CBE—Life Sciences Education
Vol. 14, 1–10, Fall 2015

Article

The Undergraduate Research Student Self-Assessment (URSSA): Validation for Use in Program Evaluation

Timothy J. Weston* **and Sandra L. Laursen**[†]

*Alliance for Technology, Learning and Society and [†]Ethnography & Evaluation Research, University of Colorado Boulder, Boulder, CO 80309

Submitted November 12, 2014; Revised April 1, 2015; Accepted April 1, 2015
Monitoring Editor: Erin Dolan

This article examines the validity of the Undergraduate Research Student Self-Assessment (URSSA), a survey used to evaluate undergraduate research (UR) programs. The underlying structure of the survey was assessed with confirmatory factor analysis; also examined were correlations between different average scores, score reliability, and matches between numerical and textual item responses. The study found that four components of the survey represent separate but related constructs for cognitive skills and affective learning gains derived from the UR experience. Average scores from item blocks formed reliable but moderate to highly correlated composite measures. Additionally, some questions about student learning gains (meant to assess individual learning) correlated to ratings of satisfaction with external aspects of the research experience. The pattern of correlation among individual items suggests that items asking students to rate external aspects of their environment were more like satisfaction ratings than items that directly ask about student skills attainment. Finally, survey items asking about student aspirations to attend graduate school in science reflected inflated estimates of the proportions of students who had actually decided on graduate education after their UR experiences. Recommendations for revisions to the survey include clarified item wording and increasing discrimination between item blocks through reorganization.

Undergraduate research (UR) experiences have long been an important component of science education at universities and colleges but have received greater attention in recent years, as they have been identified as important ways to strengthen preparation for advanced study and work in the science fields, especially among students from underrepresented minority groups (Tsui, 2007; Kuh, 2008). UR internships provide students with the opportunity to conduct authentic research in laboratories with scientist mentors, as students help design projects, gather and analyze data, and write up and present findings (Laursen *et al.*, 2010). The promised benefits of UR experiences include both increased skills and greater familiarity with how science is practiced (Russell *et al.*, 2007). While students learn the basics of scientific methods and laboratory skills, they are also exposed to the culture and norms of science (Carlone and Johnson, 2007; Hunter *et al.*, 2007; Lopatto, 2010). Students learn about the day-to-day world of practicing science and are introduced to how scientists design studies, collect and analyze data, and communicate their research. After participating in UR, students may make more informed decisions about their future, and some may be more likely to decide to pursue graduate education in science, technology, engineering, and mathematics (STEM) disciplines (Bauer and Bennett, 2003; Russell *et al.*, 2007; Eagan *et al.* 2013).

While UR experiences potentially have many benefits for undergraduate students, assessing these benefits is challenging (Laursen, 2015). Large-scale research-based evaluation

CBE Life Sci Educ September 2, 2015 14:ar33
DOI:10.1187/cbe.14-11-0206

Conflict of interest: The authors are involved in the development and dissemination of URSSA. However, this study is not an endorsement of the use of the survey to the exclusion of other assessment instruments. No financial remuneration is received by either author for the use of the survey.

Address correspondence to: Timothy J. Weston (westont@colorado.edu).

of the effects of UR is limited by a range of methodological problems (Eagan *et al*., 2013). True experimental studies are almost impossible to implement, since random assignment of students into UR programs is both logistically and ethically impractical, while many simple comparisons between UR and non-UR groups of students suffer from noncomparable groups and limited generalizability (Maton and Hrabowski, 2004). Survey studies often rely on poorly developed measures and use nonrepresentative samples, and large-scale survey research usually requires complex statistical models to control for student self-selection into UR programs (Eagan *et al*., 2013). For smaller-scale program evaluation, evaluators also encounter a number of measurement problems. Because of the wide range of disciplines, research topics, and methods, common standardized tests assessing laboratory skills and understandings across these disciplines are difficult to find. While faculty at individual sites may directly assess products, presentations, and behavior using authentic assessments such as portfolios, rubrics, and performance assessments, these assessments can be time-consuming and not easily comparable with similar efforts at other laboratories (Stokking *et al.*, 2004; Kuh *et al.*, 2014). Additionally, the affective outcomes of UR are not readily tapped by direct academic assessment, as many of the benefits found for students in UR, such as motivation, enculturation, and self-efficacy, are not measured by tests or other assessments (Carlone and Johnson, 2007). Other instruments for assessing UR outcomes, such as Lopatto's SURE (Lopatto, 2010), focus on these affective outcomes rather than direct assessments of skills and cognitive gains.

The size of most UR programs also makes assessment difficult. Research Experiences for Undergraduates (REUs), one mechanism by which UR programs may be organized within an institution, are funded by the National Science Foundation (NSF), but unlike many other educational programs at NSF (e.g., TUES) that require fully funded evaluations with multiple sources of evidence (Frechtling, 2010), REUs are generally so small that they cannot typically support this type of evaluation unless multiple programs pool their resources to provide adequate assessment. Informal UR experiences, offered to students by individual faculty within their own laboratories, are often more common but are typically not coordinated across departments or institutions or accountable to a central office or agency for assessment. Partly toward this end, the Undergraduate Research Student Self-Assessment (URSSA) was developed as a common assessment instrument that can be compared across multiple UR sites within or across institutions. It is meant to be used as one source of assessment information about UR sites and their students.

