Ellen Watson, BRANDON UNIVERSITY, watsone@brandonu.ca
Sheryl Gares, UNIVERSITY OF ALBERTA, sgares@ualberta.ca
Brian Rempel, UNIVERSITY OF ALBERTA, brempel@ualberta.ca

Is the Course Working? An Account of Our Development of an Instrument to Measure the Science Attitudes and Skills of Undergraduate Students Outside of Science Disciplines

ABSTRACT

After a redesign of our school year structure, our science team developed an introduction to science course focused on teaching science to non-science majors early in their post-secondary studies. The goal of this course was not to prepare students for further pursuit of science degrees; instead, we wanted to equip them with the skills and attitudes necessary to understand the scientific world in which we live. Consequently, we wondered whether these skills and attitudes were being met throughout the course; was the course working? When searching the literature, we did not identify any instrument that simultaneously and concisely measured general science skills and attitudes. Given this gap and based on our desire to measure science skills and attitudes for non-science majors at our campus, this research team developed Augustana Interdisciplinary Scientific Literacy Evaluation (AISLE) in order to provide a measurement of students' science skills and abilities in a general science course at the post-secondary level. However, as we would come to know, this process was not as simple as might seem. The purpose of this paper is to provide an account of the development and validation of the AISLE for those who wish to use the instrument or for others in the SoTL community looking to develop similar tools. We also offer an account of using the AISLE in our course to measure students' science skill and attitude development. In the end, our STEM-based instructional team learned that what appeared to be straight forward assessment development, was, in fact, a far more involved and complicated process.

KEYWORDS

instrument development, science skills, science attitudes, post-secondary, general science

INTRODUCTION

In the wake of a schedule redesign at our campus, our science team developed a new general science course. Designed to be taken by non-science majors (i.e., undergraduates outside of science disciplines), this course exposed students to various ideas and concepts across the traditionally defined sciences (i.e., biology, chemistry, and physics). As we saw it, this course was a chance to teach critical and creative thinking through the lens of science to those who may be less interested in pursuing higher studies in the field. As stated by SoTL scholars, Strzalkowski and Sobhanzadeh (2023), "the ability to use scientific reasoning to make personal decisions or to use scientific knowledge to appreciate natural phenomena or cultural events more fully should not be limited to scientists or

science students" (91). We concur with these scholars, and we sought to develop a course where those who did not intend to pursue science beyond our course could still use science skills appropriately and have positive attitudes toward science.

Scharff and colleagues (2023) identified the development of critical and creative thinking as one of the grand challenges for SoTL scholars. Science, by its very nature, requires critical and creative thinking, yet, as critiqued by Kopacz and Handlos (2021), many general science courses stress the mastery of science content. Focusing solely on content mastery is problematic since critical thinking is highly influenced by students' skills in and attitudes about science (Darmaji et al. 2020). With this in mind, we wanted to challenge the norm in teaching post-secondary science (i.e., mastery of content) and develop a general science course aimed at promoting skills and attitudes—those elements connected to critical thinking. We believed that our course addressed this challenge, but we wanted a way to determine whether we were addressing the skills and attitudes vital to critical thinking in science. Beyond traditional course assessments, we wondered whether our course had any impact on students' science skills and attitudes. Thus began our search for an instrument that might shed light on this wondering.

In response to the argument that much of the SoTL literature focuses on teaching, Manarin et al. (2021) called scholars to focus efforts on student learning. To investigate students' learning, our group sought an instrument that might provide insight into students' skills and attitudes in science. Science skills, to these researchers, were conceptualized as those skills necessary for undertaking scientific investigations, analyzing data, and making conclusions. Instruments aimed at measuring science skills have become increasingly popular, particularly in post-secondary contexts. Originating in the field of physics, science skills are often measured using concept inventories. One of the most popular and earliest developed concept inventories was the "Force Concept Inventory" (Hestenes, Wells, and Swackhamer 1992). Today, concept inventories can be easily found beyond physics in other subject areas such as biology (e.g., D'Avanzo 2008) and chemistry (e.g., Krause et al. 2004; Pérez García et al. 2016). Unfortunately, concept inventories are highly subject-specific, and this is problematic for those of us teaching general science[1] courses.

As we aimed to develop critical thinkers in science, we wanted our students to leave the course with appropriate science attitudes. Our team defines science attitudes as one aspect of the affective domain of science—specifically, the feelings and values held by an individual about science (Mao et al. 2021). Students' science attitudes have been shown to impact the way they interact with, and think about, science (Villafañe and Lewis 2016). Attitudes were particularly important to our instructional team because we were teaching a first-year, general science course aimed at non-science majors; we wanted students to leave our course with positive science attitudes since they were unlikely to take another post-secondary science course. There are various tools for measuring students' science attitudes in the education literature (e.g., Blalock et al. 2008; Cary, Wienhold, and Branchaw 2019; Halloun and Hestenes 1998; Supardi, Istiyono, and Setialaksana 2019). However, these instruments are designed to be administered alone, and we sought to explore students' skills and attitudes about science together in a succinct format. We wanted a way to capture students' learning and development in their skills and attitudes but needed a resource that was appropriate for a general science audience. Further, a concise instrument could provide insight into both science skills and attitudes.

The purpose of this article is to provide an account of our experiences in developing this instrument and offer advice to those considering the development of similar survey-style instruments. Given the lack of an identified instrument for our purposes, we developed a new instrument aimed at measuring students' science skills and attitudes in a general science course, designed for non-science

majors. Further, as general science courses are quite common in undergraduate education since the post-secondary system continues to diversify students' knowledge, particularly in the first year, we provide a description of one possible measure others may use to capture their students' learning of science skills and attitudes.

Context

In 2017, the academic calendar at Augustana Campus, a Canadian Liberal Arts and Sciences campus of the University of Alberta (from this point called the campus) changed significantly to introduce a block-session structure. The block courses were designed to be intensive, immersive, and experiential. Our instructional team decided to capitalize on this restructuring to develop an introductory science course specifically aimed at non-science majors on our campus.

Before the implementation of the new structure in 2017, non-science students tended to register in introductory environmental science and biology courses, and almost none of our non-science students registered in physical or mathematical sciences disciplinary courses. Further, all the introductory science courses offered by the department were designed to primarily serve those disciplinary majors. Anecdotally, our non-science students also reported that they were attracted most to introductory environmental science because it did not require a laboratory component, whereas introductory biology, chemistry, and physics courses did require laboratory time and, thus, were perceived as being more work. The twin factors of students shying away from physical or mathematical sciences, as well as avoiding courses with a laboratory component, meant that many non-science students taking their minimum six credits of science needed for a BA degree ended up with a narrow understanding of science. Further, the number of BA students registered in introductory environmental science and biology courses put enrollment pressures on those courses; in many years, those courses were at maximum capacity as students scrambled to register in the limited number of seats available.

A team of science faculty designed and offered a three-week course aimed at developing the science skills and attitudes of students not majoring in a science discipline. Our course objective was to provide students with a brief, experiential introduction to three branches of science (biology, chemistry, and physics) in a short time (a three-week block course). Survey courses for non-scientists can be found at many institutions, but we were not aware of many courses for non-scientists that include a laboratory component in a compressed format, despite the centrality of experimental work in science. We designed the course such that one-third of the class began in each of the different disciplines, and the order in which the different sciences were encountered was irrelevant. The experiments aimed to introduce students to one or two key experimental techniques in each discipline (e.g., microscope use in biology) as well as a few of the foundational concepts in that discipline that arise from the laboratory experiments (e.g., use of moles in chemistry).[2]

Our objective in designing this course was to invite non-scientists to experience one of the intrinsic ways in which scientists learn about the world: through laboratory experimentation. This course was designed to help non-science learners fulfill their degree distribution requirements, rather than to act as an entry point to science, technology, engineering, or mathematics (STEM) programs. Therefore, our overarching goals were to make sure learners enjoyed the lab experience and developed a general appreciation for modes of thought common in STEM disciplines rather than focusing on content knowledge. Some introductory lab techniques and disciplinary knowledge were necessary, but we primarily focused learning on developing general science skills (e.g., graphing and numeracy) and attitudes (e.g., relying on experimental evidence).

This course intended to increase science skills and positively impact science attitudes for non-science majors. Historically, short post-secondary courses have been shown to have little impact on science beliefs and attitudes (Abd-El-Khalick and Lederman 2000; Kite et al. 2021; Konnemann et al. 2018), but we were hopeful that the inclusion of a laboratory component and increase in science skills might influence both. As previously shown (e.g., Chowning 2022; Hunter, Laursen, and Seymour 2006), involving non-scientists in the practice of science can impact their knowledge of and beliefs about science. Given the unique design and intentions of the course—the inclusion of a laboratory experience in a general sciences course for non-science-majors—we sought a way of assessing whether this course was indeed improving students' science skills and attitudes.

INSTRUMENT SEARCH AND DESIGN

To assess whether we achieved our goal of increasing science skills and attitudes, we decided to survey students at the beginning and end of the course. In the literature, there are tools used to measure particular science skills or concepts in the form of concept inventories, of which a large number are validated and published in each of biology (see Cary, Wienhold, and Branchaw 2019; D'Avanzo 2008), chemistry (Krause et al. 2004; Pérez García et al. 2016), and physics (e.g., Hestenes, Wells, and Swackhamer 1992; Maloney et al. 2001; Thornton and Sokoloff 1998). Similarly, several excellent tools for measuring students' attitudes toward science have been published (e.g., Adams et al. 2006; Halloun and Hestenes 1998; Supardi, Istiyono, and Setialaksana 2019). However, no science attitude surveys or concept inventories described in the literature aligned directly with our objectives. Further, our course was neither aimed at science students nor content-focused, which was the case with many identified surveys and inventories. We sought a survey that was appropriate for non-science students in a general science survey course.

