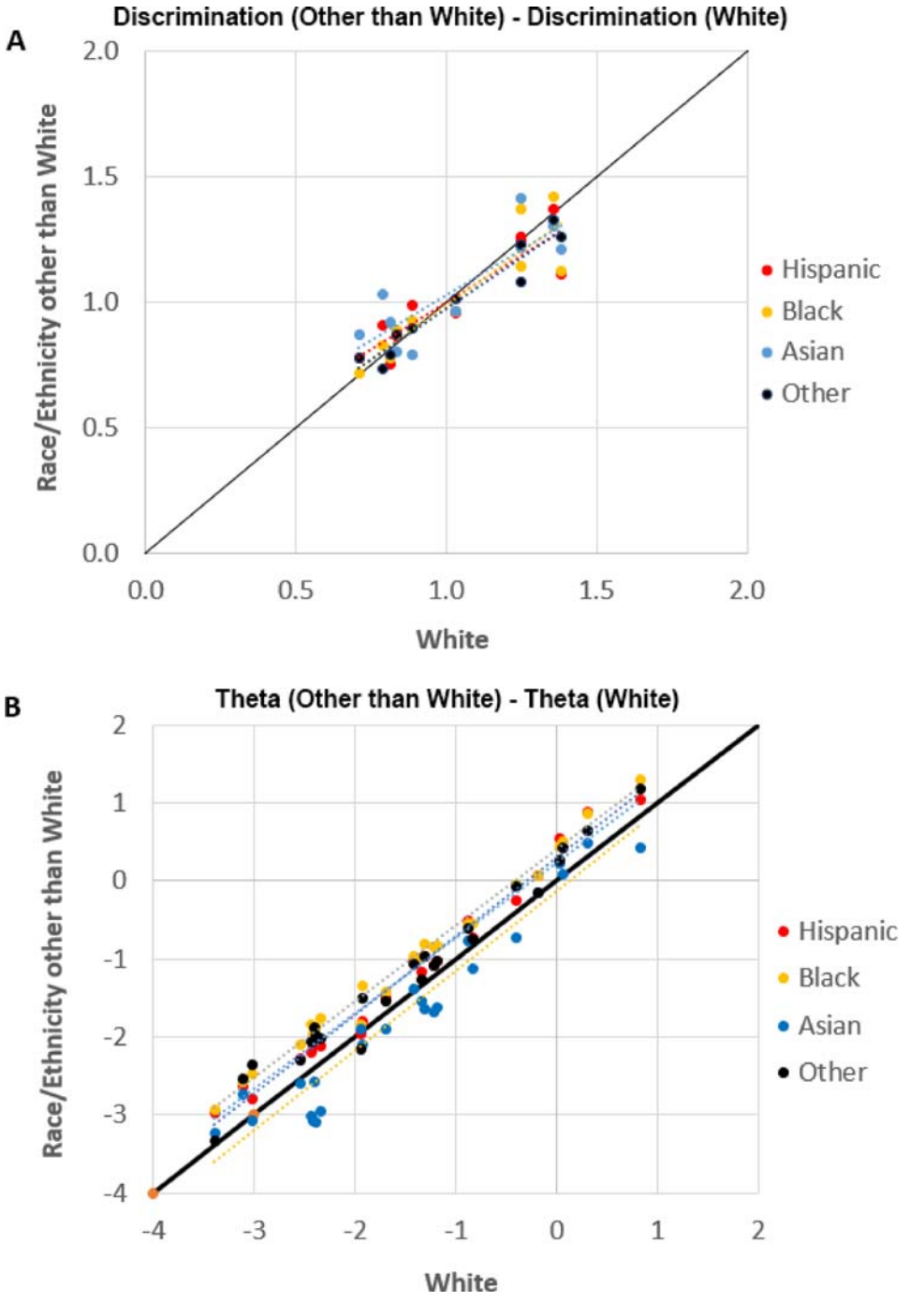

Figure 5. Measurement invariance of the Common Instrument by race/ethnicity

science engagement is essential for identifying and developing educational strategies that can improve science literacy, performance, and persistence throughout life. We discuss how the Common Instrument's research findings are meaningful for educational practice and describe how the use of the Common Instrument can inform future research, practice, and policy.

