Title: An Evaluation Capacity Building Toolkit for Principal Investigators of Undergraduate Research Experiences: A Demonstration of Transforming Theory into Practice

Author:

Audrey S. Rorrer^{1*}

University of North Carolina at Charlotte

9201 University City Blvd.

Charlotte, NC 28223 USA

Audrey.rorrer@uncc.edu

Abstract:

This paper describes the approach and process undertaken to develop evaluation capacity among the leaders of a federally funded undergraduate research program. An evaluation toolkit **was** developed for Computer and Information Sciences and Engineering² Research Experiences for Undergraduates³ (CISE REU) programs to address the ongoing need for evaluation capacity among principal investigators who manage program evaluation. The toolkit was the result of collaboration within the CISE REU community with the purpose being to provide targeted instructional resources and tools for quality program evaluation. Challenges were to balance the desire for standardized assessment with the responsibility to account for individual program contexts. Toolkit contents **included** instructional materials about evaluation practice, a standardized applicant management tool, and a modulated outcomes measure. Resulting benefits from toolkit deployment **were** having cost effective, sustainable evaluation tools, a community evaluation forum, and aggregate measurement of key program outcomes for the national program. Lessons learned included the imperative of understanding the evaluation context, engaging

¹ * Corresponding Author. Audrey.rorrer@uncc.edu. Tel +1 704 687 8648. Fax +1 704 687 3516

² Computer and Information Sciences and Engineering, CISE

³ Research Experiences for Undergraduates, REU

stakeholders, and building stakeholder trust. Results from project measures are presented along with a discussion of guidelines for facilitating evaluation capacity building that will serve a variety of contexts.

Key Words: Evaluation Capacity Building, Toolkits, Participatory Evaluation, Undergraduate Research Experiences

1. Introduction

The current zeitgeist in the field of evaluation purports that building evaluation capacity is important, necessary, and valued (Cousins & Earl, 1992; Patton, 2008; Johnson & Weiss, 2011). It is important to develop the capacity of professionals who are leading programs because many have no training in evaluation research and lack the requisite skills to conduct an effective and meaningful assessment of outcomes. Not only are evaluation skills necessary, but building capacity empowers principal investigators to conduct consistent and valid assessments, provides understanding from the unique perspectives and contexts of each project, and strengthens generalizability of collective findings across programs with similar goals (McDavid & Hawthorn, 2006; Patton, 2008).

Evaluation capacity building, ECB, is an activity separate from conducting evaluation that is done at both the organizational and individual level (Labin, Duffy, Meyers, Wandersman, & Lesesne, 2012), and by definition must be sustainable with appropriate support tools (Preskill & Boyle, 2008; Stockdill, Baizerman, & Compton, 2002). A common thread across the ECB literature is that in order for capacity building efforts to be successful, they must be collaborative, participatory, and supportive of an evaluative culture (Labin, Duffy, Meyers, Wandersman, & Lesesne, 2012). Labin, et al. (2012) note that while there is broad agreement within the evaluation community about the definition of ECB, there is an ironic lack of discussion regarding the process of evaluation capacity building.

Common barriers to ECB have been identified as “timing, lack of alignment, and information overload” (p.21, Johnson & Weiss, 2011). These barriers are particularly pertinent within the context of federally funded undergraduate research programs. Federally funded programs have been criticized for their dearth effective evaluation practice and dissemination of evaluation (Clifford & Millar, 2008;

Shavelson & Towne, 2002), yet there is an expectation of credible evaluations across such programs (Frechtling, Mark, Rog, Thomas, Frierson, et al., 2010).

To address the barriers to and need for evaluation capacity within a subset of National Science Foundation (NSF) Research Experiences for Undergraduates program in computer and information sciences and engineering fields, an online evaluation toolkit was developed to provide targeted instructional resources and tools for quality evaluation at the project level and aggregate program level. The development and use of digital evaluation support tools is not unique, and is quite common among national multi-site evaluations (Campbell, Townsend, Shaw, Karim, Markowitz, 2015). For example, the National Science Foundation's ADVANCE: Increasing the Participation and Advancement of Women in Academic Science and Engineering Careers grant program provides a website instructing institutional awardees on specific data to be collected, reported, along with reporting standards and timelines (http://www.advance.vt.edu/measuring_progress/tag_toolkits.html).

The CISE REU evaluation toolkit (the toolkit) is unique in its scope, feasibility, and adaptability to other REU programs and domain content. The toolkit aligns with the Joint Committee on Standards for Educational Program Evaluation Standards (Yarbrough, Shulha, Hopson, & Caruthers, (2011) in the incorporation of core principles of utility, feasibility, propriety, accuracy, and accountability. The rationale and approach to the development of the toolkit is presented in this paper along with the implications for evaluation theory and practice.

2. Background

2.1. Overview of CISE Research Experiences for Undergraduate Program

The National Science Foundation (NSF) sponsors roughly 700 programs for undergraduate research project sites with 7,000 college student participants, commonly referred to as Research Experiences for Undergraduates, or REUs, in scientific disciplines at universities each year. The NSF REU program is organized into domains where identified REU sites are funded to manage iterative cohorts of undergraduate students who engage in research projects, most typically during the summer. Within the domain of Computer and Information Sciences and Engineering (CISE), there are between 50-

70 REU sites with up to 700 undergraduate participants each year. The lifecycle of these site projects is typically a three-year period with an annual iterative cycle of recruiting undergraduate student participants during the fall and spring semesters for each upcoming summer research experience. While CISE REU sites share common goals, each site has unique research project objectives, for example, one site may focus its research projects on artificial intelligence and another on developing educational computer games. A clear challenge in building evaluation capacity in this domain is in balancing the need for individual site context with the need for broader, generalizable program impact (Hardre, Slater, & Nanny, 2010; McDonald & Hawthorn, 2006; Patton, 2008).