The current research examines the validity of the URSSA in the context of its use as a self-report survey for UR programs and laboratories. Because the survey has been taken by more than 3400 students, we can test some aspects of how the survey is structured and how it functions. Assessing the validity of the URSSA for its intended use is a process of testing hypotheses about how well the survey represents its intended content. This ongoing process (Messick, 1993; Kane, 2001) involves gathering evidence from a range of sources to learn whether validity claims are supported by evidence and whether the survey results can be used confidently in specific contexts. For the URSSA, our method of inquiry focuses on how the survey is used to assess consortia of REU sites. In this context, survey results are used for quality assurance and comparisons of average ratings over years and as general indicators of program success in encouraging students to pursue graduate science education and scientific careers. Our research questions focus on the meaning and reliability of "core indicators" used to track self-reported learning gains in four areas and the ability of numerical items to capture student aspirations for future plans to attend graduate school in the sciences.

URSSA: DEVELOPMENT AND SURVEY USE

The content validity of surveys and other measures is bolstered by a research-based development process, and the creators of the URSSA (Hunter *et al*., 2009) followed a multistepped and empirical approach to survey development (Groves *et al*., 2004; Blair *et al*., 2013). Working from basic research on student and faculty experiences with UR, the survey was pilot tested with students and revised before it gained wider use.

Basic Research, Piloting, and Revision

The basic research underpinning the URSSA is described in detail in the book *Undergraduate Research in the Sciences: Engaging Students in Real Science* (Laursen *et al*., 2010). Researchers conducted a large-scale qualitative longitudinal study at four institutions exploring the benefits of UR. Seventy-six students were interviewed after their UR, after graduation, and during a follow-up interview 2–3 yr after graduation. A sample of 80 faculty members were also interviewed about working with UR students. Students and faculty were asked about the benefits and learning gains experienced during their UR activities. Students who did not participate in UR were also interviewed as a comparison group.

The URSSA survey was modeled on the SALG (Student Assessment of Their Learning Gains) instrument (see www.salgsite.org), with an emphasis on student reports of their own learning gains in cognitive, behavioral, and affective areas. Developers created a survey blueprint (Groves *et al*., 2004), wrote items, and reviewed drafts of the survey with an advisory board. Students piloted the survey and were asked to discuss their interpretation of questions during "think-aloud" interviews (Willis, 2005). Further revisions were then based on student feedback.

Developers examined a preliminary round of results from a pilot version of the survey using both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), to learn whether the intended structure of the survey reflected empirical student responses. The developers also analyzed item functioning to look for "ceiling" or "floor" effects and for any items that were collinear or redundant. The early round of factor analysis found items forming intended blocks in both EFA and CFA analyses, with a handful of "orphan" items not fitting into the intended structure; these items were removed from the survey. Other items were removed after they showed very high correlations with other related items or high "nonapplicable" responses. For these earlier CFA analyses, the root-mean-square error (RMSEA) fit indices for a sample of 904 students started out with higher values nearing 0.08, indicating poor fit; removal of items made fit closer to the accepted standard of 0.06. The current CFA analysis was conducted to test fit on the new version of the survey, which included altered item wording and reflected the removal of previous items.

URSSA Survey

The resulting survey is available for free public use at www.salgsite.org, the Web platform developed for the SALG classroom instrument. The site allows for free, anonymous, online administration of the survey, with survey administrators receiving summarized and item-by-item results.

The full survey instrument template contains 134 items, grouped in 17 blocks. A core group of items, organized in four categories, ask students to rate how much they have gained in Skills, Thinking and Working Like a Scientist, Personal Gains, and Attitudes and Behaviors as a Researcher. (See the Supplemental Material for the first four sections of the survey; the full survey can be found at: www.colorado.edu/eer/research/documents/URSSA_MASTER_reviewCopy.pdf.) Because the URSSA is editable, most users prepare shorter versions of the survey that fit the needs of their individual programs.

If the survey is administered online, these four core groups of questions are "locked" and cannot be altered or deleted by those administering the survey. The remaining items are optional and can be deleted, moved, or edited to customize the survey for a given UR program. Those using the survey with a group of programs can lock additional questions they do not wish individual program directors to edit.

Additional, optional question blocks ask students to rate their satisfaction with their UR programs and their components, report their research activities, and rate their likelihood of future educational and career pursuits. Other groups of questions ask students to rate their reasons for choosing a specific UR program and to provide demographic information about themselves. These questions can be edited by individual program administrators to fit the context of their programs.

Intended Use and Limitations

The URSSA is a self-report survey instrument intended for use by UR program administrators for program-level evaluations of student outcomes. Because the items ask for self-report of understandings, skills, and attitudes, the survey is not intended to be used as a proxy for direct assessments or tests of individual ability. It is instead intended as a broad indicator of progress in these domains for groups of students.

The assessment is also primarily intended for organizations with multiple laboratories in which undergraduates work in internships or REU settings. Because of the typical size of UR programs, keeping student responses anonymous becomes difficult, because students may be easily identified from demographic information or the content of their comments on open-ended questions. Because of this, the use of the URSSA with groups of fewer than 10 students is not advised.

The URSSA purposefully only has a post version. We adopted the retrospective gains model for the survey (Hill and Betz, 2005) due to the observation that most students entering UR for the first time could not make realistic assessments of their own abilities simply because they had little or no exposure to the scientific methods and content found within UR experiences. An exception to this could be questions (before and after) about likelihood of continuing into graduate school or entering a scientific career.