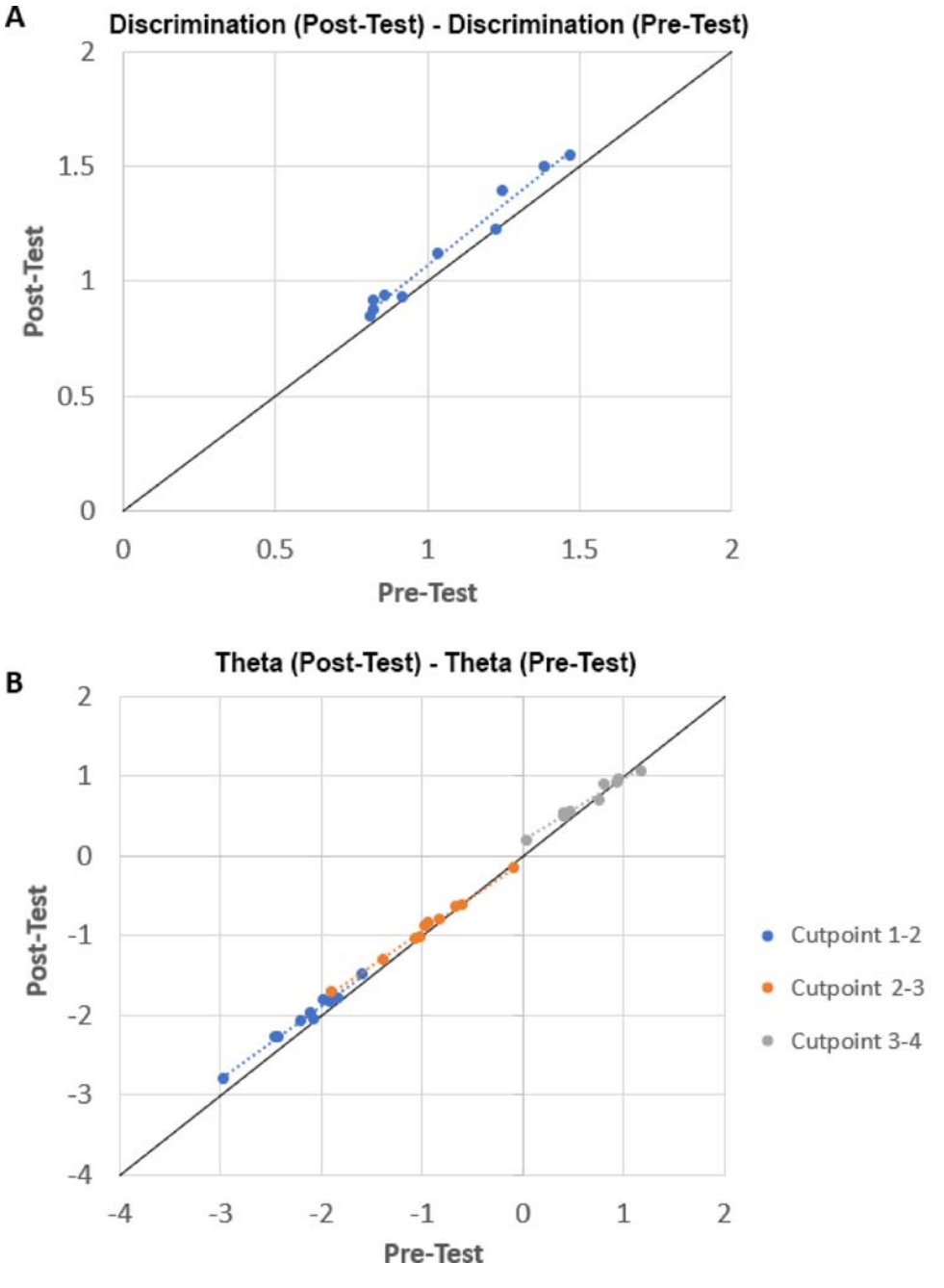

Figure 6. Measurement invariance of the Common Instrument by time (pretest to posttest administration)

Translating Common Instrument research into practice

While our research questions focused on the technical aspects of the Common Instrument project – psychometric questions that require advanced statistical methods to answer – it is important to highlight how a methodological issue can be a relevant educational issue when it comes to communicating and promoting youth science engagement. In other words, the concerns of researchers are shared by educators, and there are lessons that can be translated both ways. The present study

answered three research questions that are necessary to ensure the Common Instrument is psychometrically strong so that it can provide accurate, consistent, and useful information for OST practice and policy. Specifically, to be certain that the Common Instrument could accurately measure youth science engagement, several psychometric tests were conducted to ensure it measures the intended concept, that the items return the same result when repeated, that the items capture the full range of individual experiences, and to ensure the items can be used with different populations of youth.