2.2. Evaluation Challenges for CISE REU Project Sites

The NSF set accountability standards and provides general evaluation guidelines via a program evaluation handbook (Frechtling, et al., 2010) available on the agency website. However, financial support for REU sites is directed toward supporting the undergraduate students and their research, with little support funding for evaluation implementation. Most REU sites are led by a single Principal Investigator, who manages student participant recruitment and selection, research laboratory preparations, and general site administration, which includes student housing arrangements and stipend payment procedures. Site project evaluation is not only limited in financial support, but also in expertise, both of which are commonly noted challenges in evaluation literature (Degroff, Schooley, Chapel, Poister, 2010; Worthen & Schmitz, 1997).

Prior to the development of the CISE REU evaluation toolkit, the focus of most site evaluations consisted of reporting the number of student participants, descriptions of their research projects and any research project dissemination. Student outcomes and overall program goals were not evaluated with consistency. Instead, each site project was left to internally develop evaluation strategies and techniques that were employed by project staff with little to no background in evaluation theory and practice, as indicated in the group discussions held prior to and during the development of the toolkit. Thus, CISE REU sites across the nation were reinventing methods and instruments for assessing project outcomes to meet the evaluation demands of the funding agency.

These circumstances present the challenges of and need for implementing effective, efficient, and valid project evaluations that not only address the individual site projects, but that also enable the aggregate program evaluation across CISE REU sites. In a study of computer science faculty about undergraduate research experiences, Matzen and Alrifai (2012) found that there was great variation in pedagogy, objectives, and project implementation despite the similar objectives of developing student research skills, preparation for graduate school and professional development. These findings suggest variation in project evaluation, especially given that no specific CISE REU site evaluation guidelines existed prior to the development of the toolkit discussed herein. The needs assessment that initiated the toolkit development, discussed below, confirmed that many variations of project evaluations were occurring across the CISE REU program.

Several key challenges are inherent to the CISE REU community, and are indicative of cluster evaluations. Cluster evaluation means that the evaluation “involves multiple sites in which project-level evaluation occurring at each site is conjoined to a greater or lesser degree with an over-arching, cross-site evaluation of some ‘cluster’ of these projects at the national, state, or initiative level” (p. 300, Worthen & Schmitz, 1997). Noted challenges indicative of cluster evaluations that the CISE program face include conceptual clarity, separation of evaluation authority and responsibility, alleviating project level anxiety, tracking longitudinal results, and sufficiency of evidence (Worthen & Schmitz, 1997). Within the CISE REU community, there was considerable lack of consistency of impact and outcomes measures, low evaluation expertise, and a level of concern and trepidation at the project level evaluation requirements. These issues together compound the problem of satisfactory evidence. The intent of the Toolkit is to identify common indicators across the program sites that when examined across the sites in a collective fashion would hold more significance, more fully engage the stakeholders, and alleviate the burden of evaluation for those PIs without access to evaluation expertise.

To address these challenges, an ECB process, as prescribed by Volkov and King’s checklist (2007), was executed. The process for establishing a climate open and ripe for evaluation capacity

development, the plan, the infrastructure of tools, support for and use of the tools are described herein.

The overall logic model is presented in Figure 1.

Figure 1. CISE REU Evaluation Toolkit Logic Model

Context	Input	Output	Outcomes	
			Short Term	Long Term
Need for domain specific evaluation instruction and support	Needs Assessment to determine common indicators across program	Online assessment toolkit	Shared, domain specific assessment resources	Formative & summative feedback loop across the NSF division
Need for cost-effective, adaptable <i>project</i> assessment tools	Assessment Work Group	A la Carte Student Survey	Aggregate descriptive data on student applicants and participants	Ability to determine if program targets are met and inform budget decisions
Need for sustainable, aggregate <i>program</i> assessment	Annual Faculty PI Meeting Discussions	Common Application	Accessible, cost-effective assessment tools for project monitoring	Ability to determine program merit based upon consistent student outcomes measurement

2.3. Assessment Work Group for CISE REU Project Sites

The CISE REU community convenes annually in a meeting of the Principal Investigators (PIs) to accomplish two primary goals. The first goal is to welcome new PIs to the community and orient them to the practice of managing and assessing their projects. The second goal is to engage in a collective conversation about the merits and impacts of the overall program. In 2009, the community was organized into several topical interest groups referred to as Working Groups. Meeting attendees elected to join the group of their choice upon registration to attend the meeting. During the 2009 annual meeting, the Assessment Working Group identified two primary challenges to address. The primary objective **was** the practical need for project leaders, or PIs, to have common resources for evaluation. The resources needed to be accessible and feasible to use, i.e. cost-effective and easy to implement quickly. Another objective **was** to understand the broader impact and outcomes from the national program, which required that the

community establish a common language from which to discuss project outcomes collectively. The group recognized the practical needs of the community by targeting useful tools, and **was** also cognizant of the broader impact of enabling two distinct contexts to be considered, that of the local sites and that of the larger collective of the national program. The tools **were** designed for voluntary, modular deployment to account for project contexts, and provided a consistent language across the national program for evaluation of student outcomes.

