CBE—Life Sciences Education
Vol. 15, 1–13, Summer 2016

Article

Multilevel Assessment of Middle School Students' Interest in the Health Sciences: Development and Validation of a New Measurement Tool

William L. Romine,[†]* Michele E. Miller,[‡] Shawn A. Knese,[†] and William R. Folk[§]

[†]Department of Biological Sciences and [‡]Department of Microbiology and Immunology, Wright State University, Dayton, OH 45435; [§]Department of Biochemistry, University of Missouri, Columbia, MO 65211

Submitted February 13, 2015; Revised February 19, 2016; Accepted February 20, 2016
Monitoring Editor: Erin Dolan

Using the context of a 2-wk instructional unit focused on eye and vision health, we developed and validated a multilevel measure of middle school students' interest in science and health careers. This survey contained three subscales positioned differently with respect to curricular content. Interest in Vision Care was most related, but less transferrable to other contexts. Interest in Science was most general, and Interest in Healthcare was positioned between the two. We found that, with two exceptions, items fitted well with validity expectations and were stable across a 2-wk intervention. Further, measures of interest in science, health, and vision-care careers were shown to be reliable and valid. We found that ease of facilitating change across the intervention was generally greater in subscales closely related to the curricular context but that the average magnitude of change in Interest in Healthcare and Interest in Science was not significantly different. We discuss use of these measures in informing instructional efforts and advise that changes in students' perceptions of how science and healthcare relate should be considered in longitudinal analyses.

INTRODUCTION

Given the importance of fostering interest in science in our classrooms, it may seem surprising that interest has not received emphasis in the new vision of science literacy (Next Generation Science Standards Lead States [NGSS], 2013). In light of the absence of accountability for facilitating students' interest and increased pressure to teach to content-heavy high-stakes tests (Jorgenson and Vanosdall, 2002), it may be tempting to ignore the responsibility to focus instruction on promoting interest in science and healthcare. We argue that interest is important in its own right. However, even those in favor of exclusive focus on cognitive growth can appreciate the importance of student interest in supporting conceptual understanding (Alexander *et al*., 1995; Sadler *et al*., 2013). When students are interested in a topic, they pay closer attention in class (Hidi *et al*., 2004), are better at problem solving (McLeod and Adams, 1989), display improved information processing (Pintrich and Schrauben, 1992; Tobias, 1994; Schraw and Lehman, 2001) and increased conceptual understanding (Nieswandt, 2007), and have a greater tendency to pursue science, technology, engineering, and mathematics (STEM) professions (Silvia, 2001).

In this study, we focus on middle school students' interest in science and health careers. Motivation for this focus comes out of the need to promote health literacy in our schools (Joint Committee on National Health Education Standards, 2007) and to address the current shortage in the health workforce (Kelley *et al*., 2004). When students reach middle school, they confront new opportunities and challenges, together with cognitive and social changes (Adams and Berzonsky, 2003), and are required to navigate larger, more diverse learning environments (Hill and Chao, 2009). Middle school is thus a key window of time to support

CBE Life Sci Educ June 1, 2016 15:ar21
DOI:10.1187/cbe.15-02-0034
*Address correspondence to: William L. Romine (romine.william@gmail.com).

students' interest and engagement (Ryan and Patrick, 2001) in health careers through classroom activities.

Assessment of Interest in Science at the Middle School Level

There have been several efforts to assess interest at the middle school level. A portion of this research has focused on situational interest—students' reactions to present circumstances—including peer and parental influences. George and Kaplan (1997) analyzed data collected from the National Educational Longitudinal Study of 1988 using a structural equation modeling methodology. From these data, they found that middle school children are more likely to have a positive attitude toward science if their parents are involved in their school activities and encourage their children in science (George and Kaplan, 1997). Using the Simpson-Troost questionnaire, another study found that during middle school, parents start to be less involved in their child's learning, which could account for students losing interest in science at this age (Atwater *et al*., 1995). Research has also explored instructor influence and classroom environment. Talton and Simpson (1986) developed a survey about the emotional climate of the science classroom, which had an internal consistency reliability of 0.54; the physical environment of the science classroom, which had a reliability of 0.52; and the instructor's influence, which had a reliability of 0.57. The Science Interest Survey (SIS), which was one of the first Rasch-validated surveys of interest in science education, yielded measures with a reliability of 0.72 and adequate construct validity with respect to the Rasch model (Lamb *et al*., 2012).

Several studies have also explored measurement of personal interest—students' dispositions, which are developed through their life experiences. One of the first was the Test of Science Related Attitudes (TOSRA), which had an internal consistency reliability of 0.78 and test–retest reliability of 0.82 (Fraser, 1978). Another early test was the Attitude Toward Science in School Assessment (ATSSA), which had a reliability of 0.94 (Germann, 1988). However, both of these instruments focused exclusively on the degree to which students enjoy science, and at 70 questions long, the TOSRA may lack suitability for time-constrained classroom environments. More recently, research has focused on assessing interest in careers in science (Gibson and Chase, 2002), and two assessments, the STEM Semantic Survey and STEM Career Interest Questionnaire, moved this focus to interest in STEM. Both of these assessments had reliabilities of 0.78–0.94 (Tyler-Wood *et al*., 2010). Another recent assessment is the STEM Career Interest Survey (STEM-CIS). STEM-CIS is a 44-item survey administered to middle school students to gauge their interest in STEM careers, showing alpha values of 0.77, 0.85, 0.89, and 0.86 for the science, math, technology, and engineering subscales, respectively (Kier *et al*., 2014). Most recently, the Student Interest in Technology and Science (SITS) survey was developed and validated for both the high school and college levels. Validated through Item Response Theory and Rasch methods, the SITS measures interest in learning science and technology, and careers, with an internal consistency above 0.8 (Romine *et al*., 2014; Romine and Sadler, 2014). As of now, there is no assessment of personal interest developed in the context of health and medicine.

Assessment of Interest in Health Careers

While research on personal and situational interest in science at the middle school level has been extensive, interest in health careers at this level has received comparatively little attention. One previous assessment compared middle school students' perceptions of an ideal career and a nursing career using 17 parallel items on a five-point Likert scale (Cohen *et al*., 2004). The reliability of the assessment was 0.84 for an ideal career and 0.81 for nursing (Cohen *et al*., 2004). Todaro *et al*. (2013) showed that personal health experiences are a positive predictor for interest in health careers. Todaro and colleagues used the Personal Health Experiences scale, which had an internal consistency reliability of 0.72.

In a pre–post study by Goldsmith *et al*. (2014), students were given a survey about interest in healthcare careers and knowledge about health professions in the context of a career education program. The survey showed a significant increase in students' knowledge of physician's assistant and pharmacy careers compared with their knowledge before the program. There was also an increase in students' interest in health careers in general. Alcaraz and colleagues (2008) gave middle school students a 10-item questionnaire and then distributed magazines they wrote about different health careers, self-assessments, career-planning information, and resources for more information. The middle school students were then given the same questionnaire postintervention. The questionnaire showed increased awareness of four out of the nine careers covered in the magazines. Neither of these studies validated their questionnaires.