First, this study establishes validity for the Common Instrument. By engaging practitioners in the research, the measure is likely more practically useful and contextually valid, thus facilitating the communication of evidence-based information to promote youth engagement in science. Practitioner involvement was necessary at all phases of the work to ensure the survey could be flexibly applied across learning settings, including both inside and outside of school as well as across schools and programs to promote the sharing of meaningful results. Practitioner and youth input also helped to ensure that the language was appropriate for elementary to high school-age students. Practitioner feedback also ensured that the survey was consistent with informal learning goals, while being as brief as possible.

Statistically speaking, we found that the Common Instrument had good model fit and captured a global measure of science engagement, which includes cognitive, behavioral, and affective dimensions of science engagement. This is consistent with the continuum view of engagement measurement described by Sinatra et al. (2015):

> … it is difficult to differentiate the dimensions from one another. Cognitive engagement is often described in a manner that intersects with emotional and behavioral engagement. Emotional engagement likely includes cognitive and behavioral elements. Motivation appears to be an aspect woven through each of the dimensions … When researchers measure one dimension of engagement, it is likely their measure is at some level reflecting other dimensions as well. (pg. 8)

While it is valuable to have measures that can parse dimensions of engagement, the Common Instrument project's goal was to create a short tool that is practical for OST science settings, and therefore its global nature (unidimensionality) is both welcome and unsurprising. Studies indicate that a global factor of science engagement is an advantage when considering the practical uses of the instrument because it provides a richer characterization of youth engagement than research on a single dimension and avoids requiring many survey questions that would be impractical to administer, given the time constraints of schools or OST programs (Fredricks et al., 2004). Moreover, Wang et al. (2016) note that ' … a global measure of student engagement may be sufficient for testing policy relevant questions related to the outcomes of STEM engagement' (p. 24). Additionally, surveys that include a few items that tap different but related concepts have more predictive strength than surveys that focus on one narrow concept (Fredricks et al., 2004), meaning that the Common Instrument may provide more useful information on which youth are more likely to continue engaging in science than many measures that focus more narrowly on specific aspects of engagement.

Second, it is important for educators and the OST field to know if youth representing different demographic backgrounds are being engaged equally, especially to broaden participation in science and promote persistence in science among underserved and underrepresented youth (Bell et al., 2009). Previous studies of engagement reviewed by Sinatra et al. (2015) have found individual and developmental differences in the measurement of engagement, which make it very important to ensure that these factors are taking into consideration in the development of measures of engagement. Examination of the results (including IRT characteristics and measurement invariance) by gender, race and ethnicity, and grade showed that the Common Instrument is precise in estimating true scores and that all items of the measure are similarly understood. This is another strength of the Common Instrument, especially because few self-report measures exist for usage in informal/OST science learning settings that can provide valid and reliable data across different demographic

populations, especially by gender, race and ethnicity, and grade. It is important to test for differential item functioning to ensure that any statistical differences detected as a function of gender, race/ethnicity, or grade/age are not due to inherent survey bias, but rather reflect actual differences in ratings between different populations of youths. The literature suggests that there are differences in science engagement, as well as other related science outcomes, by age, gender, socioeconomic status, and race/ethnicity; however, many of these conclusions have been drawn based on the use of measures that have not been tested for measurement invariance or even construct validity. The present study confirms that the Common Instrument can be used to provide a valid and reliable means of assessing similarities and differences in science engagement between different populations (e.g. by age, gender, race).

Third, it is important that Common Instrument items are interpreted similarly over time because programs and initiatives are seeking ways of evaluating the effectiveness of different strategies and interventions on youth science engagement. Stability analyses based on pre-test and post-test responses demonstrated that item characteristics, including discrimination and difficulty, are stable over time – meaning items are interpreted similarly before and after science learning experiences. This evidence supports the use of the Common Instrument to detect change in response to intervention, using traditional pretest-posttest designs. In longitudinal assessment research, there is often a tradeoff between obtaining data of sufficient breadth and complexity and the need for a brief, simple assessment. The Common Instrument provides an efficient alternative to more time-consuming rating procedures (such as observational tools or teacher ratings of youth), which is essential for programs with limited time, staffing, and resources.

