Improving Undergraduate Education Through Community Building and Adoption of an Evidence-based Practice to Evaluate Undergraduate Research

Evaluation Report: Prepared by Bridget Zimmerman, Project Evaluator

Date: February 12, 2020

Background

In 2008 a university-funded project at SUNY-Buffalo State was initiated to develop and field-test a method for measuring student learning and related outcomes in the university's summer undergraduate research program, in which students work one-on-one with faculty mentors. While many valuable studies had already been done on the impact of undergraduate research on its participants, these studies focused mainly on student and faculty perceptions of the value of undergraduate research. The SUNY- Buffalo State evaluation project aimed to go beyond those studies by measuring specific student knowledge and skill outcomes and the project employed an evaluation method designed to accomplish this goal. The new evaluation method was designed by Jill Singer, Buffalo State and Daniel Weiler, Daniel Weiler Associates. The Buffalo State field test of this evaluation method was implemented successfully at that campus for six years. Then, with an award from the National Science Foundation's WIDER program, the method – now called EvaluateUR – was tested successfully at several other campuses and eventually scaled up to include 44 programs on 37 campuses across the country that fully implemented the new method. The NSF grant for this project concluded in December 2019. EvaluateUR is currently on track to continue as a fee-based project open to college and university enrollment by application.

EvaluateUR Goals

EvaluateUR set out to accomplish the following objectives:

1. Introduce undergraduate students conducting research projects under the guidance of faculty mentors to a wide range of knowledge and skills they would need for graduate studies and/or the workplace, including but not restricted to knowledge specifically related to their field of study.
2. For these students, obtain statistically valid and reliable measures of their achievement on each of these knowledge and skill outcomes.
3. Help students to improve their understanding of their academic strengths and weaknesses and what learning strategies have or have not worked well for them – i.e, their metacognitive skills.
4. Provide research students' faculty mentors with tools to enlarge their understanding of the academic strengths and weaknesses of their students, as well as new insights into the teaching and learning process.
5. Ensure that the evaluation method does not create an undue burden on students, mentors or campus administrators of undergraduate research programs.
6. Devise practical tools to make it easy for campus administrators to oversee the evaluation and report its findings for all participating students.

7. Widely disseminate information about the evaluation method and its results.
8. Enlarge the number of campuses utilizing the evaluation method, to test whether the project could be expanded beyond its original few participants.

EvaluateUR Design

Over a period of five years, EvaluateUR implemented a project whose key features were designed to meet each of the goals listed above:

1. Introduce students to a wide range of knowledge and skills they would need for graduate study and/or the workplace. When the project began at SUNY Buffalo State, faculty from diverse disciplines identified a range of student competencies that would be critically important both for graduate studies and the workplace, eventually proposing 11 outcome categories reflecting desirable student knowledge and skills. Each of the 11 outcome categories was further delineated by specific, measurable student behaviors, leading to a total of 35 outcome components that would constitute the basis for assessments. With minor refinements, these outcome categories and their respective components, shown in the table below, were used by EvaluateUR at all participating campuses.

Table 1. EvaluateUR Outcome Categories and Components

EvaluateUR Assessment Survey Outcome Categories and Components	
Outcome Category	**Outcome Component**
Communication:	Uses and understands professional and discipline-specific language.
	Expresses ideas orally in an organized, clear, and concise manner.
	Writes clearly and concisely using correct grammar, spelling, syntax, and sentence structure.
Creativity:	Displays insight about the topic being investigated.
	Shows ability to approach problems from different perspectives
	Uses information in ways that demonstrate intellectual resourcefulness.
	Effectively connects multiple ideas/approaches
Autonomy:	Demonstrates an ability to work independently and identify when guidance is needed.
	Accepts constructive criticism and uses feedback effectively.
	Uses time well to ensure work gets accomplished.
	Sets and meets project deadlines.
Ability to Deal with Obstacles:	Is not discouraged by setbacks or unforeseen events and perseveres when challenges are encountered.
	Shows flexibility and a willingness to take risks and try again.
	Trouble-shoots problems and searches for ways to do things more effectively.
Intellectual Development:	Recognizes that problems are often more complicated than they first appear.
	Approaches problems with an understanding that there can be more than one right explanation or even none at all.

	Displays accurate insight into the limits of his/her own knowledge and an appreciation for what isn't known.
Critical Thinking and Problem Solving:	Challenges established thinking when appropriate.
	Looks for the root causes of problems and develops or recognizes the most appropriate corrective actions.
	Recognizes flaws, assumptions and missing elements in arguments.
Practice and Process of Inquiry:	Demonstrates ability to formulate questions and hypothesis within the discipline.
	Demonstrates ability to properly identify and/or generate reliable data.
	Shows understanding of how knowledge is generated, validated and communicated within the discipline.
Nature of Disciplinary Knowledge:	Shows understanding of the way practitioners think within the discipline (e.g., as an earth scientist, sociologist, artist . . .) and view the world around them.
	Shows understanding of the criteria for determining what is valued as a contribution in the discipline.
	Shows awareness of important contributions in the discipline and who was responsible for those contributions.
	Reads and applies information obtained from professional journals and other sources.
	Is aware of professional societies in the discipline.
Content Knowledge and Methods:	Displays knowledge of key facts and concepts.
	Displays a grasp of relevant research methods and is clear about how these methods apply to the research project being undertaken.
	Demonstrates an appropriate mastery of skills needed to conduct the project.
Ethical Conduct:	Recognizes that creating, modifying, misrepresenting or misreporting data including omission or elimination of data/findings or authorship is unethical.
	Behaves with a high level of collegiality and treats others with respect.
Career Goals:	Is clear about academic and/or professional/work plans.
	Is aware of how research skills relate to academic and/or professional/work plans.

