OPEN ACCESS

Edited by:
Ida Ah Chee Mok,
The University of Hong Kong,
Hong Kong

Reviewed by:
Kathryn Holmes,
Western Sydney University,
Australia
Veronica Catete,
North Carolina State University,
United States

*Correspondence:
Bobby Habig
bhabig@amnh.org

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: 23 April 2020
Accepted: 10 November 2020
Published: 09 December 2020

ORIGINALRESEARC H
published:09 December 2020
doi:10.3389/feduc.2020.55480 6

Citation:
Habig B (2020) Practical Rubrics for Informal Science Education Studies:
(1) a STEM Research Design Rubric for Assessing Study Design and a (2)
STEM Impact Rubric for Measuring Evidence of Impact.
Front. Educ. 5:554806. doi:
10.3389/feduc.2020.554806

Practical Rubrics for Informal Science Education Studies: (1) a STEM Research Design Rubric for Assessing Study Design and a (2) STEM Impact Rubric for Measuring Evidence of Impact

Bobby Habig [1,2]*

[1] American Museum of Natural History, New York, NY, United States, [2] Department of Biology, Queens College, City University of New York, Flushing, NY, United States

Informal learning institutions, such as museums, science centers, and community-based organizations, play a critical role in providing opportunities for students to engage in science, technology, engineering, and mathematics (STEM) activities during out-of-school time hours. In recent years, thousands of studies, evaluations, and conference proceedings have been published measuring the impact that these programs have had on their participants. However, because studies of informal science education (ISE) programs vary considerably in how they are designed and in the quality of their designs, it is often quite difficult to assess their impact on participants. Knowing whether the outcomes reported by these studies are supported with sufficient evidence is important not only for maximizing participant impact, but also because there are considerable economic and human resources invested to support informal learning initiatives. To address this problem, I used the theories of impact analysis and triangulation as a framework for developing user-friendly rubrics for assessing quality of research designs and evidence of impact. I used two main sources, research-based recommendations from STEM governing bodies and feedback from a focus group, to identify criteria indicative of high-quality STEM research and study design. Accordingly, I developed three STEM Research Design Rubrics, one for quantitative studies, one for qualitative studies, and another for mixed methods studies, that can be used by ISE researchers, practitioners, and evaluators to assess research design quality. Likewise, I developed three STEM Impact Rubrics, one for quantitative studies, one for qualitative studies, and another for mixed methods studies, that can be used by ISE researchers, practitioners, and evaluators to assess evidence of outcomes. The rubrics developed in this study are practical tools that can be used by ISE researchers, practitioners, and evaluators to improve the field of informal science learning by increasing the quality of study design and for discerning whether studies or program evaluations are providing sufficient evidence of impact.

Keywords: informal science education, museum education, research design, STEM, rubric design, evidence-based outcomes, out-of-school time

INTRODUCTION

Informal science education (ISE) programs can be important vehicles for facilitating interest in science, technology, engineering, and mathematics (STEM) (National Research Council, 2009; Young et al., 2017; Habig et al., 2018). Indeed, in the last few decades, multiple studies and evaluations have reported evidence that involvement in informal, out-ofschool time (OST) STEM programs is linked to participants' awareness, interest, and engagement in STEM majors and careers (e.g., Fadigan and Hammrich, 2004; Schumacher et al., 2009; Winkleby et al., 2009; McCreedy and Dierking, 2013). Because many of these studies and evaluations vary considerably in how they are designed and in the quality of their designs, it is often difficult to gauge whether the outcomes reported are supported with sufficient evidence (Institute for Learning Innovation, 2007). This is particularly a dilemma for studies of ISE programs because participation is voluntary and thus more fluid than formal programs, and there is also considerable variation in the number of contact hours between programs and among participants (Institute for Learning Innovation, 2007; National Research Council, 2009). Therefore, a continuous challenge faced by the ISE community is how to gauge whether and to what extent the outcomes reported by studies and evaluations are supported by evidence. To address this problem, the goal of this study was to develop user-friendly rubrics that can be used to assess research designs and STEM outcomes of ISE studies. These rubrics, in turn, can be used to discern whether individual research studies or program evaluations are providing sufficient evidence supporting claims such as increased awareness, interest, and engagement in STEM majors and careers.

Over the past decade, there have been multiple initiatives carried out by several national agencies, including the National Research Council, the United States Department of Education, and the National Science Foundation, with the goal of identifying characteristics of high-impact STEM programs (e.g., U.S. Department of Education, 2007; What Works Clearinghouse, 2008; National Research Council, 2011, 2013). Many of these agencies have established criteria for assessing a range of STEM outcomes including the mastery of twenty first Century Skills (National Research Council, 2011), the implementation of Next Generation Science Standards (National Research Council, 2014), and the impact of teacher professional development programs on student achievement (Yoon et al., 2007). Additionally, the Committee on Highly Successful Schools or Programs for K-12 STEM Education established criteria for identifying the effectiveness of STEM-focused schools and associated student outcomes (National Research Council, 2011). In recent years, various agencies have turned their attention to ISE programs. For example, the U.S. Department of Education (2007), in a report by the Academic Competitive Council, advanced as a national goal to improve awareness, interest, and engagement of STEM careers in the context of informal education. Additionally, the Institute for Learning Innovation (2007) was charged with assessing the quality and strength of evidence of STEM outcomes of ISE programs and of its participants. Further work has also been completed by several national agencies outlining criteria that can be used to identify effective OST STEM projects (Friedman, 2008; National Research Council, 2010, 2014, 2015; Krishnamurthi et al., 2014). Overall, there have been many strides made by science education stakeholders in terms of identifying characteristics of high-quality studies and for recommending criteria for assessing program outcomes. However, for studies of informal science projects, there remains a need for developing accessible methodologies for assessing the effectiveness of STEM interventions and the strength of the evidence of ISE research outcomes.

Due to the unique characteristics of informal learning environments, it is quite challenging to develop evidence-based criteria to assess whether a research study or evaluation has achieved specific outcomes (National Research Council, 2010). Informal science education, defined as *voluntary* participation in science during out-of-school time hours, typically occurs after school, on weekends, and during the summer in a variety of settings including but not limited to museums, zoos, universities, and, non-profit organizations (Blanchard et al., 2020). By design, ISE programs are voluntary, inquiry-based, and emphasize choice learning (National Research Council, 2009). Thus, for many programs, the random assignment of participants to treatment and control groups, often considered the gold standard in research design (What Works Clearinghouse, 2008), is often logistically infeasible, potentially upsetting to learners, and may jeopardize the validity of certain studies (National Research Council, 2009, 2015). An additional challenge is that the durations of many OST experiences are short-term making it difficult to measure program impact especially if the evidence of effects occur downstream of the experience (National Research Council, 2015). Furthermore, because many ISE OST programs are designed with the specific intent of differentiating from formal school programs, program leaders often avoid the administration of written assessments (National Research Council, 2015). Lastly, the Academic Competitive Council (U.S. Department of Education, 2007) recognizes that informal learning experiences are highly individualized, complex, and multifaceted, and suggest that due to the modest scale of many of these programs, they may not warrant a costly assessment approach. Therefore, the application of criteria typically used to assess the effectiveness of outcomes of formal science programs and participants (e.g., What Works Clearinghouse, 2008) might not be feasible for assessing the effectiveness of research studies or program evaluations designed to measure outcomes of participants of ISE programs. Promisingly, in recent years, many ISE stakeholders have developed rigorous research designs that are alternatives to random control trials including the employment of mixed methods and triangulation designs (Institute for Learning Innovation, 2007; Flick,

2018a,b). Nonetheless, the unique nature of ISE programs must be accounted for when developing methodologies for assessing the effectiveness of STEM interventions in an informal setting.

Despite the inherent challenges of assessing the effectiveness of research studies and program evaluations of ISE participants, knowing whether the outcomes reported by these studies are supported with sufficient evidence is important for several reasons. First, there is an economic justification for gauging what works and what doesn't work because many ISE institutions, non-profit organizations, and governmental agencies invest considerable monetary and human resources to support informal STEM education initiatives (Wilkerson and Haden, 2014). Therefore, to convince funding agencies, policy makers, and the public at-large that investing in ISE OST programs is important, the ISE community needs to show that these programs are helping young people to persist in STEM (U.S. Department of Education, 2007; Wilkerson and Haden, 2014). Second, the use of user-friendly tools can serve as a guide to inspire researchers to design rigorous studies that maximize evidence-based outcomes (Institute for Learning Innovation, 2007; Panadero and Jonsson, 2013). Consequently, program leaders can more confidently use the information from these research studies to revise, redesign, and continuously improve their programs. Lastly, as more programs provide evidence of high impact, the ISE community can extract program design principles from highly effective programs and where appropriate, ISE OST program leaders can adopt and adapt these principles across institutions (Klein et al., 2017).

One area of interest by STEM stakeholders related to assessment is how and to what extent ISE programs augment participants' STEM major and STEM career outcomes (e.g., Cuddeback et al., 2019; Chan et al., 2020). Indeed, over the past decade, a major goal set forth by the National Research Council and the United States Department of Education is to inspire and motivate students to consider a STEM pathway (U.S. Department of Education, 2007; National Research Council, 2011, 2013, 2015). In the context of informal education and outreach, the Academic Competitiveness Council (U.S. Department of Education, 2007) identified increased awareness, interest, and engagement in STEM majors and careers as priorities; each outcome is defined below:

- *STEM major awareness*: increased knowledge and awareness of the various STEM disciplines available as fields of study at institutions of higher education
- *STEM major interest:* increased curiosity, motivation, and attention toward a STEM discipline as a focus of study at an institution of higher education
- *STEM major engagement*: a formal commitment to a STEM discipline as a focus of study in an institution of higher education
- *STEM career awareness*: increased knowledge and understanding of various STEM professions
- *STEM career interest*: increased curiosity, motivation, and attention toward STEM professions
- *STEM career engagement*: employment in a STEM profession.

In response to this challenge, many ISE programs have provided outreach and programming specifically designed to augment students' awareness, interest, and engagement in a STEM pathway. Moreover, hundreds of studies and evaluations have been carried out to assess participants' outcomes. Unfortunately, a tool to assess research design and to test whether these studies are supported with sufficient evidence is lacking, hence the focus of this study.

Our understanding of ISE OST programs is derived from two forms of published knowledge—studies that have been published in peer-reviewed journals and studies that are the result of internal and external program evaluation (National Research Council, 2015). Peer review is considered the cornerstone of academic research because research methods and findings are subject to critical examination by experts within a discipline. Peer-reviewed studies are essential for answering scholarly questions and for communicating meaningful research (Gannon, 2001). As an alternative to peer review, internal and external program evaluations are also valuable for documenting program outcomes and for informing stakeholders on how to improve program design. Of these two forms of evaluation, external evaluations are often preferred by stakeholders because during the evaluation process, a program of interest is subject to independent analysis by an objective third party; however, external evaluations are typically more expensive than internal evaluations (U.S. Department of Education, 2007; National Research Council, 2015). Of the many types of program evaluations, the two most common conducted in ISE research are formative and summative. Formative evaluations typically occur during the developmental stage of a program and are particularly informative for improving program design. Summative evaluations are conducted after the completion of a program and are useful for assessing whether the program outcomes align with the project goals and objectives (Institute for Learning Innovation, 2007). The choice to conduct peerreviewed research or an internal or external evaluation depends on many factors including available financial resources, the nature and duration of the program, and the goals of the stakeholders. Regardless of which form of published knowledge is selected, it is critical that research studies and program evaluations are rigorously designed to ensure the validity of research outcomes.