Using the Common Instrument to inform practice, research, and policy

The answers to our research questions show that the Common Instrument provides a common conceptualization and measure of youth science engagement, which can be used to support educational practice in several ways. The Common Instrument, which relies on direct input from youth, was designed to make youth partners in the design and implementation of OST programming (or interventions), promoting youth agency in the process. It was important to practitioners involved in this work to align the measure with youth development practices that are foundational to many OST programs (Noam & Shah, 2014). Additionally, the Common Instrument can be used to inform practitioners of which activities or interventions are most engaging, as well as signal to practitioners which groups may be more likely to disengage from science or opt out of future science learning opportunities before the program ends. In this way, the Common Instrument supports continuous improvement efforts of OST science programs, networks, and initiatives. Aggregating data collected during everyday practice, using common assessments like the Common Instrument, can provide strong evidence to guide future directions for research and policy.

Continuous improvement refers to the process of making plans to improve learning experiences (such as preparing curriculum and teaching strategies), checking the return on such efforts (such as examining student test scores, quality indicators, or student survey results), and adjusting educational practices (such as revising curriculum or enrolling in professional development opportunities) to make educational experiences incrementally better for students (Park et al., 2013; Reisner, 2004). Data can act as a catalyst for improvement in this continuous feedback loop, and the idea is that data collected using the Common Instrument will expedite efforts to increase science engagement among youth by providing feedback to programs and networks, even before programming ends.

At the program level, the Common Instrument can help promote science engagement among youth by providing programs with more trustworthy data that will enable them to gauge whether specific choices or investments made – such as in specific curricula, resources, training, or professional development – changed levels of science engagement among participating youth. The measure provides youth participants with a voice to shape or re-shape program-related decisions

that will make the experience more meaningful and relatable to youth. Studies have shown that student self-report surveys are most useful in the continuous improvement process when educators take a reflective, formative, student-centered approach (Allen & Noam, 2016; Noam et al., 2017). For instance, the survey can be used proactively to inform practice at the start of the program, module, or school year, which would allow educators to determine the students' level of engagement with science from the outset. The survey can also be used to track changes in science engagement over time to determine whether changes made during a program or over the school year had any effects on youth. The tool is also short enough to be bundled with other measures or indicators important for focused evaluation and research. Implementing the Common Instrument in a continuous improvement process can signal to educators – both inside and outside of school and at the start and end of programming – when adjustments to materials or practices are needed to enhance the science engagement among children and youth.

At a systems level, there are a number of national initiatives with a mission to increase the quality and quantity of STEM learning opportunities for youth – including hundreds of communities affiliated with state afterschool networks and STEM learning ecosystems – that have adopted continuous improvement models that rely on common measures to aggregate and communicate findings (Mott Foundation and STEM Next, 2018; Noam et al., 2017; Traill & Traphagen, 2015). Embedding common measures of STEM program quality and youth and educator outcomes is an essential ingredient for the success of complex collective impact initiatives and systems because, according to Parkhurst and Preskill (2014):

> To truly evaluate their effectiveness, collective impact leaders need to see the bigger picture—the initiative's many different parts and the ways they interact and evolve over time … rather than attempting to isolate the effects and impact of a single intervention, collective impact partners should assess the progress and impact of the changemaking process as a whole … rather than use performance measurement and evaluation to determine success or failure, collective impact partners should use the information they provide to make decisions about adapting and improving their initiative' (p. 17)

To assess the impacts of one such collective impact initiative – the STEM Learning Ecosystems Community of Practice (SLECoP) – Traphagen and Traill (2014) offered a logic model that relies on a common vision to measure progress for the initiative to advance opportunities for young people to succeed: 'Shared vision, priority outcomes, common language and agreed-upon measurements are needed for ecosystem cultivation' (p. 2). Considering that many communities are balancing their own local evaluation goals and interests with those of regional and national initiatives, a common vision around measurement requires flexibility, brevity, and precision. The Common Instrument provides a brief but comprehensive measure focused on a topic prioritized by practitioners – science engagement – that can easily be embedded into existing research and evaluation plans. For example, combining the Common Instrument with other measures, such as interest, could help further distinguish among youth who could benefit from more meaningful engagement to begin developing their interest. Common measures, such as the Common Instrument, that are widely used in practice support the OST field's goal to aggregate data that will inform practitioners, researchers, policymakers, and other keys stakeholders of local and national trends in youth science engagement.

