Method Note

American Journal of Evaluation
2024, Vol. 45(4) 592-609

DOI: 10.1177/10982140231152281
journals.sagepub.com/home/aje
Sage

A Protocol to Assess Contextual Factors During Program Impact Evaluation: A Case Study of a STEM Gender Equity Intervention in Higher Education

Suzanne Nobrega[1], Kasper Edwards[2], Mazen El Ghaziri[1], Lauren Giacobbe[1], Serena Rice[1], and Laura Punnett[1]

Abstract

Program evaluations that lack experimental design often fail to produce evidence of impact because there is no available control group. Theory-based evaluations can generate evidence of a program's causal effects if evaluators collect evidence along the theorized causal chain and identify possible competing causes. However, few methods are available for assessing competing causes in the program environment. Effect Modifier Assessment (EMA) is a method previously used in smaller-scale studies to assess possible competing causes of observed changes following an intervention. In our case study of a university gender equity intervention, EMA generated useful evidence of competing causes to augment program evaluation. Top-down administrative culture, poor experiences with hiring and promotion, and workload were identified as impeding forces that might have reduced program benefits. The EMA addresses a methodological gap in theory-based evaluation and might be useful in a variety of program settings.

Keywords

impact evaluation, theory-based evaluation, higher education, case studies, qualitative methods

Introduction

A major challenge in evaluating the effectiveness of institutional change efforts is the lack of a control group or counterfactual, that is, a sample of the experience that would have occurred without the intervention. The randomized evaluation is often infeasible for institutional interventions

[1] University of Massachusetts Lowell, Lowell, MA, USA
[2] Technical University of Denmark, Lyngby, Denmark

Corresponding Author:
Suzanne Nobrega, University of Massachusetts Lowell, One University Avenue, Falmouth Hall 305, Lowell, MA 01854, USA.
Email: Suzanne_Nobrega@uml.edu

because of practical concerns such as blinding or randomizing at the organization level into intervention and control conditions. Other designs are subject to confounding, meaning the failure to identify competing causes of the desired outcome(s). Another concern is whether features of the program context may serve as unidentified effect modifiers, either enhancing or interfering with the intended effect of the program. The standard methods for program evaluation and intervention research lack a way to incorporate these contributing factors into a more rigorous measurement of effectiveness, especially without an experimental design.

In response to these drawbacks, theory-based evaluation approaches to program evaluation—including contribution analysis, outcome mapping, and Realist evaluation (Better Evaluation)—attempt to generate evidence at each step along a program's causal chain to make a plausible claim of attribution for the observed program outcomes. (e.g., Kane et al., 2017; Mayne, 2001, 2008, 2012; Pawson & Tilley, 1997). To make a strong claim, evaluators should also identify and account for the effects of factors extraneous to the program itself which could have influenced the observed outcomes (Davidson, 2000; Mayne, 2012). Competing causes are forces independent of the program (e.g., policies, technology, personnel, participant characteristics) that can directly produce the program's intended outcomes. Contextual factors are characteristics of the setting (e.g., other interventions, new management, cutbacks) that can contribute positively or negatively to the program's intended outcomes. Evaluators and researchers who overlook either of these contributing factors run the risk of attribution bias, that is, drawing erroneous conclusions (either over- or under-estimation) regarding a program's effects. Erroneous conclusions can spur improper decisions by program leaders and funding agencies to discontinue support for meritorious and valuable programs or to continue supporting programs that fall short of benefitting their intended recipients. Even within experimental or quasi-experimental program designs, evidence of contextual factors that influence program success can help program leaders make decisions about where, when, and how to deliver a program to maximize its benefits (Guerin et al., 2021).