2. Obtain statistically valid and reliable measures of student achievement on each of these knowledge and skill outcomes. EvaluateUR likewise used and refined an assessment instrument first devised for use at Buffalo State for gathering data on student outcomes. The instrument used a five-point Likert scale indicating the extent to which a student had displayed the outcome component being assessed. (The assessment scale rubric ranges from "always" to "never.") Beginning with a "baseline assessment" before research begins and followed by two additional assessments (at the mid- and end-points of the student's research project) students scored themselves on each outcome component, and their research mentors, using the same instrument, also independently scored their students. This regimen provided the equivalent of pre, mid, and post assessment scores. On each of these occasions, students and mentors compared their assessments and discussed the reasons for the scores they had each assigned. By using identical outcomes categories and their respective explanatory components, as well as an identical scale and scale rubric for all assessments, EvaluateUR conducted assessments according to explicit and uniform standards across varied disciplines

and across different student-faculty pairs. These and related features of the EvaluateUR design were meant to overcome the usual objections to assessments of student outcomes based on subjective faculty judgments, which typically rely on disparate standards without common assessment parameters across disciplines or among different faculty. Orientation materials, together with live and online participant meetings organized and led by the EvaluateUR Principal Investigator, helped to minimize inter-rater reliability issues, so as to further strengthen the validity of the assessments across institutions and disciplines

3. Help students to improve their understanding of their academic strengths and weaknesses and what learning strategies have or have not worked well for them – i.e, their metacognitive skills. Metacognition encourages learners to be *aware* of what they are doing and why (i.e., self-monitoring) and then use that awareness to make intentional adjustments to their learning strategies (i.e., self-regulation) in order to learn more effectively. EvaluateUR stressed to participating students and mentors that the assessment scores were less important than the conversation that followed the assessments, at which time the student and mentor shared their rationales for assigning particular scores and discussed the reasons for differences, if any, in their perceptions. These conversations were aimed at providing students with new insights into their thinking processes and learning strategies. In this way, the evaluation approach sought to collect reliable data on specific student outcomes and also contribute directly to student learning.

4. Provide research students' faculty mentors with tools to enlarge their understanding of the academic strengths and weaknesses of their students as well as new insights into the teaching and learning process. The project's focus on parallel student and mentor assessments followed by a conversation about differences in their perceptions was designed not only to provide students with fresh insights, as described above, but to help their faculty mentors gain a new appreciation for the ways in which students saw – and perhaps misunderstood – their academic strengths and weaknesses, and to spur new thinking about how existing teaching-learning practices might be modified to support other undergraduate research experiences. Additionally, EvaluateUR assessments and processes were designed to provide an example of ways in which students in their regular classes might not be aware of the competencies and skills they should strive to achieve, and how they might also be less than clear about their academic strengths and weaknesses.

5. Ensure that the evaluation method does not create an undue burden on students, mentors or campus administrators of undergraduate research programs. The EvaluateUR assessment instruments were devised to be completed by students and faculty mentors in less than 20 minutes. Assessment frequency was designed to constitute a marginal additional time commitment beyond the time ordinarily set aside for student-mentor meetings (or to coincide with those meeting times). Within a few minutes of their assessments being completed and submitted electronically, students and mentors received an automated score report for use as the basis for comparing notes and discussing assessment differences. Students and mentors also received automated messages reminding them to complete necessary implementation steps in a timely manner, as well as access to a "dashboard" on the project website that showed where they were in the process.

6. Devise practical tools to make it easy for campus administrators to oversee the evaluation and report its findings for all participating students. The project's online dashboard was designed to help administrators keep track of the implementation progress of all participating

student-mentor pairs and a system of automated reminder messages was designed to help administrators ensure that participants were staying on schedule in completing their implementation steps. In addition, administrators had access to a web-based EvaluateUR statistical tool called EZStats. EZStats automatically generated for each outcome component the assessment scores entered by students and mentors, descriptive statistics, response frequency tables, pre- to mid-, mid- to post- and pre- to post- mean score gains (or declines) and percent change. For institutions with 20 or more pairs of completed mentor-student assessments, a paired sample t-test provided significance testing outcomes, including effect sizes. The scores were reported by EZStats in a format that was readily understandable by a lay reader and immediately usable for publication. The online dashboard, automated messages and EZStats reports were all managed by the project website at the Science Education Resource Center (SERC), at Carleton College, Minnesota. This website was originally created and managed at the SUNY Buffalo State campus. It was moved to SERC in order to serve project participants beyond Buffalo State and to add features as the project design was refined following feedback from participating campuses. In 2019 and early 2020 the project posted several new publications and videos to its website in order to further orient campus undergraduate research administrators and others to the materials, mission, and outcomes of EvaluateUR. This material included a new online publication and video on EZStats ("EZStats Guide to Statistics") to support the automated EZStats report and site administrator reporting needs.

7. Widely disseminate information about the evaluation method and its results. The project design called for dissemination by CUR through its website and other communications as well as presentations about the project to a wide variety of audiences with an interest in student learning, undergraduate research and evaluation. Presentations were made, for example, to the CUR Undergraduate Research Director's conference (June 2019); Geological Society of America annual meeting (September 2019), AAC&U: Transforming STEM Higher Education conference (November 2019, NSF Community College Undergraduate Research Experience (URE) Summit (November 2019), and the American Geophysical Union meeting (December 2019).

8. Enlarge the number of campuses utilizing the evaluation method, to test whether the project could be expanded beyond its original few participants. The project design was originally tested at the Buffalo State campus. Given its success there, EvaluateUR was funded by NSF-WIDER and the project recruited five additional campuses to test whether the design could be implemented successfully elsewhere. The project design then called for refining project procedures based on feedback from participating campuses, recruiting additional participants in a second phase of the project's expansion, and repeating this process for a third and final expansion to additional campuses. This phase-in plan was intended to test the feasibility of implementing the design on a large scale under widely different campus circumstances.