Limitations and directions for future research

Our findings need to be interpreted cautiously within the context of study's strengths and limitations. First, this study was limited to a sample of convenience, with an over-representation of youth residing in urban areas who are considered economically at-risk, limiting the generalizability of the findings to other youth populations and contexts. Follow-up studies will examine whether the findings are robust across different demographic populations of youth and across different learning settings, including formal classroom settings, to extend validity findings as well as to develop nationally representative norms. Second, now that construct validity has

been established for the Common Instrument, further work is needed to examine different types of validity, including convergent and divergent validity, based on studies that collect data from other measures of science-related outcomes (e.g. measures of science career interest or identity) and data from multiple informants (e.g. teachers, parents), as well as predictive validity, such as performing regression analyses to examine whether science engagement predicts academic performance or selection of science-related majors or careers. Third, because the present study focused primarily on science engagement, it is not clear whether levels of engagement are STEM domain-specific. Future work will use technology, engineering, mathematics, and computer science-specific variants of the Common Instrument to understand if youths' perceptions of engagement differ across these four STEM domains (Wang et al., 2016). Fourth, it was difficult to separate out the different dimensions of science engagement (i.e. social, emotional, cognitive). Conceptual clarity is important for the definition and measurement of constructs in research because it allows for the precise study of antecedents and outcomes. However, as described above, a survey that covers interrelated concepts has more predictive strength than surveys that focus on one narrow concept, especially in practice where brevity is more desirable. Fifth, more studies are needed to examine engagement longitudinally, in response to intervention, and in conjunction with other fields of study, such as cognitive and affective neuroscience, to understand how learning and engagement are linked with brain function and development (Wang & Degol, 2014). Sixth and lastly, we acknowledge that self-report is an imperfect methodology that is limited by such factors as social desirability bias or dependence on recall. However, youth development is a cornerstone of the many youth-serving organizations that participated in the creation of the Common Instrument. Thus, to understand the experience of youth participating in science programming, and to provide youth with agency to give voice to their program participation experiences, it is essential to obtain information from the participants themselves. To provide a more comprehensive understanding of outcomes, we strongly encourage research studies that bring together multiple sources of information from multiple perspectives, such as those of educators/teachers, family members, and program evaluators.

Conclusions

High-quality measures are needed to improve the implementation of OST science programming. This study describes the validation of a common instrument of science engagement – the Common Instrument – that can be used to measure, communicate, and help promote levels of youth science engagement across the OST field. The Common Instrument has implications for future science education practice, policy, and research. Foremost, we have shown that it is possible to develop an instrument capable of measuring students' science engagement with a collaborative team of practitioners and researchers. This combination of voices allowed for a tight set of questions that can serve as a short screen of self-reported science engagement. At this stage in its development, the measure appears to be an economical way to summarize a great amount of useful information. The use of the Common Instrument in the field will allow for rapid data collection of data across programs, communities, cities, and states, and will make it possible to create a large, national database with continuous updates, as new information becomes available.

Acknowledgements

This research was supported by grants from the Noyce Foundation and STEM Next Opportunity Fund. We would like to acknowledge Ron Ottinger and Dr. Cary Sneider for their guidance and support during the development and validation of the Common Instrument. We would also like to thank participating programs, their students and staff, especially ExpandED Schools (formerly The After-School Corporation, TASC), Techbridge, 4-H, BuildIT, Project LiftOff, Coastal Studies for Girls, and the Maine twenty-first Century Community Learning Centers. The views expressed in this article are ours and do not represent those of the granting agency or affiliated institutions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by grants from the Noyce Foundation and STEM Next Opportunity Fund.