Purpose of the Research

While personal interest and interest in health careers have been measured separately, there is currently no measure for personal interest in health careers. We describe the development and validation of such an instrument and illustrate how modern frameworks for educational assessment can be used to develop effective instrumentation for measuring students' interest in the health sciences. Specifically, we describe the application of two important methodological developments in science education assessment: Rasch modeling and multilevel assessment. We demonstrate how these frameworks can be applied to measuring middle school students' affective growth through development and validation of the Assessment of Interest in Medicine and Science (AIMS), a three-subscale tool for measuring student interest in science and health careers at the middle school level. We first sought to investigate the validity and reliability of the AIMS by answering this question and addressing three subquestions:

1. What is the construct validity and reliability of the AIMS subscales?
   a. How well do the data support a three-subscale model for the AIMS?
   b. How well do items on the AIMS conform to behavior that would be predicted for a well-designed rating scale item?
   c. To what extent do AIMS subscales retain their functioning across a 2-wk intervention?

Next, as assessments are always a work in progress, we sought to answer

2. What revisions of items on the AIMS can improve its measurement properties for future research on students' interest in science and health careers?

Finally, per the common need to measure change that is both statistically significant and transferrable to other contexts, we wished to understand how a subscale's tendency to change relates to its overlap with the curricular intervention through the question

3. Are the elements of interest that are positioned more closely to the curriculum more prone to change than elements that are more generalizable across multiple contexts?

METHODS

The Assessment Triangle for Affective Measurement

We consider the cognition–observation–interpretation (COI) assessment triangle (Glaser *et al*., 2001) among the most useful and widely applicable models describing elements that make up a coherent assessment system. It states that assessment development begins with a framework for how students think (cognition). The researcher must then develop a strategy for observing students' thinking (observation). Finally, a model for interpreting the data needs to be utilized (interpretation). Use of the assessment triangle for affective assessment could be criticized on the grounds that it was developed primarily with cognitive assessments in mind (hence use of the word "cognition"). However, we believe that since "cognition" and "affection" are both elements of conceptual change (Treagust and Duit, 2008), it is a small but important step to extend the cognition–observation–interpretation (COI) triangle to the "affection–observation–interpretation" (AOI) triangle to facilitate assessment of affective conceptual change.

Personal Interest: An Affective Framework

We first define our affective framework—what we mean by "interest." While the term "interest" can have a variety of meanings, we focus on assessing *personal interest*, which relates to a student's general disposition about a topic such as science (Alexander and Jetton, 1996). This can be contrasted with *situational interest*, which is largely defined by a student's reaction to present circumstances and tends to change readily with a students' present classroom environment (Nieswandt, 2007).

As instructors, we see manifestations of these two constructs daily in our classrooms. We see students who are naturally curious and tend to express positive dispositions toward science, along with students who express a sense of dread when asked to explore scientific topics. These represent personal interest, general dispositions toward science that have formed over the students' lifetimes. A question like "I enjoy reading about science" would elicit a student's personal interest, since enjoyment of reading about science is reflective of his/her disposition toward science. In contrast, a question like "My science teachers make science interesting" (Lamb *et al*., 2012) elicits situational interest, since it is reflective of a student's reaction to his/her present classroom environment.

A student's general liking of science does not prevent the student from feeling bored when asked to spend an hour completing worksheets; nor does a general dread of science keep a student from feeling engaged and interested during the course of an inquiry-based science activity. These manifestations of situational interest are often largely independent of a student's general disposition toward science. While situational interest is important—we want to foster positive engagement in our classrooms—we consider personal interest to be a more important target for instruction, given its enduring nature and resistance to change (Alexander and Jetton, 1996). Targeting a construct that is resistant to change can be inconvenient when moderate-to-large effect sizes are desired. However, the changes that are facilitated are more enduring. This is an important consideration, as the ultimate decision regarding career choice is distant to a middle school student. Efforts to interest students in particular health careers should therefore target personal interest despite its resistance to change.

Multilevel Assessment

One limitation that pertains to all assessments is the closeness or distance from the intervention context under which they are used, introducing what we call *location bias*. Specifically, assessments aligned closely with instruction will be more sensitive in detecting change, but changes will be less generalizable (Ruiz-Primo *et al*., 2002). Because high reported effect sizes are important to the ultimate perceived success of an educational intervention, it is often beneficial to align assessments closely with instruction. However, this can result in nonmeasurable inflation of type 1 error rate, since generalizable gains are likely to be overestimated. At the extreme, attempting to tie assessments closely with the instructional context can result in overalignment: "When outcome measures are closely aligned with or tailored to the intervention, the study findings may not be an accurate indication of the effects of an intervention" (What Works Clearinghouse, 2014, p. 17). Overalignment is problematic, because effect sizes cannot be generalized outside the specific context of a study. Results from overaligned surveys have questionable validity, because it is impossible to know the extent to which measured growth within instruction will be carried outside the instructional context, and it is therefore impossible to draw meaningful comparisons of growth across multiple instructional contexts.

On the flip side, instruments like the ATSSA, TOSRA, SIS, and Simpson-Troost questionnaires reviewed previously are written in more general contexts that could be applied to various interest-related instructional goals. This reduces the likelihood of inflated effect sizes, while increasing the likelihood that measured growth will be applicable to students' future learning activities. However, effect sizes given by these instruments are more conservative, and therefore less powerful, than those derived from instruments more closely aligned with instruction. How do we rationalize the conflict between the practical need for sensitivity and the epistemic need for generalizability? Multilevel assessment gives us a solution.

An effective way to negotiate location bias is to measure students at a variety of locations with respect to the curriculum of instruction. Multilevel assessment, or assessing students at multiple distances from a curriculum, was introduced as a way to give a broader, more complete perspective on how

students are impacted by instruction (Ruiz-Primo *et al*., 2002). In the STEM education community, use of multilevel assessment has been discussed almost exclusively in the context of how curricular innovations facilitate cognitive gains. With respect to location bias, assessing close to the curriculum trades generalizability for sensitivity; moving farther away from the curriculum sacrifices sensitivity for generalizability (Ruiz-Primo *et al*., 2002). Several studies have utilized multilevel assessment for measurement of cognitive growth associated with STEM curricular innovations (Ruiz-Primo *et al*., 2002; Klosterman and Sadler, 2010; Sadler *et al*., 2013, 2015). However, multilevel assessment has not been applied to affective constructs like interest. This study is the first to apply this important framework to affective assessment.

Development and Administration of the AIMS

The AIMS utilized a four-level Likert scale (0 = strongly disagree [SD], 1 = disagree [D], 2 = agree [A], and 3 = strongly agree [SA]). Questions were developed using item-wording structures from the ATSSA survey (Germann, 1988), the Science Opinion Survey (Allen *et al*., 1999), and the Student Interest in Science and Technology (SITS) survey (Romine *et al*., 2014; Romine and Sadler, 2014). To accommodate the inherent conflict between the enduring nature of personal interest and the need to detect change in response to relatively short interventions, we integrated a three-level multilevel assessment framework. In light of the intervention focused on the eye and vision used in this study, the three distances for this particular study were Interest in Science (distal), Interest in Healthcare (proximal), and Interest in Vision Care (close).

We administered the AIMS to grade 6–8 students in the context of a 2-wk intervention focusing on exploring and understanding vision. The intervention contained a variety of inquiry-based activities connected to the physical and life sciences disciplinary core ideas within the Next Generation Science Standards (NGSS Lead States, 2013). These included dissection of a cow eye, building and testing models of different types of eyes and discussing their evolutionary history, exploring vision disorders and how they are corrected, addressing misconceptions about blindness, exploring coping strategies used by people who are blind, and discussing careers in eye and vision care.