Findings

EvaluateUR tested an innovative design for evaluating undergraduate research in a way that could reliably measure specific knowledge and skill outcomes while also contributing directly to student learning. The project succeeded in meeting all its objectives and thereby demonstrated the feasibility of its design in a wide variety of college and university settings.

In 2019 EvaluateUR was implemented in 44 programs at 37 participating institutions. (One large institution had eight different departments participating, which brought the total number of active EvaluateUR programs to 44.) Four institutions dropped out because they felt that EvaluateUR would be too much work or require too much writing from their students or decided that three assessments would be too many for a four-week research program. The number of student-mentor pairs at these institutions ranged from 1 to 82, with an average of 18 student-mentor pairs. Altogether, the 2019 EvaluateUR cohort included s 781 mentors and 799 students as part of 799 student-mentor pairs. (A few mentors mentored more than one student.) The assessment completion rate from pre to mid assessment was 94%, from mid to post assessment 90%, and from pre to post of assessment 85%.

Although these assessment completion rates are quite good, it is instructive to look at the reasons given by mentors for student-mentor pair failures to complete all three assessments:

- The student-mentor pair got behind on the project and didn't have time;
- The assessments were too frequent for the length of the research project;
- The assessments took time away from research;
- The assessments were too long'
- The assessments did not apply to their discipline;
- Mentors did not need a tool to improve their mentoring;
- Mentors already meet with their students, so the discussions would have been redundant.

Some mentors also said they had wanted better support from their Undergraduate Research Directors during and after the project orientation meeting, particularly clearer expectations, deadlines, and scheduling.

Students were generally very compliant; most viewed the EvaluateUR requirements as homework. It was expected as part of the research experience, so they did it.

Over the course of the project, the project team made minor changes to a few of the assessment components in order to clarify wording, streamlined the pre-research reflection questions, and refined some of the email messages that are automatically sent to mentors and students. The project team also made a number of improvements to its user support on the EvaluateUR website. The website was overhauled to provide a clearer and more user-friendly interface. New graphics were employed throughout the website to provide a consistent graphic standard. The website updated the Roadmap, Timeline, Orientation Guide, Resources and EZStats Guide, and added several videos to support learning about and using EvaluateUR.

Assessment outcomes. All EvaluateUR assessment components saw statistically significant positive growth for both students and mentors when tested by a paired sample t test (alpha =.05). Further, the Cohen's d effect size showed medium and large magnitudes of effect for almost all components for both mentors (Table 2) and students (Table 3). The medium and large effect size suggest that the significance is not just due to chance or large sample size but to genuine impacts on student outcomes. It can be concluded with great confidence that the EvaluateUR evaluation model successfully measured objective student growth as measured by student self-reports and mentor feedback.

Table 2 Results of Paired Sample t-test of Mentor EvaluateUR Pre and Post Assessment Scores

Paired Sample t test - Mentor	t	df	Sig. (2-tailed)	Cohen's d Effect Size*
Uses and understands professional and discipline-specific language.	-21.192	640	0.000	0.84
Expresses ideas orally in an organized, clear, and concise manner.	-16.483	640	0.000	0.65
Writes clearly and concisely using correct grammar, spelling, syntax, and sentence structure.	-16.677	640	0.000	0.66
Displays insight about the topic being investigated.	-19.981	640	0.000	0.79
Shows ability to approach problems from different perspectives	-20.311	640	0.000	0.80
Uses information in ways that demonstrate intellectual resourcefulness.	-18.381	640	0.000	0.73
Effectively connects multiple ideas/approaches	-19.564	640	0.000	0.77
Demonstrates an ability to work independently and identify when guidance is needed.	-18.206	640	0.000	0.72
Accepts constructive criticism and uses feedback effectively.	-17.947	640	0.000	0.71
Uses time well to ensure work gets accomplished.	-15.730	640	0.000	0.62
Sets and meets project deadlines.	-14.087	640	0.000	0.56
Is not discouraged by setbacks or unforeseen events and perseveres when challenges are encountered.	-18.035	640	0.000	0.71
Shows flexibility and a willingness to take risks and try again.	-18.086	640	0.000	0.71
Trouble-shoots problems and searches for ways to do things more effectively.	-20.477	640	0.000	0.81
Recognizes that problems are often more complicated than they first appear.	-19.994	640	0.000	0.79
Approaches problems with an understanding that there can be more than one right explanation or even none at all.	-18.217	640	0.000	0.72
Displays accurate insight into the limits of his/her own knowledge and an appreciation for what isn't known.	-18.631	640	0.000	0.74
Challenges established thinking when appropriate.	-21.784	640	0.000	0.86
Looks for the root causes of problems and develops or recognizes the most appropriate corrective actions.	-21.316	640	0.000	0.84
Recognizes flaws, assumptions and missing elements in arguments.	-9.946	640	0.000	0.39
Demonstrates ability to formulate questions and hypothesis within the discipline.	-10.072	640	0.000	0.40
Demonstrates ability to properly identify and/or generate reliable data.	-9.677	640	0.000	0.38
Shows understanding of how knowledge is generated, validated and communicated within the discipline.	-10.143	640	0.000	0.40
Shows understanding of the way practitioners think within the discipline (e.g., as an earth scientist, sociologist, artist . . .) and view the world around them.	-8.977	640	0.000	0.35

Shows understanding of the criteria for determining what is valued as a contribution in the discipline.	-25.285	640	0.000	1.00
Shows awareness of important contributions in the discipline and who was responsible for those contributions.	-25.088	640	0.000	0.99
Reads and applies information obtained from professional journals and other sources.	-22.018	640	0.000	0.87
Is aware of professional societies in the discipline.	-24.068	640	0.000	0.95
Displays knowledge of key facts and concepts.	-23.648	640	0.000	0.93
Displays a grasp of relevant research methods and is clear about how these methods apply to the research project being undertaken.	-23.956	640	0.000	0.95
Demonstrates an appropriate mastery of skills needed to conduct the project.	-23.371	640	0.000	0.92
Recognizes that creating, modifying, misrepresenting or misreporting data including omission or elimination of data/findings or authorship is unethical.	-13.292	640	0.000	0.52
Behaves with a high level of collegiality and treats others with respect.	-10.086	640	0.000	0.40
Is clear about academic and/or professional/work plans.	-15.164	640	0.000	0.60
Is aware of how research skills relate to academic and/or professional/work plans.	-16.352	640	0.000	0.65
*Effect size: .2 small (pink), .5 medium (orange), .8 large (green)				
alpha = .05				