ORCID

Patricia J. Allen http://orcid.org/0000-0001-8753-9938

References

Afterschool Alliance. (2014). *America after 3pm: Afterschool programs in demand*. http://www.afterschoolalliance.org/AA3PM/

Allen, P. J., & Noam, G. G. (2016). *STEM learning ecosystems: Evaluation and assessment findings*. The PEAR Institute: Partnerships in Education and Resilience. http://stemecosystems.org/wp-content/uploads/2017/01/STEMEcosystems_Final_120616.pdf

Azevedo, F. S. (2015). Sustaining interest-based participation in science. In K. A. Renninger, M. Nieswandt, & S. Hidi (Eds.), *Interest in mathematics and science learning* (pp. 281–296). American Educational Research Association. doi:10.3102/978-0-935302-42-4_16.

Baker, F. (2001). *The basics of item response theory. ERIC clearinghouse on assessment and evaluation*. University of Maryland.

Bell, P., Lewenstein, B., Shouse, A. W., & Feder, M. A. (Eds.) (2009). Learning science in informal environments: People, places, and pursuits. *National Research Council*, *4*(1), 113–124. https://doi.org/10.1179/msi.2009.4.1.113

Ben-Eliyahu, A., Moore, D., Dorph, R., & Schunn, C. D. (2018). Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. *Contemporary Educational Psychology*, *53*, 87–105. https://doi.org/10.1016/j.cedpsych.2018.01.002

Chittum, J. R., Jones, B. D., Akalin, S., & Schram, ÁB. (2017). The effects of an afterschool STEM program on students' motivation and engagement. *International Journal of STEM Education*, *4*(1), https://doi.org/10.1186/s40594-017-0065-4

Dabney, K. P., Tai, R. H., Almarode, J. T., Miller-Friedmann, J. L., Sonnert, G., Sadler, P. M., & Hazari, Z. (2012). Out-of-school time science activities and their association with career interest in STEM. *International Journal of Science Education, Part B*, *2*(1), 63–79. https://doi.org/10.1080/21548455.2011.629455

Dorph, R., Cannady, M., & Schunn, C. (2016). How science learning activation enables success for youth in science learning. *Electronic Journal of Science Education*, *20*(8), 49–85.

Eccles, J. S. (2016). Engagement: Where to next? *Learning and Instruction*, *43*, 71–75. https://doi.org/10.1016/j.learninstruc.2016.02.003

Eccles, J., & Wang, M. T. (2012). Part I commentary: So what is student engagement anyway? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 133–145). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_6

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation: Preschoolers' search for explanatory information. *Child Development*, *80*(6), 1592–1611. https://doi.org/10.1111/j.1467-8624.2009.01356.x

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Fredricks, J. A., Bohnert, A. M., & Burdette, K. (2014). Moving beyond attendance: Lessons learned from assessing engagement in afterschool contexts. *New Directions for Youth Development*, *2014*(144), 45–58. https://doi.org/10.1002/yd.20112

Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, *43*, 1–4. https://doi.org/10.1016/j.learninstruc.2016.02.002

Fredricks, J. A., Hofkens, T., Wang, M.-T., Mortenson, E., & Scott, P. (2018). Supporting girls' and boys' engagement in math and science learning: A mixed methods study: SUPPORTING GIRLS' AND BOYS' ENGAGEMENT IN MATH AND SCIENCE. *Journal of Research in Science Teaching*, *55*(2), 271–298. https://doi.org/10.1002/tea.21419

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763–782). Springer US. doi:10.1007/978-1-4614-2018-7_37.

Fredricks, J. A., Naftzger, N., Smith, C., & Riley, A. (2017). Measuring youth participation, program quality, and social and emotional skills in after-school programs. In N. L. Deutsch (Ed.), *After-school programs to promote positive youth development* (Vol. 1, pp. 23–43). Springer International Publishing. https://doi.org/10.1007/978-3-319-59132-2_3

Frejd, J. (2019). When children do science: Collaborative interactions in Preschoolers' discussions about animal diversity. *Research in Science Education*, https://doi.org/10.1007/s11165-019-9822-3

Fuller, K. A., Karunaratne, N. S., Naidu, S., Exintaris, B., Short, J. L., Wolcott, M. D., Singleton, S., & White, P. J. (2018). Development of a self-report instrument for measuring in-class student engagement reveals that pretending to engage is a significant unrecognized problem. *PLOS ONE*, *13*(10), e0205828. https://doi.org/10.1371/journal.pone.0205828

Gibson, H. L., & Chase, C. (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Science Education*, *86*(5), 693–705. https://doi.org/10.1002/sce.10039