The aim of this study was to develop user-friendly rubrics to assess research design and to gauge whether a research study or program evaluation provided sufficient evidence to support specific claims (e.g., increased awareness, interest, and engagement in STEM majors and careers). To accomplish this goal, first I reviewed what experts consider to be evidence of high-quality research design and evidence of impact. Based on these research-based recommendations, I created a

STEM Research Design Rubric and a STEM Impact Rubric tailored specifically for quantitative, qualitative, and mixed methods studies and evaluations of ISE OST programs. Second, I tested these rubrics for user-friendliness, reliability, and validity. Based on feedback from STEM researchers and practitioners, I made revisions to the rubrics when appropriate. Lastly, I assessed specific ISE OST studies and evaluations using these rubrics and provided case studies that illustrate how these tools can be used to evaluate research design and evidence of impact. Through this process, I developed practical tools that can be used by ISE researchers, STEM practitioners, and other stakeholder to assess the effectiveness of STEM interventions and evidence of research outcomes for both research studies and evaluations.

THEORETICAL FRAMEWORK

The *theory of impact analysis* is described as a "rigorous and parsimonious" framework for mapping the assessment of impacts (Mohr, 1995 p. 55). The theory, which stems from the seminal work of Campbell and Stanley (1963) and later refined by Mohr (1995), considers three basic experimental designs: (1) experimental; (2) quasi-experimental; and (3) retrospective. In a true *experimental design*, the researcher sets up one or more subjects to receive the treatment (participation in the program) and another group in which one or more subjects do not receive the treatment (the control). As a benchmark, the control and treatment groups are assigned randomly, and an adequate number of subjects are assigned to participate in the study. As a best practice, the theory of impact analysis recommends when possible, the use of larger treatment and control groups in order to help increase the sensitivity of a study. For example, the Institute for Learning Innovation (2007), while assessing the impact of different program evaluations of various ISE programs, defined an adequate sample size as a minimum of 50 subjects. Two examples of a true experimental design are the pre-test, post-test design with random assignment and the post-test only design with random assignment (**Table 1**). The second design considered in the theory of impact analysis is the *quasi-experimental design*. In the quasi-experimental design, researchers set up intervention and comparison groups, but the groups are *not* assigned randomly. If the study is designed so that comparison groups are closely matched in key characteristics, evidence suggests that a quasi-experimental study can yield strong evidence of the intervention's impact. Examples of quasiexperimental studies include post-test only designs, comparative change, and comparative time series (**Table 1**). Lastly, in a *retrospective design*, also known as *ex post facto* ("after the fact"), the subjects that received treatment were not assigned by the experimenter. For example, if some youth register for an ISE program and others do not, the selection of participants was not assigned by the researcher; rather, the participants underwent self-selection.

In alignment with the theory of impact analysis, the U.S. Department of Education (2007) proposed a "Hierarchy of Study Designs for Evaluating the Effectiveness of a STEM Educational Intervention" consisting of three hierarchical levels—(1) experimental designs (randomized controlled trials), (2) quasi-experimental designs (well-matched comparison groups), and (3) other designs (e.g., pre/post studies, comparison groups without careful matching). In this hierarchy, a welldesigned randomized controlled trial is the preferred method; quasi-experimental designs are preferred when experimental designs are not feasible, and other designs are considered when the first two designs are not feasible. Thus, according to the theory of impact analysis, the best designed studies are those that exhibit internal validity (i.e., studies that can make a causal link between treatment and outcomes) and those that exhibit external validity (studies that are generalizable to other programs and populations). Threats to internal validity include non-equivalent comparison groups, non-random assignment of subjects, and confounded experimental treatments (Fuchs and Fuchs, 1986).

Threats to external validity include the timing of the study, the setting in which it occurred, the study subjects, and treatment conditions (Mohr, 1995). The theory of impact analysis thus describes a framework for reducing threats to internal and external validity and for developing experimental designs that measure the extent to which a study provides evidence of impact.

One limitation of the theory of impact analysis is that it might underestimate the power of qualitative studies. Qualitative studies are important because they emphasize depth of understanding and provide rich information on how participants have interpreted their STEM experience (Diamond et al., 2016). Because qualitative research relies on open-ended questions and in-depth responses, researchers often use this approach to identify patterns and to develop emerging themes (Jackson et al., 2007). According to Creswell and Poth (2018), high quality qualitative studies share common characteristics including the employment of rigorous methodological, data collection, and data analysis protocols, and the incorporation of one of the five following approaches to qualitative inquiry: (1) a narrative study, (2) a phenomenological study, (3) a grounded theory study, (4) an ethnographic study, and (5) a case study (**Table 1**). When qualitative studies are rigorously designed, they can be effective means for probing how ISE OST experiences impact participants' STEM pathways (U.S. Department of Education, 2007; National Research Council, 2015). One way to include qualitative research in an impact analysis is to allow for the combination of quantitative and qualitative data, which is the basis of the theory of triangulation (Greene and McClintock, 1985; Creswell and Poth, 2018).

The *theory of triangulation* is based on the idea that the collection of multiple sources of quantitative and qualitative data helps to increase study validity allowing researchers to gain a more complete picture of participants' outcomes (Denzin, 1970; Ammenwerth et al., 2003; Flick, 2018a,b). The concept of triangulation can be traced to Campbell and Fiske (1959), who developed

the idea of "multiple operationalism," which argues that the use of multiple methods ensures that the variance reflects the study's outcomes and not its methodology. Triangulation has also been described in the literature as convergent methodology, convergent validation, and mixed methods (Hussein, 2009). The triangulation metaphor stems from military and navigational strategy where multiple reference points are used to identify an object's exact position (Smith, 1975). By applying basic principles of geometry, multiple perspectives allow for greater accuracy. Similarly, researchers can improve the accuracy of their interpretations by collecting multiple sources of data assessing the same phenomenon (Jick, 1979). Together, the *theory of impact analysis* and the *theory of triangulation* are helpful lenses for informing the development of a STEM Research Design Rubric and a STEM Impact Rubric.

METHODS

The primary objectives of this study were to develop a userfriendly STEM Research Design Rubric and a STEM Impact Rubric for ISE researchers, evaluators, and practitioners. To do

TABLE 1 | Study designs for assessing the effectiveness of a STEM intervention.

Quantitative study design	Examples	Diagram	Advantages/Disadvantages
		X = STEM intervention C = comparison group T = treatment group R = random assignment O = outcome measure/evidence NA = not applicable	
Experimental design (randomized controlled trials)	1. Pre-test, post-test designwith random assignment 2. Post-test only with randomassignment 3. Solomon four group design(Solomon, 1949)	1. T(R): OXO 2. C(R): OO 3. T(R): XO C(R): O T(R): OXO C(R): OO T(R): XO C(R): O	1. Reduces threats to internal validity/doesn't control for effect of pre-test 2. Controls for pre-test effects/doesn't measure change over time 3. Strongest quantitative design for reducingthreats to validity
Quasi-experimental designs (well-matched comparison groups)	1. Pre-test, post-test designwith comparison group 2. Post-test only withcomparison group 3. Time series withcomparison group	1. T: OXO C: OO 2. T: XO C: O 3. Example: T: OXXOXXO C: OOO	1. Reduces threats to internal validity/doesn't control for effect of pre-test 2. Assesses change over time/non-random assignment increases threats to validity 3. Assesses longer term change/non-randomassignment increases threats to validity
Other quantitative designs	1. Pre-test, post-test designwithout comparison group 2. Post-test only withoutcomparison group 3. Time series withoutcomparison group	1. T: OXO 2. T: XO 3. Example: T: OXXOXXO	1. Assesses change over time/lack of control increases threats to validity 2. Provides a snapshot/lack of control increases threats to validity 3. Assesses longer term change/lack of controlincreases threats to validity
Qualitative study design	**Examples**	**Description**	**Advantages/Disadvantages**
Qualitative design	1. Narrative study 2. Phenomenological study 3. Grounded theory study 4. Ethnographic study 5. Case study	1. Researcher extracts themesfrom narratives of one or more individuals 2. Researchers study severalindividuals with shared experiences to analyze a phenomenon of interest 3. Researcher extracts datafrom interviews of ∼20–60 individuals and uses systematic coding to develop a unified theoretical explanation 4. Researcher extracts themesby describing and interpreting patterns of a shared culture of group 5. Researchers conduct anin-depth analysis of one or multiple cases	1. Allows for in-depth exploration/resource intensive, potential observer bias 2. Provides a deep understanding of phenomenonexperienced by multiple individuals/resources intensive, potential observer bias 3. Systematic approach to data analysis/exhaustiveprocess, potential observer bias 4. Development of a complex, exhaustive description of a culture of group/resource intensive, potential observer bias 5. Greater depth of analysis/limited generalizability
Mixed methods design	Uses one or more quantitative and qualitative design	Researchers collect, analyze, and integrate quantitative and qualitative data	Allows for triangulation of data, which counteracts disadvantages of individual designs/may be difficult to interpret if there are conflicting outcomes

so, first I identified national agencies and STEM governing bodies comprised of experts specifically charged with identifying criteria indicative of high-quality STEM research and study design. This was accomplished by searching documents in the Center for the Advancement of Informal Science Education repository (informalscience.org) and by contacting three distinguished ISE researchers (two university-based ISE researchers and one research director at a large ISE institution) for recommended sources of information. Through this process, seven agencies were identified as potential sources of information: (1) National Research Council; (2) United States Department of Education; (3)

National Science Foundation; (4) Institute of Learning Innovation; (5) Afterschool Alliance; (6) The PEAR Institute: Partnerships in Education and Resilience; and (7) What Works Clearinghouse. Second, I used Google Scholar and informalscience.org to identify and extract publications from these agencies with a specific focus on documents that provide research-based recommendations on how to assess research design and evidence of impact from studies and evaluations of STEM participants. Accordingly, the criteria that I identified to develop the STEM Research Design Rubric and STEM Impact Rubric were based on a synthesis of recommendations from published reports from multiple governing bodies. Importantly, these criteria stemmed from the counsel of leading ISE researchers, statisticians, and policymakers based on our current knowledge of best practices for research design and assessment. Thus, to ensure content validity, I adopted researchbased recommendations from ISE stakeholders because these experts specifically considered the unique characteristics of ISE programs when making recommendations (e.g., Institute for Learning Innovation, 2007; U.S. Department of Education, 2007; Friedman, 2008; National Research Council, 2009, 2015), and I used the theories of impact analysis and triangulation to inform this process. Lastly, based on these recommendations, I developed rubrics that researchers, STEM practitioners, and other stakeholders can use to assess research design and whether a research study or program evaluation has provided sufficient evidence to support a claim made about a particular STEM outcome.

During the design of these rubrics, I constructed a rating scale for the STEM Research Design Rubric and another for the STEM Impact Rubric. For the STEM Research Design Rubric, the rating scale was divided into four different levels: (1) a rating of one was indicative of a study or evaluation in which there was a *weak research design*; (2) a rating of two was indicative of a study or evaluation in which there was an *adequate research design*; (3) a rating of three was indicative of a study or evaluation in which there was a *strong research design*; and (4) a rating of four was indicative of a study or evaluation in which there was an *exemplary research design*. For the STEM Impact Rubric, the rating scale was also divided into four levels: (1) a rating of one was indicative of a study or evaluation in which there was *little or no evidence* of impact; (2) a rating of two was indicative of a study or evaluation in which there was *moderate evidence* of impact; (3) a rating of three was indicative of a study or evaluation in which there was *strong evidence* of impact; and (4) a rating of four was indicative of a study or evaluation in which there was *exemplary evidence* of impact. A four-point rubric was applied because this rating scale is considered the gold standard in rubric design (Phillip, 2002) and is commonly applied by STEM stakeholders in a myriad of contexts (e.g., What Works Clearinghouse, 2008; Singer et al., 2012; The PEAR Institute: Partnerships in Education and Resilience, 2017).