The AIMS was validated through three cycles of data collection. The first version of the AIMS contained 32 items targeting interest in science, healthcare, and vision care as well as related STEM constructs such as engineering and mathematics. This was administered to 70 middle school students before and after the intervention in a rural midwestern school. The instrument was revised and reduced to 26 items based on reliability analyses, and again administered pre–post to 20 middle school students in a different rural midwestern school. During these first two trials, groups of common items were identified through principal components factor analysis, and an internal consistency definition of reliability (Cronbach's alpha) was used to quantify the precision by which each level of interest was measured.

Results from the second trial indicated that the 26-item assessment yielded reliable measures of students' personal interest according to our three-level multilevel assessment framework. However, data from a larger sample evaluated using Rasch modeling were needed to arrive at more detailed conclusions regarding validity of items and the survey as a whole. Therefore the final 26-item instrument was administered to 247 students taught by two instructors at an urban midwestern middle school. Of these students, 226 (107 male and 119 female) took the pretest and 212 (99 male and 113 female) took the posttest. A total of 194 were measured at both time points.

Validity of the Measurement Model, and Stability across Time

Before a survey is used for data collection in intervention studies, it should have a well-defined measurement structure that does not change over time. Ordinal confirmatory factor analysis (CFA) with LISREL 8.80 was used to quantify the extent to which items measured their hypothesized construct consistently across time. Based on data from the first two validation cycles, nine items (S1–9) were hypothesized to measure interest in science, 11 items (H1–11) to measure interest in healthcare, and six items (V1–6) to measure interest in vision care. Factor loadings were allowed to vary freely between pre- and posttests to allow detection and evaluation of changes in factor loadings across time. Goodness of fit was evaluated using the root-mean-square error of approximation (RMSEA). We used a RMSEA below 0.06 to indicate close fit with the data (Hu and Bentler, 1995).

In addition to comparison of factor loadings, the presence of differential item functioning (DIF) across time was evaluated at the 0.05 alpha level using ordinal logistic regression. For detecting DIF, the probability of a student's response on an item is expressed as a function of the student's interest measure, time, and the interaction between interest measure and time. The presence of uniform DIF, measured by a significant time parameter, indicates that the item's agreeability changes across time. The presence of nonuniform DIF, measured by a significant interest-by-time interaction parameter, indicates that the item's ability to discriminate between students of high- and low-interest changes across time (Swaminathan and Rogers, 1990).

Validation with Respect to the Rasch Measurement Framework

The Rasch model gives science educators an approach to test and survey validation that resembles how scientists calibrate their machines in the lab—when validity is questionable, data are collected by the machine and evaluated against an objective standard that is accepted as true by the field. The Rasch model imposes a philosophical criterion for how clean, correct, and useful data should look, stating that the probability of a student choosing a correct response should be proportional only to the difference between the difficulty of the item and the student's ability (Wright and Stone, 1979). From the perspective of assessing interest, this statement can be restated in terms of a student's interest level and item's difficulty (or "disagreeability" under the framework of interest). The Rasch model defines a data-independent scale, and the quality of the data produced by the survey is evaluated based on fit with the model and the resulting scale.

The Rasch rating scale model (Andrich, 1978) was used to provide a criterion for construct validity of the AIMS items and rating scales and to quantify the reliability of student

and item measures along the scale. Rasch models specify an objective data-independent criterion with which to evaluate data. Students placed higher on the Rasch scale have a tendency to express higher interest, and items placed higher on the Rasch scale are more difficult to agree with. To illustrate, if a student's level of interest is above the "disagreeability" location of the item, then the Rasch model will predict that this student will express high interest on that item. On the other hand, if a student's level of interest is below the item's location, he/she is predicted to express a low level of interest on that item. If the student's location matches the item's agreeability location perfectly, then that student is predicted to have a 50% chance of showing high interest on that item.

To evaluate validity of the interest scales, we looked at the model residuals (or leftover variance after the Rasch model is fitted to the data). If a single Rasch model explains the data adequately, there should be no detectable pattern in model residuals. We used principal components analysis (PCA) on Rasch residuals as a means to detect such a pattern if it existed. Simulation studies on polytomous data such as those in this study suggest that a first eigenvalue around or below 2 indicates no detectable pattern in the residuals, and thus no significant presence of underlying constructs unaccounted for by the model (Linacre and Tennant, 2009). Mean-squares infit and outfit measures were used to evaluate how well individual items and their underlying rating scales fit with the Rasch rating scale model. While these have expected values of 1.0, fit values between 0.5 and 1.5 indicate that an item generates data that are useful for measurement (Wright *et al*., 1994).

We used the characteristic curve for each rating scale (Boone *et al*., 2011) to determine whether or not having four levels (SD, D, A, SA) on the rating scale was necessary for measuring students' interest on each subscale. We refer the reader to Linacre (2005) for visual examples of well-behaved and irregular rating scales. In a well-behaved rating scale, meaning that all four levels of agreement are unique and useful, the characteristic curve for each rating scale would show that the probability of observing a student at each respective level of agreement would be maximized at a specific region along the Rasch scale. Students on the low end of the Rasch scale should be predicted to select "strongly disagree." Students situated higher should be predicted to select "disagree." Those higher yet should be predicted to select "agree," and students at the top of the Rasch scale should be predicted to select "strongly agree." Irregular rating scales have characteristic curves showing that students are not likely to select the middle categories at any level. In this case, we can conclude that the four-level rating scale is unnecessary and that the scale can be reduced to two or three levels in future research.

Analysis of Instrument Sensitivity

While the outcomes of multilevel assessment have been addressed in the context of pre–post gains (Sadler *et al*., 2013, 2015), we decided to take this a step further and look at absolute change in addition to evaluating gains. In the context of multilevel assessment, it is informative to quantify the *sensitivity* of measures with respect to distance from the intervention unit. Sensitivity was defined as the average magnitude of change observed in students across the 2-wk intervention. Within-subjects multivariate analysis of variance (MANOVA) was used to test the null hypothesis ($\alpha = 0.05$) of no difference in scale sensitivity within students between Interest in Vision Care, Interest in Healthcare, and Interest in Science, using the F statistic calculated from Wilk's lambda. Assuming that the null hypothesis was rejected, post hoc contrasts of changes between curricular distances using the Bonferroni adjustment were conducted ($\alpha = 0.05$). Before MANOVA comparisons were made, the three Rasch subscales were equated to a common scale ranging between 0 (no interest) and 10 (highest interest) to aid interpretation of differences in sensitivity between levels.

RESULTS

Construct Validity and Reliability of the AIMS Subscales

The hypothesized measurement structure of the AIMS, containing three moderately correlated dimensions (interest in science, healthcare, and vision care) adequately fitted the data (RMSEA = 0.058). In addition, we found little difference in factor loadings between pre- and posttests (Table 1). The loading of item H2 onto interest in healthcare careers

Table 1. Standardized factor loadings of items onto interest measures

	SciPre[a]	SciPost[a]	HealthPre[b]	HealthPost[b]	VisPre[c]	VisPost[c]
S1	0.70	0.74	—	—	—	—
S2	0.34	0.42	—	—	—	—
S3	0.77	0.79	—	—	—	—
S4	0.67	0.78	—	—	—	—
S5	0.43	0.39	—	—	—	—
S6	0.66	0.71	—	—	—	—
S7	0.68	0.71	—	—	—	—
S8	0.65	0.72	—	—	—	—
S9	0.61	0.63	—	—	—	—
H1	—	—	0.66	0.63	—	—
H2	—	—	0.39	0.27	—	—
H3	—	—	0.55	0.64	—	—
H4	—	—	0.80	0.75	—	—
H5	—	—	0.77	0.72	—	—
H6	—	—	0.51	0.53	—	—
H7	—	—	0.68	0.75	—	—
H8	—	—	0.85	0.81	—	—
H9	—	—	0.76	0.78	—	—
H10	—	—	0.71	0.75	—	—
H11	—	—	0.78	0.77	—	—
V1	—	—	—	—	0.66	0.73
V2	—	—	—	—	0.70	0.74
V3	—	—	—	—	0.68	0.59
V4	—	—	—	—	0.75	0.71
V5	—	—	—	—	0.76	0.80
V6	—	—	—	—	0.66	0.71

[a]SciPre and SciPost indicate pre- and posttests for Interest in Science.
[b]HealthPre and HealthPost indicate pre- and posttests for Interest in Healthcare.
[c]VisPre and VisPost indicate pre- and posttests for Interest in Vision Care.

underwent the largest change, from a value of 0.39 on the pretest to 0.27 on the posttest.