2.4. CISE REU Evaluation Toolkit Goals and Objectives

The overarching goals for the toolkit are to facilitate REU site project evaluations, provide site specific accommodations, and to provide overall evaluation of the national program. The objectives are to create and disseminate an online resource for conducting REU site evaluations within the CISE domains at the local level, and to also provide summative aggregate outcomes to the national community. The toolkit contents include: evaluation practice instructional materials, an inventory of evaluation tools and techniques employed within the community, a standardized program applicant management tool, and a validated modular outcomes instrument. The toolkit provides benefits to the CISE REU community by providing cost-effective, sustainable evaluation tools, a community forum for localized project evaluation, and a national program evaluation to assess overall program outcomes.

A crucial component of evaluation capacity building is determining the overall climate and context within the community being served (Fitzpatrick, 2012; Preskill, 2008; Stockdill, Baizerman & Compton, 2002). Therefore, the first task of the Assessment Working Group was to propose the toolkit rationale and its components to the entire CISE REU community, and to solicit input and participation in using its resources. The plan was well received by the community of faculty, and the discussion among the group indicated that the toolkit would deliver relevant and useful tools and resources. As McDonald, Rogers, & Kefford (2003) note, ECB should align with the supply and demand that is extant within the community. In this case, the proposal delivered to the CISE REU community during its annual meeting was greeted with enthusiasm, believed to be due to the flexibility of the tools. The toolkit and its resources are both optional and adaptable. **Ongoing discussions throughout the CISE community have**

consistently indicated that mandating tools is undesirable. The sites have that have been operational for many funding cycles have established instruments and appropriate evaluation expertise to conduct program evaluation; therefore mandating tool use, e.g. a survey instrument, would disrupt their long-term student outcomes research.

3. Methods

3.1 Needs Assessment

A needs assessment was conducted in two phases⁴. Phase one determined what explicit and implicit goals and outcomes were being measured in each project. Phase two investigated how these were being assessed across the projects. The results indicated that project site evaluations employed evaluation methodologies incongruent with stated goals, used invalid and unreliable instruments, often developed in-house by project staff who were not trained in evaluation. However, goals pertaining to student outcomes were shared, despite the variety of labels that were used across the project sites to discuss them. An overall timeline of the toolkit development and implementation is provided below in Figure 2.

Figure 2. Toolkit Development Timeline

3.2. Define Common Indicators and Measures

The broad goals of the NSF REU program are to engage undergraduate students in meaningful

⁴ [Faculty], [Institution] CISE REU Project Needs Assessment, Phase I and Phase II, 2009; <https://reu.uncc.edu/cise-reu-toolkit>.

research and to attract a diverse pool of people into scientific careers. The stated goals of the CISE REU community are to recruit, retain and prepare undergraduate students for advanced degree programs and research in particular. The needs assessment found common indicators across all CISE REU project sites that reflect four themes of student self-efficacy, intentions to obtain advanced degrees in scientific disciplines, prevalence of help-seeking behaviors, and positive attitudes towards scientific research. The CISE REU community gravitated toward these main indicators, with the belief that measuring research knowledge was evident in individual research project outcomes, i.e. student research project findings reported in individual papers. Indicators were developed to show the overall level of REU success, and refer to the broad program outcomes intended by NSF.

The common indicators are specified by primary research questions as directed by the funding agency and by the community at large. The questions are: what is the demographic picture of undergraduate students who apply to and participate in CISE REU projects? In addition, what changes take place in project participants over the course of their undergraduate research experiences, as follows: a) to what degree do students believe they are capable of conducting computing research (i.e. computing self-efficacy); b) what are their attitudes towards computing research?; c) do students plan to seek graduate degrees?; and d) to what degree do students' possess academic help seeking skills? In order to be able to address the primary research questions of the project site evaluations and the collective program evaluation, two instruments were developed, a common application for undergraduates interested in CISE REU participation, and a modular survey instrument for undergraduate students who were selected and actually participated in a summer experience.

3.3 Common Application

All project sites are invited to utilize the Common Application as the primary means of collecting undergraduate student applicants. The Common Application, as it is referred to on the toolkit, is designed as a means of collecting descriptive information about the national pool of CISE REU applicants to determine the overall demographic of undergraduate students seeking summer research experience in CISE domains. The items were derived from an analysis of all existing site applications, which were

found to have similarity across sites. Based upon the costs of existing online applicant management systems, free Google software management tools were chosen. An online application form was created through Google Forms that included the general applicant items common across all sites, and additional items could be included as sites desired to tailor to their respective needs. For example, all sites asked applicant grade point averages, gender, ethnicity, and research interests. However, some sites asked applicants to indicate housing or lab preferences. Each site that chooses to employ the Common Application received a site specific link to the Google Form, to which access was shared with the site Principal Investigator and the toolkit researcher, who is able to populate an aggregate spreadsheet of all site applicants. Site applicants were provided consent at the initiation of an application, and can opt out of sharing their information with the aggregate pool provided to the NSF. Individual sites managed their applicants locally, including the decision to accept applicants. The aggregate applicant pool was utilized only as a descriptive tool to provide collective information to the NSF and CISE REU community about the characteristics of the national applicant pool.

3.4 A la Carte Survey

The A la Carte Survey was developed through the needs assessment conducted at Humboldt State University. Four scales for measuring the common indicators were originally recommended for consolidation into one survey instrument to collectively measure student participant outcomes in a repeated measures design, i.e. pre-test, post-test conducted at the beginning and end of student program participation. Additional scales have been added based upon interest expressed among the PI community. The intention of the survey was to provide project sites with access to an instrument with reliable and valid constructs to measure the common indicators. The survey was designed to be modular, in that a site may opt for some, if not all, of the construct measures, e.g., a site may wish to measure only three of the common indicator constructs, rather than all of them. The common indicator constructs were: self-efficacy, intent to attend graduate school, attitudes towards computing disciplines, academic help-seeking/coping skills, research skills, leadership skills, and scientific identity. Demographic items are also

measured and the post-survey includes additional items assessing mentoring satisfaction and program satisfaction.