Administrators from large programs with multiple REU sites use the URSSA for program assessment and development efforts. One example is the NSF's BIO-REU program (http://bioreu.org), a consortium of all funded REU sites in the life sciences; BIO-REU leaders have used the URSSA for the past 4 years, with encouragement from their NSF program officers. Currently, 2463 students from the BIO-REU program have taken the survey. Program leaders use the survey to evaluate the program through descriptions of student demographics; frequency of research activities; ratings of satisfaction with mentors, facilities, and activities; and self-reported changes in future plans around graduate school and scientific careers due to REU participation. They also compare composite ratings between programs and across years based on average scores on the four core sections of the URSSA. Open-ended items allow leaders to gather more substantive information about programs and enable individual lab directors to gather feedback and improve program implementation at their own REU sites.

URSSA Constructs

A range of benefits in skills and affective areas were uncovered by the previous interview study (Laursen *et al.*, 2010). The taxonomy of benefits derived from UR experiences provided the basis for the design and constructs underlying the four core areas of the URSSA. Benefits identified by students and faculty included cognitive, affective, and skills-based outcomes. Examples of benefits include better understanding of the scientific process, such as mastery in formulating research questions and identifying limitations of research methods and designs. Students identified benefits in attitudinal areas, such as gaining independence in conducting their own research and feelings of efficacy and belonging as they worked with mentors and peers. Affective benefits also included personal/professional gains, which students reported as personal gains but are recognized by their faculty research advisors as important professional development, such as increased confidence and comfort while conducting research.

These categories from the basic research were formalized in the four constructs of the current version of the URSSA. These include: Thinking and Working Like a Scientist, Personal Gains Related to Research Work, Skills, and Attitudes and Behaviors.

Thinking and Working Like a Scientist

The Thinking and Working Like a Scientist construct focuses on understandings of the process of scientific research and the nature of scientific knowledge. One of the desired outcomes from UR is the ability to engage in scientific practices and understand the process of answering research questions with experiments and observations (Russell *et al.*, 2007). Items in this area describe global understandings and skills in the research cycle, such as formulating research questions, answering questions with data, application of scientific knowledge and skills, and identifying limitations of research methods and designs.

Personal Gains Related to Research Work

The Personal Gains Related to Research Work section of the survey is meant to assess affective characteristics of confidence, comfort, and general self-efficacy with conducting research and working on a research team and in a lab. In the original UR study, students identified these affective areas as an important noncognitive outcome of their research experience, and faculty connected these developments to students' readiness to take on the role of scientist or budding professional. In other studies, this outcome has likewise been identified as related to socialization and enculturation as a scientist (Carlone and Johnson, 2007). Items in this block ask students to rate their confidence in their ability to do research, comfort in working collaboratively with others, and ability to work independently.

Skills

The Skills section is the closest to the traditional academic outcomes associated with more direct assessments. While the specific skills for science laboratory work vary dramatically across disciplines, common skills were identified in the earlier UR study (Laursen *et al*., 2010) and are included in URSSA items rating student gains in writing scientific reports or papers, making oral presentations, and conducting observations in the lab or field. The interview study indicated these skills resemble those that students may have developed in other settings, such as course work, in contrast with intellectual dispositions and abilities emphasized in the Thinking and Working Like a Scientist block.

Attitudes and Behaviors as a Researcher

The Attitudes and Behaviors as a Researcher component of the URSSA focuses on attitudes and behaviors linked to working in a scientific community and feelings of creativity, independence, and responsibility around working on scientific projects. This section of the survey lists benefits around practicing more authentic scientific inquiry versus more rote or "cookbook" science labs (Seymour *et al*., 2004). Items in this block ask students to report gains in engaging in real-world science research, feeling part of a scientific community, and feeling responsible for a project. In the interview study, these items were linked to students' developing identities and status as scientists.

Satisfaction

Satisfaction items are included in this analysis as a comparative baseline for the four core URSSA constructs. An underlying rationale for the URSSA (as well as the SALG before it) is that students' self-reports on their learning fundamentally differ from traditional ratings of satisfaction with an instructor's performance. "Learning gains" questions instead focus on student perceptions of their own learning from their course or lab experiences. However, during development, stakeholders in the URSSA wanted some traditional satisfaction questions that would help them monitor mentoring; facilities; program-wide activities, such as career seminars or field trips; and overall climate. These items include questions about students' relations with mentors and peers, plus satisfaction with the overall UR experience.

VALIDATION

As of May 1, 2014, 3671 students had taken the URSSA across the United States and Canada. These students came from public and private research universities, smaller universities, and colleges and research institutes. In this study, we focused on the URSSA survey's use as a tool for assessing learning gains and aspirations for students enrolled in REU consortium programs such as the BIO-REU program described earlier.

Questions included:

1. Do blocks of items in the fixed part of the survey represent related but separate domains?

The URSSA contains four blocks of questions based on categories of benefits uncovered in prior research. We use averages within these blocks to create composite variables ("core indicators") for comparing students between years and among groups. We wanted to know whether these categories represent separate (yet related) domains or whether categories can be collapsed into a more parsimonious structure.

2. How do student ratings of satisfaction relate to URSSA learning gain variables?

The URSSA contains questions asking students to rate aspects of their research experiences. A central question for both the URSSA and the SALG is the meaning of gains scores versus traditional ratings of satisfaction. This is an important question because core URSSA survey items are meant to ask students to report on their own gains in cognitive and affective domains, while satisfaction items are traditional external ratings of teachers, mentors, or facilities or other services.