After designing prototypes of the STEM Impact Rubric, I recruited eight ISE stakeholders to review the STEM Impact Rubric and to participate in a focus group. All eight participants are ISE educators; five are active STEM researchers with a PhD; three hold a Masters in a STEM-related field. Of the eight focus group participants, one is a director of research at one of the largest informal science education institutions in the world; two are postdoctoral fellows both with a background in informal science education and mixed methods research. The remaining five participants are program managers at various ISE institutions; three hold Master's in museum science education, one of these three has a background in statistics and two in anthropology. The other two participants are directors of a high school science research mentoring program; both have PhDs in Biology. All eight focus group members have published ISE research, and in addition to their status as STEM stakeholders, are also intended users of the rubrics developed in this study.

The focus group members were provided with rubrics and asked to conduct an initial review and to test out the rubrics on a study or evaluation on the STEM outcomes of ISE participants. Each volunteer was given 1 week to review the document, and then invited to a focus group meeting to provide their feedback. During the focus group, I conducted a semi-structured group interview that included the following questions for discussion: (1) What are your impressions of the rubrics? (2) Do you think that this tool is user-friendly for ISE stakeholders (researchers, evaluators, practitioners)? Why or why not? (3) What changes or tweaks would you make so that these rubrics are more userfriendly? The focus group discussion was also guided by the theories of impact analysis and triangulation as the focus group participants used the hierarchy of study designs and mixed methods triangulation as a lens for assessing and practicing the rubrics (Denzin, 1970; Mohr, 1995; Ammenwerth et al., 2003; U.S. Department of Education, 2007; Flick, 2018a,b). While the focus group members did not have overlap with the experts who helped inform the initial rubric design, their feedback was important for improving the categories and metrics used for each rubric. Based on the recommendations of the focus group, I revised the rubrics and next tested the tools for reliability and validity.

To test for reliability, I worked with a graduate research assistant and together we used the STEM Research Design Rubric to rate 25 ISE studies (i.e., a combination of peer-reviewed studies, program evaluations, and conference proceedings) independently. Additionally, we used the STEM Impact Rubric to rate 47 outcomes independently. To assess consistency across raters (reliability), I calculated inter-rater agreement using both percent agreement (Lombard et al., 2002) and Cohen's kappa (Cohen, 1968). Percent agreement for the STEM Research Design Rubric was 92.0% and for the STEM Impact Rubric 80.9%. Cohen's kappa (k) for the STEM Research Design Rubric was 0.89, and for the STEM Impact Rubric 0.74. These measures were indicative of substantial

agreement (Cohen, 1968). Lastly, to increase content specifically charged with identifying criteria indicative of high-quality STEM research and from focus group participants consisting of experienced ISE STEM researchers and practitioners.

Study Title: _____

Authors:

Use this worksheet to help you complete the STEM Research Design Rubric for quantitative studies.

1A. Does the study design include a *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group (e.g.: pre-test/post-test design with comparison; post-test only design with comparison)? If yes, go to 1B. If no, go to 2A.

1B. Is the study design grounded in a theoretical framework? If yes, go to 1C. If no, go to 2A.

1C. Does the sample size consist of a minimum of 50 subjects per group? If yes, this is an *exemplary research design* (rubric score:
4). If no, go to 2A.

2A. Does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasiexperimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g. pre-test/post-test, post-test only, and/or time series design(s) without comparison)? If yes, go to 2B. If no, go to 3A.

2B. Is the study design grounded in a theoretical framework? If yes, go to 2C. If no, go to 3A.

2C. Does the sample size consist of a minimum of 40 subjects? If yes, this is a *strong research design* (rubric score: 3). If no, go to 3A.

3A. Does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasiexperimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g. pre-test/post-test, post-test only, and/or time series design(s) without comparison)? If yes, go to 3B. If no, go to 4.

3B. Is the study design grounded in a theoretical framework? If yes, go to 3C. If no, go to 3C.

3C. Does the sample size consist of a minimum of 25 subjects? If yes, this is an *adequate research design* (rubric score: 2). If no, go to 4.

4. This study is a *poor research design* (rubric score: 1).

FIGURE 1 | STEM Research Design Rubric worksheet for quantitative studies.

validity of our rubric, we incorporated triangulation methods (Creswell and Miller, 2000; Creswell and Poth, 2018). Because the most valid assessments stem from "the collective judgment of recognized experts in that field" (Baer and McKool, 2014, p. 82), during triangulation, I extracted and synthesized data from multiple sources including recommendations from national agencies and STEM governing bodies comprised of experts

RESULTS Focus Group Results

Four key recommendations were made by the focus group. First, the focus group recommended that I design six distinct rubrics separated into two categories: (1) three *research design* rubrics, one for quantitative studies, one for qualitative studies, and one for mixed methods studies, and (2) three evidence of *impact* rubrics, one for quantitative studies, one for qualitative studies, and one for mixed methods studies. Initially, I provided the focus group with three rubrics that combined *research design* and evidence of *impact* together. However, the focus group unanimously agreed that *research design* and evidence of *impact* are distinct criteria warranting separate analysis. Thus, based on this recommendation, in
the final iteration, I developed six distinct rubrics to assess quantitative, qualitative, and mixed methods studies: three STEM Research Design Rubrics and three STEM Impact Rubrics. Second, to make the rubrics more user friendly, the focus group recommended that instead of presenting the four levels of evidence vertically, I should present these criteria horizontally. This recommendation was adopted in the final draft. Third, because some researchers might have difficulties using each rubric as a stand-alone tool, the focus group suggested that I develop a worksheet with a key to help guide researchers, practitioners, and evaluators through the assessment process when using each rubric. This recommendation was also adopted (see **Figures 1–6**). Lastly, there was some

debate between focus group members on whether a research design *not* grounded in theory should automatically be rated as an *adequate research design* (a score of 2 out of 4). The focus group did not reach consensus with respect to this question and instead recommended that if I retain this rating in the final draft, then I should also emphasize in the Discussion that these rubrics should be viewed as heuristics for researchers, practitioners, and evaluators and that in some cases, depending on the goals of a particular study, it might be appropriate to alter criteria.

A STEM Research Design Rubric for Quantitative Studies

A STEM Research Design Rubric for quantitative studies (**Table 2**) was designed in accordance with research-based recommendations from ISE experts (Institute for Learning Innovation, 2007; U.S. Department of Education, 2007; Friedman, 2008; National Research Council, 2009, 2015). In alignment with the theory of impact analysis (Mohr, 1995), I found that a quantitative study or evaluation with evidence of an *exemplary research design* is one that includes a control (comparison) and treatment group (program participants) selected randomly (experimental) (U.S. Department of Education, 2007; National Research Council, 2013, 2015). Alternatively, because of the unique nature of ISE programs, I found that it might be more appropriate to reference a comparison group that is not a strict control (National Research Council, 2009). Therefore, I also found that a quantitative study or evaluation that provides evidence of an *exemplary research design* may alternatively include a well-matched comparison group (control) and treatment group (program participants) selected non-randomly (quasi-experimental) (U.S. Department of Education, 2007; National Research Council, 2013, 2015). ISE experts also recommend that a quantitative study or evaluation with evidence of *exemplary research design* should be grounded in a theoretical framework (Institute for Learning Innovation, 2007) and consist of a sample size of fifty or more subjects per group (Institute for Learning Innovation, 2007; Diamond et al., 2016). Based on additional recommendations from ISE experts, I also developed criteria indicative of

Study Title: _____

_____ Authors: _____

_____

Use this worksheet to help you complete the STEM Impact Rubric for quantitative studies. Before completing the rubric, you need to identify the STEM outcome that you are assessing in the study (e.g. STEM career interest). Because some studies use multiple criteria to assess an outcome, you may need to repeat this process for multiple instruments and then average the results at the end.

List all criteria used to assess the outcome of interest (e.g. survey question 1, survey question 2, open-ended question 1, open-ended question 2, etc.)

1. Is there a statistically significant difference between comparison (control) and treatment (program participants)? If yes, this is an example *exemplary evidence* of impact (rubric score: 4). If no (or if the study did not include a treatment and comparison group), then go to 2.

2. Is there a statistically significant difference between pre- and post-survey OR did a minimum of 75% of participants indicate higher than median (e.g. very likely; likely) outcome on post-program survey. If yes, this is an example of *strong evidence* of impact (rubric score: 3). If no, then go to 3.

3. Did 40-75% of participants indicate a higher than median (e.g. very likely; likely) outcome on postprogram surveys OR was a STEM outcome maintained in comparisons of pre- and post-surveys? If yes, then this is an example of *moderate evidence* of impact (rubric score: 2). If no, then go to 4.

4. This is an example of *little or no evidence* of impact (rubric score: 1).

Repeat this process for each criterion and then average the results to attain your STEM Impact Rubric score.

FIGURE 2 | STEM Impact Rubric worksheet for quantitative studies.

strong, adequate, and *weak* research design for quantitative studies and evaluations (**Table 2**). Lastly, as recommended by the focus group, I developed a STEM Research Design Worksheet for quantitative studies

Study Title: _____

Authors: _____

Use this worksheet to help you complete the STEM Research Design Rubric for qualitative studies.

1A. Does the study design include a *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study**?** If yes, go to 1B. If no, go to 2A.

1B. Is the study design grounded in a theoretical framework? If yes, go to 1C. If no, go to 2A.

1C. Does the sample size consist of a minimum of 20 subjects per group? If yes, this is an *exemplary research design* (rubric score: 4). If no, go to 2A.

2A. Does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group*? If yes, go to 2B. If no, go to 3A.

2B. Is the study design grounded in a theoretical framework? If yes, go to 2C. If no, go to 3A.

2C. Does the sample size consist of a minimum of 15 subjects? If yes, this is a *strong research design* (rubric score: 3). If no, go to 3A.

3A. Does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group*? If yes, go to 3B. If no, go to 4.

3B. Is the study design grounded in a theoretical framework? If yes, go to 3C. If no, go to 3C.

3C. Does the sample size consist of a minimum of 10 subjects? If yes, this is an *adequate research design* (rubric score: 2). If no, go to 4.

4. This study is a *poor research design* (rubric score: 1).

FIGURE 3 | STEM Research Design Rubric worksheet for qualitative studies.

as a key to guide researchers, practitioners, or evaluators through the process of assessing research design quality (**Figure 1**).

A STEM Impact Rubric for Quantitative Studies

A STEM Impact Rubric for quantitative studies (**Table 3**) was designed in accordance with research-based recommendations from ISE experts (Institute for Learning Innovation, 2007; U.S. Department of Education, 2007; Friedman, 2008; National Research Council, 2009, 2015). Based on these recommendations, a study outcome indicative of *exemplary evidence* of impact needs to report a statistically significant difference between comparison and treatment groups (U.S. Department of Education, 2007; National Research Council, 2013, 2015). Therefore, a study that measured STEM career interest (or some other outcome of interest) and found that program participants exhibited significantly higher interest than a well-matched comparison group would be an example of an outcome indicative of *exemplary evidence* of impact. Based on additional recommendations from ISE experts, I also developed criteria indicative of *strong, moderate,* and *little or no evidence* of impact for quantitative studies and evaluations (**Table 3**). Lastly, as recommended by the focus group, I developed a STEM Impact Worksheet for quantitative studies as a key to guide researchers, practitioners, or evaluators through the process of assessing the impact of specific outcomes (**Figure 2**).