Modular Survey Scale Origins. Self-efficacy, attitudes towards computing and academic help seeking/coping have been shown to correlate with academic retention in computing disciplines (Dahlberg, Barnes, Buch, and Rorrer, 2011). Actual graduate school enrollment is a lagging indicator for REU site projects, and are not able to be tracked in the project lifecycle. However, research has shown that behavioral intentions are predictive of actual behaviors (Muchinsky, 2006), and therefore the intention to attend graduate school scale was developed. Computing self-efficacy scale was developed on the basis of Bandura's (1997) theory, and the intention to attend graduate school items were developed from Ajzen's theory of planned behavior (Ajzen, 2000). The computing attitudes scale and academic help seeking/coping scales were modifications of exiting subscales with reported reliability demonstrating strong psychometric properties (see Hoegh & Moskal, 2009 and Struthers, et al., 2000 respectively). Research skills items were adapted from the Survey of Undergraduate Research Experiences developed by the Howard Hughes Medical Institute (Lapatto, 2004). Leadership and scientific identity scales were adapted from the work of Chemers, Zurbriggen, Syed, Goza & Bearman (2011) to address minority students in undergraduate science disciplines. The mentoring items were adapted from work in the medical community (Berk, Berg, Mortimer, Walton-Moss, & Yeo (2005). Table 1 displays the scales, sample items, Cronbach alpha coefficients from the first year of deployment for the respective scales.

Table 1. A la Carte Survey Scales, Sample Items, and Reliability Coefficients

Common Indicator Subscales	Sample Items	Original Source	Cronbach's Alpha Coefficients
Computing Self-Efficacy: 15 items (2010)	I can formulate a research problem; I can understand computing research literature	Developed based upon Longitudinal Assessment of Engineering Self-Efficacy (LAESE)	.901
Intent towards Graduate School: 12 items (2010)	I plan to apply to graduate school; It is expected of me that I will apply to graduate school.	Based upon Ajzen's Theory of Planned Behavior (2000)	.903

Computing Attitudes: 12 items (2010)	I like to use computer science to solve problems; Knowledge of computing will allow me to secure a good job.	Hoegh & Moskal, (2009)	.860
Academic Help-Seeking/Coping: 7 items (2010)	When I do poorly on an important exam, I typically....I drop out of the class I'm doing poorly in; I try to come up with a strategy about what to do.	Struthers, Perry, & Menece, (2000)	.602
Research Skills : 26 items (2014)	I know about: ethics in scientific research; application of the scientific method.	SURE III; Lapatto (2004)	.915
Leadership Skills: 11 items (2014)	I know how to be a good team member; I know how to encourage team performance.	Chemers, Zurbriggen, Syed, Goza & Bearman (2011)	.918
Scientific Identity: 17 items (2014)	I have come to think of myself as a scientist; I have a strong sense of belonging to the community of scientists	Chemers, Zurbriggen, Syed, Goza & Bearman (2011)	.908
Mentoring Satisfaction- Post Only: 11 items (2014)	My mentor was: approachable; supportive & encouraging.	Berk, Berg, Mortimer, Walton-Moss, & Yeo (2005)	.957

The A la Carte survey currently includes a total of 111 items measuring the 8 common indicators on a 5 point Likert-type scale; an additional set of demographic items capture gender, ethnicity, level in school, current college major, and overall program satisfaction information. A principal component analysis was conducted via SPSS to observe the patterns of item correlations, to thereby reduce the number of items used to measure the observable factors. Factor (survey items) loadings below .4 are generally considered low and were omitted. A total of 33 items loaded below .4 and were dropped from the survey. Demographic items (13 items) and program satisfaction items (9 items in post survey) were not included in the factor analysis and remain in the modulated survey. Cronbach alpha reliability

coefficients of the survey constructs are reported in Table 1, revealing acceptable levels. However, the reliability coefficient level for academic help-seeking and coping is moderately low; because this scale is widely used in Engineering education settings, it is viewed acceptable to the CISE community as a measure student behavior.

Survey Methodology. A quasi-experimental repeated measures assessment methodology was used. An online survey was administered to REU site participants during the first week of the REU experience (pre survey). The same survey was administered during the final week of the REU experience, with additional program evaluation items (post survey; 8-10 weeks after pre survey). All CISE REU PIs were invited to participate in the study during the annual CISE REU PI Meeting held annually in March. In addition to the study information session, PIs received email reminder invitations to participate in the study. The primary researcher sent the pre survey web link to the participating PIs along with informed consent and an email invitation to be distributed to their site participants during the first week of the REU. PIs also provided the start and end dates of their site programs and total number of student participants to ensure the appropriate timing of the pre and post administration for each site and allow for response rate calculations. The survey included informed consent to participate in the study and a request for student permanent email addresses, enabling the researcher to administer the post survey directly to REU participants at the conclusion of the REU site experiences. Human subjects research compliance was obtained from the researcher's institution prior to the start of the study. Each site participating in the A la Carte survey received their respective student response data in anonymous form for individual site program analysis. The student data package included syntax codes for establishing the eight constructs, and a code sheet of the items and response keys.