We did not find any validated instruments that could be used to simultaneously evaluate both skills and attitudes. Our team intended to quickly gain perspective on students' science attitudes and skills with one concise, combined document; we worried that using two separate (and often lengthy) surveys might contribute to survey fatigue and take unnecessary class time to complete. To meet this need, we decided to create the Augustana Interdisciplinary Scientific Literacy Evaluation (AISLE), a tool that could be used to quantitatively measure students' abilities to apply science skills to a problem and to reason like a scientist.

In this article, the design and calibration of the Augustana Interdisciplinary Scientific Literacy Evaluation (AISLE) is described. The specific questions we address herein are:

- Does AISLE provide a valid and reliable representation of students' skills in science?
- Does AISLE provide a valid and reliable representation of students' attitudes toward science?

For those interested in instrument design, we highly suggest using a systematic framework to guide your development. In this study, we applied DeVellis' (2017) eight steps to developing measurement scales in the creation of this instrument. These steps include: (1) determine what you want to measure, (2) generate an item pool, (3) determine the format for measurement, (4) have items reviewed by experts, (5) consider including validation items, (6) administer items to a development sample, (7) evaluate the items, and (8) optimize scale length. To meet the first step, our team sought to measure students' science skills and science attitudes.

In our design, we focused less on establishing the reliability for any individual item and instead focused on broader coverage of different potential skills and attitudes. The goal was not to design a tool hyper-focused on a few specific items, but rather one that was broadly applicable across many items. To generate an item pool (DeVellis' step two), after reading the aforementioned studies of science skills and science attitudes, our team of STEM faculty collaboratively generated topics that

would be the focus of the questions probing science skills on AISLE, a list of ideas and skills that are common across many scientific disciplines (e.g., interpretation of graphs, experimental design). For each potential topic, we designed a multiple-choice question to test that idea specifically or the potential idea was discarded if no satisfactory question could be created. Through discussion with other STEM faculty, the list of items was iteratively narrowed down to a final group of 10 (see Table 1) that included many important science skills for which proficiency was amenable to being tested using a multiple-choice question.

Table 1. Summary of target science skills and attitudes in AISLE

Science skill	Science attitude
Graph interpretation	Instruments and measurements
Estimation and checking the reasonableness of results	Trust in the scientific process
Relating theory to observation	Mathematical models and experiments
Proposing an experiment to test a hypothesis	Disagreement in science
Measurement units	Ethics in science
Simplifying models and limiting cases	Serendipity in science
Numeracy	Scientific theories
Uncertainty in measurements and interpretation	Science and the media
Types of variables	Scientific thinking
Inductive reasoning	Who are scientists

We followed similar process to develop a list of items that were categorized as science attitudes. In deciding what to label as a science attitude, we initially generated a list of simple phrases or attributes (e.g., relying on experimental evidence to derive a conclusion, how disagreements in science are resolved). For each phrase, the authors attempted to design a multiple-choice question to test that idea specifically. The objective was to identify key elements of how practicing scientists think, construct their knowledge, and understand the entire science process. Again, through discussion with other STEM faculty, we pruned the list of phrases and questions to a final set of 10 which can be seen in Table 1.

For DeVellis' (2017) third step, researchers are tasked with determining the format for the assessment. AISLE was developed to be twenty multiple choice questions with ten questions designed to measure non-disciplinary science skills (e.g., graphing, estimation) and ten questions designed to measure science attitudes (e.g., scientific method, relying on experimental evidence). We used multiple choice style questions for ease of scoring and to assist with students' abilities to efficiently complete the AISLE. For a complete list of the AISLE questions and answers (along with the answer key and scoring information), see Appendix. Our team designed AISLE in such a way that participants selected the best response to questions about either a science skill or attitude, with the potential answers having been categorized by five scientific experts (science faculty members holding a PhD in a scientific discipline). Each question had five possible responses: one "best answer" which scored two points, two "intermediate answers" which scored one point each, and two "poor answers" which scored zero points each (Adams and Wieman 2011; Bass, Drits-Esser, and Stark 2016). Finally, science skills-based questions were assigned odd numbers 1–19 using an online randomness generation tool, and science attitude-based questions were assigned random even numbers 2–20. This organization made calculation simple but avoided too much focus on attitudes or skills at any given point in the survey. The maximum AISLE score was 40 points.

Instrument analysis

As practicing scientists who espouse positivist approaches to defining science skills and attitudes, we designed and validated this survey from a similar paradigm. Positivist research paradigms assume that there is one truth which can be known, and this truth is independent of context and observer (Miller-Young and Yeo 2015). That is, for the purposes of this survey, we assumed that these constructs (i.e. science skills and attitudes) can be described (and agreed upon) by experts. With the assistance of other scientists, a survey was developed that knowledgeable experts believed to best describe science skills and attitudes. We, the authors, recognize that positivist positionality is contentious in educational research, and we do not claim this is the only way to describe skills and attitudes. However, as we sought to measure these aspects—the idea of measurement itself being questioned in educational research (Watson 2020)—we followed the tradition of instrument development in educational psychology and applied a priori definitions to the constructs of science skills and attitudes. This survey provides one way to consider both, but we caution readers from using this as the sole descriptor of either construct.

Our instructional team presented our results of the survey at an educational conference and were asked about the validity and reliability of the survey; at the time, we did not have an answer. When working in a positivist (and quantitative) research paradigm, researchers are primarily concerned with having a valid and reliable survey. As summarized by Wilson-Doenges (2015) "reliability and validity are paramount to researchers across fields to ensure that information gathered and reported is measuring what the researcher wants to measure and measuring it well" (49). Hence, to help our fellow SoTL scholars who may wish to venture into the world of quantitative instruments, we describe our account of determining the validity and reliability of the AISLE below.

Face validity

As suggested by DeVellis (2017) in step four, experts reviewed this instrument. In our context, we considered experts to be faculty members in science. Following the initial stages of question and response generation, we administered the entire set of questions and possible answers to a panel of five STEM faculty for a further round of feedback. We later learned that this is called achieving face validity. Face validity is an indication of whether, at a surface level, a measure seems relevant and appropriate for what it is purporting to measure. At this stage, wording was altered to improve clarity, references to sample scientific disciplines (e.g., "Biologists have proposed a theory. . . .") were checked to ensure reasonability, and answers were further refined to distinguish the two-point (or correct answer), one-point (or partially correct), and zero-point (or incorrect) choices more clearly. We collected discussion feedback and used it to further refine the AISLE questions and responses into their final form. DeVellis' step five suggests considering the inclusion of validation items[3]; the team decided not to include validation items to keep the survey at an appropriate length.

Calibration data collection

In step six, DeVellis suggests forming calibration data by administering the instrument to a development sample. As we aimed to provide our readers with an account of our instrument development, we devote our analysis section to how we explored validity and reliability of the AISLE with this calibration data. The results of our course area topic for another paper. To collect calibration data, we administered the initial draft of the AISLE instrument to several sample groups at Augustana Campus as well as one group studying at the main campus of the University of Alberta; students were also asked to comment on their experience of the AISLE. The target group of students for AISLE was those registered in a course designed for students in majors or programs outside the natural sciences.

To match the calibration group as closely as possible to the future test group, we used students as the closest available comparison groups when possible.

Having a clear sense as to how groups of first year students score on AISLE is of particular importance for the future use of AISLE as a pre- and post-test in its intended context. As shown in Table 2, AISLE data were collected for first-year students not registered in a BSc program at the beginning of their first year (titled first-year arts, although not all students were registered only in BA programs), first-year students registered in a BSc program in their first year of university (titled first-year science), and second to fourth year BSc students (titled upper year science). For the first-year arts groups, we deemed it necessary to use students at another campus of the same university in order to avoid the possibility that any of the students involved in the calibration process might subsequently take the course and become part of the future test data for AISLE. We included only those surveys that were completed with no blank responses in the data analysis.

As the AISLE is designed to capture science skills and attitudes, and our course aimed to improve science skills and attitudes, it could be the case that there is overlap in the ideas taught in the course and those assessed on the AISLE. This would make sense since we intended to use this assessment (at least in part) as a measure of whether students' skills and attitudes were impacted throughout the course; we wanted to assess what we teach. To minimize the possibility that our team would be teaching specifically to this assessment, we separated the AISLE design from the course design, but we recognize there could be some overlap in the expected course skills and attitudes and the course content. Data reported in this section—which we call calibration data—were collected within the first quarter of each semester to try to establish reliability and validity with baseline scores.

ANALYSIS AND FINDINGS

Given the lack of interval data[4] and since there is a possibility of a zero score on this instrument and an inconsistent ratio between responses (i.e., the inclusion of multiple partially correct responses), we deemed the use of traditional validation approaches (i.e., factor analysis) used on multi-scale instruments inappropriate. We turned to correlational analyses to determine validity and reliability.