A STEM Research Design Rubric for Qualitative Studies

A STEM Research Design Rubric for qualitative studies (**Table 4**) was designed in accordance with research-based recommendations from ISE experts (Institute for Learning Innovation, 2007; U.S. Department of Education, 2007; Friedman, 2008; National Research Council, 2009, 2015). In alignment with the theory of impact analysis (Mohr, 1995), I found that a qualitative study or evaluation (e.g., narrative study; phenomenological study; grounded theory study; ethnographic study) that provides evidence of *exemplary research design* is one that includes a control (well-matched comparison) and treatment group (program participants) selected randomly (experimental) or non-randomly (quasi-experimental) (U.S. Department of Education, 2007;

National Research Council, 2013, 2015). ISE experts also recommend that a qualitative study or evaluation with evidence of *exemplary research design* should be grounded in a theoretical framework (Institute for Learning Innovation, 2007) and consist of a sample size of twenty or more subjects per group (Creswell and Poth, 2018). Based on additional recommendations from ISE experts, I also developed criteria indicative of *strong, adequate,* and *weak* research design for qualitative studies and evaluations (**Table 4**). Lastly, as recommended by the focus group, I developed a STEM Research Design Worksheet for qualitative studies as a key to guide researchers, practitioners, or evaluators through the process of assessing research design quality (**Figure 3**).

A STEM Impact Rubric for Qualitative Studies

A STEM Impact Rubric for qualitative studies (**Table 5**) was designed in accordance with research-based recommendations from ISE experts (Institute for Learning Innovation, 2007; U.S. Department of Education, 2007; Friedman, 2008; National Research Council, 2009, 2015). Based on these recommendations, a qualitative study indicative of *exemplary evidence* of impact may demonstrate this in one of two ways: (1) a study or evaluation in which researchers identify the outcome of interest in the treatment but not the control as an emerging theme or (2) a study or evaluation in which a minimum of 75% of data (e.g., interview data, response to open-ended question) is indicative of the outcome of interest in the treatment but not the control; this difference needs to be statistically significant. Based on additional recommendations from ISE experts, I also developed criteria indicative of *strong, moderate,* and *little or no evidence* of impact for qualitative studies and evaluations (**Table 5**). Lastly,
as recommended by the focus group, I developed a STEM Impact Worksheet for qualitative studies as a key to guide researchers, practitioners, or evaluators through the process of assessing the impact of specific outcomes (**Figure 4**).

Study _____ Title:

Authors: _____

Use this worksheet to help you complete the STEM Impact Rubric for qualitative studies. Before completing the rubric, you need to identify the STEM outcome that you are assessing in the study (e.g. STEM career awareness). Because some studies use multiple criteria to assess an outcome, you may need to repeat this process for multiple instruments and then average the results at the end.

List all criteria used to assess the outcome of interest (e.g. open-ended question 1, open-ended question 2, emerging themes, etc.)

1. Was the outcome of interest identified as an emerging theme during qualitative analyses for the treatment but not comparison group OR did a minimum of 75% of the data (e.g. responses to interview data; responses to open-ended questions) indicate that the program impacted an outcome of interest in the treatment but not control, and was there a statistically significant difference between these two groups? If yes, this is an example of *exemplary evidence* of impact (rubric score: 4). If no (or if the study did not include a treatment and comparison group), then go to 2.
2. Was the outcome of interest identified as an emerging theme during qualitative analyses OR did a minimum of 75% of data (e.g. interview data; responses to open-ended questions) indicate that the program impacted the STEM outcome of interest? If yes, this is an example of *strong evidence* of impact (rubric score: 3). If no, then go to 3.
3. Did 40-75% of the data (e.g. interview data; responses to open-ended questions) indicate that the program impacted STEM outcome of interest? If yes, this is an example of *moderate evidence* of impact (rubric score: 2). If no, then go to 4.
4. This is an example of *little or no evidence* of impact (rubric score: 1).

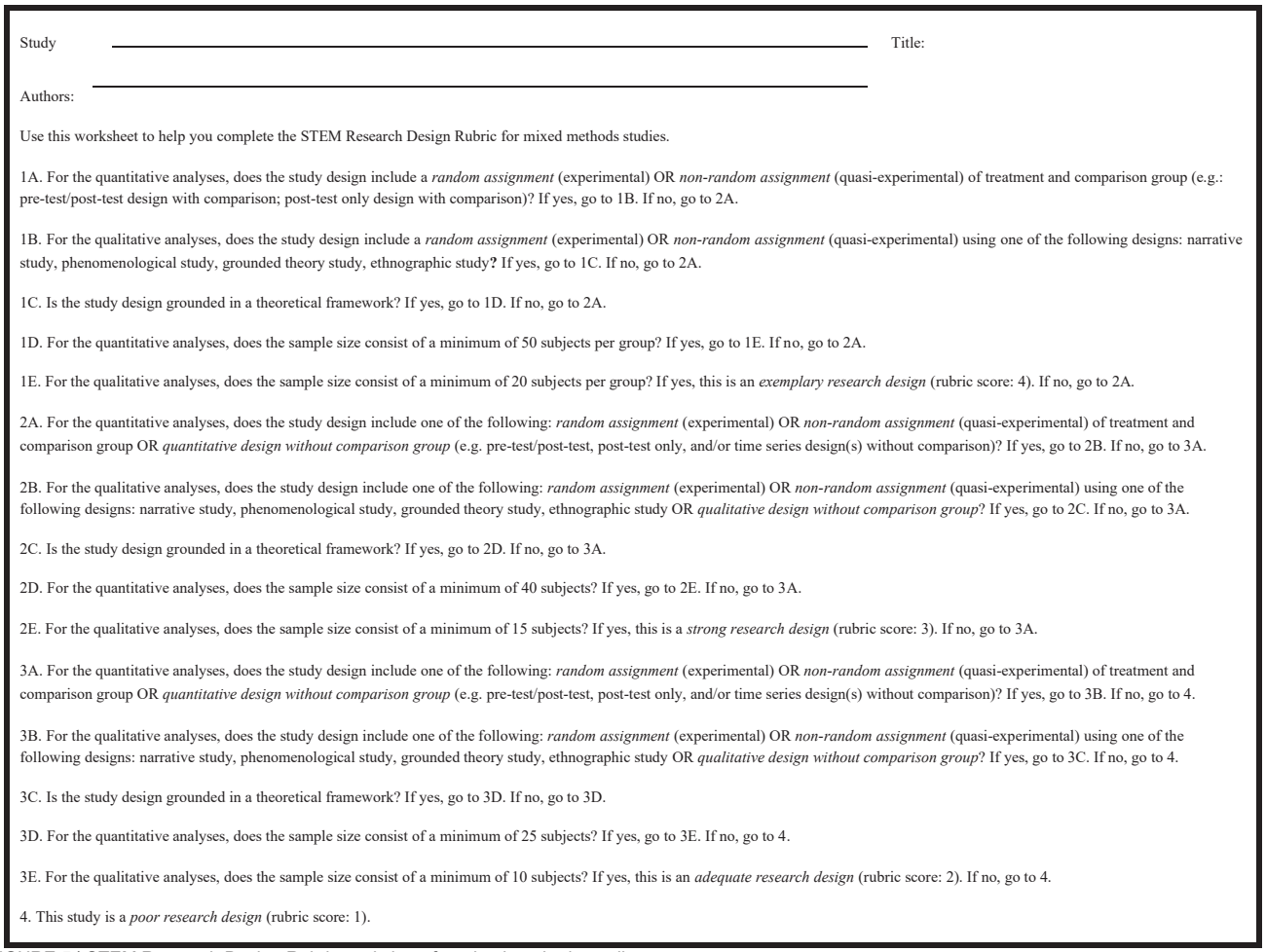
Repeat this process for each criterion and then average the results to attain your STEM Impact Rubric score.

FIGURE 4 | STEM Impact Rubric worksheet for qualitative studies.

A STEM Research Design Rubric for Mixed Methods Studies

A STEM Research Design Rubric for mixed methods studies (**Table 6**) was designed in accordance with research-based recommendations from ISE experts (Institute for Learning Innovation, 2007; U.S. Department of Education, 2007; Friedman, 2008; National Research Council, 2009, 2015). In alignment with the theories of impact analysis (Mohr, 1995) and triangulation (Campbell and Fiske, 1959), I found that a mixed methods study (i.e., a study or evaluation that incorporates both quantitative and qualitative analyses) provides exemplary evidence of quality research design if the study or evaluation meets the benchmarks for exemplary evidence described for quantitative (**Table 2**) and qualitative (**Table 4**) analyses. In terms of evidence of *exemplary research design*, the study or evaluation must also be grounded in a theoretical framework (Institute for Learning

Innovation, 2007) and the sample size needs to be comprised of fifty or more subjects per group for the quantitative analysis and twenty or more subjects per group for the qualitative analysis (Institute for Learning Innovation, 2007; Diamond et al., 2016; Creswell and Poth, 2018). Based on additional recommendations from ISE experts, I also developed criteria indicative of *strong, adequate,* and *weak* research design for mixed methods studies and evaluations (**Table 6**). Lastly, as recommended by the focus group, I developed a STEM Research Design Worksheet for mixed methods studies as a key to guide researchers, practitioners, or evaluators through the process of assessing research design quality (**Figure 5**).



Study _____ Title:

Authors: _____

Use this worksheet to help you complete the STEM Research Design Rubric for mixed methods studies.

1A. For the quantitative analyses, does the study design include a *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group (e.g.: pre-test/post-test design with comparison; post-test only design with comparison)? If yes, go to 1B. If no, go to 2A.

1B. For the qualitative analyses, does the study design include a *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study**?** If yes, go to 1C. If no, go to 2A.

1C. Is the study design grounded in a theoretical framework? If yes, go to 1D. If no, go to 2A.

1D. For the quantitative analyses, does the sample size consist of a minimum of 50 subjects per group? If yes, go to 1E. If no, go to 2A.

1E. For the qualitative analyses, does the sample size consist of a minimum of 20 subjects per group? If yes, this is an *exemplary research design* (rubric score: 4). If no, go to 2A.

2A. For the quantitative analyses, does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g. pre-test/post-test, post-test only, and/or time series design(s) without comparison)? If yes, go to 2B. If no, go to 3A.

2B. For the qualitative analyses, does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group*? If yes, go to 2C. If no, go to 3A.

2C. Is the study design grounded in a theoretical framework? If yes, go to 2D. If no, go to 3A.

2D. For the quantitative analyses, does the sample size consist of a minimum of 40 subjects? If yes, go to 2E. If no, go to 3A.

2E. For the qualitative analyses, does the sample size consist of a minimum of 15 subjects? If yes, this is a *strong research design* (rubric score: 3). If no, go to 3A.

3A. For the quantitative analyses, does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g. pre-test/post-test, post-test only, and/or time series design(s) without comparison)? If yes, go to 3B. If no, go to 4.

3B. For the qualitative analyses, does the study design include one of the following: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group*? If yes, go to 3C. If no, go to 4.

3C. Is the study design grounded in a theoretical framework? If yes, go to 3D. If no, go to 3D.

3D. For the quantitative analyses, does the sample size consist of a minimum of 25 subjects? If yes, go to 3E. If no, go to 4.

3E. For the qualitative analyses, does the sample size consist of a minimum of 10 subjects? If yes, this is an *adequate research design* (rubric score: 2). If no, go to 4.

4. This study is a *poor research design* (rubric score: 1).

FIGURE 5 | STEM Research Design Rubric worksheet for mixed methods studies.