Results from logistic regression indicate the presence of small DIF effects across time on three items. Uniform DIF was detected on items S1 ($b = -0.48$, $\chi^2(1) = 4.56$, $p = 0.03$, OR = 0.62) and S5 ($b = -0.46$, $\chi^2(1) = 6.34$, $p = 0.012$, OR = 0.63), suggesting that these items were slightly more agreeable at the beginning of the intervention. Nonuniform DIF on item H2 ($b = 0.267$, $\chi^2(1) = 5.41$, $p = 0.02$, OR = 1.31) suggests the item's ability to discriminate between students of high and low interest decreased after the intervention.

The Rasch model corroborated findings from CFA suggesting scales that measure a single construct. First eigenvalues from PCA on standardized residuals with respect to the Rasch model sat at 1.12, 1.13, and 1.08 items of variance for interest in science, healthcare, and vision care, respectively. These sit well below the value of 2 that is used to indicate a unidimensional scale (Linacre and Tennant, 2009). Further, the three dimensions were moderately correlated (0.41–0.66), suggesting expected moderate dependency between the scales.

Item locations (Figure 1) were measured with a reliability of 0.99, 0.97, and 0.97 for interest in science, healthcare, and vision care, respectively. Locations for the nine items measuring interest in science ranged from –1.99 to 0.85 logits. Ranges of locations for the 11 items measuring interest in healthcare and vision care were narrower, at –0.90 to 0.54 and –0.79 to 0.55 logits, respectively. We found no significant differences between distributions of student and item measures, indicating that the AIMS is well targeted to middle school students. This is corroborated by Rasch person-measurement reliabilities of 0.87, 0.90, and 0.86 for interest in science, healthcare, and vision care, respectively. These values are well above the minimum value of 0.50 specified by the What Works Clearinghouse (2014) and indicate that scales have sufficient precision for comparisons of individual middle school students in the context of health interventions.

Figure 1 indicates sufficient spread of item measures to quantify students' locations along the scale precisely. The easiest items on the respective scales were S2 (I enjoy doing science experiments), H6 (I like helping people get healthy), and V3 (I am interested in how the eye works). The position of these items at the bottom of the scales (Figure 1) indicates that they are relatively easy to agree with. A majority of students' measures sit above the locations of these items (Figure 1). At the other extreme, S7 (I enjoy reading about science), S8 (I would like to work in a science laboratory), H5 (I enjoy reading about doctors), H7 (I would enjoy working in a medical laboratory), H11 (I would like to work in a medical field), V1 (I want to know what it is like to be an eye specialist), and V5 (I would like to learn more about jobs in eye care) are comparatively difficult for students to agree with. Approximately one-third of the students had measures above these item locations, indicating that a majority of the students were not likely to express high interest on these items.

We found that the three rating scales for interest in science, healthcare, and vision care, respectively, were useful and displayed adequate fit with the Rasch model. Mean-squares fit values for respective levels of each rating scale (SD, D, A, SA) ranged between 0.87 and 1.18 for interest in science, 0.79 and 1.25 for interest in healthcare, and 0.83 and 1.23 for interest in vision care. Satisfactory fit indicates that our data matched the Rasch expectation that students with higher interest will tend to choose higher levels of agreement on each rating scale.

Four unique categories for each dimension's rating scale were revealed by the Rasch model (Figure 2, A–C). For example, Figure 2A illustrates that students below a logit measure of –2.12 were most likely to select a rating of "strongly disagree." Students between –2.12 and –0.24 were predicted to select "disagree," students between –0.24 and 2.36 were predicted to select "agree," and students above 2.36 were predicted to select "strongly agree." We similarly see four distinct response regions for interest in healthcare (Figure 2B) and interest in vision care (Figure 2C). That these rating scales are well-behaved and fit well with the Rasch model demonstrates the utility of each four-tiered rating scale as a measure of interest in the health sciences.

All items display satisfactory fit with the Rasch model (mean-squares fit below 1.5), except for S5 and H2 (Table 2). Item S5, measuring interest in science, is a relatively agreeable item (difficulty = –0.22; Figure 1), which displays an infit of 1.79 and an outfit of 1.77. Item H2, which measures interest in healthcare careers, is also relatively agreeable (difficulty = –0.51) and displays an infit of 1.72 and an outfit of 1.98.

Description of Problematic Items

Significant DIF across time and misfit with the Rasch model collectively suggest problematic wording. Item S5 (More time in schools should be spent doing science experiments) is worded differently from other items measuring interest in science, such as item S1 (I like learning science) or S2 (I enjoy doing experiments) in that it specifically addresses the school environment. Indeed, students who are not interested in science may nonetheless yearn for experiments as opposed to didactic lecture-based forms of teaching/learning. This item may therefore measure situational interest in addition to personal interest. Item H2 (Studying how parts of the body work is boring) is a negatively worded item. Empirical studies suggest that negative wording may not solicit the simple logical converse emotion of positive wording and may have a tendency to confuse respondents (Colosi, 2005). Our data suggest that this is also the case for the middle school students in this study. Taking the school context out of item S5 and making the wording of item H2 positive will likely further improve the measurement validity of the AIMS for measuring personal interest in the context of future research.

Efficacy of the Multilevel Assessment Structure

After translation of respective Rasch logit scales to a common scale between 0 and 10, we found significant differences in sensitivity with distance from the intervention ($\Lambda = 0.915$, $F_{2192} = 8.86$, $p << 0.001$). The close construct, Interest in Vision Care, had the highest sensitivity (M = 2.18, SD = 1.77, $CI_{95\%}$ = 1.93–2.43). The proximal construct, Interest in Healthcare (M = 1.68, SD = 1.45, $CI_{95\%}$ = 1.47–1.88), and the distal construct, Interest in Science (M = 1.70, SD = 1.43, $CI_{95\%}$ = 1.50–1.91), had significantly lower sensitivity than Interest in Vision Care at the $\alpha = 0.01$ level, but showed no significant differences between each other.