Table 2. Summary of calibration data collected

Group	Number of surveys (n)
First-year science A	26
First-year science B	110
Upper year science A	15
Upper year science B	5
Upper year science C	14
First-year arts A	20
First-year arts B	90
Total	**280**

Calibration data

In step six, DeVellis (2017) indicates that researchers must evaluate their items. Here, we present our analysis to evaluate AISLE items and scales, as well as the overall survey. We use validity and reliability to evaluate the numeric data and a thematic analysis to evaluate the items using the qualitative feedback from participants.

Validity

For our purposes, as researchers learning to develop survey instruments, we relied on an introductory definition to validity; to us, an instrument is considered valid when it accurately represents what it is supposed to measure. There is much more to this aspect and more sophisticated statistical research might go on to delineate aspects such as discriminant, construct, criterion, and/or external validity. As we seek to describe our account of learning how to develop an instrument, we are taking an introductory look into validity. Below, we describe how we came to understand our instrument as being valid, or as researching students' skills and attitudes as intended.

To determine the validity of the results of AISLE with our calibration data, we compared responses for each question to each set of questions (skills and attitudes) using a Pearson correlation. The Pearson correlation coefficient, which uses the symbol *r*, summarizes and (numerically) describes the strength and direction of a linear relationship between two variables. In this correlational analysis, we were concerned with how closely items—or questions on the survey—were related to each scale. A scale describes a construct and consists of multiple items; in this research, our two scales were skills and attitudes.

Correlations were used to determine whether these results truly connected to the other questions in the proposed area (skills and attitudes). Table 3 shows results for odd questions 1–19 (skills questions) when compared with both the skills and attitudes arrays shown on AISLE. In Table 4, the results for even questions 2 –20 (attitude questions) when compared with both the skills and attitudes arrays on AISLE are shown. In general, questions positively correlated with all areas, $r(280)=0.09$[5] to $r(280)=0.54$, and—apart from question one—correlated more strongly with their anticipated area (skills or attitudes) than the other area. All correlations had a *p*-value of $p < 0.001$; a *p*-value describes the significance of the result. The smaller the *p*-value the more likely a result is significant. A *p*-value of 0.05 or less is often considered statistically significant with a *p*-value of 0.001 or less giving the strongest cases of statistical significance ($p<0.001$ suggests that this event has a <0.1% chance of randomly occurring).

Table 3. Results for skills questions on AISLE

		Before removing items		After removing items	
Question	**Science skill**	**r to skills**	**r to attitudes**	**r to skills**	**r to attitudes**
1	Graph interpretation	0.40	0.40	Item removed	
3	Estimation and checking the reasonableness of results	0.43	0.15	0.45*	0.15
5	Relating theory to observation	0.34	0.12	0.36*	0.12
7	Proposing an experiment to test a hypothesis	0.41	0.24	0.40	0.24
9	Measurement units	0.38	0.20	0.39*	0.20
11	Simplifying models and limiting cases	0.40	0.12	0.42*	0.14*
13	Numeracy	0.45	0.13	0.47*	0.12
15	Uncertainty in measurements and interpretation	0.52	0.18	0.53*	0.17
17	Independent vs. dependent vs. constrained variables	0.49	0.24	0.48	0.22
19	Inductive reasoning	0.41	0.29	0.40	0.30*

All *r*-values shown had a *p*-value of $p < 0.001$
*Indicates an increase from values after removing items

As is common practice in instrument development, two questions were removed from AISLE to improve correlation results. For readers less familiar with quantitative research, Pearson correlation values of 0.10 are considered weak (there is a small association between the item and the scale), correlation values of approximately 0.30 are considered moderate (responses on these items are likely associated with the scale and show a moderately strong relationship with the other items in the scale), and correlational values above 0.50 are considered strong (how participants respond to these items is strongly associated with how they respond to other items in this scale). We identified and removed questions that were less correlated with the others in this area.

Table 4. Results for attitudes questions on AISLE

		Before removing items		**After removing items**	
Question	**Science attitude**	**r to skills**	**r to attitudes**	**r to skills**	**r to attitudes**
2	Instruments and measurements	0.19	0.44	0.17	0.45*
4	Trust in the scientific process	0.19	0.44	0.16	0.43
6	Mathematical models and experiments	0.09	0.30	Item removed	
8	Disagreement in science	0.29	0.48	0.27	0.50*
10	Ethics in science	0.24	0.54	0.22	0.55*
12	Serendipity in science	0.21	0.47	0.19	0.47
14	Uncertainty in measurements and interpretation	0.19	0.46	0.18	0.45
16	Science and the media	0.13	0.40	0.11	0.44*
18	Scientific thinking	0.23	0.46	0.24*	0.49*
20	Who are scientists	0.23	0.48	0.22	0.51*

All *r*-values shown had a *p*-value of $p < 0.001$
*Indicates an increase from values after removing items

We removed question one because it positively correlated with both skills and attitudes with the same strength of correlation. That is, responses to question on were equally associated with the responses to items in skills and attitudes. As we had designed question one to be associated with skills (and not attitudes), this was a confounding factor. This double correlation may have occurred for several reasons. First, it may be because it was the first question on the instrument. In future iterations, it is suggested that the order be randomized to minimize the influence of question order. Second, this question may spark uncertainty for students in their ability given its mathematical requirements. Studies have shown that students experience math anxiety when they perceive they are being assessed on their mathematical ability in science courses, often resulting in the avoidance of science (Daker et al. 2021). As this first question focuses on reading a graph, it may be that these students' answers are not consistent with other science skills given this anxiety. Third, it may be that this question assesses multiple skills, confounding the results. This question requires students to read

axes, recognize that the y-axis is time squared (as opposed to linear time), and then interpret accordingly.

As a result of a weak correlation, we removed question six from AISLE as well Question six appropriately correlated as a science attitudes question (r=0.30), but with a notably weaker correlation than those of other questions in either scale (the next closest r value was 0.40). As expected, responses to question six had a positive correlation to science attitudes. However, this correlation, $r(280) = 0.30$, $p = 0.00$, was the lowest of all the correlations to its expected scale. Also, this question performed poorly at differentiating students since 86.12% of students scored this question correctly (5.34% of students scored one mark and 9.55% of students scored incorrectly on question six). Hence, for these reasons and to optimize length as suggested by DeVellis (2017)—the removal of an attitudes question since question one was a skills question—the decision was made to remove this question from the results.

Removing questions one and six from the original data set produced an increase to the correlation of the expected area for 12 of 18 questions. Seven questions (3, 5, 9, 11, 13, and 15) more strongly correlated to the science skills array after the removal of questions one and six from the original survey. The remaining skills questions had a decrease in correlation. The correlation of question 17 to the skills array decreased from $r(280) = 0.49$ to $r(280) = 0.48$. The correlation of questions seven and 19 to the skills array decreased slightly from $r(280) = 0.41$ to $r(280) = 0.40$. Finally, all skills questions were either similarly or less correlated to the attitudes array after the removal of questions one and six except for questions 11 and 19 which increased a value of 0.02 and 0.01, respectively. In general, removing questions one and six had the desired result on the questions intended to measure students' science skills on AISLE.

Almost all questions (2, 8, 10, 16, 18, and 20) intended to measure students' science attitudes were more strongly correlated to the attitudes array after the removal of questions one and six from the data set. Question 12 had no change in correlation to the attitudes array after removing questions one and six. Question four saw a slight decrease from $r(280) = 0.44$ to $r(280) = 0.43$, as did question 14 from $r(280) = 0.46$ to $r(280) = 0.45$. All questions, except question 18, saw a decrease in their correlation to the skills array once we removed questions one and six. Question 18 saw an increase from $r(280) = 0.23$ to $r(280) = 0.24$ with the skills array after the elimination of questions one and six. In general, removing questions one and six had the desired result on the questions intended to measure students' science attitudes on AISLE.

Reliability

A reliable instrument is consistent in its measurement—it can be reproduced under similar conditions. To showcase the reliability of AISLE, we consider the scores for each of the groups, as well as the average scores for science skills and attitudes. In a reliable instrument aimed at measuring science skills and attitudes, one would expect to see upper-year science students scoring higher than first-year students. Table 5 summarizes the overall, skills, and attitudes scores for AISLE for each of these groups, noting that the overall score maximum is 36, the skills score maximum is 18, and the attitudes maximum score is 18 since we removed questions one and six.

We performed statistical comparisons of AISLE scores between the calibration groups using one-way analysis of variance (ANOVA). ANOVA tests are often used to determine whether statistical significance exists between the mean values of multiple groups. If a study is working with only two groups, a t-test could provide this information, but with more than two groups an ANOVA is preferable. We used this test to determine whether the expected differences (i.e., upper-year science scoring higher than the other groups) existed; if this is the case, it gives some[6] evidence of reliability.

When conducting ANOVA comparisons between groups, we used a "*p*-value" to determine if the differences were significant. The lower the *p*-value, the more likely we are to reject the hypothesis that there is no discernable difference between the values. In this study, that means that a low *p*-value indicates that there is a significant difference. A *p*-value of 0.05 or lower generally indicates that there is a statistical significance. We do note here that low significance does not always mean no effect; it can be heavily impacted by the number of participants.

Statistical comparison of AISLE scores between the calibration groups was performed by ANOVA using JASP software (version 0.17.1). *P*-values $\leq$ 0.05 were considered statistically significant. A statistical comparison of AISLE scores for junior science and junior arts students indicated no significant difference in science skills or attitudes scores for these two groups of students (*p* skills = 0.278, *p* attitudes = 0.121, *p* total = 0.092). However, statistical analysis suggested that senior science students' AISLE scores were significantly higher for both skills and attitudes than junior science students (*p* skills = 0.011, *p* attitudes = 0.004, *p* total < 0.001) and even higher levels of significance occurred when senior student scores were compared to junior arts students (*p* skills < 0.001, *p* attitude < 0.001, *p* total < 0.001). Thus, analysis of AISLE scores for our calibration groups support AISLE as a reliable instrument to measure acquisition of science skills and attitudes.