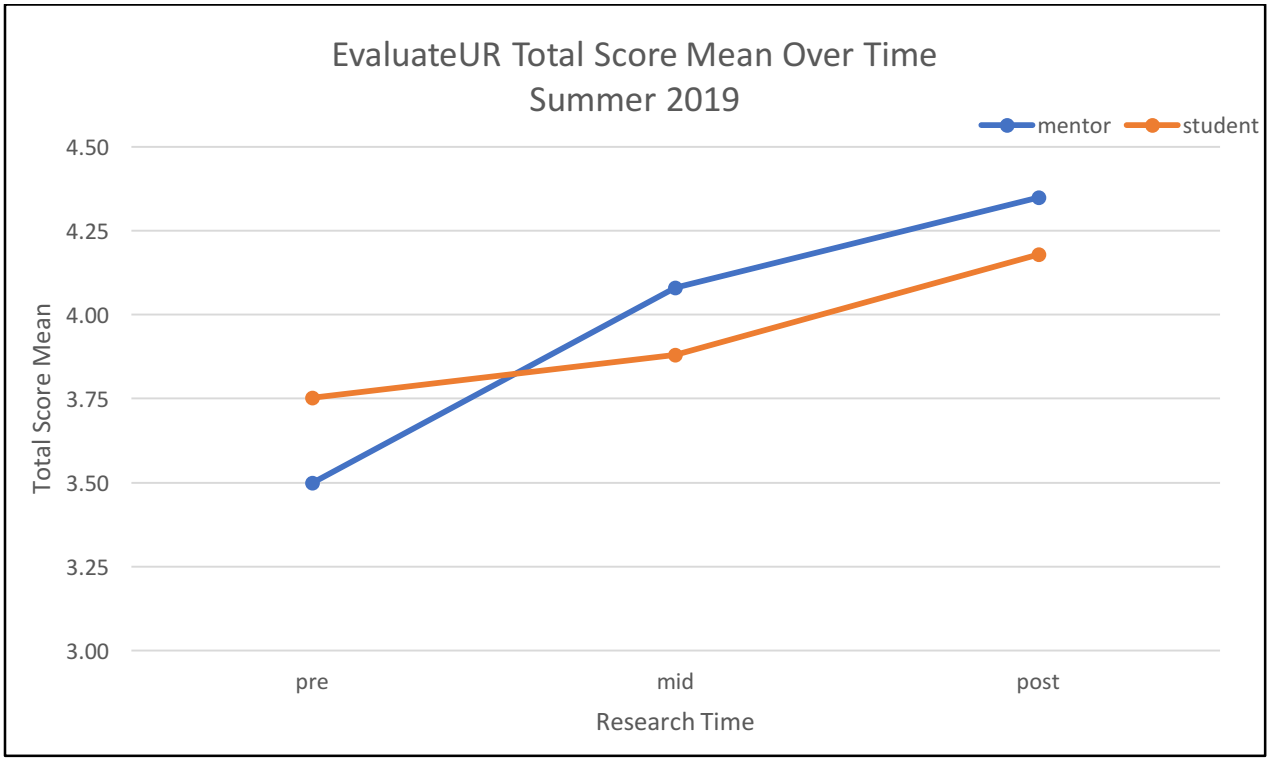
Table 3. Results of Paired Sample t-test of Student EvaluateUR Pre and Post Assessment Scores

Paired Sample t test - Mentor	t	df	Sig. (2-tailed)	Cohen's d Effect Size*
Uses and understands professional and discipline-specific language.	-13.795	630	0.000	0.55
Expresses ideas orally in an organized, clear, and concise manner.	-12.859	630	0.000	0.51
Writes clearly and concisely using correct grammar, spelling, syntax, and sentence structure.	-8.658	630	0.000	0.34
Displays insight about the topic being investigated.	-14.823	630	0.000	0.59
Shows ability to approach problems from different perspectives	-10.682	630	0.000	0.43
Uses information in ways that demonstrate intellectual resourcefulness.	-12.997	630	0.000	0.52
Effectively connects multiple ideas/approaches	-11.446	630	0.000	0.46
Demonstrates an ability to work independently and identify when guidance is needed.	-9.834	630	0.000	0.39
Accepts constructive criticism and uses feedback effectively.	-6.954	630	0.000	0.28
Uses time well to ensure work gets accomplished.	-5.925	630	0.000	0.24
Sets and meets project deadlines.	-4.963	630	0.000	0.20
Is not discouraged by setbacks or unforeseen events and perseveres when challenges are encountered.	-7.519	630	0.000	0.30
Shows flexibility and a willingness to take risks and try again.	-6.454	630	0.000	0.26
Trouble-shoots problems and searches for ways to do things more effectively.	-7.530	630	0.000	0.30
Recognizes that problems are often more complicated than they first appear.	-9.283	630	0.000	0.37
Approaches problems with an understanding that there can be more than one right explanation or even none at all.	-8.243	630	0.000	0.33
Displays accurate insight into the limits of his/her own knowledge and an appreciation for what isn't known.	-6.915	630	0.000	0.28
Challenges established thinking when appropriate.	-10.145	630	0.000	0.40
Looks for the root causes of problems and develops or recognizes the most appropriate corrective actions.	-9.602	630	0.000	0.38
Recognizes flaws, assumptions and missing elements in arguments.	-8.971	630	0.000	0.36
Demonstrates ability to formulate questions and hypothesis within the discipline.	-10.387	630	0.000	0.41
Demonstrates ability to properly identify and/or generate reliable data.	-9.099	630	0.000	0.36
Shows understanding of how knowledge is generated, validated and communicated within the discipline.	-9.326	630	0.000	0.37
Shows understanding of the way practitioners think within the discipline (e.g., as an earth scientist, sociologist, artist . . .) and view the world around them.	-10.844	630	0.000	0.43