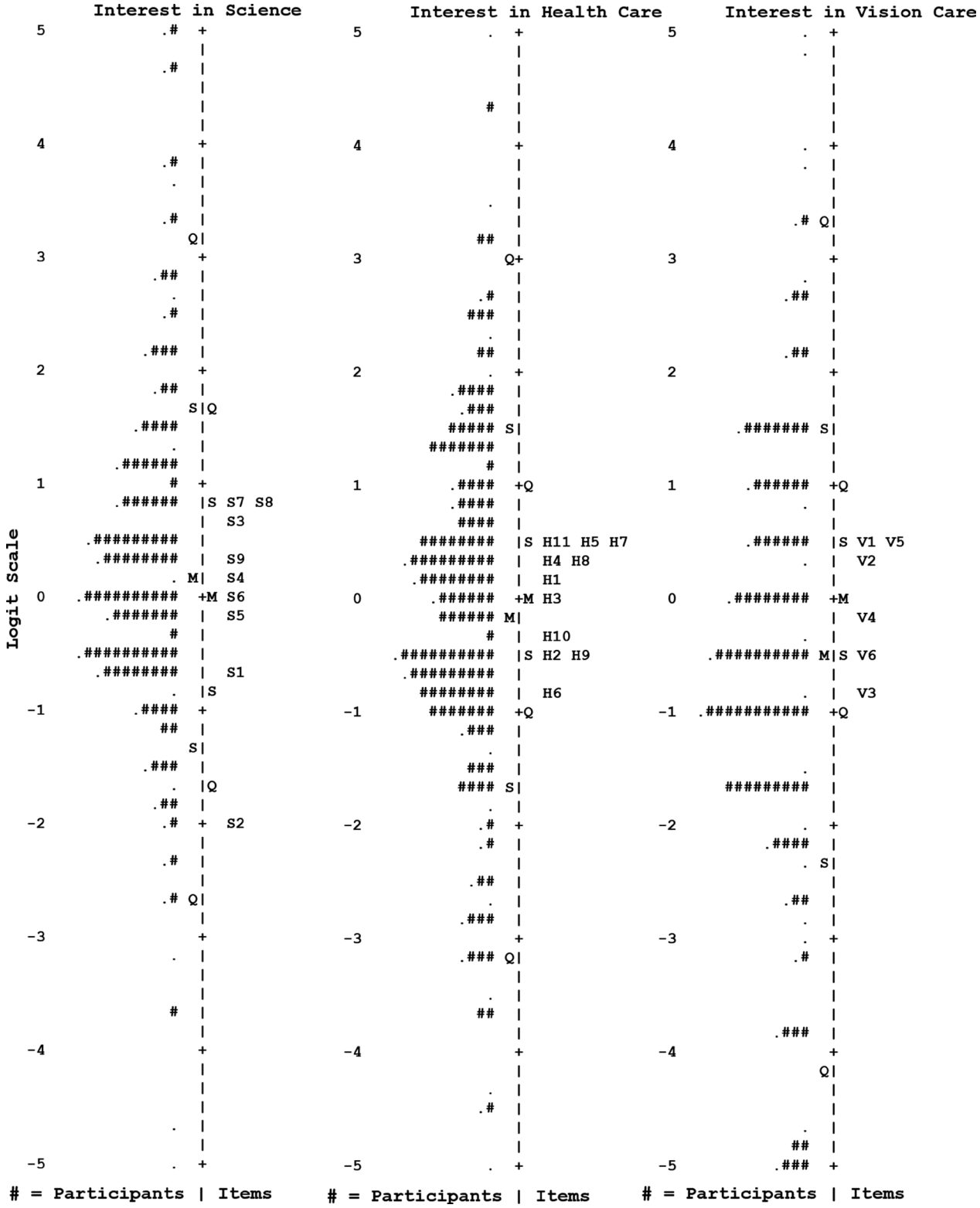
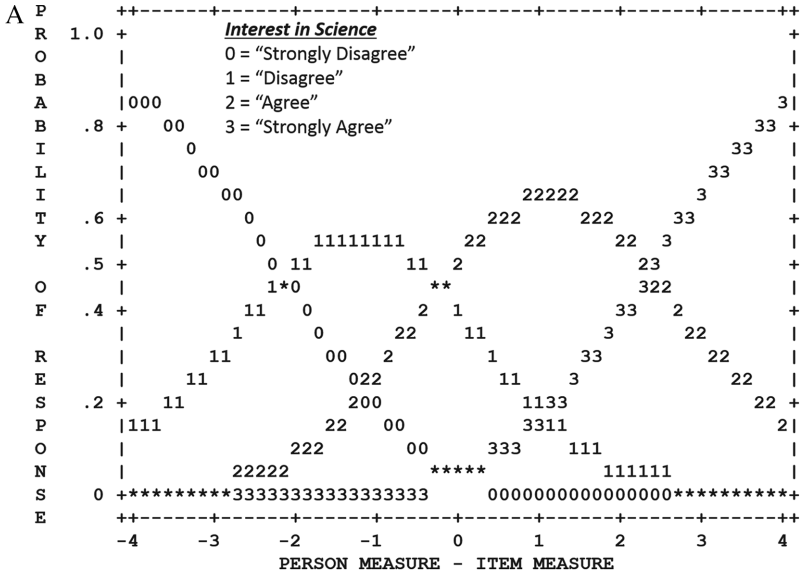


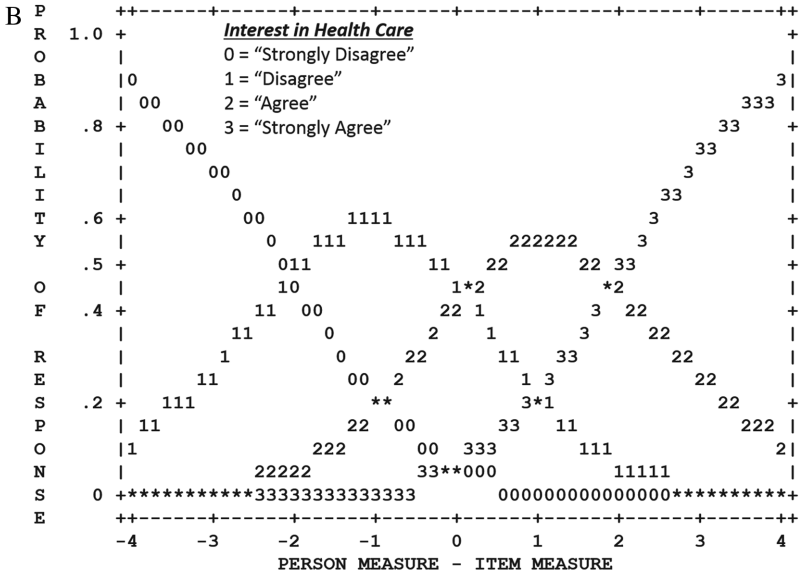
Figure 1. Person-item maps along the Rasch logit scale for Interest in Science, Interest in Healthcare, and Interest in Vision Care.

In line with results from other studies utilizing multilevel assessment in cognitive contexts (Sadler *et al*., 2013, 2015), effect sizes for pre–post changes generally increased with proximity to the intervention (science: Cohen's D = 0.07; healthcare: Cohen's D = 0.14; vision care: Cohen's D = 0.40). Much like our sensitivity measures, however, the effect sizes

A



B



C

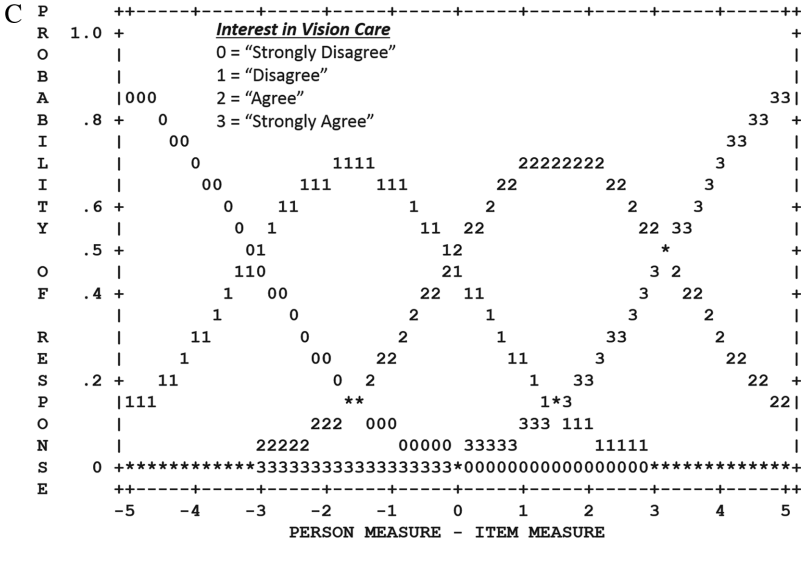


Figure 2. (A) Characteristic curve for the Interest in Science rating scale, showing four distinct response regions. (B) Characteristic curve for the Interest in Healthcare rating scale, showing four distinct response regions. (C) Characteristic curve for the Interest in Vision Care rating scale, showing four distinct response regions.