Table 5. Summary of scores on AISLE for identified groupings

Group	N	Skills score (SD)	Attitudes score (SD)	Overall score (SD)
First-year arts	110	12.77 (2.91)	13.44 (3.00)	26.22 (4.99)
First-year science	136	13.33 (2.86)	14.10 (2.40)	27.43 (4.32)
Upper year science	34	14.91 (2.49)	15.71 (1.87)	30.62 (3.56)

Note: First-year arts consists of surveys from first-year arts A & B, first-year science consists of surveys from first-year science A, & B, and upper-year science consists of surveys from upper-year science A, B, and C, all indicated in Table 2.

Student experiences of the AISLE

We offered students in the calibration data sets space to respond to one open-ended question after completing AISLE: Do you have any feedback on the survey as a whole, or on any of the specific questions? We received 55 student responses. We transcribed student responses verbatim and analyzed using thematic analysis per Braun and Clarke (2022).

In this calibration phase, many student comments (n = 21) focused on the phrasing of the questions, the time it took them to complete, and the overall structure of the AISLE. These comments came from upper-year science courses (n = 7), first-year arts courses (n = 8) and first-year science courses (n = 6). Phrasing comments were often focused on needing some clarification (e.g., "I don't understand question five at all, didn't learn this at school") or that the reading expectations were too high (e.g., "There was too much language, please use diagrams or data more because not everyone likes to read"). The reading expectation responses also commented on this being difficult for those students who may struggle with English (e.g., "This was hard to do without good English"). Further, some of these wording comments were dependent on how students had previously learned science; for example, two students commented that they did not learn the terms "independent" and "dependent" variables but that they had learned "manipulated" and "responding" variables. These 21 students identified at least one aspect of the AISLE which was difficult for them, and this may be a confounding factor influencing our statistical analysis.

Seven student comments referred to their difficulties with STEM subjects. Of these comments, four referred to struggling with mental math. Students were unsure about whether they could use a

calculator; for example, a first-year science student wrote "specify if calculator is usable" and an upper-year science student wrote "I needed a calculator, I am not good at mental math." The other two comments were from first-year arts students. Further, three first-year arts students' comments expressed frustration at STEM education; one first-year arts student said "Woohoo, STEM Education! I know nothing! Lol." Beyond the semantics and reading of the AISLE, some participants felt ill-equipped to take the AISLE.

Another five comments discussed AISLE responses as being opinion-based. All the questions identified as opinion were classified as science attitudes and these included questions 10, 16, 18, and 20. Students wondered "who is to judge which answer is best" on these questions. Further, a few comments (n = 3) indicated that these questions were difficult to answer in multiple choice and offered written explanations for how they would respond. For example, one student wrote:

> #16 is kind of an opinion. I would say go find the original article and at least try to read it, doctors can still be biased, but if you have no idea how to read articles, talking to an expert is better than blindly believing.

The science attitudes questions were particularly prone to being perceived as opinion-based, and this might have caused issues in the overall scoring of science attitudes.

Finally, many comments (n=21) expressed intrigue toward the AISLE and/or felt that it met its intended purpose. 11 comments were coded as the AISLE prompting reflection. Reflection comments ranged from labelling oneself (e.g., "I'm not that sciencey") to philosophical inquiries (e.g., "The questions made me question how well I actually know science. Does anyone actually know science?"). Some students considered how they positioned themselves with respect to science, "The survey is interesting, and the questions really made me think about how distant I've grown from science since I took it in high school. I feel like if I had to learn that stuff in class now, I might find it more interesting and useful than I did as a teenager in high school." 12 comments indicated the AISLE as a good indication of students' skills and attitudes and that they were aimed appropriately at a non-science specific audience. For example, comments such as "The questions look hard at first, but students in any field should be able to answer most questions with a bit of thinking," "it is a nice survey because for a non-science students I can answer some of the questions without using any mathematics," and "the survey is a good way of testing how much people know about science when they first begin," described the AISLE as accessible and achievable for non-science majors. Further, one participant commented they were "pleased to see questions involving literacy, ethics, and philosophy, as opposed to just science/fact-based questions." For many, the AISLE prompted reflection and intrigue while appropriately asking about the science skills and attitudes of non-science majors.

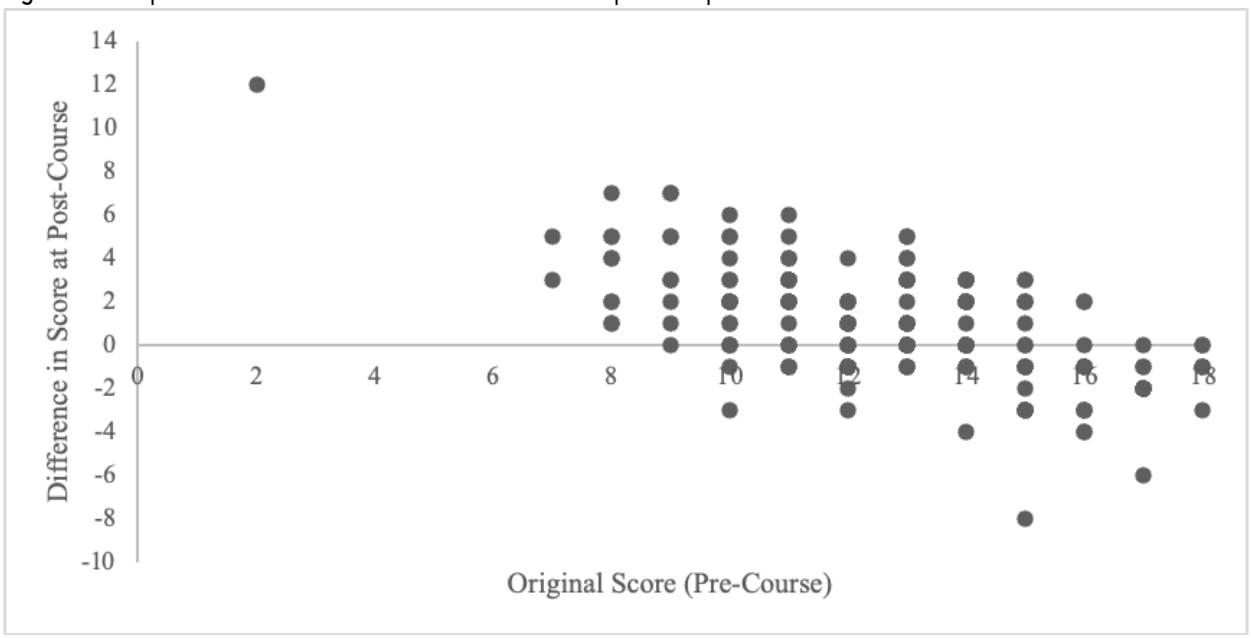
Did the course work?

After analyzing our calibration data for validity and reliability, we also investigated how our students' scores on the AISLE changed throughout the interdisciplinary science course. In this section, we present a brief analysis of five sections of our course. Each semester, students completed the AISLE at the start and end of the course. Student scores were then paired, and we analyzed for any changes in attitude and skills scores. Here we discuss the comparison of their pre- and post-course AISLE scores and the implications of them. In this analysis, we use the AISLE values without the inclusion of the original questions one and six; the maximum score on either the skills or attitudes scale in this analysis is 18.

Science skills

The average AISLE skills score before the course was 12.7 (SD = 2.8) and the average AISLE skills score after the course was 13.6 (SD = 2.4). In Figure 1, we present a graph showing the difference between students' skills scores on the pre-course and post-course AISLE. A negative difference (falling below the x-axis on Figure 1) means that the students scored lower on their science skills at the end of the course than at the start and a positive indicates that their science skills scores improved. Ideally, we would want to see a positive difference at the end of the course (unless the student had a score of 18 on the original AISLE, then a difference in score of zero would be ideal). Most students showed improvement in their AISLE scores at the end of the course. However, those students who scored highly on AISLE at the start were more likely to have a negative difference at the end of the course.

Figure 1. Comparison of students' skills scores on the AISLE pre- and post-course



Note: All 173 students are represented but data points which are repeated show as a single data point on this graph.

