Design of a Research-based Assessment for Children's Attitudes and Motivation at Chemistry Outreach and Museum Events

Christopher F. Bauer,[†]* Mary E. Emenike,[‡] Thomas A. Holme[§]

[†]Department of Chemistry, University of New Hampshire, Durham, New Hampshire 03824, United States, *Corresponding author: chris.bauer@unh.edu

[‡]Department of Chemistry and Chemical Biology, Rutgers The State University of New Jersey, Piscataway, New Jersey 08854, United States

[§]Department of Chemistry, Iowa State University, Ames, Iowa 50011, United States

ABSTRACT

Outreach activities have been an ongoing commitment of the American Chemical Society and other scientific organizations, yet formal assessment of outcomes has been infrequent. We have designed a simple, reliable, and robust format for assessing attitudinal and motivational characteristics of participants attending informal public events that promote interest in and understanding of chemistry. In final form, this 6-item survey can be distributed on the front and back of an 8.5 x 5.5 card (for pre- and post- assessment). Challenges in design included construct selection and independence, adjusting language for the grade 3-12 populations, physical form and distribution, efficiency and reliability of completion, and item performance and symmetry. Along with a description of the design process, psychometric characteristics and results of field testing at outreach events are described.

INTRODUCTION AND RELEVANT LITERATURE

Herein we describe the development of a rapid, reliable, multi-dimensional, and theoretically-grounded assessment instrument that can be used by learners as young as third grade to assess the impact of outreach, museum, or informal science learning experiences. The instrument incorporates six mental constructs: attitude, interest, perceived difficulty, efficacy, anxiety, and intelligibility. The items were adapted from previous research instruments and theoretical models of attitude and self-efficacy. The design structure allows for anonymous and unbiased pre/post comparisons of an individual's experience. Implementation may be in the form of a card that participants carry with them or in electronic format.

Valid and reliable survey instruments have been developed to assess curricular outcomes other than content learned (e.g., CSCI - Chemistry Self-Concept Inventory,[1] ASCI - Attitude to the Subject of Chemistry Inventory,[2] MCAI - Metacognitive Activities Inventory,[3] CLASS - Colorado Learning Attitudes about Science Survey[4]). Compilations of instruments such as these with stronger research grounding are being created to simplify access and broaden use (e.g., LASSO - Learning about STEM Student Outcomes).[5] These instruments, intended for college students, guided creation of the survey reported in this paper for younger learners, sampling mental constructs that would be of interest to informal science event developers. The new instrument needed to be straightforward to use in an informal setting and not interfere with participant experience.[6] Because chemistry outreach programs welcome participants of any age, the reading level needed to be appropriate for elementary and middle school aged children who might or might not have help from a parent, guardian, or teacher while completing the survey.

Public outreach intends to bring knowledge of scientific concepts and processes to “the people”. The purposes for these activities are broad.[7] They include supporting the learning of children and adults who may not have access to quality science instruction, encouraging children to consider scientific careers, raising the awareness of the general public about safety or health concerns, and showing taxpayers and voters that science plays important and often hidden roles in supporting their quality of life. Despite the visceral feeling that these purposes have value, the informal science community has been recognizing a need for more rigor in design and analysis for program evaluation.[8-13]

The scope of the challenge lies partly in the variety of informal science learning venues. Public and private institutes, such as museums and zoos, represent the most professional and stable venues for outreach. These institutes and their professional organizations are more likely to have permanent staff and mechanisms for conducting or contracting for evaluation projects. National coordination within the informal science learning community exists, for example, via the Center for the Advancement of Informal Science Education (CAISE)[7] which provides guidance and resources for evaluation. Many academic science departments and colleges offer programming for younger learners through public demonstrations, in-house multi-day science camps, and school drop-ins. These events are often

sustained by the enthusiasm and labor of undergraduate and graduate students or individual faculty members. Recent studies have begun to explore more deeply the experience and learning of these students.[14-19] Professional societies across STEM frequently sponsor outreach programs for student recruiting and public awareness. National Chemistry Day in 1987 quickly expanded to National Chemistry Week (NCW), continuing for a quarter century as an umbrella under which many events have been sponsored.[20] Outreach for STEM is also of international interest.[14]

Assessment of programs and participant outcomes has not tended to be of primary concern to many outreach providers because the substantial logistical challenges of planning and managing events consumes attention, time, and energy. In addition, events that are primarily organized by students are unlikely to have someone on the team who has expertise in, or an awareness of the need for, assessment or evaluation. Reports about many outreach programs fail to mention assessment or only describe an informal approach whereby a few questions are asked of participants as they complete an activity or event. A large fraction of the assessments are post-only, self-report, non-comparison designs.[12] Such designs do not support drawing causal inferences or identifying mechanisms by which any observed changes may have happened.[11-12] Most outreach reports focus on content learning and development of an affiliation with the host discipline (i.e. pipeline feeding).[21-29]

Over the last decade there has been a trend to incorporate stronger assessment designs and instruments in part because of sponsor expectations.[11,30] These reports and others[31] encourage measurement of outcomes beyond content knowledge gains, including emotional characteristics, which are believed to be important in developing affiliation and identity. Discussions of these reports, however, suffer from indistinct consideration of theoretical constructs, e.g., “attitudes” broadly being used to cover everything that is not content.[1] Thus, it would be valuable to create assessment tools that are structured to account for, in a fundamental way, the multidimensionality of human response.

A few reports have described more rigorous approaches, including selection and/or design of instruments, links with theory, and establishment of validity and reliability. The community of public educational institutions, such as museums and zoos, is one source of these studies. For example, after watching behaviors of specific zoo animals, visitors rated their emotional responses on eleven characteristics (curiosity, fear, respect, boredom, concern, wonder, amusement, connection, love,

attraction, sympathy, contempt) with unidirectional Likert scales from "not at all" to "very much".[32] The ratings were used in a path analysis model to assess how strongly particular animal behaviors led to affective responses and then to meaning-making. The emotional characteristics were derived from an earlier study which used bipolar scales but without identifying a rationale or theoretical justification for item selection. [33] Another example is the Museum Exhibit Skills Inventory, a behavioral observational tool used by staff to assess children in six skill areas (communication, creativity, collaboration, content, critical thinking, confidence).[34] In another study, the DoVE affective adjective checklist was developed,[35] consisting of 75 adjectives in 15 scales. It was developed through factor analysis, expert review, and reliability checks. Adult visitors to five museums rated their exit responses dichotomously (*yes I feel this, no I don't*). This post-only self-report casts an affective net to seek relationships with other visitor characteristics. The strongest link to modern cognitive motivation theory involved assessing science self-efficacy with a self-designed instrument[36] based on Bandura's work[37] in a pre/post/delayed-post design. The study was pertinent to the effects of a single museum visit.

Another source of more rigorous studies is emerging from the academic professional development community. For example, an attitude instrument originally designed for mathematics learning was used to assess attitudes and motivations of middle and high school students for working in STEM careers.[38] The outreach experience was listening to PhD biology students' elevator presentations. The instrument measured interest, intention to persist, confidence in learning, perceptions of scientists, and perceived usefulness. Interest, confidence, and usefulness are components discussed in this article. Similarly, in a STEM outreach setting, the outcomes for high school students engaged in a Biomechanics Day event were assessed.[39] Glynn's Science Motivation Questionnaire (SMQ)[40] was used along with a self-designed assessment of "attitude" using a semantic differential approach, but limited validation information was presented. The authors tried linking pre/post change for individuals but had difficulty because students did not reliably identify their separate forms. That issue was encountered and solved in the work reported here. Some questions about the applicability of the SMQ have recently been raised.[41]

In sum, past approaches have contained unresolved limitations. Most of the approaches described above were post-only, no-comparison-group approaches that required an encounter or activity to happen before being queried about it. Thus, whether the events caused a change in participant perspectives was not knowable. Further, assessment tools often were developed without theoretical justification, were adapted in unclear ways, or were not put through a rigorous process to evaluate validity and reliability.

In this article, we describe a simple, theoretically-grounded instrument to assess six facets of participant response to an informal learning experience including aspects of attitude, interest, perceived difficulty, efficacy, anxiety, and intelligibility. The instrument has been tested for construct validity and independence, checked for subtle sources of bias, and field tested for usability by different young audiences. The process of design and initial application are described in this manuscript.