Gopnik, A. (2010). How babies think. *Scientific American*, *303*(1), 76–81. https://doi.org/10.1038/scientificamerican0710-76

Grack Nelson, A., Goeke, M., Auster, R., Peterman, K., & Lussenhop, A. (2019). Shared measures for evaluating common outcomes of informal STEM education experiences: Shared measures for evaluating common outcomes. *New Directions for Evaluation*, *2019*(161), 59–86. https://doi.org/10.1002/ev.20353

Guo, J., Marsh, H. W., Parker, P. D., Morin, A. J. S., & Dicke, T. (2017). Extending expectancy-value theory predictions of achievement and aspirations in science: Dimensional comparison processes and expectancy-by-value interactions. *Learning and Instruction*, *49*, 81–91. https://doi.org/10.1016/j.learninstruc.2016.12.007

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). *Structural Equation Modelling: Guidelines for determining model fit*, *6*(1), 8.

Jipson, J. L., Callanan, M. A., Schultz, G., & Hurst, A. (2014). Scientists not sponges: STEM interest and inquiry in early childhood. In J. G. Manning, M. Hemenway, J. B. Jenson, & M. G. Gibbs (Eds.), *Ensuring STEM literacy: A national conference on STEM education and public outreach* (Vol. 483, pp. 149–156). Astronomical Society of the Pacific.

Kennedy, T. J., & Odell, M. R. L. (2014). Engaging students in STEM education. *Science Education International*, *25*(3), 246–258. https://doi.org/1044508

Krishnamurthi, A., Ottinger, R., & Topol, T. (2013). STEM learning in afterschool and summer programming: An essential strategy for STEM education reform. In T. K. Peterson (Ed.), *Expanding minds and opportunities: Leveraging the power of afterschool and summer learning for students* (pp. 133–139). Collaborative Communications Group. http://www.expandinglearning.org/expandingminds/article/stem-learning-afterschool-and-summer-programming-essential-strategy-stem

Lyon, G. H., Jafri, J., & St. Louis, K. (2012). Beyond the pipeline: STEM pathways for youth development. *Afterschool Matters*, *16*, 48–57. https://www.niost.org/pdf/afterschoolmatters/asm_2012_16_fall/ASM_2012_16_fall_6.pdf

Maltese, A. V., & Tai, R. H. (2010). Eyeballs in the fridge: Sources of early interest in science. *International Journal of Science Education*, *32*(5), 669–685. https://doi.org/10.1080/09500690902792385

Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. students. *Science Education*, *95*(5), 877–907. https://doi.org/10.1002/sce.v95.5

Martin, A. J., Way, J., Bobis, J., & Anderson, J. (2015). Exploring the ups and downs of mathematics engagement in the middle years of school. *The Journal of Early Adolescence*, *35*(2), 199–244. https://doi.org/10.1177/0272431614529365

Mott Foundation and STEM Next. (2018). *STEM in afterschool system-building toolkit*. http://expandingstemlearning.org/

Muentener, P., Herrig, E., & Schulz, L. (2018). The efficiency of infants' exploratory play is related to longer-term cognitive development. *Frontiers in Psychology*, *9*, 635. https://doi.org/10.3389/fpsyg.2018.00635

Naftzger, N., & Sniegowski, S. (2018). *Exploring the relationship between afterschool program quality and youth development outcomes: Findings from the Washington quality to youth outcomes study* (pp. 1–47). American Institutes for Research. https://raikesfoundation.org/sites/default/files/washington-quality-youth-outcomes-study.pdf

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. The National Academies Press. doi:10.17226/11463.

National Research Council. (2015). *Identifying and supporting productive STEM programs in out-of-school settings*. National Academies Press. doi:10.17226/21740.

Noam, G. G., Allen, P. J., Shah, A. M., & Triggs, B. B. (2017). Innovative use of data as game changer for afterschool: The example of STEM. In H. J. Malone & T. Donahue (Eds.), *The growing out-of-school time field: Past, present, and future* (pp. 161–175). Information Age Publishing.