A STEM Impact Rubric for Mixed Methods Studies

Lastly, a STEM Impact Rubric for mixed methods studies (**Table 7**) was designed in accordance with research-based recommendations from ISE experts (Institute for Learning Innovation, 2007; U.S. Department of Education, 2007; Friedman, 2008; National Research Council, 2009, 2015). In terms of *exemplary evidence* of impact, the mixed methods study or evaluation must demonstrate the following: (1) a quantitative analysis in which there is a significant difference between the comparison (well-matched control) and treatment (program participants) groups and (2) a qualitative analysis in which researchers identify the outcome of interest as an emerging theme in the treatment but not in the comparison (well-matched control) group or a qualitative analysis in which a

minimum of 75% of data (e.g., interview data, response to openended question) is indicative of the outcome of interest in the treatment but not the comparison (well-matched control) group; this difference needs to be statistically significant. Based on additional recommendations from ISE experts, I also developed criteria indicative of *strong, moderate,* and *little or no evidence* of impact for mixed methods studies and evaluations (**Table 7**). Lastly, as recommended by the focus group, I developed a STEM Impact Worksheet for mixed methods studies as a key to guide researchers, practitioners, or evaluators through the process of assessing the impact of specific outcomes (**Figure 6**).

Case Studies

In the next section, I present case studies demonstrating how the STEM Research Design and STEM Impact Rubrics can be used to assess research design quality and measure evidence of impact for specific outcomes. In accordance with recommendations from the Academic Competitiveness Council (U.S. Department of Education, 2007), I identified three select studies that measure one of the following: STEM major or STEM career awareness, interest, or engagement. The first case study provides an example of how to use the rubrics to assess a quantitative study; the second case study provides an example of how to assess a qualitative study; the third case study provides an example of how to assess a mixed methods study. While these case studies focus specifically on awareness, interest, and engagement, a STEM practitioner may use these rubrics to assess any outcome of interest and to compare studies that assess comparable outcomes.

Case Study 1: Stanford Medical Youth Science Program

The Stanford Medical Youth Science Program is a biomedical pipeline program for high school students. The goal of this program is to diversify participation in the health professions (Winkleby, 2007). This 5 week residential summer program includes classroom-based workshops, anatomy and pathology practicums, hospital field placements, research projects, and college readiness advisement. In 2009, Winkleby et al. (2009) published a quantitative study of the STEM outcomes of program participants. Two specific outcomes measured in this study were whether alumni of this program (1) majored in a STEM discipline or (2) engaged in a STEM career. I used the STEM Research Design Rubric (**Table 2**) to assess quality of research design and the STEM Impact Rubric (**Table 3**) to measure evidence of outcome (engagement in a STEM major; engagement in a STEM career). This process is also depicted graphically in **Figure 7A**.

In terms of research design, first I examined evidence of *exemplary research design* (**Table 2**, column 1). Since the study did not include either a random or non-random comparison group, I moved on to the second column: *strong research design*. First, I checked the first bullet point in column two. Since the study was a *quantitative design without a comparison group* (posttest only), it met the first criterion for *strong research design*. Next, I checked the second bullet point in column two. Since the study was grounded in two theoretical frameworks: (1) Cognitive Apprenticeship (Collins et al., 1991) and (2) Situated Learning (Lave and Wenger, 1991), it met the

Study _____ Title:

Authors: _____

Use this worksheet to help you complete the STEM Impact Rubric for mixed methods studies. Before completing the rubric, you need to identify the STEM outcome that you are assessing in the study (e.g. STEM career interest). Because some studies use multiple criteria to assess an outcome, you may need to repeat this process for multiple instruments and then average the results at the end.

List all criteria used to assess the outcome of interest (e.g. survey question 1, survey question 2, open-ended question 1, open-ended question 2, etc.)

1. Is the outcome of interest assessed quantitatively rather than qualitatively? If yes, go to 2. If no, go to 5.

2. Is there a statistically significant difference between comparison (control) and treatment (program participants)? If yes, this is an example *exemplary evidence* of impact (rubric score: 4). If no (or if the study did not include a treatment and comparison group), then go to 3.

3. Is there a statistically significant difference between pre- and post-survey OR did a minimum of 75% of participants indicate higher than median (e.g. very likely; likely) outcome on post-program survey. If yes, this is an example of *strong evidence* of impact (rubric score: 3). If no, then go to 4.

4. Did 40-75% of participants indicate a higher than median (e.g. very likely; likely) outcome on post-program surveys OR was a STEM outcome maintained in comparisons of pre- and post-surveys? If yes, then this is an example of *moderate evidence* of impact (rubric score: 2). If no, then go to 8.

5. Was the outcome of interest identified as an emerging theme during qualitative analyses for the treatment but not comparison group OR did a minimum of 75% of the data (e.g. responses to interview data; responses to open-ended questions) indicate that the program impacted an outcome of interest in the treatment but not control, and was there a statistically significant difference between these two groups? If yes, this is an example of *exemplary evidence* of impact (rubric score: 4). If no (or if the study did not include a treatment and comparison group), then go to 6.

6. Was the outcome of interest identified as an emerging theme during qualitative analyses OR did a minimum of 75% of data (e.g. interview data; responses to openended questions) indicate that the program impacted the STEM outcome of interest? If yes, this is an example of *strong evidence* of impact (rubric score: 3). If no, then go to 7.

7. Did 40-75% of the data (e.g. interview data; responses to open-ended questions) indicate that the program impacted STEM outcome of interest? If yes, this is an example of *moderate evidence* of impact (rubric score: 2). If no, then go to 8.

8. This is an example of *little or no evidence* of impact (rubric score: 1).

FIGURE 6 | STEM Impact Rubric worksheet for mixed methods studies.

second criterion for *strong research design*. Lastly, I checked the third bullet point in column two. Since the sample size well-exceeded the minimum of 40, it met the third criterion for *strong research design*. Thus, based on the STEM Research Design Rubric, this study was classified as a *strong research design*.

In terms of evidence of outcome (engagement in a STEM major or STEM career), first I assessed criteria for *exemplary evidence* (**Table 3**, column 1). Since there was no statistical comparison between program participants and a comparison group, I moved on to the second column (*strong evidence*). The criteria for *strong evidence* of impact included two possible outcomes: (1) statistically significant difference between pre- and post-survey or (2) minimum of 75% of participants indicate higher than median (e.g., very likely; likely) outcome on postprogram survey. Neither of these outcomes was found for STEM major engagement or STEM career engagement. Next, I moved on to the third column (*moderate evidence* of impact). A survey of alumni of the Stanford Medical Youth Science Program indicated that 57.1% were engaged in a STEM major, which met the criteria for *moderate evidence* (a favorable response by 40–75% of participants). However, only 33.1% of alumni were engaged in a STEM career; this met the criteria for *little or no evidence* (**Table 3**, column 4). Overall, this study provided evidence of a *strong research design* and *moderate evidence* of impact, in terms of engagement in a STEM major, and *little or no evidence* of impact in terms of STEM career engagement.

Case Study 2: The Source (Game Changer Chicago Design Lab, University of Chicago)

The Source is a 5 week summer program that uses alternative reality games for teaching engineering concepts to high school students (Gilliam et al., 2017). The weekly program includes 1 day of online activities off-campus and 3 to 4 days of oncampus activities focused on workshops in different STEM subject areas. Gilliam et al. (2017) published a qualitative study describing the STEM outcomes of high school participants of this program. To assess the quality of the research design of this study and to measure evidence of impact (in this case STEM career awareness), I used both the STEM Research Design Rubric for qualitative studies (**Table 4**) and the STEM Impact Rubric for qualitative studies (**Table 5**) as described below. This process is also depicted graphically in **Figure 7B**.

In terms of research design, first I examined evidence of *exemplary research design* (**Table 4**, column 1). Since the study did not include either a random or nonrandom comparison group, I moved on to the second column: *strong research design*. First, I checked the first bullet point in column two. Since the study was a *qualitative design without comparison group*, it met the first criterion for *strong research design*. Next, I checked the second bullet point in column two. Since the study was grounded in multiple theoretical frameworks, including Situated Learning Theory (Lave and Wenger, 1991), it met the second criterion for *strong research design*. Lastly, I checked the third bullet point in column two. Since 43 students were interviewed, the sample size well-exceeded the minimum of 15 subjects; thus, it met the third criterion for *strong research design*.

TABLE 2 | STEM research design rubric for quantitative studies.

STEM Research Design Rubric (Quantitative studies)

(4) Exemplary research design	(3) Strong research design	(2) Adequate research design	(1) Weak research design
• *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group (e.g., pre-test/post-test design with comparison; post-test only design with comparison) • study grounded in a theoreticalframework • minimum of 50 subjects per group	• *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g., pre-test/post-test, post-test only, and/or time series design(s) without comparison) • study grounded in a theoreticalframework • minimum of 40 subjects	• *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g., pre-test/post-test, post-test only, and/or time series design(s) without comparison) • study may or may not be groundedin a theoretical framework • minimum of 25 subjects	• *random assignment* (experimental) OR *non-random assignment* (quasiexperimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g., pre-test/post-test, post-test only, and/or time series design(s) without comparison) • study may or may not be groundedin a theoretical framework • <25 subjects or not reported

Helpful shortcuts:

x If there is no comparison group, the starting value is 3

x If there is no theoretical framework, the starting value is 2

x The sample size represents the number of subjects analyzed in the study, not the number of program participants

Rubric Score: _____

Note: In order to receive a rubric score (4, 3, 2, or 1), a study must meet the criteria of all three items in a given column.

Therefore, based on the STEM Research Design Rubric for qualitative studies, this study was classified as a *strong research design*.

In terms of evidence of outcome (in this case, STEM career awareness), first I used the STEM Impact Rubric for qualitative studies to assess criteria for *exemplary evidence* (**Table 5**, column 1). Since there was no comparison group in this analysis, I moved on to the second column and tested for *strong evidence* of impact. One criterion indicative of *strong evidence* of impact is if the *researchers identify outcome of interest as an emerging theme during qualitative analyses*. A theme that emerged from this study was "Mentoring and Exposure to STEM Professionals," which provided evidence that participants became more aware of STEM career opportunities as a result of their experiences in the program. Thus, this study provides *strong evidence* of impact. Overall, this study provided evidence of a *strong research design* and *strong evidence* of outcome, in this case, STEM career awareness.

Case Study 3: The Lang Science Program (American Museum of Natural History)

The Lang Science Program is a comprehensive 7 year program for middle school and high school students facilitated by the American Museum of Natural History. The program takes place

TABLE 3 | STEM impact rubric for quantitative studies.

Title of study or evaluation:
STEM outcome under review:
List all method(s) used to assess the STEM outcome under review (e.g., survey question 1, survey question 2, etc.):
1. _____ 2. _____ 3. _____ 4. _____ 5. _____ 6. _____ 7. _____ 8. _____ 9. _____ 10. _____

Directions:
1. Identify the STEM outcome under review (e.g., increased STEM career interest)
2. List all criteria used to assess the STEM outcome under review. For example, if there were four survey questions that evaluated STEM careerinterest, list all four questions (e.g., survey question 1, survey question 2, etc.)
3. Use the rubric to evaluate each criterion used to assess the STEM outcome under review starting from left (exemplary evidence) to right (little orno evidence)
4. Calculate the average rubric score by dividing the sum of all rubric scores by the number of criteria used to assess STEM outcomes

STEM Impact Rubric (Quantitative studies)

(4) Exemplary evidence	(3) Strong evidence	(2) Moderate evidence	(1) Little or No evidence
• statistically significant differencebetween comparison (control) and treatment (program participants) groups • If there is no comparison group,starting value is 3	• statistically significant difference between pre- and post-survey OR minimum of 75% of participants indicate higher than median (e.g., very likely; likely) outcome on post-program survey	• 40–75% of participants indicate higher than median (e.g., very likely; likely) outcome on post-program surveys OR STEM outcome is maintained in comparisons of preand post-surveys (i.e., no significant difference between preand post-assessments)	• less than 40% of participants indicate higher than median (e.g., very likely; likely) outcome on post-program surveys OR STEM outcome significantly decreases in comparisons of pre- and post-surveys OR outcomes of participants are the same or significantly lower than outcomes of comparison group

ore (sum of all rubric scores ÷ : criteria used to assess STEM outcomes): _____

TABLE 4 | STEM research design rubric for qualitative studies.