Shows understanding of the criteria for determining what is valued as a contribution in the discipline.	-13.727	630	0.000	0.55
Shows awareness of important contributions in the discipline and who was responsible for those contributions.	-12.433	630	0.000	0.49
Reads and applies information obtained from professional journals and other sources.	-11.045	630	0.000	0.44
Is aware of professional societies in the discipline.	-13.815	630	0.000	0.55
Displays knowledge of key facts and concepts.	-14.563	630	0.000	0.58
Displays a grasp of relevant research methods and is clear about how these methods apply to the research project being undertaken.	-16.151	630	0.000	0.64
Demonstrates an appropriate mastery of skills needed to conduct the project.	-16.076	630	0.000	0.64
Recognizes that creating, modifying, misrepresenting or misreporting data including omission or elimination of data/findings or authorship is unethical.	-3.048	630	0.002	0.12
Behaves with a high level of collegiality and treats others with respect.	-4.564	630	0.000	0.18
Is clear about academic and/or professional/work plans.	-6.470	630	0.000	0.26
Is aware of how research skills relate to academic and/or professional/work plans.	-4.838	630	0.000	0.19
*Effect size: .2 small (pink), .5 medium (orange), .8 large (green)				
alpha = .05				

Some mentors expressed concern over how to score a student at the pre assessment if they didn't really know the student (e.g. from previous interactions). To explore if the mentors and students were scoring students the same way, a total scale score was computed by taking the average of all 35 components at each time for both students and mentors. Total mean scores will range from 0 to 5, the same as the component score. Figure 1 shows the results. Mentors' student scores start out a little lower than those given by students at the pre research assessment then score a little higher than students at the mid and post research assessments. An Independent Sample t test was done with 'who' (mentor or student) as the grouping variable and the 'three total scale scores' as the test variables. The mentor and student pre assessment scores were significantly different $T(1522) = t(-4.799)$, $p=.001$. The mid and post scores were not statistically different. Thus, the mentor scores were indeed different than those given by students at the onset of EvaluateUR, but converged with the students' scoring at the mid and post assessment points.

Figure 1



Impacts on students, mentors and campus site administrators of undergraduate research

Surveys administered to project participants revealed the following findings:

Students. EvaluateUR student participants were introduced to a comprehensive list of competencies and skills, including but not restricted to subject area knowledge, that they would need in order to go on to graduate work and/or succeed in the workplace. The use of these project outcome categories and their explanatory components was an EvaluateUR innovation: Evaluations of student outcomes in undergraduate research had largely been confined to faculty mentor assessments of their students' subject area knowledge. And evaluations of undergraduate research programs had focused mainly on student (and faculty) perceptions of the programs' value. In a related project innovation, post-assessment student-mentor conversations contributed to the development and enhancement of student metacognitive skills, characterized by learners becoming aware of what learning strategies they are pursuing and why, and then using that awareness to make intentional adjustments to those strategies in order to learn more effectively. Data collected at different times over the course of the project show that on average, student skills and competencies in many areas improved over the course of their research projects and students said that the project assessment protocols, including their post-assessment conversations with mentors, had been very helpful. In a survey conducted in 2019 with 189 student respondents, 88 % reported that they had discussed their academic and professional strengths and weaknesses with their mentors during at least one of the three assessment meetings, with half of the respondents

indicating that they had discussed this topic at all three meetings. Nine out of 10 survey respondents said that these discussions helped them to achieve a deeper understanding of their academic and professional strengths and weaknesses and become more aware of the knowledge, skills and competencies they needed to do well in graduate school and/or the workplace. Four out of five respondents said that the discussions helped them to identify new problem-solving strategies and ways to overcome obstacles in their other academic pursuits. These conversations also helped students confirm their plans for graduate school (76%) or employment in their major field (65%). Some 80% of the respondents also said that the assessments and follow-up discussions helped them consider the ways research can contribute to their skills and knowledge in preparation for graduate school and/or employment. When asked to explain, students said, for example:

"The conversations I had with my research mentor helped me to better understand the process of experimental design. I also had the opportunity to discuss my weaknesses, such as my lack of knowledge of professional societies, so that I could become more aware of the opportunities and resources available to me. Overall, these conversations helped me learn to critically assess my progress as a researcher, and to seek out professional development opportunities."

"It made me take a step back and look holistically at the skills I was applying or lacking then taking the steps to observe where there is room for improvement."

"Helped me improve my researching techniques and skills as well as having deeper conversations with my mentor about academic goals."

"I didn't realize all of the skills I had and was improving by doing research until I completed the evaluations and talked with my mentor. This was a great help!"

"All assessments allowed my research mentor to be completely honest with me and created a prompt as to how to determine their evaluation of me. Based on their comments and my experience, I was able to improve on both my technical and soft skills as well as find parallels to using those skills outside of the lab. However, in terms of preparing for applying to graduate school, I believe there was a lot more to be discussed that wasn't".

"Conducting research helps exercise many skills that can be applied across a wide variety of fields. Such skills include creating achievable goals, problem-solving, critically thinking, and overcoming obstacles. It also gets you more acquainted with the world of academia and the process we go through to establish data and share it with others."

General Comments from students were also largely positive*:*

"I found EvaluateUR to be an incredible helpful program for creating a dialogue about how research is conducted, and how I can personally improve as a researcher and become successful in my current and future research endeavors. I am very grateful that I had the opportunity to participate in this program."