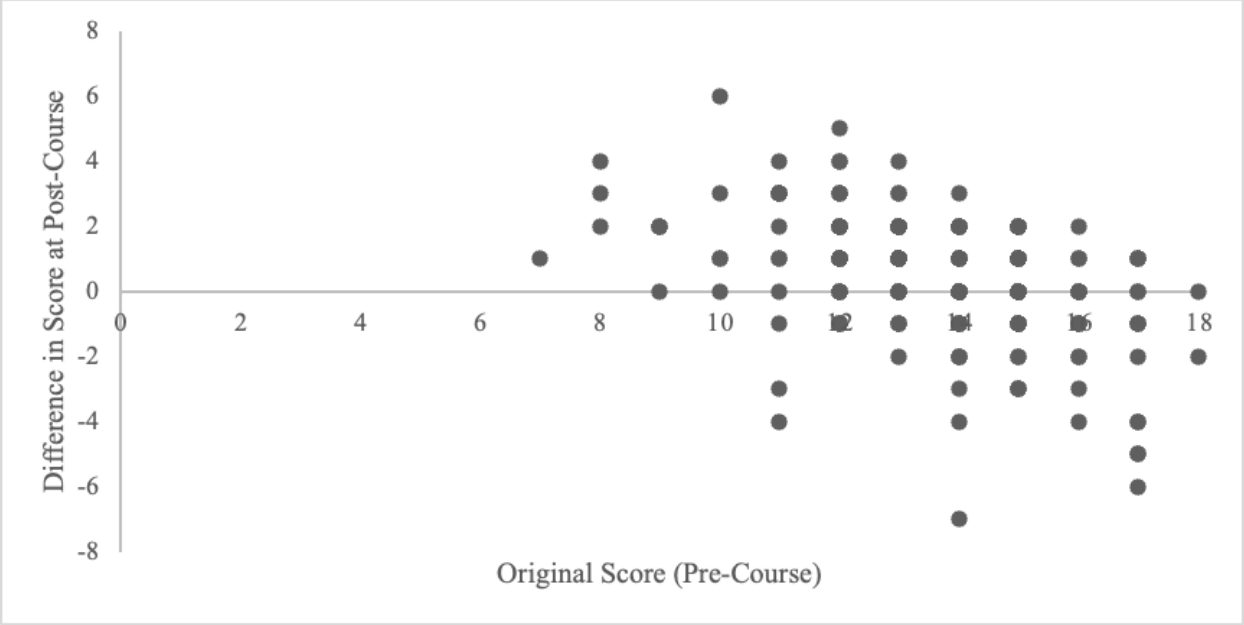
Continuing to look at students' AISLE science skills scores pre- and post-course, we used descriptive statistics—specifically, measures of central tendency—to analyze the data. Descriptive statistics do not offer comparisons (as is the case with inferential statistics), but, instead, summarize the main characteristics of the data set. Below, we include the number of students in each section (N), average difference and standard deviation (which describes how spread out the averages are), and the median (or middle number of the data set).

Science attitudes

The average AISLE attitudes score before the course was 13.6 (SD = 2.2), and the average AISLE attitudes score after was 13.9 (SD = 2.2). In Figure 2, we present a graph showing the difference between students' attitudes scores on the pre-course and post-course AISLE. Again, scores on the pre-test could range from 0 to 18. In viewing this graph, it appears students tended to have higher scores on their attitudes over their skills in general. Further, students near the middle of the scoring on the pre-course AISLE were more likely to improve at the end of the course, while students with high scores were more likely to decrease after the course. The decreases and increases were less extreme than we

saw with skills (which had a maximum increase of 12 and biggest decrease of eight points). The biggest increase in attitudes scores was six and the biggest decrease in attitudes scores was seven at the end of the course. These less extremes would indicate that the average difference score might be more reflective of this data than it was of the skills data.

Figure 2. Comparison of students' attitudes scores on the AISLE pre- and post-course



Note: All 173 students are represented but data points which are repeated show as a single data point on this graph

In Table 6, we can see that the average difference score in each section ranged from 0.44 to 1.69. We combined the data from all sections and found that the average difference from pre- to post-course was 0.9. This indicates an overall increase in students' science skills. We should note that the standard deviation values were all around 2.5 or 3, which indicates that about 70% of the data sits within 2.5–3 marks of the average. So, given a normal distribution of data for the combined values, we can assume that about 70% of the data falls between a difference of -1.76 and 3.58 for science skills scores. However, most data sets are not normally distributed. In skewed data sets, it is better to consider the median since it is less impacted by extreme outliers in the data set. In our data, the median number for most courses was zero, with one section having one and another having two as its median. Given that both our average and median are both consistently positive, this indicates that, on average, students in all sections scored better on the skills section of the post-course AISLE.

Table 6. Descriptive statistics of changes in science skills pre- and post-course

	2018	**2019**	**2020**	**2021**	**2023**	**Combined**
N	36	67	45	14	10	173
Average difference (standard deviation)	0.44 (2.78)	0.63 (2.46)	1.69 (2.55)	1.07 (3.71)	0.7 (2.26)	0.91 (2.67)
Median	0	0	2	1	0	1

In Table 7, we can see that the average difference score in each class ranged from -0.5 to 1.43. This range is smaller, but also lower, than with the skills data. We combined the data from all sections and found that the average difference from pre- to post-course was 0.30. There was an overall increase in student attitudes in all courses, except 2023, but with a smaller increase than we witnessed with science skills. We also note that the standard deviation values were all around 1.75 or 2, which indicates that about 70% of the data sits within 1.7–2 marks of the average. So, given a normal distribution of data for the combined values, we can assume that about 70% of the data falls between a difference of -1.75 and 2.35 for science attitudes scores. This range is slightly smaller than the range for science skills but also lower in value. We also considered the median in science attitudes differences. Again, the median number for most courses was zero, with one section having one and another having two as its median. Given that both our average and median are both generally positive (except in 2023), this indicates that, in general, students scored better on the attitudes section of the post-course AISLE, albeit the increase was much less than for students' skills.

Table 7. Descriptive statistics of changes in science attitudes pre- and post-course

	2018	2019	2020	2021	2023	Combined
N	36	67	45	14	10	173
Average difference (standard deviation)	0.11 (2.28)	0.22 (1.91)	0.38 (2.12)	1.43 (1.79)	-0.5 (1.79)	0.30 (2.05)
Median	1	0	0	2	0	0
Mode	2	0	0	2	0	0

DISCUSSION

Designing the instrument

The designers of this instrument were a team of scientists—not educational researchers. As experts in science, we constructed this tool to evaluate what we felt were key skills and attitudes for students to develop in a first-year interdisciplinary course. Being scientists, we used quantitative data—a form with which we are familiar—but learned that the analysis of social science quantitative data (and development of a social science instrument) required further knowledge that we had not all developed. For example, our experts in biology and physics had done correlational statistics, and we were all familiar with descriptive statistics, but our chemist had not taken any formal statistics classes. We analyzed our data the best we knew how, but when we attended a STEM education conference and were asked about the validation and reliability of our survey, it caught us a bit off guard. Hence, we enlisted our colleague from education to help with the analysis of this work.

A list of science skills that the designers and practicing scientists involved believed to be common across many or most scientific disciplines and a series of statements or concepts that were considered to describe parts of how scientific experts think, which were labelled science attitudes, drove the design of the questions for AISLE. These two lists are not exhaustive, and there are surely other skills or attitude statements that could also be included in the list of potential items.

Further, we designed the AISLE questions to test the listed skill or attitude, yet it is also possible that a given question could have an underlying concept separate from the intended skill or attitude that influences participants' ability to correctly answer the question. For example, a few participants commented on the need for a calculator to solve some questions, suggesting that they

understood certain questions as requiring a precise calculation, rather than a question requiring an understanding of relative relationships between mathematical variables as we intended. As another example, in question one on the AISLE, students' skill to read a graph is being measured. However, to properly read this graph, students need to recognize that the vertical axis is squared. Does this question truly represent their ability to read graphs, or does the question consider their attention to detail? This confounding of constructs is likely why question one did not correlate properly with either scale and was ultimately removed—however, had we not done the thorough statistical analysis in this paper, we might have continued to assume that this item was measuring what we intended. This is an aspect for further consideration on each of the individual items of AISLE and a consideration we raise for any colleagues looking to design an instrument.

Finally, we urge those considering instrument development such as this to include the use of qualitative data. Even responses to our one open-ended response gave us useful information. In future iterations of the AISLE, we plan to revisit the wording of questions and provide clear instructions regarding calculator use. We also recognize the need to address the idea of "opinion" and science attitudes; there are ways of doing and knowing in science that are accepted and are not a matter of opinion (this is called the nature of science), but there were many misconceptions among participants in all sections about this idea. Many students agreed that the AISLE measured what it purported, but, to our delight, this also prompted some philosophical reflection on science. This finding indicates that it might be worthwhile to include an instrument like this into coursework beyond a measurement in order to prompt students' reflections on their ideas and knowledge of science.

Is the AISLE valid?

Based on the statistical analysis, AISLE can produce valid and reliable results when considering individuals' science skills and attitudes. Questions on AISLE were all positively correlated to their intended areas. Odd questions (except the removed question one) positively correlated to science skills with all p-values at $p < 0.001$. Even questions (except the removed question six) positively correlated to science attitudes with all p-values at $p < 0.001$. Further, the results from calibration data showed no significant differences between students enrolled in first-year science or first-year arts classes, but there were significant differences between first-year courses and those students in upper-year science classes for science skills and attitudes and overall scores on the AISLE. This suggests results were consistent and reproducible at a similar level of education and that skills, attitudes, and overall score were improved with years of STEM education, as expected.

One key underlying assumption is that the participants' formal level of STEM education will correlate with their overall science skills and attitudes, and therefore higher levels of formal STEM education should be a proxy for higher levels of science skills and attitudes. The results presented are consistent with this assumption since students in the upper-year science courses had statistically higher scores on science skills, attitudes, and overall scores on the AISLE. Commonly, instrument validation of science skills and/or attitudes is conducted with STEM classes (e.g., Adams et al. 2006; Cary, Wienhold, and Branchaw 2019; Maloney et al. 2001), and we did not identify any instrument that tested with upper-year arts students. This could be an interesting vein of future research. Based on the calibration results, more formal STEM education correlated with higher scores on science skills and attitudes in this instrument.