RESEARCH QUESTION

To what extent can an assessment instrument for outreach events be designed to sample multiple theoretical constructs with fidelity, provide anonymous pre/post data regarding individuals to allow robust conclusions about effects of the event, and gather a complete response by school-aged children within a few minutes.

INSTRUMENT DESIGN

Overview

The intent was to design an instrument that would be useful for outreach events for children, such as provided by academic institutions or professional societies. By providing a strong grounding in theory, rigorous validity and reliability characteristics, and a format that allowed for easy administration for pre/post comparison designs, users would be able to strengthen outcome claims. Furthermore, an instrument was desired that was straightforward, inexpensive, and efficient to administer and analyze for the most likely user population. For these reasons, a semantic differential structure was chosen, in which pairs of contrasting words are placed at opposite ends of a rating

scale.[1,42] This format has the advantage over traditional Likert-scale statements that it narrows the focus for younger readers to just two words.

Design and refinement of the instrument was conducted in three phases (Figure 1) with several groups of students (3rd grade through college-level). The phases are described by setting: National Chemistry Week events; Science Week events, and Peer Leader Training sessions. This study was approved by Iowa State University's Institutional Review Board.

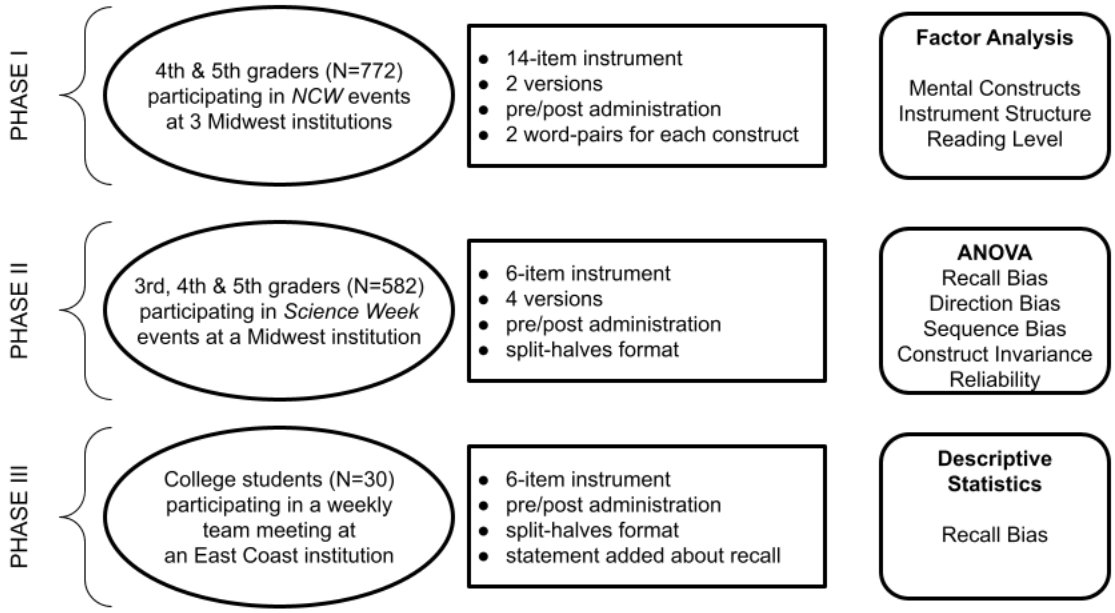

Figure 1. Overview of instrument design process

Survey Versions and Participants

In Phase 1, an initial 14-item survey was administered at NCW events in Fall 2010 at three institutions in the Midwest. Three groups were involved: One event included school children in grades 4 to 6 transported to a university for chemistry activities (N ≈ 778), another was a group of happenstance visitors (mostly grades 4 through 12) to a science event in a public setting (N ≈ 31) in another state, and the third was a group of students at a public outreach event (N ≈ 24) in a third state. The survey was printed on a half sheet of 8.5" x 11" cardstock. Each participant was given a survey card at the beginning of the event and carried it with them (see Supporting Information). The

pre survey was on one side (with the text “COMPLETE THIS SIDE FIRST” at the top) and the post survey was on the reverse (with the text “COMPLETE THIS SIDE LAST” at the top). At the bottom of the pre survey side, students were asked to circle “boy” or “girl”, circle their grade level (PK, K, 1, 2, …, 12, 13+), and write in their age. No names or other identifying information was collected.

In Phase 2 (Science Week events), four versions of a shortened, 6-item survey were generated to test for biases in physical design (recall bias, direction bias, sequence bias), as well as for construct invariance and reliability. All four versions (labeled A, B, C, and D) of the revised survey are provided in the Supporting Information. The Phase 2 versions were administered to 3rd, 4th, and 5th graders (N ≈ 700) from three elementary schools within a single district who were participating in a Science Week event at an academic institution in the Midwest. During the event, students engaged in a variety of 25-50 minute, hands-on science activities lasting a normal school day. A different grade level attended each day. Third and fourth grades engaging in two different chemistry activities one after the other. Fifth grade engaged in a single activity. Children were given the cards to complete the “pre” side at the start of their first activity, carried the card with them to the second activity, and completed the “post” side afterwards. One of the co-authors was present for the event to help manage the process. Children could ask for help reading from their teacher or parent chaperones, if needed. The final recommended format (not one of the preliminary, experimental versions) is shown in Figure 2 and provided for use by others in the Supporting Information.

In Phase 3, evidence for the significance of recall bias was obtained with the assistance of 30 second-year college students enrolled in a Peer Leader training course on teaching and learning at an institution on the East Coast. They were presented the pre-side of the survey card (Figure 2) at one point in time and at a later unannounced time asked to recall whether they had seen these words before and to report confidence in that recall.

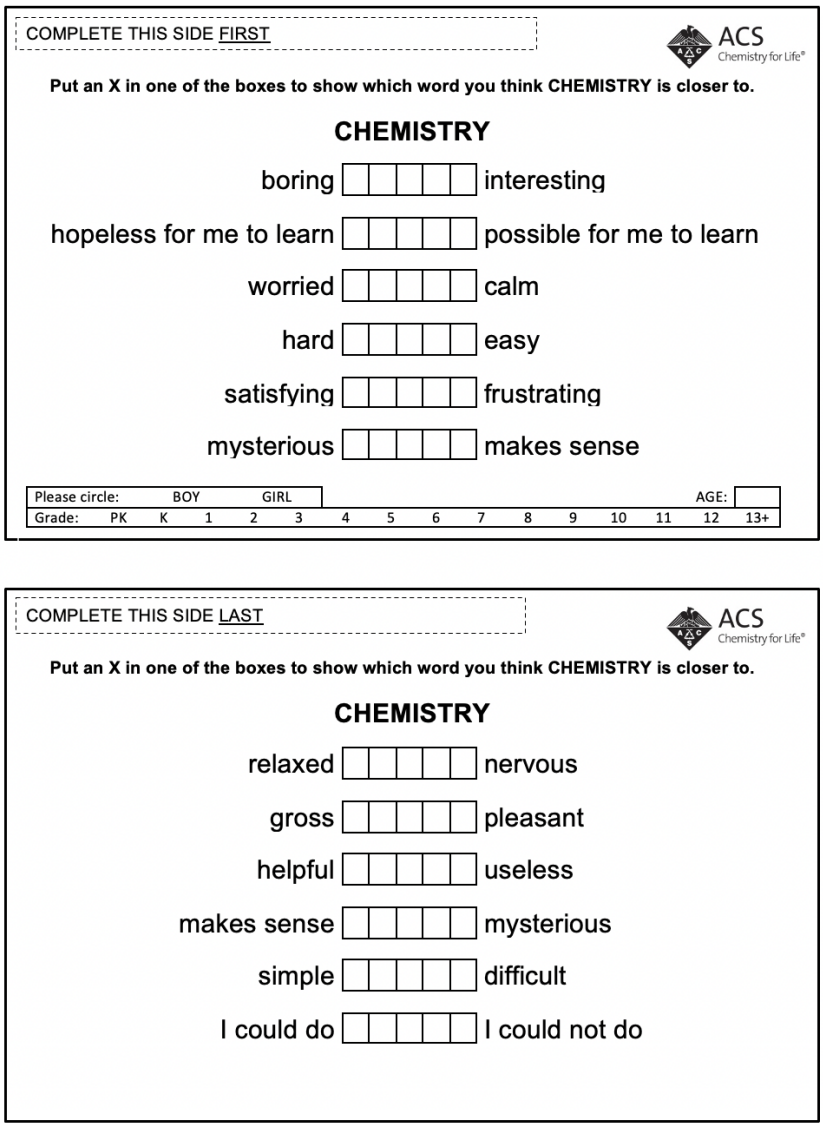

Figure 2. Final recommended version of card survey. This version, along with Versions A, B, C, and D used for hypothesis testing, is provided in the Supporting Information.