Table 2. Rasch measures for items on the AIMS

Interest	Item	Difficulty[a]	SE	Infit	Outfit	PtBis	Statement (SD, D, A, SA)
Science							
	S1	–0.63	0.07	0.67	0.66	0.64	I like learning science.
	S2	–1.99	0.10	1.29	1.17	0.34	I enjoy doing experiments.
	S3	0.71	0.07	0.70	0.70	0.69	I plan to take a lot of science classes in high school.
	S4	0.23	0.07	0.81	0.82	0.63	I would like to make discoveries using science.
	S5[b]	–0.22	0.10	1.79	1.77	0.35	More time in school should be spent doing science experiments.
	S6	0.01	0.07	0.80	0.85	0.57	I like learning how to use science in my life.
	S7	0.77	0.07	0.84	0.87	0.58	I enjoy reading about science.
	S8	0.85	0.07	0.98	1.00	0.57	I would like to work in a science laboratory.
	S9	0.27	0.08	1.14	1.12	0.52	Learning about science makes me a better person.
Healthcare							
	H1	0.17	0.07	0.94	1.05	0.61	I want to learn more about careers in medicine.
	H2[b,c]	–0.51	0.09	1.72	1.98	0.31	Studying how parts of the body work is boring.
	H3	–0.05	0.07	1.00	1.06	0.56	Making discoveries in medicine would be interesting.
	H4	0.25	0.07	1.02	0.97	0.71	I would like to become a doctor or nurse someday.
	H5	0.54	0.07	0.76	0.74	0.68	I enjoy reading about doctors.
	H6	–0.90	0.08	1.15	1.25	0.47	I like helping people get healthy.
	H7	0.51	0.07	0.85	0.82	0.68	I would enjoy working in a medical laboratory.
	H8	0.30	0.07	0.74	0.71	0.76	I would like to work in a doctor's office.
	H9	–0.42	0.07	0.87	0.86	0.73	Helping a doctor or nurse over the summer would be interesting.
	H10	–0.38	0.07	1.03	1.00	0.69	Working in a hospital over the summer would be interesting.
	H11	0.51	0.07	0.87	0.83	0.73	I would like to work in a medical field.
Vision care							
	V1	0.49	0.08	0.99	0.98	0.57	I want to know what it is like to be an eye specialist.
	V2	0.37	0.08	0.89	0.89	0.60	I would like to study vision.
	V3	–0.79	0.09	1.15	1.16	0.49	I am interested in how the eye works.
	V4	-0.20	0.08	0.97	0.98	0.61	I would like to help improve people's vision.
	V5	0.55	0.08	0.79	0.80	0.64	I would like to learn more about jobs in eye care.
	V6	–0.42	0.09	1.13	1.12	0.51	I would enjoy caring for people with vision problems.

[a]Logit scale.

[b]Poor fit with Rasch rating scale model (means-squares infit or outfit > 1.5).

[c]Negatively worded item.

for changes in interest in science and healthcare are relatively small and close together, whereas the effect size for interest in vision care is considerably larger. The changes in interest in science and medical careers were nonsignificant at the 90% confidence level. Cohen's D values below 0.2 indicate that these changes also have little practical significance (Cohen, 1988). A small but statistically significant decline was found for interest in vision care ($t_{df = 193} = 3.92$, Cohen's D = 0.40). This indicates that students were slightly less inclined to want to become vision care professionals after the intervention.

DISCUSSION

Interest in STEM developed in middle school impacts motivation to learn and engage successfully in appropriate programs of study at the high school level, and previous research indicates that even relatively modest 2-wk interventions can make significant lasting impacts on middle school students' interest in science (Gibson and Chase, 2002). However, the difficulty of making significant changes in personal interest with short interventions has been documented in previous work (Romine *et al*., 2013; Sadler *et al*., 2015) as has

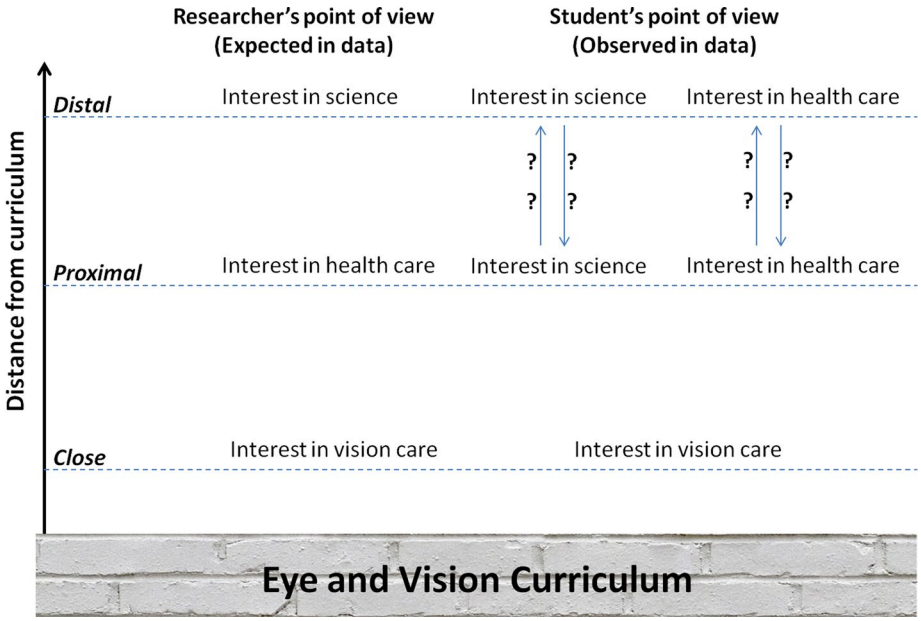


Figure 3. Hypothesized contrast between researchers' and students' interpretation of the AIMS's multilevel assessment framework.

the success of more enduring efforts in evoking significant change (Romine and Sadler, 2014). While these studies target high school and college students, respectively, our data suggest that a similar difficulty applies in the case of middle school students. Further research is needed to understand the effects of longer-lasting interventions in facilitating change in middle school students' personal interest.

The AIMS addresses the location bias inherent in assessment by measuring interest in vision care, healthcare, and science at the middle school level. Further, the AIMS is the first instrument to assess interest in science and healthcare as moderately related constructs positioned at multiple distances from an educational intervention. Hence, the AIMS can provide both sensitivity and generalizability in the context of educational interventions aimed at improving students' interest in the health sciences.