CONCLUSION

Our team of science faculty wanted to develop an instrument that could concisely and succinctly provide us with an assessment of students' science skills and attitudes throughout a first-year general science course with a laboratory component. However, when asked to defend the validity of our quantitative work, we learned that we needed a more systematic way to determine whether our instrument was producing valid and reliable results—was it measuring what we thought it was measuring? According to our statistical analysis, AISLE provided a valid and reliable description of students' science skills and attitudes after removing questions one and six. Further, in comments, many students (n = 21) indicated they thought AISLE was a useful exercise and appropriate for non-science majors. Considering our work at the micro and meso levels (Simmons 2009, 2020), we claim that the AISLE would provide one data point in our assessment of whether students' skills and attitudes were developing as a result of our course.

We hope that readers take away two lessons regarding teaching and learning. First, we anticipate that others teaching science courses to non-science majors might find AISLE useful in assessing their students' science skills and attitudes. AISLE is succinct and quickly evaluated, so it can provide useful formative feedback for both the student and the instructor. For our course, which was entirely focused on a laboratory-based experience, we saw improvement in students' science skills but little change in their attitudes. The one exception was in 2021, where we saw a much bigger improvement in attitudes (an average of 1.43) than skills (an average of 1.07). In future work, we aim to investigate why there was little change in attitudes overall as well as consider results from individual sections in light of our teaching practices and reflections on those years. Further, AISLE can prompt discussions about science and the philosophy of science (as indicated in participant comments). Using AISLE and discussing the items in class might allow instructors to address the misconception of science attitudes and approaches as opinion when, in reality, these are epistemological underpinnings that are widely accepted by the scientific community. We share our work so that others in similar contexts may draw on it.

Second, we hope that others who wish to attempt instrument development might read our process and reflections before starting what turned out to be a more complex journey than we anticipated. Naively, we thought that the development of a new instrument measuring science skills and attitudes would be a fairly straightforward process; our team consisted of experts in STEM, so we felt prepared to assess these aspects. Creating questions and seeing student responses was not overly difficult, but to claim that we had created an instrument that measured what it was designed to measure turned out to be a more complex process. Our team accomplished this task by having someone knowledgeable in instrument design in education join and advise our group. For example, we learned of DeVellis' (2017) useful framework and the systematic approaches to instrument development and validation. We encourage other SoTL scholars who may be less familiar with instrument design to reach out to your centres for teaching and learning and/or colleagues in the social sciences for collaboration in these endeavours.

AUTHOR BIOGRAPHIES

Ellen Watson, PhD (CAN), is an associate professor at Brandon University's Faculty of Education. She teaches courses in science education, curriculum, research methodologies, and other educational topics.

Sheryl Gares, PhD (CAN), is an associate professor at the Augustana Faculty of the University of Alberta, where she teaches in the integrative biology program in the Department of Science.

Brian Rempel, PhD (CAN), is an associate professor at the Augustana Faculty of the University of Alberta, where he teaches organic chemistry in the chemical and physical science program in the Department of Science.

NOTES

1. The authors recognize that how "science" is defined can shift from discipline to discipline and from individual to individual. We leave this debate to philosophers of science and, instead, define "general science" here as encompassing the fields of biology, chemistry, and physics at their broadest levels.
2. We understand that TLI readers may be interested in the development of this course, and we encourage you to reach out to the authors for that discussion. The full intention and development of this course is a subject for another paper, and we intentionally focused our discussion in this manuscript on the instrument development experience.
3. Validation items are extraneous items used to ensure participants are thoughtfully answering. This is common practice on long instruments (e.g., more than 20 questions). An example of a validation item might be writing, "answer three to this item," and if a respondent answers something other than three, the researcher can assume they have not thoughtfully completed the survey, and their results will likely be invalid.
4. Interval data is numerical and considered to have equal distances (or intervals) between each response. It requires answers to be given in the form of integers. Some examples of interval data include temperature, time, and distance; for example, each minute has the same "distance" of 60 seconds.
5. For those less familiar with statistical research, this is read as the "r-value" (or Pearson correlation coefficient value) of 280 participants is 0.09. The two statistics here indicate that the lowest Pearson correlation coefficient was 0.09 and the highest was 0.54.
6. It should be noted that any claim of reliability could be improved by looking at reproducibility. Reproducibility could mean seeing similar results between student assessments and their AISLE scores, re-testing students before any intervention, and looking for similar scores or comparing groups for similar scores across multiple sections or years with similar conditions. These measurements were not collected with our calibration data, but it is something to consider for those interested in developing their own instruments.

ETHICS

We consulted the institution's Research Ethics Board (REB) on the development and calibration of the AISLE. They determined that this project was outside the scope of the REB and was allowed to proceed.

REFERENCES

Abd-El-Khalick, Fouad, and Norman G. Lederman. 2000. "The Influence of History of Science Courses on Students' Views of Nature of Science." *Journal of Research in Science Teaching* 37 (10): 1057–95. https://doi.org/10.1002/1098-2736(200012)37:10<1057::AID-TEA3>3.0.CO;2-C.

Adams, Wendy K., Perkins, Katherine, Podolefsky, Noah S., Dubson, Michael, Finkelstein, Noah D., and Carl E. Wieman. 2006. "New Instrument for Measuring Student Beliefs about Physics and Learning Physics: The Colorado Learning Attitudes about Science Survey." *Physical Review Special Topics - Physics Education Research* 2 (1): 1–14. https://doi.org/10.1103/PhysRevSTPER.2.010101.

Adams, Wendy. K., and Carl E. Wieman. 2011. "Development and Validation of Instruments to Measure Learning of Expert-Like Thinking." *International Journal of Science Education* 33 (9): 1289–312. https://doi.org/10.1080/09500693.2010.512369.

Bass, Kristin. M., Dina Drits-Esser, and Louisa A. Stark. 2016. "A Primer for Developing Measures of Science Content Knowledge for Small-Scale Research and Instructional Use." *CBE—Life Sciences Education* 15 (2): 1–14. https://doi.org/10.1187/cbe.15-07-0142.

Blalock, Cheryl L., Michael J. Lichtenstein, Steven Owen, Linda Pruski, Carolyn Marshall, and MaryAnne Toepperwein. 2008. "In Pursuit of Validity: A Comprehensive Review of Science Attitude Instruments 1935–2005." *International Journal of Science Education* 30 (7): 961–77. https://doi.org/10.1080/09500690701344578.

Braun, Virginia, and Victoria Clarke. 2022. *Thematic Analysis: A Practical Guide*. SAGE.

Cary, Tawnya L., Caroline J. Wienhold, and Janet Branchaw. 2019. "A Biology Core Concept Instrument (BCCI) to Teach and Assess Student Conceptual Understanding." *CBE—Life Sciences Education* 18 (3):, https://doi.org/10.1187/cbe.18-09-0192.

Chowning, Jeanne T. 2022. "Science Teachers in Research Labs: Expanding Conceptions of Social Dialogic Dimensions of Scientific Argumentation." *Journal of Research in Science Teaching* 59 (8): 1388–415. https://doi.org/10.1002/tea.21760.

D'Avanzo, Charlene. 2008. "Biology Concept Inventories: Overview, Status, and Next Steps." *BioScience* 58 (11): 1079–85. https://doi.org/10.1641/B581111.

Daker, Richard J., Sylvia U. Gattas, Moriah H. Sokolowski, Adam E. Green, and Ian M. Lyons. 2021. "First-Year Students' Math Anxiety Predicts STEM Avoidance and Underperformance throughout University, Independently of Math Ability." *Nature Partner Journals, Science of Learning* 6 (17). https://doi.org/10.1038/s41539-021-00095-7.

Darmaji, Darmaji, Dwi A. Kurniawan, Astalini Astalini, Rahmat Perdana, Kuswanto Kuswanto, Muhammad Ikhlas. 2020. "Do a Science Process Skills Affect on Critical Thinking in Science? Differences in Urban and Rural." *International Journal of Evaluation and Research in Education (IJERE)* 9 (4): 874. https://doi.org/10.11591/ijere.v9i4.20687.

DeVellis, Robert F. 2017. *Scale Development: Theory and Applications,* fourth edition. Thousand Oaks, California: SAGE Publications Ltd.

Halloun, Ibrahim, and David Hestenes. 1998. "Interpreting VASS Dimensions and Profiles for Physics Students." *Science and Education* 7 (6): 553–77. https://doi.org/10.1023/A:1008645410992.

Hestenes, David, Malcolm Wells, and Gregg Swackhamer. 1992. "Force Concept Inventory." *The Physics Teacher* 30 (3): 141–58. https://doi.org/10.1119/1.2343497.

Hunter, Anne-Barrie, Sandra L. Laursen, Elaine Seymour. 2006. "Becoming a Scientist: The Role of Undergraduate Research in Students' Cognitive, Personal, and Professional Development." Science Education 91 (1): 36–74. https://doi.org/10.1002/sce.20173.

Kite, Vance, Soonhye Park, Katherine McCance, and Elsun Seung. 2021. "Secondary Science Teachers' Understandings of the Epistemic Nature of Science Practices." *Journal of Science Teacher Education* 32 (3): 243–64. https://doi.org/10.1080/1046560X.2020.1808757.

Konnemann, Christiane, Christian Höger, Roman Asshoff, Marcus Hammann, and Werner Rieß. 2018. "A Role for Epistemic Insight in Attitude and Belief Change? Lessons from a Cross-Curricular Course on Evolution and Creation." *Research in Science Education* 48 (6): 1187–204. https://doi.org/10.1007/s11165-018-9783-y.