Noam, G. G., & Shah, A. (2014). Informal science and youth development: Creating convergence in out-of-school time. *Teachers College Record*, *116*(13), 199–218. https://doi.org/1152603
OECD. (2016). *Key findings from PISA 2015 for the United States*. https://www.oecd.org/pisa/PISA-2015-United-States.pdf
Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079. https://doi.org/10.1080/0950069032000032199
Park, S., Hironaka, S., Carver, P., & Nordstrum, L. (2013). *Continuous improvement in education (p. 48)*. Carnegie Foundation for the Advancement of Teaching. https://www.carnegiefoundation.org/wp-content/uploads/2014/09/carnegie-foundation_continuous-improvement_2013.05.pdf
Parkhurst, M., & Preskill, H. (2014). Learning in action: Evaluating collective impact. *Stanford Social Innovation Review*, *12*(4), 17–19. https://ssir.org/articles/entry/evaluating_collective_impact
Perez, T., Wormington, S. V., Barger, M. M., Schwartz-Bloom, R. D., Lee, Y., & Linnenbrink-Garcia, L. (2019). Science expectancy, value, and cost profiles and their proximal and distal relations to undergraduate science, technology, engineering, and math persistence. *Science Education*, *103*(2), 264–286. https://doi.org/10.1002/sce.21490
Price, L. (2011). *Differential item and test functioning and language translation: Differential item functioning and language translation : a cross-national study with a test developed for certification*. LAP, Lambert Academic Pub.
Rabi, I. I. (1965). Science in the satisfaction of Human aspiration. In H. Jordan, & E. Kone (Eds.), *The scientific Endeavor: Centennial celebration of the national academy of sciences* (pp. 303–309). Rockefeller University Press.
Reeve, J., & Tseng, C.-M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology*, *36*(4), 257–267. https://doi.org/10.1016/j.cedpsych.2011.05.002
Reisner, E. R. (2004). Using Evaluation Methods to Promote Continuous Improvement and Accountability in After-School Programs: A Guide. *Policy Studies Associates, Inc.*
Renninger, K. A., & Hidi, S. (2016). *The power of interest for motivation and engagement*. Routledge.
Shah, A. M., Wylie, C., Gitomer, D., & Noam, G. G. (2018). Improving STEM program quality in out-of-school-time: Tool development and validation. *Science Education*, *102*(2), https://doi.org/10.1002/sce.21327
Shernoff, D. J. (2010). Engagement in after-school programs as a predictor of social competence and academic performance. *American Journal of Community Psychology*, *45*(3–4), 325–337. https://doi.org/10.1007/s10464-010-9314-0
Shernoff, D. J. (2013). Measuring student engagement in high school classrooms and what we have learned. In D. J. Shernoff (Ed.), *Optimal learning environments to promote student engagement* (pp. 77–96). Springer. http://link.springer.com/10.1007/978-1-4614-7089-2_4
Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924
Sneider, C., & Noam, G. G. (2019). The Common Instrument Suite: A means for assessing student attitudes in STEM classrooms and out-of-school environments. *Connected Science Learning*, *11*. csl.nsta.org/2019/07/the-common-instrument-suite.
Tai, R. H., Liu, C. Q., Maltese, A. V., & Fan, X. (2006). Planning early for careers in science. *Science*, *312*(5777), 1143–1144. https://doi.org/10.1126/science.1128690
Traill, S., Traphagen, K., & with Devaney, E. (2015). *Assessing the impacts of STEM learning ecosystems: Logic model template and recommendations for next steps*. Noyce Foundation. https://www.informalscience.org/assessing-impacts-stem-learning-ecosystems-logic-model-template-recommendations-next-steps
Traphagen, K., & Traill, S. (2014). *How cross-sector collaborations are advancing STEM learning*. Noyce Foundation. http://stemecosystems.org/resource-category/key-resources/
Wang, M.-T., & Degol, J. (2014). Staying engaged: Knowledge and research needs in student engagement. *Child Development Perspectives*, *8*(3), 137–143. https://doi.org/10.1111/cdep.12073
Wang, M.-T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The math and science engagement scales: Scale development, validation, and psychometric properties. *Learning and Instruction*, *43*, 16–26. https://doi.org/10.1016/j.learninstruc.2016.01.008
Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business School Press.
The White House. (2009, November 23). *President Obama launches "Educate to Innovate" campaign for excellence in science, technology, engineering & math (STEM) education*. Whitehouse.Gov. https://obamawhitehouse.archives.gov/the-press-office/president-obama-launches-educate-innovate-campaign-excellence-science-technology-en
Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015