3.5. Assessing the Toolkit

The quality of the toolkit and its subsequent tools were assessed in three primary ways. Each year at the annual meeting of PIs, a short survey was deployed to measure overall satisfaction, use, and feedback. The usage of the tools was monitored and tracked within each annual cycle. Additionally, the toolkit

website user statistics were compiled periodically, along with responses to an ongoing feedback survey embedded on the website.

4. Results

4.1 Common Application

In the five years of deployment of the Common Application tool, about a third of sites used the tool each year. An upward trend in usage is indicated over time. Table 2 displays the site participation statistics for the Common Application and A la Carte Survey.

Table 2. Common Application Usage Statistics

Year	# of CISE REU Sites	Common Applications		A la Carte Survey	
		#	%	#	%
2014	62	25	40	22	35
2013	58	26	45	30	52
2012	62	22	35	23	37
2011	60	20	33	18	30
2010	54	13	24	20	37

Usage of the Common Application shows a promising upward trend, albeit a slight decline in 2014. Follow up email inquiries to the site PIs each year indicated that the main reasons for electing not to participate in the Common Application were due to the existence of internal applications already in use, as many sites have been in operation prior to the development of the application tool. All sites that employed the Common Application noted the benefits in having an automated application, readily available and deployable upon receipt of NSF funding, thereby saving site resources. No sites have discontinued their use of the common application after initial adoption.

Because participation was limited to under half of the total sites, a clear picture of the total applicant pool is not yet possible. However, the demographics of the applicants from the participating sites provided a glimpse into the trends of the national pool of applicants. Gender and ethnic minority disparities among the applicant pool continue, mirroring the enrollment disparities widely discussed in the computing fields (Gilbert, 2006). Twenty-seven percent of the applicant pool in 2010 consisted of

women, dropping to 24% in 2012, and rising to 28% in 2013 and 2014, marking a consistent low pattern of less than a third of all applicants being women. Underrepresented minority student applicants, defined as any ethnicity other than Asian or Caucasian, have ranged between 31% and 36% from 2010-2014.

There is strong indication that targeted recruiting of women and minority group members is necessary, which while not particularly surprising to the CISE PI community, was startling when considered across the variety of sites using the application tool. The aggregate compilation of data gives voice to a rising concern within the community.

4.2 A la Carte Survey

All sites that have used the A la Carte Survey chose to deploy the full survey of eight constructs. Site participation is indicated in Table 2, with a range of between 18 and 30 sites deploying the instrument each year. Student participation has been upwards of 66% response rates; see Table 3. Response rates were estimates based upon the funding model of the REU program, which supports 10 students per site per summer; site PIs were asked to provide actual numbers of their student participants but not all complied with the request.

Table 3. Student Participation in A la Carte Surveys by Number and Response Rates

Year	Pre Survey	Post Survey	Response Rate*	% Female*	% URM*
2014	226	149	66%	42%	23%
2013	217	209	70%	38%	22%
2012	167	151	66%	37%	25%
2011	199	137	76%	37%	31%
2010	196	144	72%	30%	31%

*Calculated from Post Survey responses; estimates based upon 10 students per site (e.g., 2010: 200 students; 2011: 180 students; 2012: 230 students; 2013: 300; 2014: 220); URM is underrepresented minority ethnicity students and includes all ethnicity except Caucasian and Asian.

The survey responses were analyzed each year following the summer program. An analysis of variance was conducted to determine significant differences between pre and post assessment on the construct means. In the first year of the survey, no significant findings were obtained. Self-efficacy gains from pre to post were significant each year since 2011. Research skills increased significantly both years of measurement (2013 and 2014). Attitudes toward computing improved significantly in 2012. Intentions

to attend graduate school outcomes were inconsistent, increasing significantly in 2013, and decreasing significantly in 2014. An overview of the annual survey findings is presented below (Table 4).

Table 4. Pre and Post Survey Construct Means, Standard Deviations, and Statistical Significance

Variable	2010		2011		2012		2013		2014	
	Pre (SD)	Post (SD)	Pre (SD)	Post (SD)	Pre (SD)	Post (SD)	Pre (SD)	Post (SD)	Pre (SD)	Post (SD)
	N=196	N=144	N=199	N=137	N=167	N=151	N=217	N=209	N=226	N=149
Self-Efficacy	3.05 (.47)	3.43 (.40)	3.11 (.47)	3.48 (.39)*	3.75 (.69)	4.06 (.79)*	3.81 (.74)	4.34 (.53)*	3.96 (.92)	4.27 (.50)*
Intent to Grad School	3.10 (.46)	3.19 (.47)	3.22 (.59)	3.23 (.66)	3.93 (.75)	3.93 (.79)	3.89 (.65)	4.06 (.70)*	4.06 (.80)	3.84 (.63)*
Attitudes	3.59 (.32)	3.65 (.29)	3.67 (.36)	3.70 (.38)	3.58 (.24)	4.55 (.49)*	4.39 (.44)	4.43 (.47)	3.09 (1.01)	4.21 (.44)
Help-Seeking/ Coping	2.94 (.33)	2.98 (.37)	2.98 (.39)	2.98 (.36)	3.52 (.48)	3.48 (.51)	3.40 (.48)	3.47 (.45)	3.85 (0.66)	4.04 (.82)
Research	Not Measured						3.22(.74)	4.00 (.77)*	3.91 (.78)	3.97 (.58)*
Scientific Identity	Not Measured								4.00 (0.59)	3.79 (1.00)
Leadership	Not Measured								4.09 (1.18)	4.32 (.63)

The increases in self-efficacy, attitudes toward computing research, research skills, scientific identity, and leadership are supported in prior research (Lapatto, 2004, 2007; Laursen, Hunter, Seymour, Thiry & Melton, 2010; Russell, Hancock, & McCullough, 2007). Differences in plans to attend graduate school have been found. While many studies report increased interest in graduate school (Laursen, Hunter, Seymour, Thiry & Melton, 2010; Russell, Hancock, & McCullough, 2007), there have been several studies that have shown decreasing interest in graduate school following involvement in a research program (Laursen, Hunter, Seymour, Thiry & Melton, 2010).