Preliminary Design (Phase 1)

Affective Domain Constructs. Mental constructs that outreach designers typically intend to influence or seek information about were included in the survey. These constructs include:

- *Attitude* - refers to the tendency to approach or to avoid an attitude object (in our case CHEMISTRY).[1] This was called "emotional satisfaction" in the original instrument. [2]
- *Interest* - relates to whether the subject sparks curiosity or potential value for the respondent.[4] This was called "interest and utility" in the original instrument.[2]

- *Perceived difficulty* - one's perception of the level of challenge the subject represents, and is an indirect indicator for whether the subject seems "intellectually accessible" (as it was called in the original study).[1, 2]
- *Self-efficacy* - pertains to the sense that one could successfully engage with the subject to accomplish or do something specific with it.[37]
- *Anxiety*[43,44] – relates to nervousness or concern about the subject. For example, students might come away from a chemistry-based event afraid of the subject because of the types of demonstrations viewed or activities they engaged in.
- *Intelligibility* – refers to whether or not the subject makes sense as scientifically explainable as opposed to being mysterious or involving magical-thinking.[45-48] For example, an undesirable outcome would be if participants leave an event with the idea that science is magical or mysterious and without a rational basis.

Instrument Structure. For the semantic differential structure, respondents mark on a linear scale between two contrasting words describing a particular object (here: CHEMISTRY). In developing a new scale, dozens or more pairs of words are presented to responders because one cannot know *a priori* which words will best relate to which mental construct. Factor analysis, or some other reductionist statistical procedure, is then used to reduce the many individual items to correlated sets. In principle, where word pairs are correlated with each other, this indicates a common mental status – a mental construct – in the minds of the responders. Through inspection of the related items, some appropriate term is found to describe what those items seem to have in common.

The ASCI,[2] a semantic differential instrument for assessing attitude, was the primary source for selecting word pairs because that instrument had established validity and reliability, albeit for a college population. Literature on instruments for assessing self-efficacy and motivation were also consulted. A thesaurus was used to explore candidate synonyms and antonyms, particularly to consider words that could be used with students at primary and secondary reading levels. Previous work concerning children's use of language regarding chemistry was also a guide.[49-51]

Reading Level. To establish whether elementary through middle school children would respond to the survey items in a way that demonstrated they were distinguishing different mental constructs, an initial survey, with fourteen word pair items, was deployed at NCW events at three institutions in the Midwest. The collected data were subjected to factor analysis to determine whether the survey items performed as intended (i.e., word pairs correlating highly for a common construct and poorly between other constructs). The same word pairs appeared on both sides of the card, but in a different vertical order and reversed in left/right position. This was a preventive strategy to minimize recall bias, although we did not have direct evidence for our concern at that time. Vertical order on each side of the card was determined by random number selection. Two card versions were distributed. The alternate version contained some items that were the same as the first version, but a few items were exchanged in order to test additional word pairs. The number of cases included in the factor analysis was ~400 or ~800, depending on whether the word pair appeared in one or both forms. Results of the factor analysis guided decisions on word pairs to use and for streamlining survey form.

Design Refinement and Bias Testing (Phase 2)

Results from the first phase suggested that the form had too many items for younger children. There was also suspicion that post answers may have been influenced by pre-responses because word pairs were the same on the pre and post sides of the card (though arranged differently). Specific evidence of this issue was not observed, but it was desired to minimize the risk that post responses would be influenced by either recall of pre responses or by flipping the card over to look at the pre responses. Either situation throws doubt on pre/post differences being due solely to the event experience.

To build in stronger response blindness, split-half versions of the survey were created, putting one word pair for each construct on the front of the card (for pre responses) and the other word pair on the reverse (for post responses). In order for this to work, the pre and post word pairs had to have a reasonably strong construct relationship with each other. This means that at the pre and post time points, only a single response item per mental construct is included.

This trimming does sacrifice psychometric characteristics (losing the noise-minimizing advantage of averaging over multiple items for each construct). Given that the target audience is young children,

a short length, to which participants might give their full attention, perhaps without much adult assistance, was desirable. Furthermore, this structure minimizes data processing effort for event planners. A final advantage of the doubled-sided, paper card is that both pre and post responses are physically linked without the need to identify the respondent, a challenge mentioned by a previous author.[39] This is very helpful for studies with children where identification can be an issue. We argue that simplicity, efficiency and anonymity are important for encouraging utility.

The split-half design decision introduced additional concerns regarding the potential for response bias due to pre/post item recall, left/right alignment, and sequencing. It was important to minimize as much as possible the chance that the arrangement of items on the survey would lead to systematic bias. In other words, the goal was for any observed pre/post changes to be due to event participation and not the survey structure itself.

For the Phase 2 Science Week event, each day hosted a different grade level (3rd, 4th, 5th). On each day there were three cohorts of students starting at different times (in classroom groups), but otherwise all participated in the same activities in their grade level. We assumed that students would likely not be different based on time of participation, so we assigned the four survey forms (cards) to different time cohorts: 10 am for Card C, 11 am for Cards A and B, 1 pm for Card D. The structure of each Card version is in Table 1, indicating whether similar or different words are used on the pre- and post-surveys, and whether or not the word-pairs were presented in reverse (left/right) order.

Reliability. The comparison of Cards A and C on the pre-side, which are identical, provided a determination of two things: the reliability of items on repetition and the homogeneity of the student population. It was important to determine whether our assumption that the student cohorts were equivalent was justified.

Table 1. Survey word pairs and positioning on the four card survey forms, with abbreviated identifiers

Card Label	*Pre Side Layout*	*Pre Side Version (V)*	*Post Side Layout*	*Post Side Version*
A	First split-half	V1	First split-half reversed	V1Rev
B	Second split-half	V2	Second split-half reversed	V2Rev
C	First split-half	V1	Second split-half reversed	V2Rev
D	Second split-half reversed	V2Rev	First split-half	V1

Construct Invariance. The comparison of Cards A and C with B on the pre-side provided for assessment of construct invariance. The same six constructs are represented but with different words (the opposite split halves). Factor analysis in Phase 1 established that relationships did exist between the word pairs, but here we attempt to confirm this with a different physical format and different student population. If the construct relationship is robust, the results with Card B wording should be the same as that for Cards A and C.

A second check on construct invariance is to look at the correlation matrix for all card items pre and post together. One would expect that the word pairs belonging to the same construct (which will be on opposite sides of the card) will have larger correlations than those for different constructs.

Direction Bias. The concern is that particular directions of word pairs (left/right) might create subtle biases for unanticipated reasons. One way in which this could occur is if all adjectives align with a positive sense to one side and negative sense to the other. Thus, in these revisions of the survey, items alternate direction in a random way down the cards.

A second concern is whether it matters which specific word is on left and which is on right. Comparison of Cards B and D pre-side allow for this determination as they consist of the same words but in different left/right orientation.

Sequence Bias. Because the decision was made to use different words on the pre and post forms, an important question is whether it matters which split-half is on the pre side. It might it be possible that encountering the words in one order leads to a bias in responses that is not apparent for the opposite sequence, particularly when an event or experience occurs between the two administrations of the survey. The comparison of Cards C and D allows for this determination because the pre and post sides are direct reverses. If the order does not matter, the event effect (the post-pre difference) would be expected to have the same distribution of responses for the experience. A statistical comparison of the mean differences and distribution shapes would provide an indicator in this case. Note that comparing by subtracting Cards C and D differences and then looking for an overlap with zero was not valid because different students responded to the two different cards forms.