STEM Research Design Rubric (Qualitative studies)

(4) Exemplary research design	(3) Strong research design	(2) Adequate research design	(1) Weak research design
• *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study • study grounded in a theoreticalframework • minimum of 20 subjects per group	• *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group* • study grounded in a theoreticalframework • minimum of 15 subjects	• *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group* • study may or may not be groundedin a theoretical framework • minimum of 10 subjects	• *random assignment* (experimental) OR *non-random assignment* (quasiexperimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group* • study may or may not be groundedin a theoretical framework • <10 subjects or not reported

Helpful shortcuts:
x If there is no comparison group, the starting value is 3
x If there is no theoretical framework, the starting value is 2
x The sample size represents the number of subjects analyzed in the study, not the n participants

Rubric Score: _____

Note: In order to receive a rubric score (4, 3, 2, or 1), a study must meet the criteria of all three items in a given column.

on alternating Saturdays during the academic year and for 3 weeks during the summer for a minimum of 165 contact hours per year. The program centers on the research disciplines of the Museum—the biological sciences, Earth and planetary sciences, and anthropological sciences. As students transition from middle school to high school, they increasingly engage in authentic research projects alongside scientists and educators as well as career and college readiness workshops (Habig et al., 2018). A mixed methods study of the Lang Program centered on STEM major engagement and STEM career awareness (Habig et al., 2018). To assess the quality of the research design of this study and to

measure evidence of impact, I used both the STEM Research Design Rubric for mixed methods studies (**Table 6**) and the STEM Impact Rubric for mixed methods studies (**Table 7**) as described below. This process is depicted graphically in **Figure 7C**.

qualitative analysis, this study did not meet the criteria for an *exemplary research design*. Therefore, I moved on to the second column: *strong research design*. Since the study met the criteria for the first two bullet points in column 2 (a quantitative analysis with *non-random assignment* (quasi-experimental) of treatment and comparison group; a *qualitative design without comparison group*), I next checked the third bullet point in column 2. Since the study was grounded in two theoretical frameworks: (1) communities of practice (Lave and Wenger, 1991) and (2) possible selves (Markus and Nurius, 1986); it met the third criterion for *strong research design*. Next, I checked bullet points five and six (sample sizes). Since the sample sizes exceeded the minimum of 40 subjects for quantitative analyses and 15 subjects for qualitative analyses, this study met the fifth and sixth criteria for *strong research design*. Therefore, based on the STEM

TABLE 5 | STEM impact rubric for qualitative studies.

Title of study or evaluation:
STEM outcome under review:
List all method(s) used to assess the STEM outcome under review (e.g., survey question 1, survey question 2, etc.):
1. _____ 2. _____ 3. _____ 4. _____ 5. _____ 6. _____ 7. _____ 8. _____ 9. _____ 10. _____

Directions:
1. Identify the STEM outcome under review (e.g., increased STEM career interest)
2. List all criteria used to assess the STEM outcome under review. For example, if there were four survey questions that evaluated STEMcareer interest, list all four questions (e.g., open-ended question 1, open-ended question 2, etc.)
3. Use the rubric to evaluate each criterion used to assess the STEM outcome under review starting from left (exemplary evidence) toright (little or no evidence)
4. Calculate the average rubric score by dividing the sum of all rubric scores by the number of criteria used to assess STEM outcomes

STEM Impact Rubric (Qualitative studies)

(4) Exemplary evidence	(3) Strong evidence	(2) Moderate evidence	(1) Little or No evidence
• researchers identify outcome ofinterest as an emerging theme during qualitative analyses for treatment but not comparison OR minimum of 75% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted an outcome of interest in the treatment but not control, and there was a statistically significant difference between these two groups • If there is no comparison group,starting value is 3	• researchers identify outcome of interest as an emerging theme during qualitative analyses OR minimum of 75% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted STEM outcome of interest	• 40–75% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted STEM outcome of interest	• less than 40% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted STEM outcome of interest OR anecdotal evidence of STEM outcomes (e.g., handful of participants quotes, but no systematic analysis) OR no difference in emerging themes between treatment and comparison groups

Average Rubric Score (sum of all rubric sc ÷ number of criteria used to asses: _____

In terms of research design, first I examined evidence of *exemplary research design* by assessing the first bullet point in column 1 (**Table 6**). Since the study design met the criteria of the first bullet point (a quantitative analysis that included a comparison group), I moved on to the second bullet point in column 1. Since there was no comparison group for the Research Design Rubric for mixed methods studies, this study was classified as a *strong research design*.

In terms of evidence of outcome, first I assessed the first outcome of interest: STEM major engagement. Four quantitative criteria were used to measure STEM major engagement: (1) percentage of STEM majors; (2) percentage of STEM majors compared to college students nationally; (3) percentage of STEM majors compared to college students in New York City; (4) percentage of STEM majors compared to students who attended specialized science, mathematics, and technology high schools. Since 80.3% of alumni engaged in a STEM major, this outcome was indicative of *strong evidence* of impact (**Table 7**, column 2). Since the percentage of STEM majors of the Lang program was significantly higher than college students nationally and locally, these two outcomes were indicative of *exemplary evidence* of impact (**Table 7**, column 1). However, since there was no significant difference in the percentage of STEM majors when comparing Lang alumni to alumni of specialized science, mathematics, and technology high schools, this last outcome was indicative of *little or no evidence* of impact. After averaging the four outcomes, the mean STEM

rubric score was 3 suggestive of *strong evidence* of impact overall. For the second outcome, STEM career awareness, a qualitative analysis was used to measure evidence of impact. Based on the STEM Impact Rubric (**Table 7**, column 2), one criterion indicative of *strong evidence* of impact is if the *researchers identify outcome of interest as an emerging theme during qualitative analyses*. A theme that emerged from this study was "Discovering Possible Selves," which provided evidence that participants were exposed to and became more aware of STEM career opportunities as a result of their experiences in the program. Thus, this outcome was indicative of *strong evidence* of impact. Overall, this study provided evidence of a *strong research design* and *strong evidence* of outcomes in terms of STEM

TABLE 6 | STEM research design rubric for mixed methods studies.

STEM Research Design Rubric (Mixed methods studies)			
(4) Exemplary research design	(3) Strong research design	(2) Adequate research design	(1) Weak research design
• quantitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group (e.g., pre-test/post-test design with comparison; post-test only design with comparison) • qualitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study • study grounded in a theoreticalframework • quantitative analyses: minimum of 50 subjects per group • qualitative analyses: minimum of 20subjects per group • Note: If there is no comparisongroup, starting value is 3	• quantitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g., pre-test/post-test, post-test only, and/or time series design(s) without comparison) • qualitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group* • study grounded in a theoreticalframework • quantitative analyses: minimum of 40 subjects per group • qualitative analyses: minimum of 15 subjects per group	• quantitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g., pre-test/post-test, post-test only, and/or time series design(s) without comparison) • qualitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasi-experimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group* • study may or may not be groundedin a theoretical framework • quantitative analyses: minimum of 25 subjects per group • qualitative analyses: minimum of 10 subjects per group	• quantitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasiexperimental) of treatment and comparison group OR *quantitative design without comparison group* (e.g., pre-test/post-test, post-test only, and/or time series design(s) without comparison) • qualitative analysis: *random assignment* (experimental) OR *non-random assignment* (quasiexperimental) using one of the following designs: narrative study, phenomenological study, grounded theory study, ethnographic study OR *qualitative design without comparison group* • study may or may not be groundedin a theoretical framework • quantitative analyses: < 25 subjects per group • qualitative analyses: < 10 subjects per group • sample size not reported Rubric Score: _____

Note: In order to receive a rubric score (4, 3, 2, or 1), a study must meet the criteria of all items in a given column.

major engagement and STEM career awareness, although the former outcome varied considerably based on the analysis used to measure impact.

DISCUSSION

The purpose of this study was to develop user-friendly rubrics that can be used by ISE STEM researchers, practitioners, and evaluators to improve quality of research designs and to measure evidence of outcomes. By consulting with informal learning experts and by leveraging several sources of data from STEM governing bodies, non-profit organizations, and ISE institutions, I identified research-based recommendations on how to assess research designs and on how to measure evidence of impact. With the feedback of STEM practitioners and recommendations from STEM stakeholders, the end products were two types of rubrics for quantitative, qualitative, and mixed methods ISE studies: (1) a STEM Research Design Rubric and (2) a STEM Impact Rubric. These tools were found to be especially applicable for assessing informal learning studies designed to inspire and motivate students to consider a STEM pathway. Here, I discuss what was learned from this process, the general applicability of these rubrics, and their respective limitations.

When designing the STEM Research Design and STEM Impact Rubrics, I used the theories of impact analysis (Mohr, 1995) and triangulation (Denzin, 1970; Ammenwerth et al., 2003; Flick, 2018a,b) to inform my research. By espousing the principles of impact analysis in alignment with triangulation theory, this study allowed for ISE educators to carefully consider different hierarchical levels for designing effective studies, ranging from higher-level designs (i.e., experimental and quasi-experimental designs) to lower-level designs (pre/post studies; comparison groups without careful matching, etc.), while simultaneously considering the unique characteristics of ISE programs (National Research Council, 2009, 2010). Moreover, because many ISE studies use qualitative designs to assess participants' outcomes, the application of impact analysis theory, coupled with research-based recommendations, were particularly instructional during the process of developing rubrics for qualitative study design and impact.

Based on these results, ISE practitioners might consider the inclusion of comparison groups in qualitative study designs (Mohr, 1995). For example, the use of open-ended survey questions or semi-structured interviews that qualitatively compare program participants to well-matched comparisons are practices that will help to increase the internal validity of ISE studies (Mohr, 1995). Lastly, the theories of impact analysis and

triangulation were particularly informative when designing rubrics for mixed methods studies as these studies provide multiple sources of data, both quantitative and qualitative, and help to increase validity and paint a more complete picture of participants' STEM outcomes (Denzin, 1970; Ammenwerth et al., 2003).

The STEM Research Design and STEM Impact Rubrics developed in this study are practical tools that can be used by ISE researchers, practitioners, and evaluators to improve quasi-experimental design was applied to compare STEM major engagement outcomes between museum program participants and comparison groups. If the research team that conducted this study had access to the STEM Research Design Rubric prior to conducting this study, the authors might have opted to design their study differently. Specifically, the researchers might have matched program participants to a comparison group with shared demographic characteristics and used propensity score analysis to compare STEM major outcomes between groups

TABLE 7 | STEM impact rubric for mixed methods studies.