4.3 Online Toolkit

The online toolkit was successfully deployed to the CISE REU community. Feedback gathered at the annual meetings, via discussions and informal polls, provided resounding support for the accessibility,

utility and cost benefits to the group. The toolkit contains a library of resources tailored to the community along with detailed outlines regarding evaluation practice. The toolkit serves as a pathway for use of the Common Application and the A la Carte Survey, both of which can be deployed independently, or in conjunction with the collective research project for the NSF. The toolkit is thought to also benefit any researcher involved in research experiences for undergraduates, in that the tools and resources can be interdisciplinary. This belief has been substantiated by the expressed interest in the toolkit from an additional NSF directorate, the Chemistry REU program⁵. Website traffic on the toolkit reached 1,909 unique visitors since inception in 2012, with a slight increase in traffic in recent years. Website traffic for the first half of 2015 was more than comparable traffic in 2014, an indication that use is increasing. The traffic increase was likely due to the development of video tutorials explaining the tools and how to adopt and/or adapt them. These videos were the direct result of receiving feedback during the PI meetings, which called for simplicity in instructions and immediacy in support of the tools. Polling results during these meetings showed consistent positive opinions about the toolkit, its component tools, and the scope of services provided therein. PIs who have chosen not to use the Common Application and the A la Carte Survey indicated that their sites have well established application and measurement procedures in place, and that switching to toolkit products would be disruptive to their program administration. Among the sites that do use the tools, i.e. the Common Application and/or A la Carte Survey, all of them have continued to utilize these tools throughout the duration of their programs. This is a strong indication of the satisfaction within the community of their functionality.

5. Discussion

The survey tool enables program wide analysis of student outcomes while simultaneously providing site project measures with valid constructs. The toolkit provides suggestions for non-parametric analyses appropriate for individual site assessment while also allowing aggregate student outcomes to be measured across multiple sites, providing valuable information to the community. Aggregate analyses of student

⁵ Author was invited speaker at the Chemistry REU Principal Investigator's Workshop, July 15-17, 2015, San Antonio, TX

outcomes across sites are more powerful than by individual sites alone, and provide a comprehensive description of CISE REU sites. For example, each year statistical analyses are conducted to assess changes between pre to post attitudes, and to measure any group differences by gender, level in school and ethnicity, which are reported at the annual meeting. Significant increases have been found on several common indicators, indicating the overall impact of the program.

There is value in having operationalized, comparable metrics for a cohesive conversation within the community of CISE REU PIs. The tools provide site autonomy and flexibility while also providing consistency and efficiency. Aggregate descriptions enable confirmation of candidate and student outcomes traits, supply and demand for undergraduate research, and the ability to understand the characteristics and trends of students who apply to and participate in undergraduate research programs. Without consistency of the application and assessment, the community could not deduce trends, and are limited in the ability to discern program success or areas of needed attention. For example, viewing trends in applications over five years has shown that despite individual site recruiting efforts to attract women and underrepresented minority candidates, the percentages of these types of applicants has remained flat.

Limitations to the interpretation of the A la Carte Survey findings are confined for several reasons inherent in the study of undergraduate research programs. Responses are self-report, and thus susceptible to response bias. Causal links between student research experiences and attitudes may not be made given the lack of experimental design. These issues are prominent and pervasive in the study of undergraduate research (Lapatto, 2004 and 2007). However, Lapatto notes that the “validity of student reports seems modest but not trivial” (2007, p.305), and argues that student reports of individual attitudes and personal gains are indeed best obtained by asking students directly.

6. Lessons Learned

6.1 Understanding Context is a Continual, Reciprocal Process

Evaluation context is important, as is indicated throughout the evaluation literature (Fitzpatrick, 2012). Assessing the evaluation context allows purposeful direction of the evaluation beyond accountability, i.e. of reporting what took place and the outcomes attributable, toward an examination of

the overall process and what is learned throughout that process for continuous formative evaluation. The toolkit researcher guided the toolkit development with context in mind, and made a concerted effort to understand the context of the CISE REU community, including PIs, their sites, and the funding agency supporting each site. In following Greene's five dimensions of context (2005), setting demographics, the economics, the climate, the norms for interaction, and the political realm, were observed. The development of the toolkit was conducted collaboratively within the community for which intended, and therefore the group climate was well understood. The characteristics of the CISE REU community, including the PIs and NSF program directors, engaged in a symbiotic relationship. The toolkit was intended to supplement, not replace, project site evaluations and was devised to be flexible and adaptable, because of the awareness of the climate within the community.

The nature of context is that it continuously evolves and is reciprocal between the community being evaluated and the evaluator. Context is transformative, and as the community began to discover larger trends in student characteristics and outcomes, the conversation shifted. On one level, the context unfolded to the evaluator as the tools are developed with a focus on utility, feasibility, and accountability, and on another level, positioned the community to measure 'reciprocal learning' in evaluation (Baakan, Nunez, & Couture, 2014). The CISE community appears to be shifting into a learning organization (Preskill and Boyle, 2008).