Despite the popularity of theory-based evaluation and the importance of contextual evidence, few studies report on the assessment of competing causes in program impact evaluations (Lemire et al., 2012; Weiss, 1997). A few well-defined protocols exist to assess competing causes and contextual variables including the Relevant Explanation Finder (REF) (Lemire et al., 2012), process tracing, and causal loop diagramming (Befani & Mayne, 2014; Renmans et al., 2020). These methods rely heavily on the availability of a well-developed theory of change and require substantial resources, such as a program evaluation team with expertise in theory development and a sufficient budget to support multiple modes of data collection (Delahais & Toulemonde, 2012; Dybdal et al., 2010). A more straightforward, less resource-demanding data collection procedure would be highly useful for practical reasons. Additionally, collecting data on competing causes using an exploratory approach (i.e., not organized around an existing theory of change) has the added benefit of avoiding confirmation bias (Budhwani & McDavid, 2017; Copestake, 2014). Exploratory approaches have been used successfully in complex international development settings to identify program and nonprogram causes of outcomes reported during program impact evaluation (Copestake, 2014).

An exploratory, qualitative evaluation method that specifically emphasizes the investigation of factors outside the program logic model is the Effect Modifier Assessment (EMA) method (Edwards & Winkel, 2018), in which "effect modifiers" refer to factors extrinsic to the intervention that could amplify or dampen its effects, or even be entirely responsible for the observed changes. EMA uses focus groups to engage small groups of program beneficiaries, offering a practical strategy to augment other evaluation activities. Early versions of the EMA focus groups, previously called "chronicle workshops," were tested in research settings to evaluate participatory interventions (Poulsen et al., 2015). Subsequently, EMA has been applied primarily to evaluate small-scale interventions affecting specific work units, rather

than to large-scale organizational change efforts. For example, Edwards and Winkel (2018) used the EMA method to evaluate the impacts of ergonomics improvements in a hospital operating room; they assessed the strength of work environment contextual factors such as work intensification or adding new staff that may have undermined or amplified (respectively) the effects of the ergonomic intervention. The method has also been used in manufacturing (Edwards et al., 2020), public administration (Jørgensen et al., 2021), and healthcare settings (Edwards et al., 2020; Jørgensen et al., 2019). An added benefit is EMA's ability to seek out unintended negative consequences, which are rarely formally included in program evaluation methods.

In this study, we utilized the EMA method to assess competing causes and contextual factors relevant to a university institutional change intervention to support female faculty (FF) in science, technology, engineering, and mathematics (STEM) disciplines. The EMA method was selected because of the modest resource consumption and ability to identify contextual factors. Institutional change efforts to improve workforce diversity, equity, and inclusion (DEI) are a prime example of where competing causes and contextual factors should be identified to help explain conditions that support or interfere with the intervention's success. Enhancing the proportion of women in STEM (NSF, 2020) has been gaining increasing attention as an important goal for educational institutions and employer organizations. The impact of DEI initiatives is difficult to assess because of the complex and dynamic nature of organizational settings (Kalpazidou Schmidt & Cacace, 2017). We used the EMA to augment an ongoing evaluation of this organizational program. To further clarify the potential mechanisms of influence for the identified contextual factors and their potential strength of influence on intended program outcomes, we extended the EMA method by adding two new steps that addressed these issues.

Here we begin by introducing the general EMA method. Then we describe a case study of a university-based institutional transformation intervention aimed at improving gender equity for STEM faculty members. We describe customization of the EMA method and protocol, how we analyzed and triangulated the data on competing causes and contextual factors, the evaluation results (i.e., potential strength of influence on the intended program outcomes) and their use alongside other data collected in the broader evaluation. We conclude with reflections on the strengths and limitations of the EMA method and recommendations for future research.

The EMA Method for Impact Evaluation

EMA is a qualitative data collection method to identify and evaluate changes and events that occurred in the organization during the intervention period, both those occurring *because of* and those *independent of* the intervention. The protocol is implemented in four phases (Figure 1).

Phase 1—Sampling and Recruitment

The EMA workshop is intended for people who are beneficiaries of the intervention but are not involved in planning and/or delivering the intervention. Keeping the groups small (3–5 people) is essential to allow enough time for participants to elaborate their responses. The number of workshops needed depends upon the number of units participating in the program being evaluated and the number and characteristics of the intended beneficiaries. For example, one or two EMA workshops may suffice for a program delivered to workers in a single department, whereas more would be appropriate for a program delivered to multiple work units, either with sampling stratified on unit or selectively enrolling participants to ensure even distribution.