Title of study or evaluation:
STEM outcome under review:
List all method(s) used to assess the STEM outcome under review (e.g., survey question 1, survey question 2, etc.):
1. _____ 2. _____ 3. _____ 4. _____ 5. _____ 6. _____ 7. _____ 8. _____ 9. _____ 10. _____

Directions:
1. Identify the STEM outcome under review (e.g., increased STEM career interest).
2. List all criteria used to assess the STEM outcome under review. For example, if there were four survey questions that evaluated STEMcareer interest, list all four questions (e.g., survey question 1, survey question 2, open-ended question 1, etc.)
3. Use the rubric to evaluate each criterion used to assess the STEM outcome under review starting from left (exemplary evidence) toright (little or no evidence)
4. Calculate the average rubric score by dividing the sum of all rubric scores by the number of criteria used to assess STEM outcomes

STEM Impact Rubric (Mixed methods studies)

(4) Exemplary evidence	(3) Strong evidence	(2) Moderate evidence	(1) Little or No evidence
• quantitative analysis: statisticallysignificant difference between comparison (control) and treatment groups (program participants) • qualitative analysis: researchersidentify outcome of interest as an emerging theme during qualitative analyses for treatment but not comparison OR minimum of 75% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted outcome of interest and there is a statistically significant difference between the comparison and treatment groups • Note: If there is no comparisongroup, starting value is 3	• quantitative analysis: statisticallysignificant difference between preand post-survey OR minimum of 75% of participants indicate higher than median (e.g., very likely; likely) outcome on post-program survey • qualitative analysis withoutcomparison group: researchers identify outcome of interest as an emerging theme during qualitative analyses OR minimum of 75% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted outcome of interest	• quantitative analysis: 40-75% ofparticipants indicate higher than median (e.g., very likely; likely) outcome on post-program surveys OR STEM outcome is maintained in comparisons of pre- and post-surveys • qualitative analysis withoutcomparison group: 40-75% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted STEM outcome of interest	• quantitative analysis: less than 40% of participants indicate higher than median (e.g., very likely; likely) outcome on post-program surveys OR STEM outcome significantly decreases in comparisons of preand post-surveys OR outcomes of participants are the same or significantly lower than outcomes of comparison group • qualitative analysis withoutcomparison group: 40-75% of data (e.g., interview data; responses to open-ended questions) indicate that the program impacted STEM outcome of interest OR qualitative analysis with comparison group: no difference between groups in outcome of interest

Average Rubric Score (sum of all rubric scores ÷ number of criteria used to assess STEM outcomes): _____

the field of informal science learning. First, depending on the goals of a study, the STEM Research Design Rubric is a useful tool for designing a study and for deciding which of the three hierarchical levels of study design is most appropriate for a given study (U.S. Department of Education, 2007). In some cases, especially studies receiving external funding and that focus on long-term outcomes, it might be most appropriate to design an experimental or quasi-experimental study. For example, in the third case study presented in this paper (Habig et al., 2018), a (Rosenbaum and Rubin, 1983; Hahs-Vaughn and Onwuegbuzie, 2006). Alternatively, prior to the study, the research team might opt to randomly select students to participate in the program via lottery and then compare the outcomes between the treatment and comparison groups (e.g., Hubelbank et al., 2007). In other cases, especially when funding is limited and the study is focusing on a short-term outcome, it might be appropriate to design a study using one of the lower hierarchical levels. For example, a study of a 1 day engineering outreach event for

4th−7th grade Girl Scouts used a pre/post study design to assess whether participation in this program increased participants' awareness of engineering careers (Christman et al., 2008). Thus, for programs with short-term outcomes and/or little or no funding, it might be most appropriate to design quasi-experimental studies or more simple designs such as a pre/post study design. Practically speaking, studies that assess short-term outcomes, such as STEM major and STEM career awareness, might operate at lower hierarchical levels (e.g., pre/post study design); studies that assess intermediate STEM outcomes, such as STEM major and STEM career interest, might operate at middle hierarchical levels (e.g., quasi-experimental), and programs that assess long-terms STEM outcomes, such as STEM major and STEM career engagement, might operate at higher hierarchical levels (e.g., experimental design) (Cooper et al., 2000; Wilkerson and Haden, 2014).

Second, while the STEM Research Design Rubric is useful for informing study design, the STEM Impact Rubric is an important

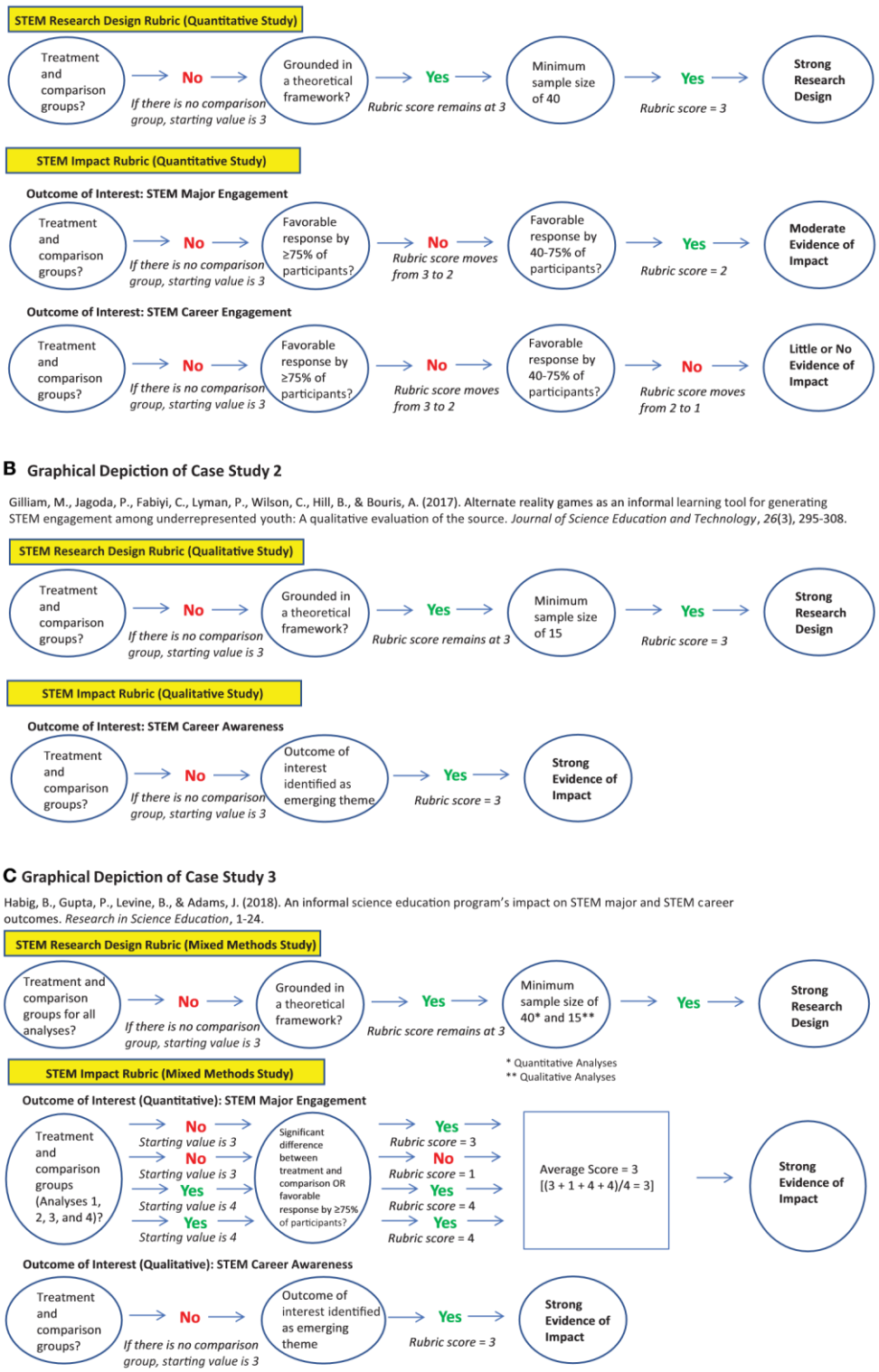


FIGURE 7 | (A) Graphical depiction of case study 1. **(B)** Graphical depiction of case study 2. **(C)** Graphical depiction of case study 3.

companion tool for helping ISE researchers, practitioners, and evaluators assess evidence of impact. Critically, the more rigorous the research design (*exemplary research design*), the more likely you can trust the study's validity (Mohr, 1995). If a study provides evidence of *exemplary research design*, then the results of the STEM Impact Rubric can be used to more confidently claim evidence of outcome. Thus, the STEM Research Design Rubric is a practical tool for developing rigorous study designs and in turn, the STEM Impact Rubric is a practical tool for providing evidence of impact. Importantly, the use of these rubrics across different ISE studies will ensure consistency in research design and measurement of impact, and when applicable, the results of these rubrics can be used to refine programs especially when there is *little or no evidence* of impact. Notably, when a STEM Impact Rubric is informed by a study with evidence of *strong* or *exemplary research design*, these results can be used for reports to government officials and funding agencies and can be potential sources for increased funding.

Finally, while the development of STEM Research Design and STEM Impact Rubrics are critical for improving the field of informal science learning and for promoting consistency between studies, it should be noted that there are several limitations of these tools. First, if a study is not designed rigorously, then we cannot reliably infer evidence of outcome. In data science, the term "garbage-in, garbage-out" is used to describe a situation in which the quality of the output is linked to the quality of the input (Rose and Fischer, 2011). Analogously, evidence of STEM outcomes (based on the STEM Impact Rubric) is linked to the quality of study design (based on the STEM Research Design Rubric). Thus, without a well-designed study, it is virtually impossible to confidently infer evidence of impact. In support, the National Research Council (2013) recommends that results from non-rigorous study designs (e.g., pre/post study design; comparison group without careful comparison) are appropriate for refining hypotheses that can be used to inform more rigorous study designs conducted in the future; however, these results should not be interpreted as conclusive evidence of impact. A second limitation is that STEM stakeholders might not always agree on the criteria used to assess research design or evidence of impact. For example, in the present study, there were some disagreements among the focus group members about whether or not a research design needs to be grounded in a theoretical framework as recommended by the Institute for Learning Innovation (2007). Thus, it is important to emphasize that these rubrics should be viewed as heuristics—tools for aiding ISE stakeholders in evaluating research design and evidence of impact—and that in some cases, depending on the goals of a particular study, it might be appropriate to alter certain criteria. Lastly, a third limitation is the challenges of applying these rubrics universally. Some ISE studies are very complex consisting of varied analyses where it might be quite difficult to assess study design and impact. For example, the third case study (Habig et al., 2018) used four different methods to assess evidence of impact with respect to STEM major engagement. In two cases, the rubric indicated that there was *exemplary evidence* of impact; in one case, the rubric indicated that there was *strong evidence* of impact, and in the final case; the rubric indicated that there was *little or no evidence* of impact. While averaging the rubric score provided a rough estimate of overall impact, the choice of an inappropriate analysis might skew the results. One of the comparison groups in this quasi-experimental study was students who attended specialized science, mathematics, and technology high schools. While there were no significant differences in STEM major engagement between participants of the informal museum program and these high school students, this analysis might not accurately measure the outcome of interest—whether participation in the ISE museum program had an impact on participants' decisions to major in STEM. Thus, it is critical that a study design is aligned to the outcome of interest and that the appropriate comparison group is considered carefully.

CONCLUSIONS

The STEM Research Design and a STEM Impact Rubrics developed in this study are potentially useful tools for ISE researchers, practitioners, and evaluators for improving study design and for assessing the effectiveness of STEM interventions. There are several possible applications for these rubrics especially in terms of areas of future research. First, in future studies, ISE researchers can measure whether large scale use of the STEM Research Design Rubric results in overall improvements in research design across the informal learning community. Second, researchers can measure whether the STEM Impact Rubric helps to inform program design and in turn, helps ISE institutions to improve specific outcomes such as persistence in STEM. More specifically, STEM researchers can assess studies yielding *exemplary evidence* of impact, extract program design principles from these studies, and share best practices across institutions. Critically, and as recommended by the National Research Council (2013), researchers can also track outcomes of interest by race, ethnicity, language status, and socioeconomic status to ensure that programs are effective across different populations. Lastly, these rubrics can be used in meta-analyses to quantitatively compare research design quality and evidence of impact across studies and outcomes of interest. In summary, the large-scale application of the STEM Research Design Rubric and the STEM Impact Rubric has the potential to transform research design quality and more confidently measure evidence of outcomes.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

The author conceived, designed, and wrote the manuscript for this study.