6.2 Stakeholder Trust is Critical and Earned

Because the toolkit proposal to the CISE REU community received positive support, the toolkit researcher assumed that adoption of the tools would be more widespread across the group than it was initially. Recognizing that ECB efforts need to align with the organizational and leadership needs (Fitzpatrick, 2012; Patton, 2008; Stufflebeam, 2004), project leaders were motivated by the appeal of practical needs and having garnered endorsement from the funding agency. Initially, toolkit practice adoption was most successful among newly funded CISE REU sites. The PIs who elected to utilize the toolkit, the common application and survey were primarily newly awarded CISE REU sites, or had been struggling to find feasible means for evaluation. The PIs who opted out either had access to an affordable

internal evaluator, and/or had developed intricate systems for their internal evaluations prior to the development of the toolkit. Adoption of the toolkit practices have continued as credible results were disseminated during the annual meetings. Developing trust across stakeholders was an effort that builds upon exposure and experience within the community. The toolkit was the product of community consensus, i.e. something created by, and for the community, rather than something required for regulatory compliance purposes. The external evaluator status of the toolkit researcher provided the necessary expertise to be invited to serve the community, but to fully understand and be trusted by the community, the evaluator had to fully engage with the community not as an outsider but as a core member. Transparency of methods, accountability to the community to disclose the process and products, and to communicate regularly catalyzed and sustained this relationship.

6.3 Takeaways for the Evaluation Community At Large

Evaluation consistency across like programs can be informative to those programs, and perhaps more importantly, can support localized efforts to measure program outcomes. In establishing tools for evaluation capacity building, relationships between the evaluator and the community of stakeholders are paramount. Evaluators must be objective and external to the stakeholder group, while maintaining a close relationship. The relationship can become trustworthy through listening and understanding the group's concerns and needs. An open dialogue and ongoing discussion loop ensures that the evaluative process is organic.

7. Future Directions

The context of the CISE REU community continues to evolve as new sites are funded, and as new research questions arise. Several new studies and toolkit components are in development, as a result of the conversation among the CISE REU community, which include a study of faculty mentoring engagement and career impact, a longitudinal alumni tracker tool, and study of REU site organizational characteristics that relate to student success. The toolkit has allowed the question, are CISE REU programs effective at improving student attitudes and knowledge?, to be addressed collectively. Now, attention can be devoted to exploring the program more deeply, by exploring how faculty are impacted by

their engagement, and also investigating structural factors that contribute to student success.

8. Conclusion

Research experience program evaluation has been inconsistent because it is localized and often led by faculty with little or no exposure to evaluation theory and practice. Project staff, who are typically the faculty Principal Investigators, often find themselves ill-equipped to plan and conduct rigorous project evaluations, and yet are accountable for producing site evaluations annually. The CISE REU Evaluation Toolkit is the result a grass roots efforts to produce common approaches to evaluation and share experiences and expertise among the CISE REU Site PI group. The toolkit provides instructional resources for site PIs who may not have the evaluation expertise or funds to engage professional evaluators within the start-up timeframe, and consists of three primary components: a website containing assessment and evaluation guidelines particularly applicable to the CISE REU group, a Common Application for recruiting students, and a pre and post student outcomes survey. The consensus of the group, based upon the annual meeting discussions, is that the toolkit is an essential resource for supporting efficient and quality site assessment and evaluation.

By involving the stakeholders directly, we are able to better understand the context, to align the toolkit practices with the community needs, and to develop trust and credibility within the community. A key success of the CISE REU evaluation toolkit is that it has aligned project indicators and measures, to form a national language for measuring the projects both individually and holistically. Overall results from the aggregate student studies inform the funding agency of the overall CISE REU program outcomes and serve as a preliminary step in formative policy evaluation for the national program.

Acknowledgements

This work was initiated by Teresa Dahlberg, and supported by NSF-0851745: CISE REU Site: Computing Research for Undergraduates: Visualization, Virtual Environments, Gaming, and Mobile Robotics & Networking; NSF-1156822: CISE REU Site: Exploring Human Centered and Socially

Relevant Interactive Technologies in Computer Vision, Visualization, Pervasive Computing, Serious Games, and Social Networks, and by NSF- 1346847: CISE REU Evaluation Toolkit Expansion Project. Support for the development of the toolkit website was provided by Kashif Sharif, Scott Heggen, and Naveen Mysore, former graduate student assistants, in the Computer Science Department at the University of North Carolina's College of Computing and Informatics.

References

- Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, 52, 27-58.
- Baakan,, L.L., Nunez , J., & Couture, C. (2014). A course model for building evaluation capacity through a university – community partnership. *American Journal of Evaluation*, 35(4), 579-593.
- Bandura, A. 1997. *Self-efficacy: The Exercise of Control*. New York: W. H. Freeman and Company.
- Berk, R., Berg, J., Mortimer, R., Walton-Moss, B. & Yeo, T. (2005). Measuring the effectiveness of faculty mentoring relationships, *Academic Medicine*, 80(1), 66-71.
- Bourgeois, I., Whynot, J., & Theriault, E. (2015). Application of an organizational evaluation capacity self-assessment instrument to different organizations: Similarities and lessons learned. *Evaluation and Planning*, 50, 47-55.
- Campbell, R., Townsend, S., Shaw, J., Karim, N., & Markowitz, J. (2015). Can a workbook work? Examining whether a practitioner evaluation toolkit can promote instrumental use. *Evaluation and Planning*, 52, 107-117.
- Clifford. M. & Millar, S. B. (2008). *K-20 partnerships: Literature review and recommendations for research* (WCER Working Paper 2008-3). Wisconsin Center for Educational Research. Retrieved September 28, 2012 from http://www.wcer.wisc.edu/publications/workingPapers/Working_Paper_No_2008_03.pdf.
- Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, 14, 397-418.