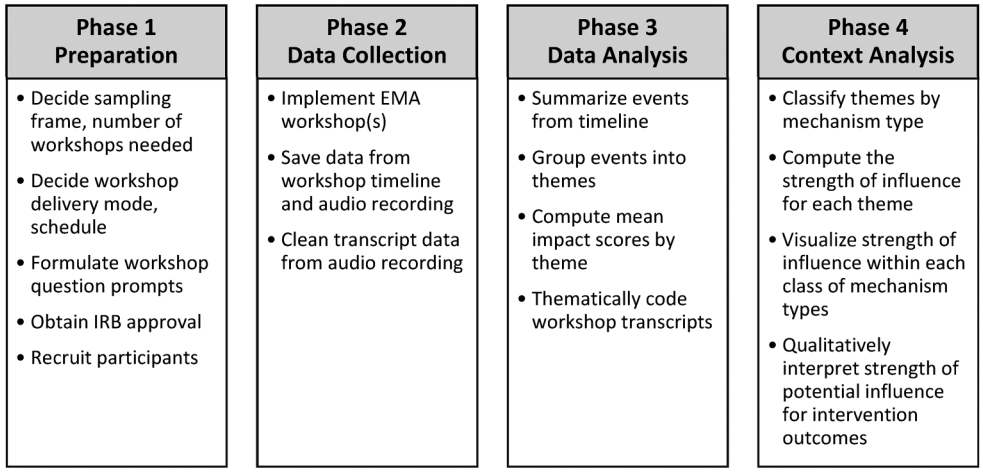

Figure 1. Phases of the effect modifier assessment evaluation protocol.

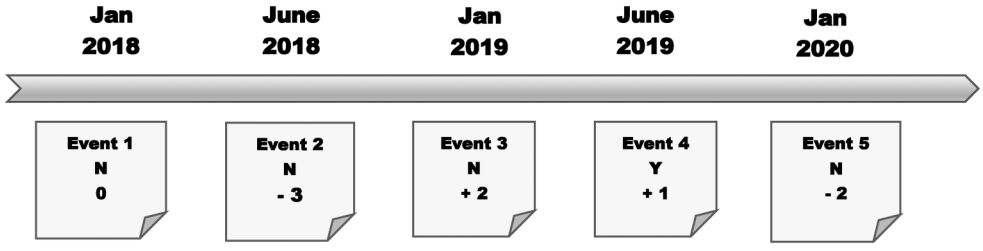

Figure 2. Example of Effect Modifier Assessment (EMA) workshop event-notes on a timeline with notation to indicate event topic and number, relatedness (Y/N) to intervention, and impact to score (range −3 to +3).

Phase 2—Data Collection (the EMA Workshop)

The EMA workshop takes about two hours to complete and is organized into two parts: Part 1 is a structured focus group interview and Part 2 is a group scoring activity of the collected interview data. The authors adapted the original published protocol (Edwards & Winkel, 2018), for virtual delivery (Nobrega et al., 2021) following a transition to remote operations during the 2020 onset of the COVID-19 pandemic. The entire workshop is audio-recorded, with participant consent.

The workshop begins with a focus group interview facilitated by a researcher who poses three questions about events and changes that have taken place during the investigated period. The questions are arranged in a hierarchy from general to specific, first about the events and changes generally, then successively narrowing in scope to the specific focus of the study. Questions used in the current study are provided in the case demonstration below.

The respondents answer a single question at a time. For each one, the facilitator asks the respondents to write associated events and changes, along with estimated month and year, that they remember on individual note cards (event-notes), without discussion. The period of silent reflection allows respondents to reflect without the influence from what others are saying or limiting their contributions to match those of others. The researcher collects and numbers each event-note one at a time from

participants. For every event-note provided, the researcher invites the person to describe it, then numbers it and places it on a timeline (Figure 2). Other participants are then invited to comment on the event-note and share their experiences. The accumulation of event-notes from all participants creates a timeline for the investigated period.