3. Are composites created from these item blocks reliable?

Because programs use composite variables as the basis for comparison for students and programs, we wanted to know whether these variables were reliable enough to be used confidently for this purpose.

4. Are ratings of likelihood of future plans for graduate school and scientific careers derived from numerical items congruent with student responses on open-ended questions?

One promised benefit of UR is that students participating will be more likely to choose to continue on to graduate school and careers in science (Eagan *et al*., 2013). The survey asks students to (numerically) rate their likelihood of pursuing graduate school compared with the likelihood before they started the UR program. The percentage of students who answer this question positively is often used as one indicator of a program's success. For comparison, we examined student responses to an open-ended question found later in the survey: "How did your research experience influence your thinking about future career and graduate school plans?" We wanted to know whether the answers students gave to the textual question about changing their minds about attending graduate school matched ratings for the similar numerical item.

METHODS

CFA Methods

To best represent the range of students who have taken the URSSA, we created a sample drawn randomly from demographic blocks of data of all students who took the survey

Table 1. How much did you gain in the following areas as a result of your most recent research experience?[a]

	Thinking and Working Like a Scientist		Personal Gains
Q1	Analyzing data for patterns	Q9	Confidence in my ability to contribute to science
Q2	Figuring out the next step in a research project	Q10	Comfort in working collaboratively with others
Q3	Problem-solving in general	Q11	Confidence in my ability to do well in future science courses
Q4	Formulating a research question that could be answered with data	Q13	Ability to work independently
Q5	Identifying limitations of research methods and designs	Q14	Developing patience with the slow pace of research
Q6	Understanding the theory and concepts guiding my research project	Q15	Understanding what everyday research work is like
Q7	Understanding the connections among scientific disciplines		
Q8	Understanding the relevance of research to my course work		
	Skills		**Attitudes and Behaviors**
Q16	Writing scientific reports or papers	Q28	Engage in real-world science research
Q17	Making oral presentations	Q29	Feel like a scientist
Q18	Defending an argument when asked questions	Q30	Think creatively about the project
Q19	Explaining my project to people outside my field	Q31	Try out new ideas or procedures on your own
Q20	Preparing a scientific poster	Q32	Feel responsible for the project
Q21	Keeping a detailed lab notebook	Q33	Work extra hours because you were excited about the research
Q22	Conducting observations in the lab or field	Q34	Interact with scientists from outside your school
Q23	Using statistics to analyze data	Q35	Feel a part of a scientific community
Q24	Calibrating instruments needed for measurement		
Q25	Understanding journal articles		
Q26	Conducting database or Internet searches		
Q27	Managing my time		

[a] Question numbers correspond to factor model in Figure 1. Gains scale: 1 → 5 = no gains → great gains.

from 2010 to 2014.[1] The sample for assessing factor analysis contained 506 students.

CFA (Harrington, 2009) examines whether a model of hypothesized factors fits the structure of empirical student responses. We created factor models using the software AMOS 22.0 (Arbuckle, 2011). The theorized model follows the structure of the four blocks of items in the survey with factors corresponding to the four correlated item blocks: Thinking and Working Like a Scientist, Personal Gains, Skills, and Attitudes and Behaviors (see Table 1). Additionally, we also wanted to know the nature of the relationship between gains variables and student ratings of satisfaction (Satisfaction) with their mentors and other aspects of the REU.

We compared four different models to examine whether alternative arrangements of survey items better fit the data. As noted, the four-factor "survey" model reflects the hypothesized structure of the survey in four item blocks. Three alternative models were hypothesized a priori from examining the correlation and covariance matrices for the data before conducting the analysis. One of these was based on the high correlation of items within and between the first three blocks of items, which suggested they might be better represented by a single factor representing generalized learning gains, thus yielding a two-factor model. A one-factor model was used as a baseline comparison. Finally, a five-factor model incorporated items about satisfaction to answer a different question about the correlation between factors. This model retains the first four factors but adds the Satisfaction items, which were also completed by all students taking the survey. We added Satisfaction items to examine whether learning gains items were discrepant given the different underlying meanings of the constructs. Three items from the original survey were omitted (Q23, Q24, Q34) due to large numbers of students choosing the "nonapplicable" option. Omitted items asked students to rate their skills in analyzing statistical data, calibrating instruments used for measurement, and interacting with scientists.

We used a maximum likelihood estimation procedure after checking the data to confirm they did not violate assumptions for underlying normality of the observed variables. Critical ratios for both skewness and kurtosis for individual items and multivariate normality were less than accepted cutoffs (Jackson *et al*., 2009). The small amount of missing data from items were imputed in the maximum likelihood estimation procedures. Model fit was assessed with the RMSEA, comparative fit index (CFI), and chi-square statistics, using commonly held standards for fit (Li-tze Hu and Bentler, 1999; Marsh *et al*., 2004). We also assessed the degree of correlation among latent factors in the model using estimates generated by AMOS.

Assessing Reliability

We assessed the internal reliability of composite variables with Cronbach's alpha, a commonly used measure. We assessed invariance of the coefficient on four samples from the BIO-REU program 2010–2014. We used the "rule of thumb" standard for a reliable composite of $\alpha > 0.8$ (Knapp, 1991).

[1] A sample was used to 1) avoid inflating chi-square estimates for model fit occurring with very large samples and 2) trim demographic proportions caused by outliers at individual sites, for example, large numbers of women at one site.