Recall Bias. Students might remember “pre” choices when marking “post” choices and be cued by the words or positions being the same; they might then adjust “post” choices purposely or

subconsciously. This is clearly a risk with the two-sided card because the pre-survey result is in the hands of the participants. Because the left-right order of the word-pairs on Phase 2 survey versions were switched, recall bias could not be assessed directly. Therefore, a separate experiment was set up (Phase 3) in which college students were presented the front side of Card C during a class session (six word pairs) with no other instructions other than to "mark an X in a location between each word pair" and the admonition not to go straight up/down vertically (see Supporting Information). They were not prompted to consider a context while responding, and nothing was mentioned about a subsequent task. Without warning, an hour later, they were presented seven word pairs in a different vertical placement. Two word pairs were repeats of pairs on the front side in the same left-right orientation. The other pairs were the parallel word pairs as established by factor analysis. These latter pairs were all reversed in direction compared to the original list. The expectation was that word pairs that were simply repeated would be recalled accurately by more people and that confidence in these judgments would be higher. Whereas, for the parallel and reversed word pairs, simple recall was not possible and, thus, lower accuracy and lower confidence was expected.

RESULTS

Results are presented for each phase of development, eventually resulting in a recommended survey format.

Phase 1: Factor Analysis (14-item survey format]

Factors were extracted via Maximum Likelihood procedure with varimax rotation. Solutions were forced to extract two to five factors. Separate analyses were run for the pre-event data and post-event data. The data were inspected to determine whether or not the items that were expected to behave similarly did in fact do that. This was accomplished by first observing whether the paired items loaded strongly on a single factor (factor loadings greater than 0.5), and whether, as the number of factors extracted was forced to decrease, the item pairs continued to load together on a common factor. Table 2 lists each item pair categorized as:

- "robust and exclusive" when loadings were greater than 0.6 for just one factor
- "robust but broad" when loadings were 0.2 to 0.5 on more than one factor
- "non-robust" when loadings were not consistently on a single factor

Pearson correlation between the items is also listed in Table 2. The pattern of loadings and correlations reported here suggests that elementary students interpret the meaning and word relationships in a manner similar to college students reported previously (ASCI).[2]

The factor analysis informed the culling of some constructs and/or word-pairs from the instrument. The *safe/dangerous* item, for example, tended to load with calm/worried, indicating that "Fear" might not be distinguishable from anxiety for elementary students; therefore, this pair was removed. The *out of/in control* item did not correlate well with other items and was also eliminated. The distribution of responses on the *real/pretend item* showed nearly all individuals responding strongly to the "real" side of the scale both before and after the event, suggesting a ceiling effect for this item. The thought was that *mysterious* and *pretend* might be interpreted the same way by children, but the data do not support that. Consequently, *real/pretend* was dropped. However, the *makes sense/mysterious* item showed a spread of responses, which suggested that this scale might have utility for monitoring whether an experience with science moved participants away from magicality and more toward rationality. This item was labeled "intelligibility".

Responses from the hundreds of students at the Phase 1 NCW events using the earliest version of the surveys are shown in two complementary ways illustrated in Figures 3 and 4. Figure 3 shows how the tested word pairs do or do not align, supporting the factor structure data in Table 2. Note that the range of individual responses is typically wide (2 units on the scale). The mean response scores show movement pre to post in a desirable direction on nearly all items except *real/pretend* and *helpful/useless*, for which pre-post means were unchanging. For example, *simple/difficult* showed a substantial change in a positive direction, as did *easy/hard.* Because those two items belonged to the same construct, similar movement was expected. On the other hand, the adjectives *real/pretend* and *mysterious/makes sense* were intended to align, but *real/pretend* showed no difference pre-post and was not found to be robust in the factor analysis with *mysterious/makes sense*.

Table 2. Robustness of semantic differential word/pairs in trial with pre-college students

Putative construct	Word pairs	Factor Analysis	Correlation (pre, post)	Instrument Development Decision
Perceived difficulty	simple/difficult easy/hard	Robust and exclusive	0.86, 0.85	retained both
Self-efficacy	possible for me to learn/ hopeless for me to learn I could do/I could not do	Robust and exclusive	0.67, 0.69	retained both
Interest/utility	helpful/useless interesting/boring	Robust and exclusive	0.71, 0.72	retained both
Anxiety	calm/worried relaxed/nervous	Robust and broad	0.50, 0.58	retained both
Attitude	satisfying/frustrating pleasant/gross	Robust and broad	0.71, 0.68	retained both
Intelligibility	real/pretend makes sense/mysterious	Non-robust	0.06, 0.09	removed: ceiling effects retained
Fear	safe/dangerous out of control/in control	Non-robust	0.31, 0.37	removed: redundant with anxiety removed: little correlation to any constructs

Figure 3. Display of mean pre- and post-responses from pre-college students (N=770) at three National Chemistry Week outreach events in the Midwest. Error bars are plus/minus one standard deviation for individual responses. Matched word pairs are adjacent to each other.

Figure 4 shows another way to visualize the student responses. In this display, the response difference (post minus pre) is shown for two item pairs (anxiety: *nervous/relaxed* and *worried/calm*). On this item, the tendency for the population as a whole is toward less anxiety as a result of the

chemistry experience. Figure 4 also shows that the decrease for girls was a bit greater than that for boys, although this representation does not provide information about the relative starting points for girls and boys on the pre-survey. We did not analyze further the data for sex differences.

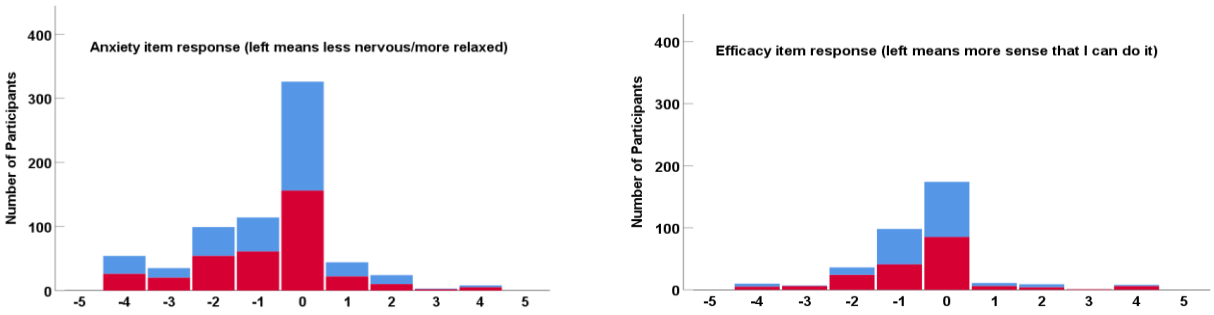

Figure 4. Amount of change in response to efficacy and anxiety items after outreach event (post) versus before the event (pre). Blue bar represents self-identified boys and red bars represent self-identified girls.

Ultimately, eleven word pair items were retained for six constructs: attitude, interest, perceived difficulty, efficacy, anxiety, and intelligibility. For intelligibility, the same word pair was included on both pre- and post-sides of the instrument (makes sense/mysterious) because a good alternative word pair was not discovered in the Phase 1 studies. With these six constructs identified, a new more concise version of the survey was created, and potential biases for that structure were explored in Phase 2 trials by administering multiple versions of the revised instrument.

Phase 2: Direction Bias, Construct Invariance, and Sequence Bias (6-item survey format)

Table 3 shows the number of students responding to the four different survey forms (cards) pre and post, along with demographic breakdown by grade level and sex. It should be noted the students are different individuals in the statistical comparisons between forms A, B, C, and D. Different forms were deployed at different times during the day (with students arriving in three cohorts for 10 am, 11 am, and 1 pm). The number of students using each form was even across the three grades levels (each participating on a different day).

Table 3. Administration conditions for distributed surveys at multi-day outreach event. Structure refers to pre/post version forms as defined in Table 1.