After having answered all three questions and discussed all events, participants assist in a scoring activity. Participants score each event-note based on their perceptions along three dimensions: (1) whether or not the event was part of the intervention being evaluated (yes/no); (2) the direction of impact, if any, to intended intervention beneficiaries (positive, neutral, negative); and (3) strength of impact (scale 0–3). Intervention-relatedness scoring was done as a group, whereas impact scores were collected by individuals affected by the specific events. This allows subsequent data analysis to identify "effect modifiers" or confounders, that is, events that were not part of the intervention and had either a positive or negative impact on the studied outcome, whether directly or indirectly.

Phase 3—Event-Note Analysis

EMA data analysis generates: event counts for the two main categories: intervention and nonintervention (modifier) events, categorized by direction of impact score (e.g., number of intervention and nonintervention events with positive, neutral, and negative impact scores); thematic analysis and coding of event-notes, utilizing the audio-recorded data and selecting illustrative quotes; and computing mean impact scores of events within each theme.

The results of these analyses allow evaluators to summarize the events that were salient to participants during the time of the intervention; the degree to which participants were aware of events related and unrelated to the intervention; and whether these were perceived as having a favorable, unfavorable or neutral impact on the intervention outcome and the size of impact.

Phase 4—Context Analysis

In the present study, the authors added two data analysis steps to facilitate the interpretation of whether the recalled events might have explained or influenced the intervention impacts.

Explanation Type Classification. Themes identified in Phase 3 are analyzed for their relationship with the goals of the intervention and divided into related and unrelated to the goals of the intervention. Each theme is classified according to the type of influence on intervention outcomes, inspired by Lemire et al.'s (2012) classifications: (1) primary explanation (i.e., themes part of the intervention and related to the intervention goals); (2) rival explanation or "competing cause" (i.e., themes not part of the intervention and related to the intervention goals); (3) influencing or contextual factor (i.e., themes not part of the intervention that amplify or dampen its effect on the goals); or (4) unrelated (i.e., themes not part of the intervention and unrelated to the goals). The primary explanation type is identical to intervention category but the nonintervention (modifier) category is further

Table 1. Sample Layout for Results of the Analysis.

Theme	Mechanism Type	Intervention goals	No. event-notes	Impact (mean)	Strength of influence
Theme name	Primary explanation/ competing cause/influencing factor/unrelated	Planned intervention outcome 1, 2, 3			

divided into explanation types (2), (3), and (4) to support further analysis. The strength of influence for each theme is assessed by multiplying the number of event-notes by mean impact score (Table 1).

Analysis and Interpretation of Outcomes Based on Explanation Type. Here each theme is analyzed based on explanation type. The analysis interprets major or minor sources of competing causes and contextual influencing factors, that is, themes that could have impacted the intervention goals but that were not part of the intervention. The analysis considers the number of event-notes in a theme and theory to analyze and interpret the effect.

Case Demonstration

This case demonstration describes the application of the EMA method as part of a broader evaluation being conducted of faculty DEI intervention in a North American university. The study aimed to answer the following research questions: Could the EMA method, which had been used successfully to identify competing causes in evaluations of smaller scale interventions, be useful for assessing competing causes for an institution-level intervention? What kinds of contextual factors, if any, could be identified as relevant for DEI in this university setting using the EMA workshop?

Case Setting

The university was carrying out an institutional transformation program called, making women academics valued and engaged in STEM (WAVES) (University of Massachusetts Lowell, n.d.), to increase equity for women in academic STEM careers and foster a supportive institutional culture. The program's three goals were to disrupt microaggressions, provide mentoring, and promote equity and accountability. The microaggression campaign was the most visible program component; the campaign included awareness-raising campaign materials, targeted bystander training, and climate survey feedback sessions (Bond & Haynes-Baratz, 2022; Haynes-Baratz et al., 2021). The mentoring activities included leadership development and networking lectures. The equity accountability activities involved consultation with academic departments to enhance equity in hiring protocols, goal setting, sharing of best practices, and faculty service.