We found that the sensitivity of the AIMS was greatest when measuring students' interest in vision care (closest to the intervention), which we expected in light of the multilevel assessment framework. However, we were surprised to find no significant difference in sensitivity between measures of interest in healthcare and science, which we labeled a priori as "proximal" and "distal," respectively. This lack of sensitivity difference between proximal and distal levels conflicts with the expectation that sensitivity should decrease significantly as we move away from the intervention (Ruiz-Primo *et al*., 2002). This observation provides an admonition that the functioning of an assessment depends not on how the researcher interprets the survey but upon the interpretation by the study's participants (Figure 3). Lack of sensitivity difference between adjacent levels defined as "proximal" and "distal" may be reflective of how middle school students view the relationship between science and healthcare. From the student's perspective, careers in science and medicine may be equidistant from the intervention (Figure 3). A student who sees science and healthcare as highly related may show similar growth on both constructs, while a student who sees them as completely distinct may show gains that are quite distinct in magnitude. It is worth noting that a major goal of the Next Generation Science Standards (NGSS Lead States, 2013) has been to facilitate students' understandings of both the connectedness among the various STEM disciplines, including health, forensics, and medicine, and the distinctions between these careers as subdisciplines of STEM.

This possible distinction between our and middle school students' views of the AIMS's multilevel structure does not compromise validity of the individual scales, given that the three scales measure their constructs unidimensionally. However, if a multilevel assessment framework is used in longitudinal analyses of these constructs, it is useful to keep in mind that the interaction between interest in science and healthcare that students perceive may change as students gain more exposure to specific topics in these fields. This may affect interpretation of questions on the AIMS. We would expect this type of change to be especially visible in intensive STEM-focused learning environments like those offered by STEM schools (Olszewski-Kubilius, 2009) and in schools developing enduring instructional sequences with high fidelity to the NGSS. If used over such an extensive time period as required to evoke significant changes in interest in science and health careers, the multilevel assessment framework behind the AIMS may make its holistic interpretation of scores multidimensional, measuring interest in health careers *and* understanding of the relationship between science and healthcare. In statistical terms, we may expect to see a time-based dependency of the covariance between interest in science and healthcare. In the context of extended interventions, we therefore recommend use of the AIMS with an additional question or series of questions aimed at understanding how students position the fields of science and healthcare relative to each other, which may help contextualize change over time at different distances from the curriculum.

Limitations of the AIMS

We show that the AIMS provides useful measures of interest in health-related careers for middle school students. However, survey measures of personal interest come into their own in extensive pre–post and longitudinal (Romine and Sadler, 2014) designs associated with evaluating effects of curricular interventions or learning environments. Just as interest changes over a sufficient period of time, students change in other ways, too, which may affect how they manifest their interest. The AIMS does not take these external factors into account directly. With respect to instrument validity, changes that affect students' interpretation of items, as discussed previously, are of primary concern.

Another limitation of the AIMS is that it is only validated for middle school students. We hypothesize that the AIMS may retain its validity for high school students. However, we suggest that using the AIMS in studies reaching into college and elementary contexts may require a thorough validity analysis for those particular age groups.

Implications for Middle School Health Contexts

Implementing the suggestions for improvement described above will only improve the measurement validity of the AIMS. Due to its multilevel framework, practitioners and researchers can use the AIMS as a meaningful measure of personal interest in the context of a variety of health-related interventions for middle school and beyond. Interest in Science, which we defined as the most distal measure of interest, is meaningful across any intervention focused on science instruction aimed at changing personal interest. Similarly, Interest in Healthcare provides useful information for any intervention focused on health careers. Interest in Vision Care is meaningful in the context of the intervention in this study, but this close measure of interest will need to be adapted for the specific instructional context.

Career awareness and planning is becoming an important part of middle school counseling programs around the nation (American School Counselor Association, 2014), and extracurricular programs aimed at increasing interest in health careers are increasing in popularity. For example, Future Health Professionals (HOSA) is an international organization designed to promote interest in health professions by sponsoring afterschool activities, competitions, and career-path advising. Area Health Education Center (AHEC), a national organization, advances the goal of recruitment, training, and retention of healthcare workers in the United States. With this commitment toward promoting health careers in middle and high school comes the need for efficient and valid measurement tools aimed at understanding how these efforts facilitate student interest. While tracking students impacted by these programs through college and into their careers may provide the best summative measure of the extent to which these efforts meet their aims, we believe periodic formative measurement with valid survey measures is essential to maximizing the impact of these programs.

Measures of personal interest can inform a variety of efforts focused on engaging students in health careers. At the school level, impacts of peer-based health-career counseling approaches, including those facilitated by extracurricular organizations like HOSA and AHEC, can be explored. Peer-based approaches have shown promise for facilitating constructive health behaviors (Posavac *et al*., 1999), and the same may hold true in the context of facilitating interest in health careers. It may also be interesting to measure the differential impacts of diverse school-based health career counseling and intervention structures. Measures of interest can also be combined with other metrics such as mathematics and science grades and aptitude test scores to support schools' efforts to direct students toward successful and meaningful career paths by addressing multiple aspects of conceptual change together.

In a study of doctors in Finland, Hyppölä and colleagues (1998) found that 22% of doctors stated that if they had to do it over, they would choose a different profession. The greatest dissatisfaction was expressed by students who entered medicine due to its academic rigor, whereas the highest satisfaction was expressed by doctors who entered the profession due to enjoyment of human interaction. Further, Dinsmore *et al*. (2001) found that a portion of medical students experience negative emotions such as fear and disgust when assigned to undertake cadaver dissection in their anatomy class.

Informal feedback from interviews with participating students suggest that the cow eye dissection elicited similar emotions from many students, which may account for the small but statistically significant reduction in interest in vision care observed in this study. This said, informal feedback from interviews with both teachers and students suggest consistently that the cow eye dissection is overwhelmingly students' favorite activity in the curricular set. This suggests that activities like dissection may lead to gains in situational interest (not measured directly in this study) while helping some students realize that a career in medicine may not fit well with their personalities. As with any data, changes in dispositions must be considered in context, and a validated survey like the AIMS can provide a useful complement to other types of data in helping educators understand how instruction is impacting students' dispositions. By using the AIMS in conjunction with learning experiences, educators can use valid data to make informed evaluations of the impact their instruction is having on student interest, deduce possible reasons for the measured effects, and proceed with instruction accordingly.

ACKNOWLEDGMENTS

This work was supported by the Missouri Foundation for Health (12-0463-WFD-12).

REFERENCES

Adams GR, Berzonsky MD (eds.) (2003). Blackwell Handbook of Adolescence, Malden, MA: Blackwell.

Alcaraz KI, Kreuter MW, Davis KL, Rogers VL, Samways TW, Bryan RP (2008). Increasing awareness of and interest in public health and cancer control careers among minority middle school students. Public Health Rep *123*, 533–539.

Alexander PA, Jetton TL (1996). The role of importance and interest in the processing of text. Educ Psychol Rev *8*, 89–121.

Alexander PA, Jetton TL, Kulikowich JM (1995). Interrelationship of knowledge, interest, and recall: assessing a model of domain learning. J Educ Psychol *87*, 559.

Allen NL, Carlson JE, Zelenak CA (1999). The 1996 NAEP Technical Report, Report No. NCES-1999-452, Washington, DC: National Center for Education Statistics.

American School Counselor Association (2014). Why Middle School Counselors? www.schoolcounselor.org/school-counselors-members/careers-roles/why-middle-school-counselors(accessed 21 October 2014).

Andrich D (1978). A rating formulation for ordered response categories. Psychometrika *43*, 561–573.

Atwater MM, Wiggins J, Gardner CM (1995). A study of urban middle school students with high and low attitudes toward science. J Res Sci Teach *32*, 665–677.