"It seemed like a lot of work when it was explained during the orientation but now that I have done it, it wasn't too much work at all. This program is helpful and I got to talk to my professor more because of it."

A few students felt that EvaluateUR procedures diverted them from lab time they needed:
"I think that this is valuable, but only if taken seriously. My professor and I did not discuss this much due to how much time it took to work on our research each day. If we would have sat down more and talked about it, it would have been more valuable."
"I believe the assessments should be shorter as it took valuable time away from actually completing the research project."

Mentors. EvaluateUR enabled faculty mentors to observe changes in student competencies and skills using an identical assessment scale and rubric at several different stages of student research. They could therefore more easily identify the academic strengths and weaknesses of the students they mentored, enabling them to focus their mentoring efforts more productively. They could also more easily see areas where students might be over- or under- estimating their competencies and could help students gain new insights into their academic strengths and weaknesses and the relative efficacy of their learning strategies. A survey of mentor reactions to their project experiences provided responses that illustrate these findings. One hundred and thirty-six mentors responded to the survey from 22 different institutions. Most (71%) mentors had worked with one student; 21% mentored two students, and about eight percent had mentored thee or more students. When the mentors were asked about how much time they spent per week on EvaluateUR activities the range of responses was between 30 minutes and 2 hours per week with the average saying that they spent about one hour per week.

Nearly all (96%) of the respondents said they believed that there is value in undergraduate research even if the students do not accomplish their research goals.

"Research is a matter of discovery and exploration. Sometimes a "failed" project, or one that deviates substantially from what was initially envisioned, will be more instructive."

"There is a tremendous amount of learning to be gained just from the act of carrying out research. One of the major values that students usually learn is that their experiments are likely to fail. Learning how to deal with failure and overcoming it is a critical life skill."

"While it is desirable for students to achieve their research goals, they should learn many usable skills even if the results of the research are not great. For instance, how to formulate a potential solution to an open-ended problem, how to work independently, how to know when to ask for assistance, how to manage time well, how to talk about a technical topic to someone who is unfamiliar with it, how to prepare reports/presentations, etc.

"EvaluateUR was a great tool for discussing areas in which my student could improve without sounding entirely negative. It was a useful platform in that regard because it started the conversation that needed to be had."

"I used the pre- mid- and final assessment comparisons as jumping off points to discuss with my student where he needed to improve. The comparisons were extremely helpful in seeing how the student's perception of their performance was quite different from how I

saw it. It is also possible that the student was indeed mastering some of the outcomes but was not clearly communicating them to me as a mentor."

"There is a lot of value in these evaluations. Some of the most useful is providing opportunities to discuss and share our experiences as mentors and empowering students to probe these topics with their mentors and others in the lab. They help provide better feedback by reminding us of the broader impact, influence and discussion points we can use in shaping young scientists."

Faculty perspectives on how they used EvaluateUR to provide feedback to their student(s) were very positive. Seventy percent said EvaluateUR influenced the way in which they spoke to their students and 87% said that the EvaluateUR discussions helped the student gain a better understanding of their academic and professional strengths and weaknesses. Open-ended questions also reflected this positive feedback.

"The discussions encourage both of us to think about how the experience doing research provided learning opportunity and professional growth."

"The students became much more reflective about their strengths and weaknesses and this program helped them identify strengths and weaknesses they had not previously thought of."

"My student liked the discussions and helped him realize qualities he had that he did not realize before."

"My student was not confident in any of his abilities other than computer science and this program helped him realize what he was also good at and what he needed to work on."

"This helped my student, who just finished her first year of college, realize what skills she had to develop during her next few years at this university."

However, EvaluateUR feedback was found less "necessary" when mentors met with their student(s) daily or worked side by side in a lab with the student(s)

"We had these kinds of conversations but not necessarily connected to the EvaluateUR assessments. I think we were already having those conversations daily, and EvaluateUR simply documented them in a slightly different way."

"We communicate in-person all the time basis, not with an online tools like this."

"Our discussions were mainly extensions of topics already covered in our weekly meetings to some degree."

"The software might be useful if you do not have regular discussions. The students were in my office for much of their project - so feedback was constant."

According to 70% of the mentor survey respondents, EvaluateUR also had a positive influence on the way they spoke to their students about the range of outcomes that go beyond research results:

"I didn't think about many of these aspects of doing research until I read them in EvaluateUR, but all of these are good points to bring up with students."

"The EvaluateUR evaluation forms reminded us to think about the bigger research picture. Our project had a large applied component, so it was good for us to reflect more on how the work fit in with research in the area of virtual reality. If we had had more time, I would have had the student doing even more of this."

In addition, about nine out of ten respondents thought that EvaluateUR discussions helped their student researcher(s) gain a better understanding of their academic and professional strengths and weaknesses, with 42% believing it had a "considerable" or "fairly large" influence and 45% saying that it helped to "some extent." (Many of the mentors who said it only helped to some extent said they were already discussing the topics in EvaluateUR or they had a previous history with the student and had already covered many of these topics.):

"These discussions were necessary for student development. While it didn't change the content of our discussions, it did provide a pointed, rigorous opportunity to discuss key issues which made it difficult for my student to disregard points raised."

"The students I selected to do mentored undergraduate research with me had previous research experience prior to EvaluateUR. In that time and under my guidance before EvaluateUR we took necessary steps to improve their overall strengths. I feel for the most part students in my lab exceeded expectations when it came to questions and topics related to EvaluateUR."

"The survey's points of emphasis provided further confirmation of my thoughts, however most of the ideas were not new conversation points with my students."

About three-fourths of the survey respondents said that EvaluateUR assessments and related student discussions added value to their student's research experience:

"It added terminology to the discussion of growth and progress. Neutral tool was nice way to start the conversation about growth areas and relating student self-assessment to my assessment was a low-stakes way to start a conversation that could seem scary/critical to the student."