Phase 1—Sampling and Recruitment

Recruitment methods for this study were described previously (Nobrega et al., 2021). Briefly, email invitations were sent by the Deans' offices of the colleges of engineering and natural sciences to all full-time tenured, tenure-track, and teaching faculty members (approximately 200) to participate in a two-hour focus group to discuss the working climate at the university. The email provided a link to an electronic enrollment survey for volunteers to indicate consent and to supply their email address, preferred focus group dates, gender, academic department, job title, and length of employment at the university. Volunteers were arranged into three groups to include and ensure diverse representation of job tenure, academic discipline, and rank in each workshop.

Faculty participants self-reported their gender as 50% female and 50% from the colleges of sciences and engineering, respectively (Table 2). Half of the participants held the rank of assistant professor; 57% were teaching faculty; and 28% were tenured. Overall, job tenure was 8 years (range 1–25 years).

Table 2. Characteristics of Study Participants (*N*=14).

Characteristic	Number (%)	Department	Number (%)
Gender			
Male	7 (50%)	Mechanical engineering	3 (21%)
Female	7 (50%)	Biology	2 (14%)
Position		Biomedical engineering	2 (14%)
Assistant professor	7 (50%)	Chemistry	2 (14%)
		Mathematical science	2 (14%)
Associate professor	5 (36%)	Chemical engineering	1 (7%)
		Computer science	1 (7%)
Professor	2 (14%)	Plastics engineering	1 (7%)
Career track			
Teaching faculty	8 (57%)		
Tenured	4 (28%)		
Tenure eligible	2 (14%)		

Phase 2—Data Collection (the EMA Workshop)

Based on the Making WAVES goals, three question prompts for Part 1 of the workshop elicited information on the workplace climate and faculty DEI (Nobrega et al., 2021). The first question was broadly framed to stimulate participants' thinking about changes at the university during the 2 years coinciding with the intervention period: "What important changes or events have you noticed related to your workplace? When I say, 'changes or events,' I mean something that occurred that had an impact on the work situation here (e.g., your department, college or the broader university)." The second question asked about changes related to workplace climate generally: "What important changes or events related to your work climate occurred during this time period?" We defined work climate as the general character of the organizational environment as perceived by those who work within it (APA, 2022). The third question asked specifically about changes related to the goals of the Making WAVES initiative: "What important changes/events related to supporting faculty diversity and equity occurred during this time period?"

In Part 2 of the workshop, for each event-note, we requested participants' scoring of relatedness to the intervention (yes/no) and estimation of the direction and strength of the impact of the event on faculty members. We asked participants to estimate the direction and strength of impact, first for faculty generally, and then specifically for female faculty.

Phase 3—Event-Note Analysis

Event-Note Counts. All events recorded in the timeline were tallied and arranged with their impact scores into two categories: events related to the investigated intervention and those that were not (i.e., "effect modifiers"). The number of event-notes in each intervention and effect modifier category was summed. Frequencies of impact (positive, neutral, or negative) for event-notes were computed for STEM general and female faculty separately.

The participants generated 118 event-notes over the course of the three EMA workshop sessions (Figure 3). One-fourth of the event-notes were designated by participants as related to the intervention and three-fourths as not related to the intervention (i.e., "modifiers," or other organizational changes or events at the departmental or institutional level).

The frequency distributions in Figure 3 indicate that participants overwhelmingly (90%) perceived the changes related to Making WAVES as having a positive impact to faculty work environment, and that Making WAVES impacts benefitted male and female faculty equally. By contrast, participants