Kopacz, Dawn M., and Zachary Handlos. 2021. "Less is More: Course Redesign and the Development of an Atmospheric Science Process Skills Assessment." *International Journal for the Scholarship of Teaching and Learning*15 (2). https://doi.org/10.20429/ijsotl.2021.150212.

Krause, Stephen, Birk, James, Bauer, Richard, Jenkins, Brook, and Pavelich, Michael J. 2004. "Development, Testing, and Application of a Chemistry Concept Inventory." In *34th Annual Frontiers in Education,* 103–7. Savannah, GA: IEEE. https://doi.org/10.1109/FIE.2004.1408473.

Maloney, David P., Thomas L. O'Kuma, Curtis J. Hieggelke, and Alan Van Heuvelen. 2001. "Surveying Students' Conceptual Knowledge of Electricity and Magnetism." *American Journal of Physics* 69 (S1): 12–23. https://doi.org/10.1119/1.1371296.

Manarin, Karen, Christine Adams, Richard Fendler, Heidi Marsh, Ethan Pohl, Suzanne Porath, and Alison Thomas. 2021. "Examining the Focus of SoTL Literature—Teaching and Learning?" *Teaching and Learning Inquiry* 9 (1): 349–64. https://doi.org/10.20343/teachlearninqu.9.1.23.

Mao, Peipei, Zhihui Cai, Jinbo He, Xinjie Chen, and Xitao Fan. 2021. "The Relationship between Attitude toward Science and Academic Achievement in Science: A Three-Level Meta-Analysis." *Frontiers in Psychology* 12. https://doi.org/10.3389/fpsyg.2021.784068.

Miller-Young, Janice, and Michelle Yeo. 2015. "Conceptualizing and Communicating SoTL: A Framework for the Field." *Teaching & Learning Inquiry* 3 (2): 37–53. https://doi.org/10.20343/teachlearninqu.3.2.37.

Pérez García, Marilú, Cynthia J. Luxford, Theresa L. Windus, and Thomas Holme. 2016. "A Quantum Chemistry Concept Inventory for Physical Chemistry Classes." *Journal of Chemical Education* 93 (4): 605–12. https://doi.org/10.1021/acs.jchemed.5b00781.

Scharff, Lauren, Capocchiano, Holly, Chick, Nancy, Eady, Michelle, Friberg, Jen, Gregory, Diana, Loy, Kara and Maurer, Trent. 2023. "Grand Challenges for SoTL #1." *International Society for the Scholarship of Teaching and Learning*. https://issotl.com/grand-challenges-for-sotl/gc-sotl-1/.

Simmons, Nicola. 2009. "Playing for SoTL Impact: A Personal Reflection." *International Journal for the Scholarship of Teaching and Learning* 3 (2). https://doi.org/10.20429/ijsotl.2009.030230.

Simmons, Nicola. 2020. "The 4M Framework As Analytic Lens for SoTL's Impact: A Study of Seven Scholars." *Teaching & Learning Inquiry* 8 (1): 76–90. https://doi.org/10.20343/teachlearninqu.8.1.6.

Strzalkowski, Nicholas, and Mandana Sobhanzadeh. 2023. "Views and Value of an Undergraduate General Education on Advancing Student Attitudes and Engagement with Science." *Imagining SoTL* 3 (2): 89–119. https://doi.org/10.29173/isotl687.

Supardi, Rafsanjani, Edi Istiyono, and Wirawan Setialaksana. 2019. "Developing Scientific Attitudes Instrument of Students in Chemistry." *Journal of Physics Conference Series* 1223: Paper 012025. https://doi.org/10.1088/1742-6596/1233/1/012025.

Thornton, Ronald K. and David R. Sokoloff. 1998. "Assessing Student Learning of Newton's Laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula." *American Journal of Physics* 66 (4): 338–52. https://doi.org/10.1119/1.18863.

Villafañe, Sachel M., and Jennifer E. Lewis. 2016. "Exploring a Measure of Science Attitude for Different Groups of Students Enrolled in Introductory College Chemistry." *Chemistry Education Research and Practice* 17 (4): 731–42. https://doi.org/10.1039/c5rp00185d.

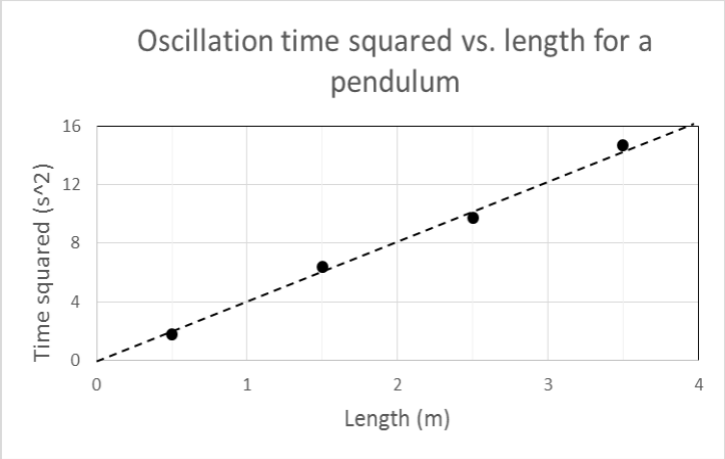
Watson, Ellen. 2020. "The Slippery Business of Measuring Beliefs Lessons from a Failed Attempt at Developing an Instrument to Measure Teachers' Epistemic Beliefs about Physics Knowledge." *Electronic Journal for Research in Science & Mathematics Education* 24 (2): 119–40. https://ejrsme.icrsme.com/article/view/20294.

Wilson-Doenges, Georjeanna. 2015. "The State of Scale Validation in SoTL Research in Psychology." In *A Compendium of Scales for the Use of Scholarship of Teaching and Learning*, edited by Rajiv S. Jhangiana, Jordan D. Troisi, Bethany Fleck, Angela M. Legg, and Heather D. Hussey, 44–54. Society for the Teaching of Psychology. https://teachpsych.org/ebooks/compscalessotp.

APPENDIX

AISLE (Augustana Interdisciplinary Scientific Literacy Evaluation)

1. The four dots on the graph below indicate the data from an experiment in which pendulums of various lengths were constructed and their oscillation time was measured. When the oscillation time is squared, the graph suggests a simple straight line pattern for the data, as shown by the dashed line.



Based on this graph, what length of pendulum would we expect to have an oscillation time of 2 seconds?

- **A.** About 8 meters
- **B.** About 1 meter
- **C.** About 0.5 meters
- **D.** Unknown: there is no connection between the pendulum's oscillation time and length
- **E.** Unknown: none of the four measurements had an oscillation time of 2 seconds

2. Imagine that you are trying to test a hypothesis for a chemical reaction, and you cannot directly observe the reacting molecules with your senses. Currently, there is no instrument that can directly observe individual molecules. What is the best way of going about testing your hypothesis?

- **A.** Use an instrument to measure a property of the molecules that is known to be directly related to the chemical reaction, even if this is not something you can directly observe with your senses.
- **B.** Find a chemical reaction that is not the one you are studying but has some similarities and can be observed with your senses. Test your hypothesis on this related chemical reaction.
- **C.** Because this chemical reaction cannot be observed with your senses, there is no way to test your hypothesis.
- **D.** Wait for someone to develop a new instrument that will let you directly observe the chemical reaction with your senses.
- **E.** Devise a mathematical formula to describe the chemical reaction. Use this formula to develop a computer simulation to test your hypothesis.

3. Would it be possible to put 100 oranges in a bathtub?

- **A.** Probably not — a row of about 20 oranges would already be about as long as the bathtub.
- **B.** It doesn't matter. Science is only concerned with ideas that have immediate practical

applications in the real world.
- **C.** It is impossible to determine this because every orange and every bathtub are different.
- **D.** It is impossible to determine this without doing an experiment.
- **E.** Probably — a rough estimate can indicate that the volume of a bathtub is more than the volume of 100 oranges.

4. Imagine a hypothetical scenario in which a research scientist announces a discovery that will radically alter their research field. This will also have an impact on our entire society. How will you know that this discovery is genuine and not fraudulent?
- **A.** Many of the experts in the field suspect that this discovery is genuine.
- **B.** Scientists are fundamentally honest people. They don't stand to gain anything resulting from their new theory.
- **C.** Other scientists can independently replicate the discovery and report similar results.
- **D.** A majority of the general public believe the discovery is genuine.
- **E.** The most important scientist in the field agrees that this discovery is genuine.

5. Biologists have proposed a theory that sunflowers have a stronger structure because their seeds are usually arranged in a spiral pattern with a total number of spirals belonging to the Fibonacci sequence: 1, 2, 3 (=1+2), 5 (=2+3), 8 (=3+5), etc. If a sunflower with 21 spirals is found, how does this relate to the theory?
- **A.** It contradicts the theory, therefore the theory must be rejected.
- **B.** It proves that the theory is correct.
- **C.** It contradicts the theory, but the theory may still be correct.
- **D.** It supports the theory, but it doesn't mean that the theory must be correct.
- **E.** It has no relation because mathematics cannot be used to describe living things.

6. Which statement below best describes the role of mathematics in scientific experimentation?
- **A.** A mathematical model can be used to suggest an experiment, but not the other way around.
- **B.** Mathematics is not real and therefore has no connection to science.
- **C.** An experiment can be used to suggest a mathematical model, but not the other way around.
- **D.** Mathematics and science are the same thing: all science is applied mathematics.
- **E.** A mathematical model can be used to suggest an experiment OR an experiment can be used to suggest a mathematical model.