Form	A	B	C	D	Total
Time	11 am	11 am	10 am	1 pm	NA
Structure	V1/V1R	V2/V2R	V1/V2R	V2R/V1	NA
Total Surveys Collected	108	117	229	257	711
Pre Surveys	105	114	205	219	643
Post Surveys	102	114	196	223	635
Matched Pre-Post	102	113	175	192	582
Grade (%) 3|4|5	38.9|33.3|27.8	36.8|35.9|27.4	29.7|46.3|24.0	30.7|30.0|39.3	32.6|36.7|30.7
Sex (%) Boy|Girl |Not Provided	40.7|50.0|9.3	47.9|48.7|3.4	44.1|43.7|12.2	40.5|39.7|19.8	42.9|44.0|13.1

The checks for bias in format involved comparing the mean pre-score results for each word pair for the four card versions by analysis of variance (Supporting Information). Posthoc statistical tests for paired differences were accomplished using Dunnet's T3 test (does not assume equal variances) with $p < 0.05$ as the significance criterion. Table 4 shows mean values with indications of where significant differences were found. Figure 5 shows the data graphically.

Table 4. Mean of student responses for each survey item on each card form. Note that the Form Label order in the table is purposely ACBD because Forms A and C are the same. Common superscripts (a,b) indicate mean values that are non-significantly different (p<0.05).

Form	Interest/Utility	Efficacy	Anxiety	Difficulty	Attitude	Intelligibility
A	1.34^{a}	4.38^{ab}	4.08^{b}	2.46^{b}	1.96^{ab}	3.52^{a}
C	1.37^{a}	4.50^{b}	4.18^{b}	2.68^{b}	2.05^{b}	3.46^{a}
B	1.49^{a}	4.27^{a}	4.02^{b}	2.05^{a}	1.70^{a}	3.55^{a}
D	1.54^{b}	4.23^{a}	3.82^{a}	2.46^{b}	1.92^{ab}	3.32^{a}

Reliability. Table 4 and Figure 5 show that the means for Forms A and C, which are identical on the pre-side, are not significantly different for any of the six items. The average difference between items on the two cards is 0.1 units on the response scale. This magnitude of difference is the same as for the *makes sense/mysterious* item which has the same words pre and post. These observations indicate that the response variation between student cohorts is negligible. As well, the small difference indicates that item reliability is high.

Construct Invariance. The comparison of Forms A and B (pre-side have different words, same construct) provides a means for verifying the parallelism between matching word pairs. Note that

adjective pair orientation in terms of positive/negative sense is the same, as is the vertical position on the card. In Table 4 and Figure 5, B is not different from A except for the Difficulty item and the Attitude item. For the former, B is about 0.5 units lower; for the latter, about 0.3 units lower. This may indicate a different language/word perception between *simple/difficult* and *easy/hard*: That students were attracted more to respond "easy" over "simple" and to "pleasant" rather than "satisfying". It could. however, be that this group of students was primed by something that they experienced prior to their activity. Further review was necessary (described below) before deciding what to do, if anything, about this slight discrepancy. The results overall (Figure 5) suggest, however, that to a large extent, students respond consistently to the different word pairs from the same construct.

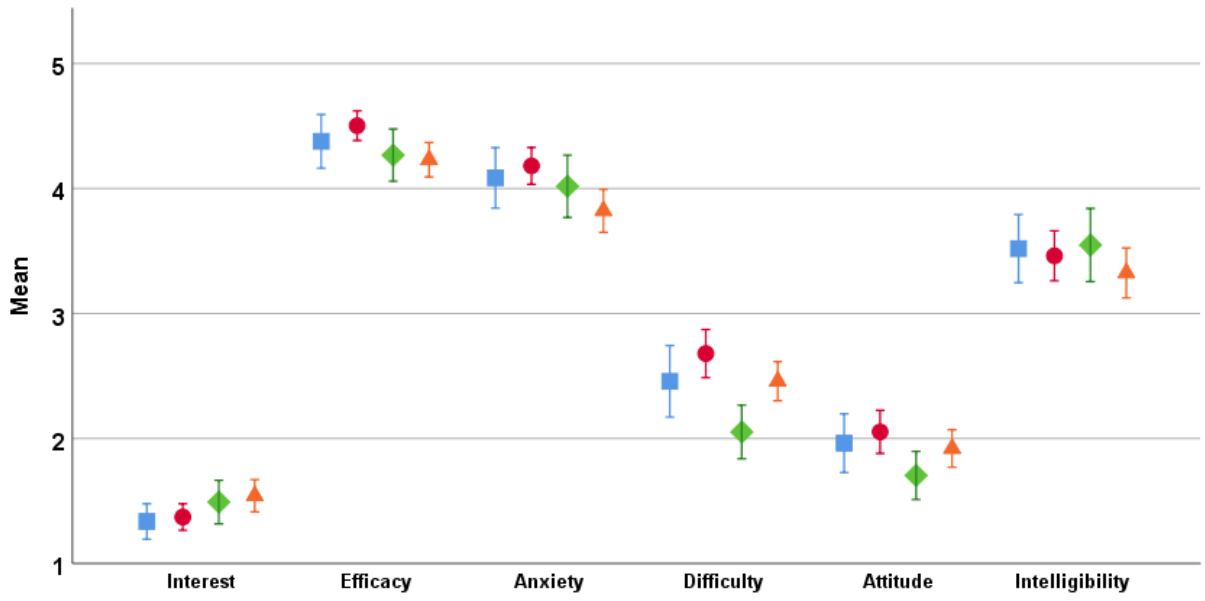

Figure 5. Mean pre-scores of student responses for each of the six survey items for the four different card forms. In each cluster, left to right is form A, C, B, D, respectively. This is not alphabetical in order to make comparisons easier. Error bars represent 95% confidence interval for mean.

A second check on construct invariance was the extent to which item response patterns correlated with each other. We expected that correlations across constructs would be weak if respondents were perceiving them as different ideas. The Phase 2 Science Week student responses were screened to

eliminate anyone who did not provide a response to all six survey items on pre and post sides of the cards, and anyone who chose the maximum rating on all scales. This removed 65 cases and retained 517. The correlation between the pre and post responses for the *same* construct ranged from 0.19 to 0.27 for five of the constructs, all significant at $p<0.01$. The construct of Intelligibility had a correlation of 0.38, but this item used the same words pre and post, likely leading to a somewhat higher relationship. The 30 *between-construct* correlations were smaller, ranging from 0.02 to 0.20, with many being non-significant. Thus, school-aged children appear to respond to the six different constructs as independent entities, a confirmation of the original factor analysis results.

Direction Bias. Forms B and D have opposite pre structure (i.e., same words, but reversed and in different order). Except for the *easy/hard* (difficulty) item (0.4 units different), Forms B and D (having both population and structural differences) showed means that differed by 0.1 - 0.2 scale units. This is slightly larger than the 0.1 scale unit difference for Forms A and C (with only population differences). Students completing Form B, prior to the event, rated chemistry as somewhat easy (mean = 2 on scale of 5). Students completing Form D, presented in reverse order, rated chemistry neutral (2.5 on scale of 5). It was noted above that on Form B this was the item most different from the A and C forms. Thus, there may be a slight increase in noise because of the direction and placement of items, but it does not suggest there is any substantial uncontrolled bias, with possible exception of the *easy/hard* item. This suggests that the D form direction of *hard/easy* would be a better choice than the B form direction because of the apparent alignment with Forms A and C, avoiding whatever the problem is with the *easy/hard* item.

Response means were disaggregated by grade level (not shown) to determine whether anything unique might have happened between the grade level groups (present on different days). All three grade levels reported a more extreme response on Form B vs Forms A, C, and D. This suggests that the unusually low response for Form B may be because of the word order as opposed to some factor related to the event or prior experience.

In this analysis, one could challenge the interpretation of statistical difference in terms of lack of normality of distributions. Figure 6 shows response patterns for all four card versions for the efficacy item (*hopeless/possible to learn* or *could/could not do*) and for the intelligibility item (*makes*

sense/mysterious). The efficacy item shows a skew response. The intelligibility item shows use of Likert scale neutral and extreme positions (1,3,5), with lower use of the intermediate options (2,4). The ANOVA estimate of Type I error is robust when data are not normally distributed, particularly when the sample is large and comparison groups about the same size, which is our situation here. Consequently, the lack of normality does not compromise the ability to make these comparisons. To provide some compensation, the *post hoc* comparison was accomplished with Dunnett's metric, which allows for non-equal variances.