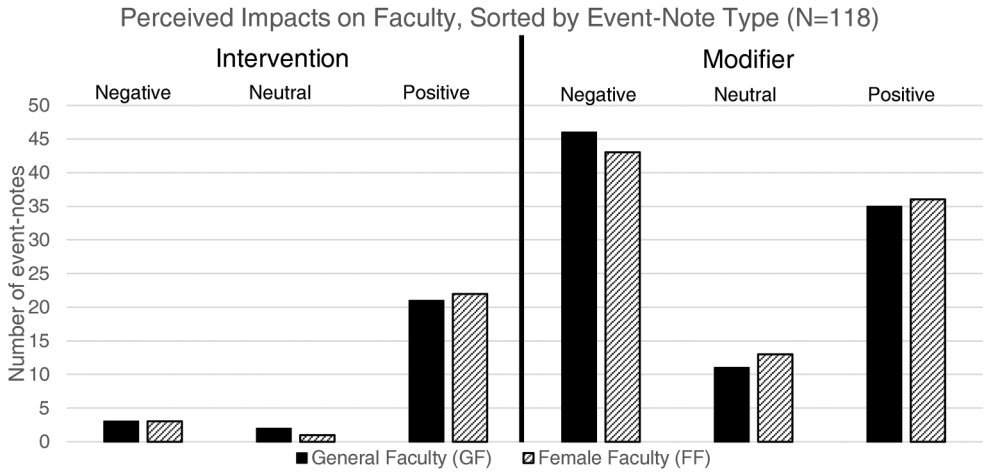

Figure 3. Frequency of intervention and modifier events by direction of impact, for general and female faculty.

scored the "modifier" event-notes in about equal numbers in positive and negative impact categories, with the negative slightly outweighing the positive. Similar to the intervention impact scores, the modifier impact scores did not vary much between impacts on general and female faculty. However, participants scored events in the negative modifier category slightly less often for female faculty compared with general faculty.

Thematic Analysis and Coding of Event-Notes Based on the Audio Recorded Data. Transcripts from the workshops were produced by the video conferencing audio recording software and exported as text files. These were cleaned manually by checking against the video recordings. Corrected transcripts for each workshop were imported to Nvivo 12 (Version 12.6; QSR, 2019) for analysis.

Thematic analysis began by reviewing the timeline data and audio transcripts. All four researchers who had participated in the sessions reviewed the data individually and generated ideas for grouping event-notes into themes. We then convened to discuss and decide on groupings and together plotted every event-note into one of the themes, refining the theme labels and event-note groupings until consensus was achieved. Transcript data were used to verify the assignment of event-notes to specific theme categories, to evaluate the potential influence of the themes, and to select quotations illustrating their effects on the work environment. In a few instances, faculty participants had different perceptions of how an event-note impacted the work environment. For example, some faculty participants reported an improved work environment after facility renovations, whereas others experienced it negatively. In such instances where an event-note contained both positive and negative scores, it was divided into two separate event-notes so that all event-notes in each theme had a single direction of impact. Examples of event-note groupings for some of the most salient themes, along with illustrative quotes from the focus group transcripts, are provided in Table 3.

Assessment of Event Impact on Work Environment. To assess the level of impact specifically for general and female faculty, we computed a mean score for all event-notes within each theme. Mean impact scores were computed in Microsoft Excel (rounding to the closest integer) using a rating scale of Minor (+1 or −1), Medium (+2 or −2), or Major (+3 or −3). The research team met with the intervention team members to validate the preliminary theme groupings. Some events that had been classified by participants as not part of the intervention (i.e., modifiers), such as revised hiring practices,

Table 3. Event Note Groupings by Theme, by Category (Intervention vs Modifier), and Illustrative Quotes.