Table 2. Model fit statistics

	Thinking and Working Like a Scientist	Personal Gains	Skills	Attitudes	Satisfaction
Thinking and Working Like a Scientist	1				
Personal Gains	0.70	1			
Skills	0.68	0.71	1		
Attitudes and Beliefs	0.54	0.51	0.37	1	
Satisfaction.	0.51	0.47	0.27	0.64	1

Note: Correlations based on five-factor model; values may differ for four-factor model.

Qualitative Ratings

Coding qualitative textual data followed best practices from Creswell (2013) for coding for thematic content. We coded the question: "How did your research experience influence your thinking about future career and graduate school plans?" We first developed codes for this item on a separate data set of 834 students from the BIO-REU program in 2012; we then used the same coding rubric on the data set of 762 valid responses from 2013. Themes for coding remained constant between years, as did percentages of responses in each code. Responses fell into four major categories:

1. Students who said they were much more likely to go to graduate school after attending the REU,
2. Students who said the REU confirmed their existing plans to go to graduate school,
3. Students who said they received valuable information about graduate school but had not made a decision about their future, and
4. Students who said they decided not to attend graduate school after their REU, or students who already had decided on other educational paths such as medicine.

We then compared response codes with their respective numerical rating percentages for the likelihood item: "Compared to your intentions before doing research, how likely are you now to enroll in a PhD program in science, mathematics or engineering?"

RESULTS

We wanted to know whether blocks of items meant to represent related but different concepts presented this pattern of correlation for empirical student responses. The best fit for the three alternative models was the original four-factor model that reflected the structure of the survey. This model had strong standardized factor loadings for individual items in the 0.4–0.7 range. We found model fit statistics for the four-factor model of RMSEA = 0.064, CFI = 0.76, and chi-square/df = 3.0. These model fit statistics suggest the model nears but does not fully meet standards for model fit. While the model nearly meets the standards of fit for an RMSEA statistic of less than or equal to 0.06, other indicators do not meet the standards for CFI greater than 0.9 and a chi-square degrees of freedom ratio of <2.0 (Li-tze Hu and Bentler, 1999).

Table 2 presents model fit statistics for three alternative models and the five-factor model with the satisfaction variable with standards for each. See Figure 1 for factor loadings.

We found high correlations between the latent factors for Thinking and Working Like a Scientist, Skills, and Personal Gains; we saw lower correlations between these variables and the Attitudes and Beliefs variable, with the lowest correlation of $r = 0.37$ for Skills and Attitudes and Behaviors. This indicates there is a high degree of commonality among the first three factors (Table 3).

The five-factor model also provides a comparison of learning gains items with ratings for satisfaction. The highest correlation is $r = 0.64$ between Satisfaction and the Attitudes and Beliefs variable; the lowest correlation is $r = 0.27$ with the Skills variable. These comparisons suggest that Attitudes and Behaviors and the two other gains variables (Thinking and Working Like a Scientist and Personal Gains) share a great deal of variance with satisfaction ratings.

Reliability of Composite Variables

We used the averages of composite variables to track changes between years and make comparisons between REU sites and programs. Composite variables exceeded accepted standards for reliability (Table 4).

Comparison of Textual and Numerical Ratings

One perceived benefit of UR is that students may decide to enter graduate school in STEM and pursue scientific careers after participating in UR. Students rated the numerical item: "Compared to your intentions before doing research, how likely are you now to enroll in a PhD program in science, mathematics or engineering?" Typically, program directors collapse the two highest categories, "much more likely" and "extremely more likely," and use this percentage as an indicator of the likelihood of students entering graduate school after attending an REU. For our sample, 47% of students were in this "positive" category. We then coded 732 responses to the question: "How did your research experience influence your thinking about future career and graduate school plans?" with the coding scheme described above.

We found that 26% of students answering the open-ended question said they decided during or directly after their UR experience to go to graduate school; of these students, 85% answered "more likely" or "extremely more likely" on the numerical item. (We called responses in these two categories