Sequence Bias. One more potential concern is whether it matters which of the split-half word pairs is placed on the pre survey and which is placed post survey. This is a more challenging comparison because post data also includes any effects from the science activity. To compensate for the effect of the activity, the following calculation was done with Forms C and D. Form C has one split-half of the word pairs as pre, and the other split-half as post (format V1/V2R, Table 3). Form D switches this order (V2R/V1). We calculated the post minus pre differences for the students in each condition. The Form D differences were reversed to align response direction with that of Form C. The possible range of difference values thus is -5 to +5. If the pre/post word sequence does not matter, then we would expect the mean and distribution shape of the differences between pre and post to be the same between the two forms. This assumes the student population across cards is uniform (established above) and that their experiences with the chemistry activities were uniform for each group. This is a reasonable assumption as the same activity leaders were involved with each group.

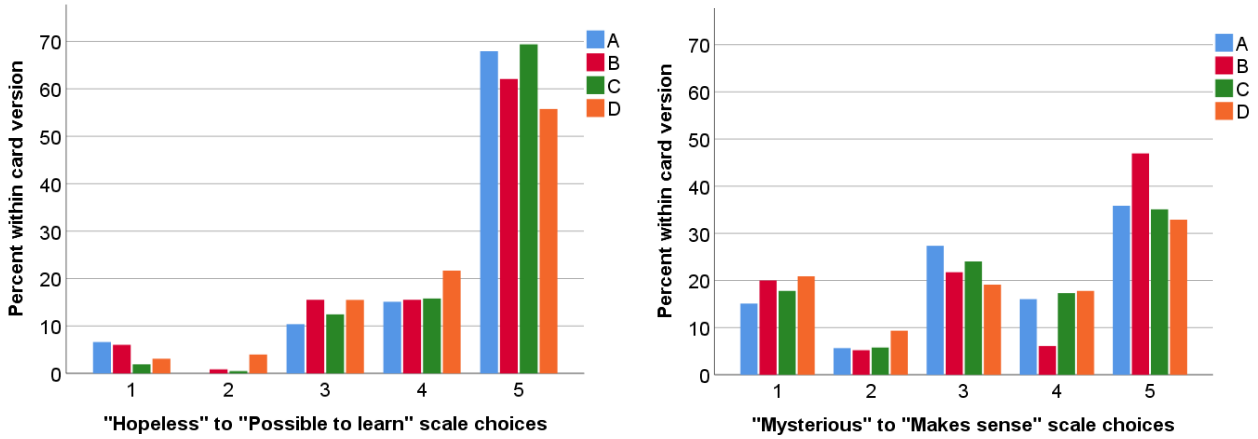

Figure 6. Pre-score response pattern for students on two survey items across four card forms.

T-test comparisons of the means were used to identify whether the two forms differed from each other. Table 5 lists the difference between post-minus-pre on Forms C and D and results of statistical tests. In some cases, the variances of the distributions were not equal, but allowing for unequal variance in the t-test, or also testing by the Kruskal-Wallis nonparametric test, led to the same conclusions. Only the anxiety item demonstrated a substantial and significant difference. Figure 7 shows the results for the anxiety item for Forms C and D. Inspection shows that Form D is skewed to the positive side, including about 20 students with extreme responses. Those extreme responses track to about twenty third grade boys who initially marked the nervous end of the scale, and after the activity marked the calm end, showing a maximum difference. It is possible that something unique occurred at the time when Form D was deployed (1 pm) vs Form C (10 am). Recall that the cards were completed by different students. It is also possible that the item sequence facilitated this extreme response. In that respect, the sequence for Form C is preferred because it did not show those extreme responses. The difficulty item (*hard/easy; simple/difficult*) was borderline significant, with Form C showing more skew. Form D was thus selected as preferred to be conservative.

Table 5. Results of comparison of card Forms C and D concerning potential bias in sequencing of survey items as pre or post. (N Form C = 179; N Form D = 205)

Items for Form C are shown here. Form D has opposite pre/post position **R = reverse scored**	**Mean difference between Forms C and D (scale -5 to 5)**	**Significance (p)**	**Form with lesser skew**
Pst: boring/interesting (R) Pre: helpful/useless	0.06	0.65	Form D
Pst: gross/pleasant (R) Pre: satisfying/frustrating	0.15	0.33	Form C
Pst: could do/could not do Pre: hopeless/possible to do (R)	0.07	0.62	Form C
Pst: relaxed/nervous (R) Pre: worried/calm	-0.76	<0.01	Form C (without outliers)
Pst: hard/easy (R) Pre: simple/difficult	-0.29	0.06	Form D

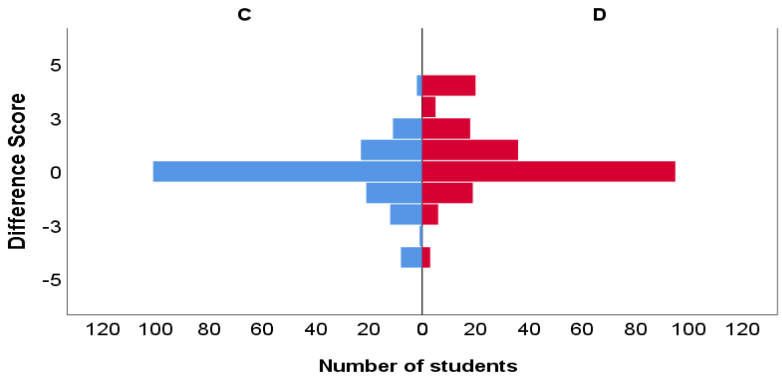

Figure 7. Back-to-back histograms for the anxiety items showing the post-pre differences (i.e. the effect of the chemistry event) for two card forms C (left) and D (right) that have the word-pairs reversed relative to each other.

Phase 3: Potential for recall bias

As described previously, college students were presented the words on Form C (pre-survey) with no context other than to choose a position between two adjectives for six different adjective pairs. An hour after initial viewing, a surprise recall test was presented, for which 2 items were identical to the pre-survey in words and left-right orientation and the other 5 items included the matching word-pair in opposite left-right orientation. The constructs on the second presentation were also in a different vertical order. When the second set of words were identical to the first set, students recalled their earlier response (on a five point scale) with 60-70% accuracy and a reported certainty of judgment of 85%. When the words presented were not the same (but from the same construct), recall accuracy was 30-40% (slightly more than chance, 1 in 5), with reported certainty of 20 to 40%. In other words, accuracy and certainty were much lower when the words were not the same, but substantial when they were the same. Thus, for adult students, recall bias is a legitimate concern. Therefore, the safer choice for this assessment design is to use different, but parallel, words for the pre and post versions of the instrument.

Final Decisions on Survey Format

The final format of the survey is provided in Figure 2. Based on the comparisons above, none of the tested survey formats tested was entirely ideal. The best combination of characteristics seemed to be Form C with the following modifications: (a) For the interest construct, the Form D front/back sequence is preferred -- *boring/interesting* (pre) and *helpful/useless* (post). (b) For the difficulty construct, the Form D front/back sequence is preferred -- *hard/easy* (pre) and *simple/difficult* (post).

At the same time, Form D avoids a potential problem noted with having the word pair *easy/hard* in this left-to-right direction.

Additional Considerations and Limitations

The need for developing this survey was conceived more than a decade ago. At time of first design, cell phone and survey technology did not support sophisticated electronic data acquisition. We have subsequently demonstrated that it is not difficult to convert the survey described here for implementation using the Qualtrics survey platform and cell phones as the response device.

One can then ask, "Why bother with the physical cards?". Particularly, why be concerned about the potential bias possible because of the fixed format of the cards because tools like Qualtrics permit randomization when items are presented in electronic formats. Randomization will help bury bias in the noise only if there are enough respondents. Many outreach programs may not attract sufficiently large numbers of people to assure that condition. In that case, it is valuable to have a tool that, to the extent possible, has been shown to minimize potential format bias. Furthermore, the physical card obviates the need for participants to have access to a cell phone (a likely situation with groups of school children), can be used when wireless service is non-existent or spotty, does not require technical savvy or help to log in, and guarantees pre/post response data linkage while maintaining absolute anonymity. Given the general increased awareness of, and interest in, providing accessible and equitable assessment tools, the concern about access to devices and wireless service is legitimate and should be considered by outreach coordinators and those administering the evaluation. Data analysis and presentation can be done via transcription of paper/pencil survey administration to a spreadsheet and simple graphic and statistical summaries provided.