FUNDING

This work was supported by National Science Foundation grant no. 1710792, Postdoctoral Fellowship in Biology, Broadening Participation of Groups Underrepresented in Biology.

ACKNOWLEDGMENTS

The author would like to thank the following individuals for providing feedback during the design of the STEM Research Design and STEM Impact Rubrics: Rachel Chaffee, Jennifer Cosme, Leah Golubchick, Preeti Gupta, Brian Levine, Maleha Mahmud, Nickcoles Martinez, Maria Strangas, Mark Weckel, and Alex Watford. The author would also like to thank Veronica Catete and Kathryn Holmes for providing commentary on previous drafts of this manuscript.

REFERENCES

Ammenwerth, E., Iller, C., and Mansmann, U. (2003). Can evaluation studies benefit from triangulation? A case study. *Int. J. Med. Inform.* 70, 237–248. doi: 10.1016/S1386-5056(03)00059-5

Baer, J., and McKool, S. S. (2014). The gold standard for assessing creativity. *Int. J. Qual. Assur. Eng. Tech. Educ.* 3, 81–93. doi: 10.4018/ijqaete.2014010104

Blanchard, M. R., Gutierrez, K. S., Habig, B., Gupta, P., and Adams, J. (2020). "Informal STEM Education," in *Handbook of Research on STEM Education,* eds C. Johnson, M. Mohr-Schroeder, T. Moore, and L. English (Routledge: Taylor & Francis) 138–151.

Campbell, D., and Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Chan, H. Y., Choi, H., Hailu, M. F., Whitford, M., and Duplechain DeRouen, S. (2020). Participation in structured STEM-focused out-of-school time programs in secondary school: linkage to postsecondary STEM aspiration and major. *J. Res. Sci. Teach*. 57, 1250–1280. doi: 10.1002/tea.21629

Christman, K. A., Hankemeier, S., Hunter, J., Jennings, J., Moser, D., and Stiles, S. (2008). Overnights encourage girls' interest in science-related careers. *J. Youth Dev.* 3, 89–101. doi: 10.5195/JYD.2008.322

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* 70, 213–220. doi: 10.1037/h0026256

Collins, A., Brown, J. S., and Holum, A. (1991). Cognitive apprenticeship: making thinking visible. *Am. Educ.* 75, 6–11.

Cooper, H., Charlton, K., Valentine, J. C., and Muhlenbruck, L. (2000). Making the most of summer school: a meta-analytic and narrative review. *Monog. Soc. Res. Child. Dev.* 65, i–v. doi: 10.1111/1540-5834.00058

Creswell, J. W., and Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theor. Pract*. 39, 124–131. doi: 10.1207/s15430421tip3903_2

Creswell, J. W., and Poth, C. N. (2018). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches,* 4th Edn. Thousand Oaks, CA: Sage Publications.

Cuddeback, L., Idema, J., and Daniel, K. (2019). Lions, tigers, and teens: promoting interest in science as a career path through teen volunteering. *IZE J.* 2019:10.

Denzin, N. K. (1970). *The Research Act in Sociology*, Chicago, IL: Aldine.

Diamond, J., Horn, M., and Uttal, D. H. (2016). *Practical Evaluation Guide: Tools for Museums and Other Informal Educational Settings*. Lanham, MD: Rowman and Littlefield.

Fadigan, K. A., and Hammrich, P. L. (2004). A longitudinal study of the educational and career trajectories of female participants of an urban informal science education program. *J. Res. Sci. Teach.* 41, 835–860. doi: 10.1002/tea.20026

Flick, U. (2018a). *Doing Triangulation and Mixed Methods,* Vol. 8. Thousand Oaks, CA: Sage Publications.

Flick, U. (2018b). "Triangulation in data collection," in *The Sage Handbook of Qualitative Data Collection*, ed U. Flick (London: SAGE Publications Ltd), 527–544.

Friedman, A. (2008). *Framework for Evaluating Impacts of Informal Science Education Projects*. Washington, DC: National Science Foundation.

Fuchs, L. S., and Fuchs, D. (1986). Effects of systematic formative evaluation: a meta-analysis. *Except. Child.* 53, 199–208. doi: 10.1177/001440298605300301 Gannon, F. (2001). The essential role of peer review. *EMBO Rep.* 2, 743–743. doi: 10.1093/embo-reports/kve188

Gilliam, M., Jagoda, P., Fabiyi, C., Lyman, P., Wilson, C., Hill, B., et al. (2017). Alternate reality games as an informal learning tool for generating STEM engagement among underrepresented youth: a qualitative evaluation of the source. *J. Sci. Educ. Technol.* 26, 295–308. doi: 10.1007/s10956-016-9679-4

Greene, J., and McClintock, C. (1985). Triangulation in evaluation: design and analysis issues. *Eval. Rev.* 9, 523–545. doi: 10.1177/0193841X8500900501

Habig, B., Gupta, P., Levine, B., and Adams, J. (2018). An informal science education program's impact on STEM major and STEM career outcomes. *Res. Sci. Ed.* 50, 1051–1074. doi: 10.1007/s11165-018-9722-y

Hahs-Vaughn, D. L., and Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *J. Exp. Educ.* 75, 31–65. doi: 10.3200/JEXE.75.1.31-65

Hubelbank, J., Demetry, C., Nicholson, M. E., Blaisdell, S., Quinn, P., Rosenthal, E., et al. (2007). "Long term effects of a middle school engineering outreach program for girls: a controlled study," in *Paper Presented at 2007 Annual Conference & Exposition, Honolulu, Hawaii*. Available online at: https://peer.asee.org/2098 (accessed July 01, 2020).

Hussein, A. (2009). The use of triangulation in social sciences research: can qualitative and quantitative methods be combined? *J. Comp. Soc. Work* 1, 1–12. doi: 10.31265/jcsw.v4i1.48

Institute for Learning Innovation (2007). "Evaluation of learning in informal learning environments," in *Paper Prepared for the Committee on Science Education for Learning Science in Informal Environments*. Available online at: https://www.informalscience.org/evaluation-learning-informal-learningenvironments (accessed July 01, 2020).

Jackson, R. L., Drummond, D. K., and Camara, S. (2007). What is qualitative research? *Qual. Res. Rep. Commun.* 8, 21–28. doi: 10.1080/17459430701617879

Jick, T. D. (1979). Mixing qualitative and quantitative methods: triangulation in action. *Admin. Sci. Quart.* 24, 602–611. doi: 10.2307/2392366

Klein, C., Tisdal, C., and Hancock, W. (2017). *Roads Taken—Long-Term Impacts of Youth Programs*. Washington, DC: Association of Science Technology Centers.

Krishnamurthi, A., Ballard, M., and Noam, G. G. (2014). *Examining the Impact of Afterschool STEM Programs.* Afterschool Alliance. Avaulable online at: https://www.informalscience.org/examining-impact-afterschool-stem-programs (accessed July 01, 2020).

Lave, J., and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content analysis in mass communication: assessment and reporting of intercoder reliability. *Hum. Commun. Res.* 28, 587–604. doi: 10.1111/j.1468-2958.2002.tb00826.x

Markus, H., and Nurius, P. (1986). Possible selves. *Am. Psychol.* 41, 954–969. doi: 10.1037/0003-066X.41.9.954

McCreedy, D., and Dierking, L. D. (2013). "Cascading influences: long-term impacts of informal STEM experiences for girls," in *Presented at 27th Annual Visitor Studies Association Conference*. Available online at: https://www.informalscience.org/cascading-influences-long-term-impacts-stem-informalexperiences-girls (accessed July 01, 2020).

Mohr, L. B. (1995). *Impact Analysis for Program Evaluation.* Thousand Oaks, CA: Sage Publications.

National Research Council (2009). *Learning Science in Informal Environments: People, Places, and Pursuits*. Washington, DC: The National Academies Press.

National Research Council (2010). *Surrounded by Science: Learning Science in Informal Environments*. Washington, DC: The National Academies Press. National Research Council (2011). *Successful K-12 STEM Education*. *Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*. Washington, DC: The National Academies Press.

National Research Council (2013). *Monitoring Progress Toward Successful K12 STEM Education*. *A Nation Advancing*. Washington, DC: The National Academies Press.

National Research Council (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.

National Research Council (2015). *Identifying and Supporting Productive STEM Programs in Out-of-School Settings*. Washington, DC: The National Academies Press.

Panadero, E., and Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: a review. *Educ. Res. Rev.* 9, 129–144. doi: 10.1016/j.edurev.2013.01.002

Phillip, C. (2002). Clear expectations. *Knowl. Quest.* 31:26.

Rose, L. T., and Fischer, K. W. (2011). Garbage in, garbage out: having useful data is everything. *Measurement* 9, 222–226. doi: 10.1080/15366367.2011.632338

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55. doi: 10.1093/biomet/70.1.41

Schumacher, M., Stansbury, K., Johnson, M., Floyd, S., Reid, C., Noland, M., et al. (2009). The young women in science program: a fiveyear follow-up of an intervention to change science attitudes, academic behavior, and career aspirations. *J. Women Minor. Sci. Eng.* 15, 303–321. doi: 10.1615/JWomenMinorScienEng.v15.i4.20

Singer, S. R., Nielsen, N. R., and Schweingruber, H. A. (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering.* Washington, DC: The National Academies Press.

Smith, H. W. (1975). *Strategies of Social Research: the Methodological Imagination*. Englewood Cliffs, NJ: Prentice-Hall.

Solomon, R. L. (1949). An extension of control group design. *Psychol. Bull.* 46, 137–50. doi: 10.1037/h0062958

The PEAR Institute: Partnerships in Education and Resilience (2017). *A Guide to PEAR's STEM tools: Dimensions of Success and Common Instrument Suite*. Cambridge, MA: Harvard University.

U.S. Department of Education (2007). *Report of the Academic Competitiveness Council*. Washington, DC.

What Works Clearinghouse (2008). *What Works Clearinghouse Evidence Standards for Reviewing Studies, Version 1.0*. Washington, DC: US Department of Education.

Wilkerson, S. B., and Haden, C. M. (2014). Effective practices for evaluating STEM out-of-school time programs. *Aftersch. Matt.* 19, 10–19.

Winkleby, M. A. (2007). The stanford medical youth science program: 18 years of a biomedical program for low-income high school students. *Acad. Med.* 82, 139–145. doi: 10.1097/ACM.0b013e31802d8de6

Winkleby, M. A., Ned, J., Ahn, D., Koehler, A., and Kennedy, J. D. (2009). Increasing diversity in science and health professions: a 21-year longitudinal study documenting college and career success. *J. Sci. Educ. Technol.* 18, 535–545. doi: 10.1007/s10956-009-9168-0

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., and Shapley, K. (2007). *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement*. *Issues and Answers Report, No. 033*. Washington, DC: U.S. Department of Education.

Young, J. R., Ortiz, N., and Young, J. L. (2017). STEMulating interest: a metaanalysis of the effects of out-of-school time on student STEM interest. *Int. J. Educ. Math. Sci. Technol.* 5, 62–74. doi: 10.18404/ijemst.61149

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.