7. Milk "spoils" due to the growth of certain types of bacteria over time. Which of the following experiments would best test the hypothesis that bacterial growth in milk increases at higher temperatures?
- **A.** Observe a bacterial population under a microscope and measure their average size and speed. Then, increase the temperature of the room and repeat the measurements.
- **B.** Prepare one fresh milk sample and measure the bacteria population every hour for 24 hours as the sample is heated from 2 degrees Celsius to 50 degrees Celsius.
- **C.** Mix a sample of fresh milk with an antibiotic, store the sample at room temperature for a week, then smell the sample to see if it has spoiled.
- **D.** Prepare 10 identical milk samples, store them each at a different temperature for 24 hours, then measure the bacteria population in each sample.
- **E.** Prepare 5 samples with 5 different kinds of milk, store the samples in a refrigerator for 24

hours, then measure the bacteria population in each sample.

8. Consider two groups of scientists studying the same phenomenon. Group A and Group B have different hypotheses and vigorously disagree with each other, to the point that some outside observers have described their conflict as a feud. What does this suggest?
- **A.** This problem is beyond the capability of scientists to understand.
- **B.** A neutral judge needs to pick which hypothesis is correct.
- **C.** The two groups of scientists need to sit down and rationally find a way to merge their ideas into a better hypothesis.
- **D.** Someone should devise an experiment that could distinguish between the competing hypotheses.
- **E.** Each group should perform experiments designed to support their hypothesis.

9. An experiment was designed to measure the temperature of a lake at various depths. In what units should these depths be measured and why?
- **A.** Inches. Smaller units lead to bigger numbers and more accurate measurements.
- **B.** It doesn't matter. So long as the measurements are precise, the chosen units can be converted to other units later on.
- **C.** Feet. Units that have the longest history are the most reliable.
- **D.** Fathoms. Uncommon units tend to discourage others from disagreeing with the results of an experiment.
- **E.** Meters. Metric system units are the most accurate units to use in any experiment.

10. Consider a scenario in which a group of scientists announce their intention to offer a service where people can clone their pets. Who should be in charge of deciding whether this service is ethically sound and why?
- **A.** All of society. A wide-scale discussion across all members of society, including a decision-making process such as a referendum is the best way to make sure that everybody's viewpoint is heard.
- **B.** The government. They are tasked with overseeing how science and society conduct themselves.
- **C.** The customers. The morality of these services should be decided by those with the ability to purchase the services.
- **D.** A panel combining citizens, scientists, ethicists, and lawyers. A group that combines input from a variety of informed viewpoints will best identify relevant issues.
- **E.** The scientists offering the service. They are the ones most familiar with the technology and have naturally thought through all of the potential consequences.

11. A forklift can carry a stack of N heavy crates. The maximum speed of the forklift (S, measured in kilometers per hour) depends on how many crates it is carrying. Based on the behavior for small and large values of N, which of the following equations might represent a good model for the maximum speed of the forklift?
- **A.** $S = \frac{20}{2+N}$
- **B.** $S = \sin(N)\sqrt{1+N^2}$
- **C.** $S = 15N$

D. $S = 25$
E. $S = 5 - N$

12. What is the role of unexpected results in scientific discovery?
- **A.** By designing a very large number of experiments, scientists rely on unexpected results to find the experiments which work as expected.
- **B.** An unexpected result means the current theories are wrong.
- **C.** An unexpected result shows that the experiment was poorly designed or executed, as it did not properly control all possible variables.
- **D.** An unexpected result means that a previously unrecognized variable is important for the system of interest.
- **E.** Because scientists carefully control experimental conditions, unexpected results cannot occur.

13. Bob's orchard has ten-thousand apple trees. Each tree produces one-hundred apples each year, on average. If a new system for watering the trees could lead to a ten percent increase in apple production, how many additional apples could the orchard produce each year?
- **A.** Ten million
- **B.** One million
- **C.** One hundred thousand
- **D.** One thousand
- **E.** Ten

14. What do scientists mean when saying they have a theory?
- **A.** They have found a mathematical way of describing a scientific phenomenon.
- **B.** They have an idea that cannot be experimentally tested.
- **C.** They have an idea that might be able to predict future experimental results.
- **D.** They have made a speculative guess as to how something works, and use of the word "theory" indicates that this guess is related to a scientific question.
- **E.** They have an explanation for a series of scientific facts and experimental observations.

15. A biologist is measuring honeybee populations in an area in order to try to predict future fruit production from blueberry bushes. If the population measurements have a margin of error of 20% then:
- **A.** There will also be a certain margin of error for the fruit production predictions, but these predictions can still be useful.
- **B.** The fruit production predictions will also have a margin of error of exactly 20%.
- **C.** The predictions will be exactly correct 80% of the time.
- **D.** The predictions will be useless because the measurements are unreliable.
- **E.** The biologist must obtain more accurate population measurements before attempting to make any predictions.

16. Consider a recent media report linking a food additive with a slight increase in the risk for cancer. Although you have no background knowledge on this topic, you are worried by what you are reading. What should your response be?
- **A.** Go read the original research article, even if you barely understand it.

B. Consult with a recognized expert, such as a medical doctor.
C. Wait for further research studies to appear in the media which either support or dispute the findings before making your decision.
D. Do nothing different, because you don't know anything about this topic.
E. Remove all food additives from your diet just to be safe.

17. A scientist wants to determine the effect of air resistance on a falling basketball. To do this, she will first measure the mass of the basketball (*m*), then drop it from various heights (*h*) while measuring the time required (*t*) to reach the ground. In this experiment,
A. The independent variable is *t,* and the dependent variable is *h*.
B. The independent variable is *t,* and the dependent variable is *m*.
C. The independent variable is *h,* and the dependent variable is *t*.
D. The independent variable is *h,* and the dependent variable is *m*.
E. The independent variable is *m,* and the dependent variable is *h*.

18. Which of the following statements do you agree with most?
A. Science is a complicated set of theories that can sometimes apply to the "real world."
B. Science is a tool used by certain groups to advance their goals.
C. Science is a process designed to test ideas about how the world works
D. Science is a huge collection of facts about various disconnected topics.
E. Science is a set of opinions about the world that may turn out to be wrong.

19. All known animals on Earth use liquid water. If scientists discover a new type of animal on Earth:
A. It will most likely use liquid water.
B. There is no reason to believe beforehand that it will use liquid water.
C. It absolutely must use liquid water.
D. It will definitely use some sort of liquid, though possibly not water.
E. It will not use liquid water, otherwise we would already have known about it.

20. Which of the following statements do you agree with most?
A. Only highly intelligent people can understand science or be a scientist.
B. Science is performed by a collection of people who are advancing a particular agenda.
C. With enough effort and time, anyone is capable of learning about science and becoming a scientist.
D. Some people are born to understand science while others do not have the ability to learn about science.
E. Nobody understands science — some people only pretend to.

AISLE Evaluation (Augustana Interdisciplinary Scientific Literacy Evaluation)

1.	Skill: graph interpretation	**A-0**	**B-2**	**C-1**	**D-0**	**E-1**
2.	Attitude: instruments and measurements	**A-2**	**B-0**	**C-0**	**D-1**	**E-1**
3.	Skill: estimation and checking the	**A-1**	**B-0**	**C-1**	**D-0**	**E-2**

	reasonableness of results					
4.	Attitude: trust in the scientific process	**A-1**	**B-0**	**C-2**	**D-0**	**E-1**
5.	Skill: relating theory to observation	**A-1**	**B-1**	**C-0**	**D-2**	**E-0**
6.	Attitude: mathematical models and experiments	**A-1**	**B-0**	**C-1**	**D-0**	**E-2**
7.	Skill: proposing an experiment to test a hypothesis	**A-1**	**B-1**	**C-0**	**D-2**	**E-0**
8.	Attitude: disagreement in science	**A-0**	**B-0**	**C-1**	**D-2**	**E-1**
9.	Skill: measurement units	**A-1**	**B-2**	**C-0**	**D-0**	**E-1**
10.	Attitude: ethics in science	**A-1**	**B-0**	**C-0**	**D-2**	**E-1**
11.	Skill: simplifying models and limiting cases	**A-2**	**B-0**	**C-0**	**D-1**	**E-1**
12.	Attitude: serendipity in science	**A-1**	**B-1**	**C-0**	**D-2**	**E-0**
13.	Skill: numeracy	**A-1**	**B-1**	**C-2**	**D-0**	**E-0**
14.	Attitude: scientific theories	**A-1**	**B-0**	**C-1**	**D-0**	**E-2**
15.	Skill: uncertainty in measurements and interpretation	**A-2**	**B-1**	**C-1**	**D-0**	**E-0**
16.	Attitude: science and the media	**A-1**	**B-2**	**C-1**	**D-0**	**E-0**
17.	Skill: independent vs. dependent vs. constrained variables	**A-1**	**B-0**	**C-2**	**D-1**	**E-0**
18.	Attitude: scientific thinking	**A-1**	**B-0**	**C-2**	**D-1**	**E-0**
19.	Skill: inductive reasoning	**A-2**	**B-1**	**C-1**	**D-0**	**E-0**
20.	Attitude: who are scientists	**A-1**	**B-0**	**C-2**	**D-1**	**E-0**