Boone WJ, Townsend JS, Staver J (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: an exemplar utilizing STEBI self-efficacy data. Sci Educ *95*, 258–280.

Cohen J (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Hillsdale, NJ: Erlbaum.

Cohen JA, Palumbo MV, Rambur B, Mongeon J (2004). Middle school students' perceptions of an ideal career and a career in nursing. J Prof Nurs *20*, 202–210.

Colosi R (2005). Negatively worded questions cause respondent confusion. Proc Surv Res Methods Sect, Am Stat Assoc, Minneapolis, MN, August 2005, 2896–2903.

Dinsmore CE, Daugherty S, Zeitz HJ (2001). Student responses to the gross anatomy laboratory in a medical curriculum. Clin Anat *14*, 231–236.

Fraser BL (1978). Development of a test of science-related attitudes. Sci Educ *62*, 509–515.

George R, Kaplan D (1997). A structural model of parent and instructor influences on science attitudes of eighth graders: evidence from NELS: 88. Sci Educ *82*, 93–109.

Germann PJ (1988). Development of the attitude toward science in school assessment and its use to investigate the relationship between science achievement and attitude toward science in school. J Res Sci Teach *25*, 689–703.

Gibson HL, Chase C (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. Sci Educ *86*, 693–705.

Glaser R, Chudowsky N, Pellegrino JW (eds.) (2001). Knowing What Students Know: The Science and Design of Educational Assessment, Washington, DC: National Academies Press.

Goldsmith C, Tran TT, Tran L (2014). An educational program for underserved middle school students to encourage pursuit of pharmacy and other health science careers. Am J Pharmac Educ *78*(9), 1.

Hidi S, Renninger KA, Krapp A (2004). Interest, a motivational variable that combines affecting and cognitive functioning. In: Motivation, Emotion, and Cognition: Integrative Perspectives on Intellectual Functioning and Development, ed. DY Dai and RJ Sternberg, Mahwah, NJ: Erlbaum, 89–115.

Hill NE, Chao RK (2009). Families, Schools, and the Adolescent: Connecting Research, Policy, and Practice, New York: Instructors College Press.

Hu LT, Bentler P (1995). Evaluating model fit. In: Structural Equation Modeling. Concepts, Issues, and Applications, ed. RH Hoyle, London: Sage, 76–99.

Hyppölä H, Kumpusalo E, Neittaanmaki L, Mattila K, Virjo I, Kujala S, Luhtala R, Halila H, Isokoski M (1998). Becoming a doctor—was it the wrong career choice? Soc Sci Med *47*, 1383–1387.

Joint Committee on National Health Education Standards (2007). National Health Education Standards: Achieving Health Literacy, 2nd ed., Atlanta, GA: American Cancer Society.

Jorgenson O, Vanosdall R (2002). High-stakes testing—the death of science? What we risk in our rush toward standardized testing and the three R's. Phi Delta Kappan *83*, 601.

Kelley MA, Angus D, Chalfin DB, Crandall ED, Ingbar D, Johanson W, Medina J, Sessler C, Vender JS (2004). The critical care crisis in the united states: a report from the profession. Chest J *125*, 1514–1517.

Kier MW, Blanchard MR, Osborne JW, Albert JL (2014). The development of the STEM career interest survey (STEM-CIS). Res Sci Educ *44*, 461–481.

Klosterman ML, Sadler TD (2010). Multi-level assessment of scientific content knowledge gains associated with socioscientific issues based instruction. Int J Sci Educ *32*, 1017–1043.

Lamb RL, Annetta L, Meldrum J, Vallett D (2012). Measuring science interest: Rasch validation of the science interest survey. Int J Sci Math Educ *10*, 643–668.

Linacre JM (2005). Dichotomous & polytomous category information. Rasch Measurement Transactions *19*, 1005–1006.

Linacre JM, Tennant A (2009). More about critical eigenvalue sizes (variances) in standardized-residual principal components analysis (PCA). Rasch Measurement Transactions *23*, 1228.

McLeod DB, Adams VM (1989). Affect in Mathematical Problem Solving: A New Perspective, New York: Springer-Verlag.

Next Generation Science Standards Lead States (2013). Next Generation Science Standards: For States, By States. www.nextgenscience.org/ (accessed 13 May 2016).

Nieswandt M (2007). Student affect and conceptual understanding in learning chemistry. J Res Sci Teach *44*, 908–937.

Olszewski-Kubilius P (2009). Special schools and other options for gifted STEM students. Roeper Rev *32*, 61–70.

Pintrich PR, Schrauben B (1992). Students' motivational beliefs and their cognitive engagement in academic tasks. In: Students' Perception in the Classroom: Causes and Consequences, ed. D Schunk and J Meece, Hillsdale, NJ: Erlbaum, 149–183.

Posavac EJ, Kattapong KR, Dew DE Jr. (1999). Peer-based interventions to influence health-related behaviors and attitudes: a meta-analysis. Psychol Rep *85*, 1179–1194.

Romine W, Sadler T (2014). Measuring interest in science and technology at the college level in response to two instructional interventions. Res Sci Educ, DOI:10.1007/s11165-014-9452-8.

Romine W, Sadler T, Presley M, Klosterman M (2014). Interest in STEM: development and validation of an instrument for measuring student interest. Int J Sci Math Educ *12*, 261–283.

Ruiz-Primo MA, Shavelson RJ, Hamilton L, Klein S (2002). On the evaluation of systemic science education reform: searching for instructional sensitivity. J Res Sci Teach *39*, 369–393.

Ryan AM, Patrick H (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. Am Educ Res J *38*, 437–460.

Sadler T, Romine W, Menon D, Ferdig R, Annetta L (2015). Learning biology through innovative curricula: a comparison of game- and nongame-based approaches. Sci Educ *99*, 696–720.

Sadler T, Romine W, Stewart P, Merle D (2013). Game-based curricula in biology classes: differential effects among varying academic levels. J Res Sci Teach *50*, 479–499.

Schraw G, Lehman S (2001). Situational interest: a review of the literature and directions for future research. Educ Psychol Rev *13*, 23–52.

Silvia PJ (2001). Interest and interests: the psychology of constructive capriciousness. Rev Gen Psychol *5*, 270–290.

Swaminathan H, Rogers HJ (1990). Detecting differential item functioning using logistic regression procedures. J Educ Measure *27*, 361–370.

Talton EL, Simpson RD (1986). Relationships of attitudes toward self, family, and school with attitude toward science among adolescents. Sci Educ *70*, 365–374.

Tobias S (1994). Interest, prior knowledge, and learning. Rev Educ Res *64*, 37–54.

Todaro A, Washington S, Boekeloo BO, Gilchrist B, Wang MQ (2013). Relationship of personal health experiences with interest in health careers among youth from an underserved area. J Allied Health *42*, 135–140.

Treagust DF, Duit R (2008). Conceptual change: a discussion of theoretical, methodological and practical challenges for science education. Cult Stud Sci Educ *3*, 297–328.

Tyler-Wood T, Knezek G, Christensen R (2010). Instruments for assessing interest in STEM content and careers. J Technol Instruct Educ *18*, 341–363.

What Works Clearinghouse (2014). WWC Procedures and Standards Handbook, version 3.0, Princeton, NJ: U.S. Department of Education, Institute of Education Sciences.

Wright BD, Linacre JM, Gustafson JE, Martin-Lof P (1994). Reasonable mean-square fit values. Rasch Measurement Transactions *8*, 370.

Wright BD, Stone MA (1979). Best Test Design, Chicago: MESA.