"Structured time for conversations enforced externally is a good way to overcome the rush that is summer where such valuable conversations might otherwise be forgotten."

Moreover, six out of 10 respondents said that EvaluateUR influenced their thinking about mentoring students who are conducting research:

"Simply put, it was useful, in the least, as a helpful reminder to periodically step outside of the details of the project to talk about the experience of research more broadly. Purposefully taking a meta-awareness on occasion rather than letting the experience "speak for itself" is certainly beneficial, and tying our experience to this evaluation system ensured that this happened with regularity."

About half of the mentor survey respondents indicated that they were "very" or "fairly" likely to apply EvaluateUR ideas to other students conducting research with them either independently or as part of a course. The mentors who responded that they did not plan to apply EvaluateUR concepts with other student researchers either seemed unaware of the goals of EvaluateUR or tended to emphasize research outcomes rather than student intellectual growth:

"My research is done for grants and papers. The students get mentoring in the process of the research. But it will not change the way to conduct research."

"Our project was very applied, and again, I'm not sure what the questions were supposed to bring out. The student did well on the project, but I'm not sure what the point of the questions was."

The same mentors may have influenced their student's impressions of EvaluateUR. Students who worked with mentors who said they were unlikely to use EvaluateUR concepts with other students tended to share their mentors' views:

"We both thought this program was silly and detracted from the research."

"Overall, I did not think that the evaluations with my mentor were of much use. The questions did not reveal anything that I already did not know and my mentor didn't seem interested in discussing it."

Notwithstanding these "dissenting" mentors and students, as previously noted, student survey respondents reported a very positive experience with EvaluateUR, and reported that discussions with their mentors helped them to identify their academic strengths and weaknesses.

Mentor criticisms of EvaluateUR mainly focused on the timing of the three assessments in relation to the length of the research experience (too frequent in too short a time) and the length of the assessment instruments. A number of mentors also said they had wanted better support from their Undergraduate Research Directors during and after the project orientation meeting, particularly clearer expectations, deadlines, and scheduling. A few respondents said that the project's assessment questions didn't seem to apply well to research of an interdisciplinary nature. On the whole, though, mentor feedback on their EvaluateUR experience was quite positive.

Undergraduate Research Directors. EvaluateUR provided undergraduate research directors with reliable evidence of the potential benefits of undergraduate research in helping students to achieve specific knowledge, skills and competencies they will need for graduate work and/or the workplace. This evidence was provided in the form of readily understandable statistical reports that could be used directly by research directors in their own reports to other interested parties. The project also provided the research directors with powerful tools to help them manage the projects on their campuses. These tools included a web-based dashboard that enabled the directors to track the implementation progress of all participating student-mentor pairs and automated messages to students and mentors reminding them of next steps they

should take on the project schedule. Research Director reactions to these project components were captured in two surveys administered in late summer 2019. The survey responses, together with findings from the student and mentor surveys, confirm that while the implementation of EvaluateUR did not proceed without the need to make some minor adjustments to project procedures, the project was implemented successfully at participating campuses.

Two surveys of undergraduate research directors garnered responses from 26 and 33 institutions participating in 2019 EvaluateUR. Twenty of these administrators reported participating in virtual training sessions provided by the Project Director and PI, Jill Singer, prior to implementing EvaluateUR on their campus. Site administrators who did not participate in these sessions prepared to implement EvaluateUR by availing themselves of online resources provided by the project (e.g., a project Roadmap and a project Timeline), met with or spoke to Singer directly, or consulted other administrators who had participated in the virtual training sessions. Several survey questions asked about problems encountered at each step of the process from pre-research reflection to the post-research assessment. Only minor problems were identified and included suggestions such as gaining the ability to modify student emails after the project has begun. Most of the problems identified were related to issues outside the control of EvaluateUR such as students not reading their emails, mentors not staying on schedule, and noncompliance.

More than two-thirds of the survey respondents said they checked their web-based project dashboard at least every few days in order to keep track of the progress of EvaluateUR student/mentor teams on their campus, with 21% saying they checked the dashboard every day. All but one respondent said the dashboard made it clear to them when student-mentor pairs were falling behind. When asked about how much time they spent on a weekly basis overseeing EvaluateUR, 58% of the responding administrators said 1-2 hours per week, 39% 3-5 hours, and 3% said 6-10 hours per week. Much of this time was spent cajoling mentors to complete their steps in a timely manner according to the site administrators.

More than half of the survey respondents indicated that they used EZStats – the automated statistical report of all EvaluateUR assessment results. Those who had not used EZStats, said they were waiting for student-mentor pairs to finish EvaluateUR or they planned to access EZStats soon. According to the respondents, EZStats information is being shared with faculty advisory boards, campus administrators, and other campus decision makers. Several site administrators specifically noted their intention to use the reports in order to demonstrate the value of undergraduate research and to maintain funding for those programs. Most responding site administrators (91%) also said that the students and mentors who participated in EvaluateUR benefited from the assessments and their follow-up conversations discussing the reasons for their respective assessment scores.

By and large, site administrators did not report having to face any significant challenges. The biggest issue site administrators reported was motivating mentors to stay on schedule and complete the assessments.

"At the debrief session faculty said the conversation about the learning opened up dialogue opportunities. They now could easily talk about something--Evaluate UR showed the differences between student perception and faculty perception which then made the conversation easier to have."

"The students said they enjoyed having these conversations with their mentors and that answering the questions helped them realize all of the skills they developed by doing research that they hadn't realized. It also helped them have a starting point to ask their mentor questions about careers and strengths and weaknesses. The faculty said they liked being able to use the scores also to bring up areas of improvement for the students and to also discuss relative strengths."

"Mentors repeatedly told me that they never would have taken time to pause and focus on these elements with their students. I am positive the tool helped deepen relationships and improve the overall experience."