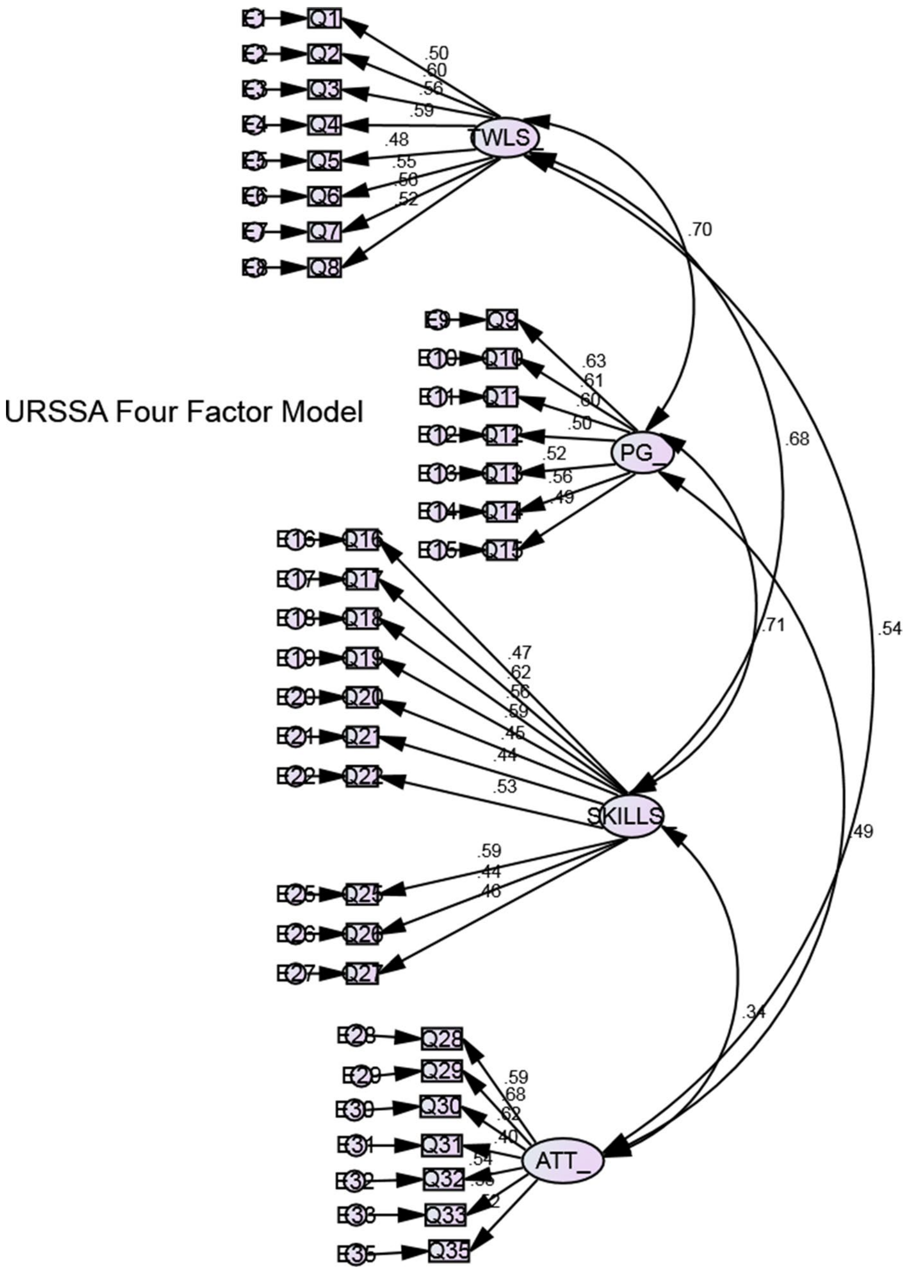

Figure 1. Factor loadings for four-factor model for the URSSA.

the "positive category.") However, of all students in the positive category, only 43% answered the item in a way consonant with the question's meaning.[2] The remaining students who provided ratings in this category were divided between those who said they gained enough information to make a decision about graduate school based on the REU (32% of positive category) and those who said that attending the REU confirmed their existing plans to continue to graduate school (22% of positive category). This result indicates that less than half of those in the positive category are interpreting the question in agreement with its intended meaning (Table 5).

DISCUSSION

The URSSA has gained use among UR sites for program assessment. As an evaluation tool, the survey allows programs to document student representation, monitor student ratings of their learning, describe student research activities, and gauge the likelihood that the UR experience has changed or confirmed students' plans to pursue STEM education and careers. Faculty members and program administrators also

[2] Total proportion of valid answers in this category is calculated by total answers in open-ended category divided by total answers in the "much more" or "extremely more" category for the numerical item.

Table 3. Correlation among factor variables

Model	RMSEA	CFI	χ^2/df
Expected value	0.06 or less	>0.9	2 or less, nonsignificant χ^2
Five factor (satisfaction model)	0.059	0.79	1699, 619, 2.7, $p < 0.0001$**
Four factor (survey model)	0.064	0.76	1418, 458, 3.0, $p < 0.0001$**
Two factor	0.073	0.68	1693, 463, 3.6, $p < 0.0001$**
One factor	0.081	0.76	3378, 464, 7.2, $p < 0.0001$**

**All chi-square significant at $p < 0.0001$.

use the URSSA to gain feedback about student interactions with their mentors and peers and their overall satisfaction with their UR experience.

We examined some aspects of the construct validity of the URSSA to answer questions about the structure of the survey, its meaning, and its use as a reliable measurement tool. Validity questions addressed outstanding questions about the meaning of the survey's core indicators and items addressing the likelihood of future activities.

Do Blocks of Items in the Fixed Part of the Survey Represent Related But Separate Domains?

The answer to this question is mixed. The four-factor model reflecting the structure of the survey is the best-fitting model among the alternatives for these items. It also nearly meets one of the three model-fit criteria but does not meet the criteria for the CFI indicator or the chi-square test. Marsh *et al*. (2004) argue that these borderline models should not be dismissed, especially if factor loadings are high, as is the case for this model, and if the fit index least affected by large numbers of degrees of freedom (as is the case in this model) is near the criteria for fit.

The main concern with structural validation of the URSSA is the high correlation between latent factor variables found in the four-factor model. Correlations between Thinking and Working Like a Scientist, Personal Gains, and Skills factors are near $r = 0.7$, making them nearly collinear (Harrington, 2009). However, the two-factor model, made by collapsing the three factors into one, showed a much worse fit than the four-factor model. Overall, this result points to separate factors that have some independent variance but mostly represent similar concepts. Any reorganization of the survey should attempt to achieve greater discrimination between these factors; such reorganization could lead to a more parsimonious and shorter survey that better represents underlying content.

Table 4. Reliability of composite variables

	Alpha (range)	Number of items
Thinking and Working Like a Scientist	0.88–0.90	8
Personal Gains	0.90–0.91	7
Skills	0.91–0.92	10
Attitudes and Behaviors	0.83–0.84	7

Values for Cronbach's alpha are calculated for each sample.

