EvaluateUR: Findings from the NSF Project

Background

In 2008 SUNY-Buffalo State initiated an effort to develop a new method for measuring student knowledge and skill outcomes in its undergraduate summer research program. where each student researcher is mentored by a faculty member. The new evaluation method was designed by Dr. Jill Singer, Distinguished Teaching Professor at SUNY Buffalo State and Mr. Daniel Weiler, Daniel Weiler Associates, an education evaluation consulting firm. An independent evaluation of the method was conducted by Ms. Bridget Zimmerman, Nautilus Evaluation, an education evaluation firm. After six years of successful implementation of the new method at Buffalo State, an award from the National Science Foundation's WIDER program supported the scaling up of the method – now called EvaluateUR – to additional college and university undergraduate research programs. In 2019, the final year of NSF funding, EvaluateUR was implemented on 37 campuses across the country. Findings from the 2019 implementation are presented here. EvaluateUR is now available as a subscription-based service.

EvaluateUR Design

Key features in the design, testing and implementation of EvaluateUR include:

Introduce students to a wide range of knowledge and skills they would need for graduate study and/or the workplace.
Students are assessed on 11 outcome categories reflecting desirable student knowledge and skills. Each of the 11 outcome categories is further delineated by specific, measurable student behaviors, leading to a total of 35 outcome components that 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.

<u>Obtain statistically valid and reliable measures of student achievement on each of these knowledge and skill outcomes.</u>

EvaluateUR refined the assessment instrument first introduced at Buffalo State for gathering data on student outcomes. The instrument uses a five-point Likert scale indicating the extent to which a student displays 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 score themselves on each outcome component, and their research mentors, using the same instrument, independently score their students. This regimen provides pre, mid, and post assessment scores. On each of these occasions, students and mentors compare their assessments and discussed the reasons for the scores they have 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 assessments are implemented according to explicit and uniform standards across varied disciplines and across different student-mentor pairs. These and related features of the EvaluateUR design are meant to overcome the usual objections to assessments of student outcomes based on subjective mentor judgments, which typically rely on disparate standards without common assessment parameters across disciplines or among different mentors. Orientation materials, together with live and online participant meetings, and online training videos and resources also help to minimize inter-rater reliability issues, so as to further strengthen the validity of the assessments across institutions and disciplines.

<u>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.</u>

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 stresses to participating students and mentors that the assessment scores are less important than the student-mentor conversation that follows the assessments, at which time the student and mentor share their respective rationales for assigning particular scores and discuss the reasons for differences, if any, in their perceptions. During these conversations, the mentor can assist the student with identifying strategies and opportunities to enhance the student’s learning process. These conversations are aimed at providing students with new insights into their thinking processes and learning strategies. In this way, the evaluation seeks to collect reliable data on specific student outcomes and also contribute directly to student learning.

<u>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.</u>

The project’s focus on parallel student and mentor assessments followed by a conversation about differences in their perceptions is designed to provide students with fresh insights, as described above. At the same time, this design can help faculty mentors gain a new appreciation for the ways in which students see – and perhaps misunderstand – 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.

Devise practical tools to make it easy for undergraduate research program directors to oversee the evaluation and report its findings for all participating students.
EvaluateUR's online dashboard is designed to help administrators keep track of the implementation progress of all participating student-mentor pairs. A system of automated messages helps to ensure that participants are staying on schedule in completing assessments and other steps in the research program. At each of the three assessments (pre, mid, post) an automated Score Report is generated to facilitate student and mentor score comparisons and discussion. In addition, administrators have access to a web-based statistical tool called EZStats. EZStats automatically generates, 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 changes. For institutions with 20 or more pairs of completed mentor-student assessments, a paired sample t-test provides significance testing outcomes, including effect sizes. The scores are reported by EZStats in a format that is readily understandable by a lay reader and immediately usable for reports. The online dashboard, automated messages, Score Reports, and EZStats reports are all managed by the Science Education Resource Center (SERC) at Carleton College.

Summary Findings from Scaling Up EvaluateUR to Serve Many Colleges/Universities

In 2019, EvaluateUR was implemented by undergraduate research directors at 37 institutions with 799 student-mentor research pairs completing all of the EvaluateUR steps. Changes in assessment scores for this group and feedback from user surveys support a number of conclusions about the method:

- EvaluateUR succeeded in introducing student participants to a comprehensive list of competencies and skills they would need in order to go on to graduate work and/or succeed in the workplace.
- All EvaluateUR assessment components saw statistically significant positive student gains and the EvaluateUR evaluation model successfully measured objective student growth.
- 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. These conversations also helped most students confirm their plans for graduate school or employment in their major field.
- Research mentors found it easier to identify the academic strengths and weaknesses of the students they mentored, enabling them to focus their mentoring efforts more productively.
- The mentors 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.

- EvaluateUR provided undergraduate research program directors with tools to help them manage their research program as well as provide them with reliable, readily understandable evidence of the potential benefits of their undergraduate research program.

Statistical Tests

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 method 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 - Student	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 about assigning a score on the pre-research assessment instrument for a student whose work they did not know from previous research interactions. To explore this concern and determine whether or not the mentors were scoring students the same or differently than students scored themselves, a total scale score was computed by taking the average of all 35 outcome components for each assessment (pre-, mid-, and end-of-research) for both students and mentors. Compared to the student scores, mentor scores on the pre-research assessment were slightly lower, increasing for both the mid-research and end-of-research assessments. Mid and post mentor scores were in alignment with student scores at mid and post times respectively.

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-research assessment scores were significantly different T(1522) = t(-4.799), p=.001. However, the mid- and end-of-research scores were not statistically different indicating that while the mentor scores were lower than those given by students at the beginning of the research experience, the student and mentor scores converged for both the mid-point and end-of-research assessments.