"A chemistry faculty member said that EvaluateUR helped her to be more intentional in her work as a faculty mentor and that it helped her to understand better how the research process is a professional and personal development process for the students involved."

"I definitely think the students benefited a lot from the experience, based on my experience in mentoring and based on student feedback. As for the mentors, I think personal attitude had much to do with the benefits they received. Some people just don't like being told what to do and some mentors don't think they need a guide so they approach the exercise a bit closed off to the benefits. I saw that in some feedback. But I also saw mentors who valued the experience and gained from it. I think ultimately there needs to be a culture shift wherein mentors recognize that it isn't a bad thing or reflect poorly on them to get help for their mentoring."

"I know many personally told me they gained utility from it. In fact, they requested the paper version of the 11 learning outcomes and respective activities to share with their colleagues."

"Depth & breadth of conversation: Without this system, such check in points would have occurred but likely not involved the level of discussion that we ended up having this summer. I found the system to be a concrete way for framing our discussions. It was useful in having me consider more carefully the suite of skills one would like to see developed in undergraduate researchers."

The second survey asked site administrators for feedback on the EvaluateUR website and the online support materials and videos provided to help orient potential and current users of EvaluateUR. The feedback was very positive and constructive. Nearly all (89% of survey respondents) thought the web-based Roadmap was helpful in explaining the overall structure and implementation of EvaluateUR and 96% felt that the Timeline was helpful in explaining the sequence of project implementation steps. All responding administrators thought that a project Orientation Guide published on the website was helpful in preparing them to run an

onsite orientation session. Respondents had suggestions for very minor improvements that were reviewed by the project team and modification to the materials were made as necessary.

"Over the 2019 summer pilot there were significant changes/updates that not only looked great but made navigating the site much easier and more intuitive."

"I think that the site is attractive and easy to navigate. The main landing page is well thought out and I like the video placement front and center. When you click on the "Learn More" button, you're taken to a very useful overview of the system that articulates the benefits for different stakeholders."

Table 4 summarizes the ways in which the project benefited students, mentors and research directors.

Table 4. Benefit of EvaluateUR on Students, Mentors and Site Administrators

Students
1. Were introduced to a comprehensive list of competencies and skills – including but not restricted to subject area knowledge – that they will need to pursue graduate work and/or succeed in the workplace 2. Were provided with regular feedback about their progress through repeated assessments and follow-up conversations with mentors 3. Obtained a realistic picture of their strengths and weaknesses across all competencies and skills they should strive to achieve 4. Developed or enhanced their metacognitive skills 5. Gained greater self-awareness and confidence as they tracked their academic growth 6. Strengthened their applications to graduate programs or resumes for entering the workplace
Mentors
1. Were able to observe their research student over an extended period of time and have multiple opportunities to familiarize themselves with student work 2. Were able to make more consistent and reliable assessments of their students' academic strengths and weaknesses 3. Were able to focus mentoring efforts on specific areas where students might need extra guidance, thereby making the research more productive
Undergraduate Research Directors
1. Received new support for campus assessment efforts 2. Received readily understood and publishable statistical analyses of assessment score data providing a highly reliable and explicit portrait of student growth in knowledge and skills across a wide range of outcomes. The data could be used to demonstrate the impact of undergraduate research to campus administrators and/or external funding source(s) 3. Were provided with evidence that could be used to present their program impact/outcomes at professional meetings and could be published in journals

Project Growth and Dissemination

The EvaluateUR growth and dissemination strategies worked well. Beginning in the first phase of the project with six campus participants, EvaluateUR successfully expanded to a second phase with 14 participating campuses, including five campuses that had successfully implemented the project in Phase I and nine new participants. The project then capitalized on this success by enrolling 39 campuses in the third and final phase of its growth plan, including 28 new participants. This success can be traced mainly to two factors: 1) The project made ongoing efforts to obtain useful feedback from participating campuses on project design and procedures and continually made project refinements in response to this feedback and 2) The project engaged in successful efforts to disseminate information about the advantages of its design through CUR and at numerous presentations at professional gatherings.

The success of EvaluateUR has also provided insights into ways in which the project's methods could be adapted to the evaluation of classroom-based research, which could potentially greatly enlarge the number of students and faculty members who could benefit from the use of these methods. A new project designed to test these ideas is now underway. In addition, the core ideas of EvaluateUR are being adapted to the evaluation of students participating in an international underwater robotics (remotely operated vehicle) competition under the auspices of the Marine Advanced Technology Education Center in Monterey, California. Finally, EvaluateUR, with its NSF-WIDER supported having ended, has now been made available on a subscription basis to college and university undergraduate research programs.

Conclusions

EvaluateUR introduced several unique and innovative evaluation ideas that had not previously been tried or widely employed. The project has succeeded in demonstrating that these ideas can be the basis for a practical approach to gathering reliable data about specific student knowledge and skill outcomes, advancing student learning and helping faculty to focus their mentoring efforts more productively. The unique features of EvaluateUR were a departure from the *practice* but not the *purpose* of most evaluations. Evaluations ordinarily do not get directly involved in the teaching/learning process, out of concern that doing so might compromise objectivity in their measures of student progress. In this tradition, evaluation is seen, not as a direct, real-time contribution to student learning, but as a potential corrective to education programs that should be either improved or abandoned. EvaluateUR shared the universal evaluation goal of enhancing student learning but the project's designers believed that for undergraduate research that goal could best be achieved by involving students and faculty together in the assessment task in order to develop and enhance student metacognitive skills, as described above. Given the manner in which the assessments were structured, EvaluateUR was able both to collect reliable data on specific student outcomes and to use this procedure as a way to contribute directly to student learning. The EvaluateUR method was designed to accomplish this without sacrificing accuracy or objectivity and without posing an undue burden for participating faculty and students. Based on the findings summarized above, these benefits have been confirmed by students, mentors, and undergraduate research directors.