CONCLUSIONS

These design tests have resulted in a robust, reliable, unbiased, and user-friendly form that is easily administered in informal settings. A strategic design decision was made to target only one survey item for each of six mental constructs. This is not typical for most research surveys because averaging multiple items increases reliability. However, the reality of administering assessments in informal learning environments with young children argues for paying more attention to time, logistical considerations, and usability for respondents and event sponsors. The physical format of the

survey was designed to increase the likelihood that it would be completed thoughtfully, would allow for matched pre/post data collection with anonymity, and would allow collection of simple non-intrusive demographic information. Printed instructions were clean and minimal. Words were selected that research suggested would have meaning for children as young as the earliest readers. Responses can be made with any writing utensil in about a dozen marks (Figure 2). The pre and post versions of the survey were printed on opposite sides of the same card, allowing pre/post data to be linked without a need to identify the respondent. In addition to using this survey in informal science settings, such as museums, this instrument is intended for use by the chemistry education community in the evaluation of outreach events to inform both the design of outreach activities and the training and mentorship of faculty, students, and staff who facilitate these activities. The survey was printed on 8.5" x 11" card stock (two per page and cut into halves) so that it would be durable in the handling that might occur between pre and post responses. A template is provided in the Supporting Information that can be used for creating front/back surveys.

At the time of first trials, assessment of outreach events was in need of improvement. This is confirmed by the reports published subsequent to our initial work and the increased interest of funding agencies.[52] Some progress has been made, but there is substantial room for having additional valid and reliable tools, such as the survey presented in this manuscript.

ASSOCIATED CONTENT

Supporting Information

Phase 1, 2, and 3 surveys design, development, and statistical comparison (doc)

Final survey template (pdf)

ACKNOWLEDGMENTS

All authors acknowledge the participants at the various outreach events who completed the surveys, as well as the outreach event coordinators and volunteers for their engagement with the public, engagement with this research study, and help administering the surveys to participants. CB acknowledges the support of the Ted Ashford Fellowship from the ACS Examinations Institute.

REFERENCES

1. Bauer, C.F. Beyond "student attitudes": Chemistry self-concept inventory for assessment of the affective component of student learning. *J. Chem. Educ.* **2005**, 82 (12), 1864-1870. DOI: 10.1021/ed082p1864
2. Bauer, C.F. Attitude towards chemistry: A semantic differential instrument for assessing curriculum impacts. *J. Chem. Educ.* **2008**, 85 (10), 1440-1445. DOI: 10.1021/ed085p1440
3. Cooper, M. M.; Sandi-Urena, S. Design and validation of an instrument to assess metacognitive skillfulness in chemistry problem solving. *J. Chem. Educ.* **2009**, 86 (2), 240-245. DOI: 10.1021/ed086p240
4. CLASS (Colorado Learning about Science Survey). https://www.colorado.edu/sei/class (accessed 01/01/2021).
5. LASSO (Learning about STEM Student Outcomes). Learning Assistant Alliance. https://www.learningassistantalliance.org/modules/public/lasso.php (accessed 01/01/2021).
6. Barriault, C.; Pearson, D. Assessing exhibits for learning in science centers: A practical tool. *Visitor Studies* **2010**, 13 (1), 90-106. DOI: 10.1080/10645571003618824
7. CAISE (Center for Advancement of Informal Science Education). https://www.informalscience.org (accessed 01/01/2021).
8. Laursen, S.; Liston, C.; Thiry, H.; Graf, J. What good is a scientist in the classroom? Participant outcomes and program design features for a short-duration science outreach intervention in K-12 classrooms. *CBE Life Sci. Educ.* **2007**, 6, 49-64. DOI: 10.1187/cbe.06–05–0165
9. Bell, D. A.; Porter, O. F. Improving access: Determining the effective parameters of academic outreach programs in STEM education. In *4th International Technology, Education and Development Conference (INTED)*, Chova, L. G.; Belenguer, D. M.; Torres, I. C., Eds.; **2010**, pp 2468-2476. Accession Number: WOS:000318805502058
10. Mujtaba, T.; Lawrence, M.; Oliver, M.; Reiss, M. J. Learning and engagement through natural history museums. *Stud. High. Educ.* **2018**, 54 (1), 41-67. DOI: 10.1080/03057267.2018.1442820
11. Fu, A.C.; Kannan, A.; Shavelson, R.J.; Peterson, L.; Kurpius, A. Room for rigor: Designs and methods in informal science education evaluation. *Visitor Studies* **2016**, 19 (1), 12-38. DOI: 10.1080/10645578.2016.1144025
12. Fu, A.C.; Peterson, L.; Kannan, A.; Shavelson, R. J.; Kurpius, A. A framework for summative evaluation in informal science education. *Visitor Studies* **2015**, 18 (1), 17-38. DOI: 10.1080/10645578.2015.1016363
13. MacPherson, A.; Hammerness, K.; Gupta, P. Developing a set of guidelines for rigorous evaluations at a natural history museum. *J. Mus. Educ.* **2019**, 44 (3), 277-285. DOI: 10.1080/10598650.2019.1585172

14. Sadler, K.; Eilam, E.; Bigger, S. W.; Barry, F. University-led STEM outreach programs: purposes, impacts, stakeholder needs and institutional support at nine Australian universities. *Stud. High. Educ.* **2018**, 43 (3), 586-599. DOI: 10.1080/03075079.2016.1185775
15. Pratt, J. M.; Yezierski, E. J. Goodwill without guidance: College student outreach practitioner training. *J. Chem. Educ.* **2019**, *96*(3), 414-422. doi:10.1021/acs.jchemed.8b00882
16. Pratt, J. M.; Yezierski, E. J. "You lose some accuracy when you're dumbing it down": Teaching and learning ideas of college students teaching chemistry through outreach. *J. Chem. Educ.* **2019**, *96*(2), 203-212. doi:10.1021/acs.jchemed.8b00828
17. Pratt, J. M.; Yezierski, E. J. College students teaching chemistry through outreach: Conceptual understanding of the elephant toothpaste reaction and making liquid nitrogen ice cream. *J. Chem. Educ.* **2018**, *95*(12), 2091-2102. doi:10.1021/acs.jchemed.8b00688
18. Santos-Díaz, S.; Towns, M. H., Chemistry outreach as a community of practice: investigating the relationship between student-facilitators' experiences and boundary processes in a student-run organization. *Chem. Educ. Res. Pract.* **2020**, 21, 1095-1109. DOI: doi.org/10.1039/D0RP00106F
19. Blatti, J. L.; Garcia, J.; Cave, D.; Monge, F.; Cuccinello, A.; Portillo, J.; Juarez, B.; Chan, E.; Schwebel, F. Systems thinking in science education and outreach toward a sustainable future. *J. Chem. Educ.* **2019**, 96(12), 2852-2862. DOI:10.1021/acs.jchemed.9b00318
20. National Chemistry Week, American Chemical Society, National Chemistry Week, https://www.acs.org/content/acs/en/education/outreach/ncw.html (accessed 01/01/2021).
21. Lam, C.; Danforth, M.; Mehrpouyan, H.; Hughes, R. Summer engineering outreach program for high school students: Survey and analysis. Presented at 2014 ASEE Annual Conference and Exposition, Indianapolis, Indiana, June **2014**. DOI: 10.18260/1-2--23074
22. Gonzalez-Sola, M.; Rosario-Canales, M. Encouraging Puerto Rican high school students to pursue a STEM career through an anatomy & neurobiology outreach. *The FASEB Journal* **2015**, 29 (S1) 693.6. DOI 10.1096/fasebj.29.1_supplement.693.6
23. Ivey, S. S.; Palazolo, P. J. Girls experiencing engineering: Evolution and impact of a single-gender outreach program. Presented at 2011 ASEE Annual Conference and Exposition, Vancouver, BC, June 2011. DOI: 10.18260/1-2--18026
24. Leas, H. D.; Nelson, K. L.; Grandgenett, N.; Tapprich, W. E.; Cutucache, C. E. Fostering curiosity, inquiry, and scientific thinking in elementary school students: Impact of the NE STEM 4U intervention. *J. Youth Dev.* **2017**, 12 (2), 103-120. DOI: 10.5195/jyd.2017.474
25. Dehipawala, S.; Sullivan, R.; Armendariz, R.; Shekoyan, V.; Tremberger, G.; Lieberman, D.; Cheung, T. Assessment of high-school engineering education outreach program employing project-based learning in astronomy and bio-optics within a college setting. In *Optics Education and Outreach V*, Proceedings of SPIE (International Society for Optics and Photonics), San Diego, CA, Aug 2018; Gregory, G.G. Ed.; **2018**, Vol 10741. DOI: 10.1117/12.2320683