Table 5. Percent in positive category with valid answers on open-ended question

Answered "much more or extremely more likely to attend graduate school"			
43% changed their minds about graduate school (valid answers)	32% gained information about graduate school	22% confirmed plans to attend graduate school	3% other

Further insight about the structure of the survey can be gained from examining individual item correlations within and between factors. We found higher correlations between items representing more global skills such as "understanding a research journal article" (from the Skills section) and the item "formulating a research question that can be answered with data" ($r = 0.47$) from the Thinking and Working Like a Scientist section. We saw the lowest correlations between specific skills such as "making a poster" and a more global skills item "formulating a research question that could be answered with data" ($r = 0.22$). This pattern of correlations suggests that the survey would discriminate more if divided into separate factors for global and cognitive skills related to research design on one hand, and more concrete and clearly defined skills on the other.

How Do Student Ratings of Satisfaction Relate to URSSA Learning Gain Variables?

An important aspect of the URSSA is its emphasis on ratings of individual learning gains versus ratings of satisfaction with people, services, or others' abilities. The patterns of correlations among Satisfaction and the URSSA core factors show commonality and correlation with some parts of the survey and not others.

We found the highest correlation between the Satisfaction and Attitudes and Behaviors ($r = 0.67$) factors. While both constructs are meant to measure attitudes, the pattern of correlation between individual items within each factor lends support to the idea that Attitudes and Behaviors items behave like satisfaction items. Satisfaction items typically call on students to rate their mentors, peers, or services, all external characteristics of the UR setting. Items within the URSSA factors that have the highest correlation with satisfaction generally correlate to overall characteristics of a specific setting versus ratings of internal abilities, traits, or understandings. For instance, we saw higher item correlations ($r = 0.4$) between "satisfaction with the overall research experience" and gains in "engaging in real world research." While students articulated this as a benefit of a UR experience during the interviews, it is possible that students perceive the responsibility for providing an authentic research setting as resting with the laboratory mentor or as inherent in the nature of the research project, and less as an individual characteristic of their own learning or its relevance. This idea is supported by a much lower correlation ($r = 0.13$) between the overall satisfaction variable and the item "try out new procedures on your own," another item from the Attitudes and Behaviors block. Here, students may perceive this item as a rating of a personal characteristic that emphasizes individual initiative and their own ability to execute specific actions, and this item is therefore less likely to provide a rating of an external characteristic of the REU setting.

The relative independence of the Skills factor with the satisfaction items supports the central validity claim that students are making ratings of their own abilities when answering items about their ability to create posters, make presentations, analyze data, or work on computers or about other specific skills. In this case, inter-item correlations with the Satisfaction factor are uniformly low, possibly because ratings of skills are more easily conceived as individual capabilities versus ratings of aspects of one's surroundings.

Are Composites Created from These Item Blocks Reliable?

Internal reliability for composites was above $\alpha = 0.9$ for the first three URSSA composites Thinking and Working Like a Scientist, Personal Gains, and Skills. The reliability for the Attitudes and Behaviors variable was lower but still high at $\alpha = 0.83$. Results are not surprising, given the high item loadings on each factor and the pattern of correlations between items within factors. The reliabilities for the composites provide supporting evidence for the use of these composites as indicators used in simple statistical comparisons over years and between groups.

Are Ratings of Likelihood of Future Plans for Graduate School and Scientific Careers Derived from Ratings Congruent with Student Responses on Open-Ended Questions?

If using a strict interpretation of the numerical ratings for the item: "Compared to your intentions before doing research, how likely are you now to enroll in a PhD program in science, mathematics or engineering?," responses do not fully represent valid agreement with the question's meaning. Less than half of the students answering the question "much more" or "extremely more likely" gave valid responses that reflect students changing their minds and wanting to attend graduate school. Other students answering the question in the same way said they received valuable information about graduate school or that attending the UR confirmed existing plans to attend graduate school.

RECOMMENDATIONS

We make three primary recommendations for revising the URSSA based on findings from the validity study. First, the core indicators for Thinking and Working Like a Scientist, Personal Gains, and Skills would benefit from reorganization of items and elimination of at least one category. Examination of items within the categories suggest that most of the items in the Thinking and Working Like a Scientist category could be construed as general cognitive skills used in the UR setting (e.g., formulating research questions). Conversely, the Skills items might be more coherent if split between specific skills, such as keeping a detailed lab notebook, and general or cognitive skills similar to those in the Thinking and Working Like a Scientist section, such as managing time and understanding journal articles.

Likewise, we also recommend examining which attitudinal items in both Personal Gains and Attitudes and Behaviors would better be construed as satisfaction ratings. These items are likely those that ask students to provide ratings about the authenticity of the research experience, sense of community, or responsibility given to students to create their own projects. It may be possible to uncover some of these distinctions through interviews with UR students by asking them their views about personal agency in these settings.

The likelihood questions would also benefit from more pilot testing and redesign. It is possible that some students are not reading or understanding the beginning of the question: "Compared to your intentions before doing research…" and are giving direct assessments of their likelihood of going to graduate school instead of gauging the effect of the REU on this likelihood. Possible options for this question include phrases such as "did you change your mind?" This question is important, given that some programs may want to use these likelihood items as indicators of program success.