- Dahlberg, T., Barnes, T., Buch, K. & Rorrer, A. (2011). The STARS Alliance: Viable strategies for broadening participation in computing. *Transactions on Computer Education, Special Issue on Broadening Participation, 11 (3)*,18.
- Degroff, A., Schooley, M., Chapel, T., Poister, T. (2010). Challenges and strategies in applying performance measurement to federal public health programs. *Evaluation and Program Planning, 33*, 365-372.
- Fitzpatrick, J. L. (2012). An introduction to context and its role in evaluation practice. In D. J. Rog, J. L. Fitzpatrick, & R. F. Conner (Eds.), *Context: A framework for its influence on evaluation practice. New Directions for Evaluation, 135*, 7-24.
- Frechtling, J., Mark, M., Rog, D., Thomas, V., Frierson, H., Hood, S., & Hughes, G. (2010). *The User-Friendly Handbook for Project Evaluation*. Arlington, VA: NSF.
- Gilbert, J. (2006). Making a case for BPC. *Computer, 39 (3)*, 83-86.
- Hardré, P. L., Slater, J. & Nanny, M. (2010). Redesigning and Aligning Assessment and Evaluation for a Federally-funded Math and Science Teacher Education Program. *Evaluation and Program Planning, 33*, 498-510.
- Hoegh, A., & Moskal, B. M. (October, 2009). Examining science and engineering students' attitudes toward computer science. Paper presented at 39th Frontiers in Education Conference, San Antonio, Texas.
- Johnson, K., & Weiss, I. R. (2011). Compulsory project-level involvement and the use of program-level evaluations: Evaluating the Local Systemic Change for Teacher Enhancement program. In J. A. King & F. Lawrenz (Eds.), *Multisite evaluation practice: Lessons and reflections from four cases. New Directions for Evaluation, 129*, 17-23.
- Labin, S., Duffy, J., Meyers, D., Wandersman, A., & Lesesne, C. (2012) A research synthesis of the evaluation capacity building literature. *American Journal of Evaluation, 33*, 307-338.
- Lapatto, D. (2004). Survey of Undergraduate Research Experiences (SURE): First Findings. *Cell Biology*

Education, 3, 270-277. Available:

<http://www.grinnell.edu/academics/areas/psychology/assessnebts/sure-iii-survey>

Lapatto, D. (2007). Undergraduate research experiences support science career decisions and active learning. *CBE - Life Sciences Education*, 6, 297-306.

Laursen, S., Hunter, A., Seymour, E., Thiry, H., & Melton, G. (2010). *Undergraduate Research in the Sciences: Engaging Student in Real-Science*. San Francisco, CA: Jossey-Bass.

LAESE Survey, AWE Assessment Project, Survey, V3.1, Available:

<https://www.engr.psu.edu/awe/default.aspx>

Matzen, R., and Alrifai, R. (2012). Defining undergraduate research in computer science: A survey of computer science faculty, *Journal of Computing Sciences in Colleges*, 27, 31-37.

McDavid, J. C., & Hawthorn, L. R. L. (2006). *Program evaluation and performance measurement: An introduction to practice*. Thousand Oaks, CA: Sage.

McDonald, B., Rogers, P., & Kefford, B. (2003). Teaching people to fish? Building the evaluation capacity of public sector organizations. *Evaluation*, 9, 9-29.

Muchinsky, P. M. (2006). *Psychology Applied to Work*, 8th Ed. Thomson Higher Education, Belmont, CA.

National Science Foundation ADVANCE Toolkit, available

http://www.advance.vt.edu/measuring_progress/tag_toolkits.html

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Preskill, H. (2008). Evaluation's second act: A spotlight on learning. *American Journal of Evaluation*, 29, 127-138.

Preskill, H., & Boyle, S. (2008). A conceptual model of evaluation capacity building: A multidisciplinary perspective. *American Journal of Evaluation*, 29, 443-459.

Russell, S. H., Hancock, M. P., and McCullough, J. (2007). Benefits of undergraduate research experiences. *Science* 316, 548-549.

- Shavelson, R. J., & Towne, L. (2002). *Scientific research in education*. Washington, DC: National Academy Press.
- Stockdill, S. H., Baizerman, M. & Compton, D. W. (2002). Toward a definition of ECB process: A conversation with the ECB literature. *New Directions for Evaluation*, 93, 7-25.
- Struthers, C. W., Perry, R. P., & Menec, V. H. (2000). An examination of the relationship among academic stress, coping, motivation, and performance in college. *Research in Higher Education*, 41, 581-593.
- Stufflebeam, D. L. (2004). The 21st century CIPP model: Origins, development, and use. In M. Alkin (Ed.), *Evaluation roots tracing theorists' views and influences*. Thousand Oaks, CA: Sage.
- Volkov, B., & King, J. A. (2007). A checklist for building organizational evaluation capacity. *The Evaluation Center*: Western Michigan University. Retrieved from <http://www.wmich.edu/evalctr/checklists/ecb.pdf>
- Worthen, B. R., & Schmitz, C. C. (1997). Conceptual challenges confronting cluster evaluation. *Evaluation*, 3, 300-319.
- Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). The program evaluation standards: A guide for evaluators and evaluation users (3rd ed.). Thousand Oaks, CA: Sage.