Intervention themes	Quotes
Theme: Making women academics valued and engaged in science, technology, engineering and mathematics (WAVES) program Event note: 50/50 lectures Event note: ADVANCE grant training	"The WAVES project on campus, which has done the bystander training, they've also done the project 50/50 lectures and … There was this change on campus to look at these [opt out] policies that would generally affect women more than men. So, you didn't have to ask to get extra time for tenure … it was assumed that you would. I think that was a good climate change."
Theme: Increasing diversity hires (pos.*) Event note: Hiring committee shift-increased awareness of hiring diverse workforce. Event note: Department personnel committee guidelines regarding diversity	"There just seems to have been … a shift in our hiring committees. We brought in all underrepresented minorities … there has been a real awareness shift and change in how we're doing our candidate evaluation … more aware of inherent biases in the hiring process."
Theme: Increasing diversity hires (neg.) Event note: Administration interference with hiring-gender, ethnicity	"We received notice that we must put this person into our candidate pool who ranks in the bottom 25%."
Modifier themes	**Quotes**
Theme: Hiring, promotion, terminations (neg.) Event note: Tenure case failure at administration Event note: Reduced civility in [promotion/tenure] committee after moving	"[This faculty member] was approved [for promotion] by every faculty level. But he was denied at every administration level. And there was a sentiment that the faculty input was not being considered…"
Theme: University 2020 strategic plan Event note: University 2020 strategic plan	"… the 2020 strategic planning started back in 2010 … if you go back and look at the report cards on the diversity …there were things put in motion back then to change the campus in a very good way. I really believe that work has had an impact."
Theme: Change in chair/dean/provost (pos.) Event note: Change in department chair Event note: New Dean-caring, responsive	"It's had a major effect on … the climate in very positive ways as far as having a real push towards diversity and inclusion and … working towards questions of equity and for our students and for the faculty."
Theme: Top-down administrative culture Event note: Increasing bureaucracy	"[Administration is] hierarchical, top down, non-consultative. … it feels disempowering, it makes me less motivated and [like] I cannot change things."
Theme: Facility renovations/relocations (pos.) Event note: Opening of [new] hall Event note: Moving to [newly renovated] hall	"… it was great to be able to get my own teaching space … private place to store equipment and things like that."

*Themes containing event-notes that were assessed by participants as having positive (pos.) or negative (neg.) effect on work environment were split as two themes.

were reclassified based on information from the intervention team's knowledge of events and intervention activities that would not have been visible to workshop participants. Additional themes were suggested by the intervention team and added to the final list. Once themes were finalized, event-note counts and score means were recalculated.

Overall, the research team identified 21 themes, of which 11 were unique topics, from the 118 event-notes generated by faculty focus group respondents (Table 4). The most salient changes mentioned by faculty were DEI efforts related to the Making WAVES program (total of 27 event-notes), followed by turnover in administrative leadership; events related to general hiring/promotion/termination; facility

Table 4. Workshop Themes Classified by Mechanism Type (Primary MW Intervention, Competing Cause, Influencing Factor, Unrelated), Event-Notes per Theme, Mean Level of Impact to Work Environment, and Strength of Influence on MW Outcomes (n = 118 Event-Notes).

Theme	Number of event-notes[+]	Level of impact on WE*	MW outcomes	Strength of influence on MW outcomes**
Total - All event classes	118			6
Total - Primary explanation	**27**			**39**
Making WAVES program	9	+ 2	A, B, C	18
Microaggression training	9	+ 2	A	18
Increasing diversity hires: Pos.**	6	+ 2	C	12
Increasing diversity hires: Neg.	3	− 3	C	−9
Total - Competing cause	**19**			**7**
Hiring, promotion, termination: Pos.	10	+ 2	A, B, C	20
Hiring, promotion, termination: Neg.	8	− 2	A, B, C	−16
University 2020 strategic plan - diversity	1	+ 3	A, B, C	3
Total - Influencing factors	**38**			**−19**
Change in chair/dean/provost: Pos.	16	+ 1	A, C	16
Change in chair/dean/provost: Neg.	3	− 2	A, C	−6
Workload changes: Pos.	1	+ 1	B	1
Workload changes: Neg.	9	− 2	B	−18
Increased student enrollment	4	− 2	B	−8
Resources for faculty: Fewer	3	− 2	B	−6
Resources for faculty: More	2	+ 1	B	2
Total - Unrelated	**34**			**−21**
Top-down administrative culture	11	− 2		−22
Facility renovation, office relocation: Pos.	9	+ 2		18
Facility renovation, office relocation: Neg.	4	− 2		−8
New department/degree program: Pos.	3	+ 1		3
New department/degree program: Neg.	4	− 3		−12
Union activity: Pos.	1	+ 2		2
Union activity: Neg.	2	− 1		−2

WAVES = women academics valued and engaged in science, technology, engineering and mathematics; GF = general faculty; FF = female faculty; WE = work environment.