CONCLUSIONS

We examined some aspects of the validity of the URSSA. Findings suggest that the four core components of the survey represent separate but highly related constructs for self-reported cognitive skills and affective learning gains derived from the UR experience. Averages across these item blocks form reliable but correlated composite measures. Additionally, some parts of the survey, especially those related to affective areas, are highly related to ratings of satisfaction with the research experience. The pattern of correlation among individual items suggests that items on which students rate external aspects of their environment (such as gains in conducting authentic scientific work) are more like satisfaction ratings than items that directly ask about student skills attainment. Finally, the survey item asking about student aspirations to attend graduate school in science reflected inflated estimates of the proportions of students who had actually decided on graduate education after their UR experience. In response to these findings, we recommend revisions that will increase discrimination between item blocks and clarify item meaning.

ACKNOWLEDGMENTS

The final version of the URSSA assessment instrument was developed under NSF grant CHE-0548488, Mapping the Extent and Character of the Undergraduate Research Experience at Public Research Universities: An Exploratory Study, funded by the NSF Divisions of Chemistry and Undergraduate Education, the Biological Sciences Directorate, and the Office of Multidisciplinary Affairs.

Validity research was conducted under NSF grant DBI-1052683, Collaborative Research: Evaluation of Biology REU Site Programs, funded by the NSF Division of Biological Infrastructure, Research Experiences for Undergraduates Sites. We acknowledge the valuable contributions of the BIO-REU assessment coordinators Janet Branchaw and Julio Soto to this research and the work of Anne-Barrie Hunter, Heather Thiry, and Elaine Seymour in the development of the survey.

REFERENCES

Arbuckle JL (2011). IBM SPSS Amos 22 User's Guide, Crawfordville, FL: Amos Development.

Bauer KW, Bennett JS (2003). Alumni perceptions used to assess undergraduate research experience. J High Educ *74*, 210–230.

Blair J, Ronald FC, Blair EA (2013). Designing Surveys: A Guide to Decisions and Procedures, Thousand Oaks, CA: Sage.

Carlone HB, Johnson A (2007). Understanding the science experiences of successful women of color: science identity as an analytic lens. J Res Sci Teach *44*, 1187–1218.

Creswell JW (2013). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, Sage.

Eagan MK, Hurtado S, Chang MJ, Garcia GA, Herrera FA, Garibay JC (2013). Making a difference in science education: the impact of undergraduate research programs. Am Educ Res J *50*, 683–713.

Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R (2004). Survey Methodology, Hoboken, NJ: Wiley.

Frechtling J (2010). The 2010 User-Friendly Handbook for Project Evaluation, Arlington, VA: National Science Foundation.

Harrington D (2009). Confirmatory Factor Analysis, Oxford University Press.

Hill LG, Betz DL (2005). Revisiting the retrospective pretest. Am J Eval *26*, 501–517.

Hu Li-tze, Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling *6*, 1–55.

Hunter A-B, Laursen SL, Seymour E (2007). Becoming a scientist: the role of undergraduate research in students' cognitive, personal, and professional development. Sci Educ *91*, 36–74.

Hunter A-B, Weston TJ, Laursen SL, Thiry H (2009). URSSA: evaluating student gains from undergraduate research in science education. Counc Undergrad Res Q *29*, 315–19.

Jackson D, Gillaspy JA Jr, Purc-Stephenson R (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. Psychol Methods *14*, 6.

Kane MT (2001). Current concerns in validity theory. J Educ Measurement *38*, 319–342.

Knapp TR (1991). Focus on psychometrics. Coefficient alpha: conceptualizations and anomalies. Res Nursing Health *14*, 457–460.

Kuh G (2008). High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter, Washington, DC: American Association of Colleges and Universities.

Kuh GD, Jankowski N, Ikenberry SO, Kinzie J (2014). Knowing What Students Know and Can Do: The Current State of Learning Outcomes Assessment at U.S. Colleges and Universities. Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Laursen S, Seymour E, Hunter AE, Thiry H, Melton G (2010). Undergraduate Research in the Sciences: Engaging Students in Real Science, Hoboken, NJ: Wiley.

Laursen SL (2015). Assessing undergraduate research in the sciences: the next generation. Counc Undergrad Res Q *35*(3), 9–14.

Lopatto D (2010). Undergraduate research as a high-impact student experience. Peer Rev *12*, 227–30.

Marsh HW, Hau K-T, Zhonglin W (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struct Equ Modeling *11*, 320–341.

Maton KI, Hrabowski FA III (2004). Increasing the number of African American PhDs in the sciences and engineering: a strengths-based approach. Am Psychol *59*, 547.

Messick S (1993). Validity. In: Educational Measurement, 3rd ed., ed. R Linn, Washington, DC: American Council on Education/Macmillan, 13–103.

Russell SH, Hancock MP, McCullough J (2007). Benefits of undergraduate research experiences. Science *316*, 548–549.

Seymour E, Laursen SL, Hunter AE, Deantoni T (2004). Establishing the benefits of research experiences for undergraduates in the sciences: first findings from a three-year study. Sci Educ *88*, 493–534.

Stokking K, Schaaf M, Jaspers J, Erkens G (2004). Teachers' assessment of students' research skills. Br Educ Res J *30*, 93–116.

Tsui L (2007). Effective strategies to increase diversity in STEM fields: a review of the research literature. J Negro Educ *76*, 555–581.

Willis GB (2005). Cognitive Interviewing: A Tool for Improving Questionnaire Design, Thousand Oaks, CA: Sage.