26. Feldhausen, R.; Weese, J. L.; Bean, N. H. Increasing student self-efficacy in computational thinking via STEM outreach programs. Presented at 49th ACM SIGCSE Technical Symposium on Computer Science Education, Baltimore, Maryland, February 2018. DOI: 10.1145/3159450.3159593
27. Wyatt, B. N.; Schram, M.; St. Maurice, M. Are you a scientist? Exploring science identity in a structural biology outreach program. *The FASEB Journal* **2018**, 32 (S1) DOI: 10.1096/fasebj.2019.33.1_supplement.454.25
28. Brown, A. R.; Egan, M.; Lynch, S.; Buffalari, D. Neuroscience and education colleagues collaborate to design and assess effective brain outreach for preschoolers. *J. Undergraduate Neurosci. Educ.* **2019**, 17(2), A159-A167. MEDLINE:31360132
29. Pedrozo-Acuna, A.; Favero, R. J. Jr.; Amaro-Loza, A.; Mocva-Kurek, R. K.; Sanchez-Peralta, J. A.; Magos-Hernandez, J. A.; Blanco-Figueroa, J. An innovative STEM outreach model (OH-Kids) to foster the next generation of geoscientists, engineers, and technologists. *Geosci. Commun.* **2019**, 2, 187-199. DOI: 10.5194/gc-2-187-2019
30. OERL (Online Evaluation Resource Library), Division of Research, Evaluation and Communication, Directorate for Education and Human Resources, National Science Foundation. Faculty Development Instrument 8B. https://oerl.sri.com/instruments/fd/teachwork/instr102.html (accessed 01/01/2021).
31. Haywood, B. K.; Besley, J. C. Education, outreach, and inclusive engagement: Toward integrated indicators of successful program outcomes in participatory science. *Public Underst. Sci.* **2014**, 23 (1), 92-106. DOI: 10.1177/0963662513494560
32. Luebke, J. F.; Watters, J. V.; Packer, J.; Miller, L. J.; Powell, D. M. Zoo visitors' affective responses to observing animal behaviors. *Visitor Studies*, **2016**, 19 (1), 60-76. DOI: 10.1080/10645578.2016.1144028
33. Myers, O. E., Jr.; Saunders, C. D.; Birjulin, A. A. Emotional dimensions of watching zoo animals: An experience sampling study building on insights from psychology. *The Curator*, **2004**, 47 (3), 299-321. DOI: 10.1111/j.2151-6952.2004.tb00127.x
34. Braswell, G.S. Creation and validation of an observational tool to assess children's domain-general skills at museum exhibits. *Visitor Studies*, **2016**, 19 (2), 211-234. DOI: 10.1080/10645578.2016.1220190
35. Packer, J.; Ballantyne, R.; Bond, N. Developing an instrument to capture multifaceted visitor experiences: The DoVE adjective checklist. *Visitor Studies*, **2018**, 21 (2) 211-231. DOI: 10.1080/10645578.2018.1553925
36. Gutwill, J. P. Science self-efficacy and lifelong learning: Emerging adults in science museums. *Visitor Studies*, **2018**, 21 (1), 31-56. DOI: 10.1080/10645578.2018.1503875
37. Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.* **1977**, 84 (2), 191–215. DOI: 10.1037/0033-295X.84.2.191

38. Kompella, P.; Gracia, B.; LeBlanc, L.; Engelman, S.; Kulkarni, C.; Desai, N.; June, V.; March, S.; Pattengale, S.; Rodriguez-Rivera, G.; Ryu, S. W.; Strohkendl, I.; Mandke, P.; Clark, G. Interactive youth science workshops benefit student participants and graduate student mentors. *PLOS Biology*, **2020**, 18 (3), e3000668-e3000668. DOI: 10.1371/journal.pbio.3000668

39. Teeter, S. D.; Husseini, N. S.; Cole, J. H. Assessing changes in attitudes toward engineering and biomechanics resulting from a high school outreach event. *J. Biomech.* **2020**, 103, 109683 DOI; 10.1016/j.jbiomech.2020.109683

40. Glynn, S. M.; Brickman, P.; Armstrong, N.; Taasoobshirazi, G. Science Motivation Questionnaire II: Validation with science majors and nonscience majors. *J. Res. Sci. Teach.* **2011**, 48 (10) 1159–1176 DOI: 10.1002/tea.20442

41. Komperda, R.; Hosbein, K. H.; Phillips, M. M.; Barbera, J. Investigation of evidence for the internal structure of a modified science motivation questionnaire II (mSMQ II): a failed attempt to improve instrument functioning across course, subject, and wording variants. *Chem. Educ. Res. Pract.*, **2020**, 21, 893-907. DOI: 10.1039/D0RP00029A

42. Rosenberg, B.; Navarro, M.A. Semantic differential scaling. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation.* Frey, B.B., Ed.; SAGE Publications, Inc. Thousand Oaks, CA, **2018**.

43. Bowen, C. Development and score validation of a chemistry lab anxiety instrument. *Ed. Psych. Meas.* **1999**, 59 (1), 171-187. DOI: 10.1177/0013164499591012

44. Mallow, J.V. *Science Anxiety: Fear of science and how to overcome it.* Elsevier North Holland Inc., 1981.

45. Blanquet, E.; Picholle, E. Science or Magic? Reactions of 5-Year-Old Pupils to a Counterintuitive Experiment. In *Bridging Research and Practice in Science Education: Selected Papers from the European Science Education Research Association (ESERA) 2017 Conference,* Dublin, Ireland. McLoughlin, E., Finlayson, O.E., Erduran, S., Childs, P.E. Eds.; Springer International Publishing: Cham **2019**, pp. 91-104. ISBN 978-1-873769-84-3

46. Chinn, C. A.; Brewer, W. F. Knowledge change in response to data in science, religion, and magic. In *Imagining the Impossible: Magical, Scientific, and Religious Thinking in Children* Johnson, C.N., Rosengren, K.S., Harris, P.L., Eds.; Cambridge University Press: Cambridge, **2000**; 334-371.

47. Sless, D.; Shrensky, R. Conversations in a landscape of lcience and magic: Thinking about science communication. In *Science Communication in Theory and Practice*; Stocklmayer, S.M., Gore, M.M., Bryant, C., Eds.; Springer Netherlands: Dordrecht, **2001**; 97-105.

48. Subbotsky, E. Magical thinking in judgments of causation: Can anomalous phenomena affect ontological causal beliefs in children and adults? *Br. J. Dev. Psychol.,* **2004***, 22*(1), 123-152. DOI:10.1348/026151004772901140

49. Emenike, M.E. What is a Chemical? Fourth-Grade Children's Categorization of Everyday Objects and Substances. Ph.D. Dissertation, Miami University, Ohio. 2010. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=miami1283196816

50. Gilbert, J. K.; Osborne, R. J.; Fensham, P. J., Children's science and its consequences for teaching. *Science Education* **1982,** *66* (4), 623-633.

51. Osborne, R.; Freyberg, P., *Learning in Science: The Implications of Children's Science*. Heinemann: Auckland, 1985.

52. National Science Foundation. Advancing Informal Science Education program. https://www.nsf.gov/pubs/2020/nsf20607/nsf20607.htm (accessed 03/14/2021)