[+]Bold denotes total number of event-notes for the class (category) of themes.

*WE level of impact scores: minor (±1), medium (±2), or major (±3).

**Defined as the product of "number of event notes" and "strength of influence" data.

***Only theme with the difference between GF and FF.

renovations and office moves; and perceived intensification of top-down administrative culture. Changes collectively related to overall job resources and demands (e.g., 19 event-notes among themes of workload, resources, student enrollment) were also widely acknowledged by faculty across all focus groups.

The vast majority (over 90%) of the event-notes related to the Making WAVES intervention were reported as having a positive impact for all STEM faculty, regardless of gender. For example, one

participant stated, "I can see how the university is making strides towards creating a more equit[able] environment." Although infrequent, faculty did acknowledge unintended negative consequences of MW gender equity policies, namely a loss of faculty autonomy when administrators override faculty rankings of candidates for hire. While both the WAVES program and the microaggression trainings were scored as a medium positive impact (+2) regardless of gender, increasing diversity hires was viewed as having a slightly more positive impact for female faculty (+3) than for general faculty (+ 2).

The themes that were most often scored negatively were top-down administrative culture changes (11 event-notes on increasing bureaucracy and centralization that reduced faculty autonomy and convenience); workload increases (9 event-notes); and poor experiences of faculty hiring, promotion, and terminations (8 event-notes). Other than MW-related themes, the most frequent positive themes included changes in administrative leadership (16 event-notes on improvements because of a new department chair, dean, or provost); positive experiences with faculty hiring, promotion, and termination (10 event-notes); and positive experiences with facility renovations and office moves (9 event-notes on more pleasant physical environment or new space with better proximity to colleagues). Other themes, less frequently mentioned, were changes in resources, increases in student enrollment, and new degree programs.

Phase 4—Context Analysis

Explanation Type Classification. The identified themes from Phase 3 were classified in Phase 4 according to four potential types (mechanisms) of influence on MW outcomes: Primary Explanation, Competing Cause, Influencing Factor, and Unrelated. Themes in the Phase 3 intervention category were classified as Primary Explanation. For the remaining themes in the Phase 3 modifier category, the research team assessed the relevance of the *effects on work environment and DEI* for the event-notes on the intended outcomes of the MW program. These were then classified according to explanation types (2), (3) and (4). We used participants' statements from transcripts associate with event-notes within each theme, to theorize ways in which the stated events or changes could have influenced (positively or negatively) one or more of the stated MW outcomes (i.e., disrupting microaggressions, establishing new mentoring approaches, and promoting equitable policies and procedures (Table 4, Column D). Discrepancies in codes were discussed and reconciled in the research team. If the events in the theme could have *directly* produced any of the MW outcomes, then the theme was classified as the "Competing Cause" type. If the events in the theme could have *indirectly* modified (amplified or dampened) any of the MW outcomes, then the theme was classified as the "Influencing Factor" type. If no potential for influence could be theorized, then the theme was classified as the "Unrelated" type.

Finally, we summed the total number of event-notes by explanation type of *mechanism* of influence (Table 4: totals rows). We then computed *strength* of influence (Table 4: Column E) for each theme as the product of the number of event-notes (Column B) and the level of impact to work environment (Column C). These results were graphed (Figure 4) to visualize the relative importance of the identified contextual factors for MW intervention outcomes.

Interpretation of Outcomes by Explanation Type. Overall, participants perceived events related to the MW intervention (i.e., primary explanation) as the strongest influence (strength score +39, mostly positive) for MW outcomes. Although some Competing Cause events were identified, their potential influence on MW outcomes appeared to be modest when taking into account of the countervailing forces between positive and negative impacts to work environment (strength +7). In contrast, the contribution of events in the Influence Factor category may have had a greater dampening action on MW outcomes, based on the sizable number of events rated to have a negative work environment impact (strength -19). The theoretical link between these events and MW outcomes makes it plausible

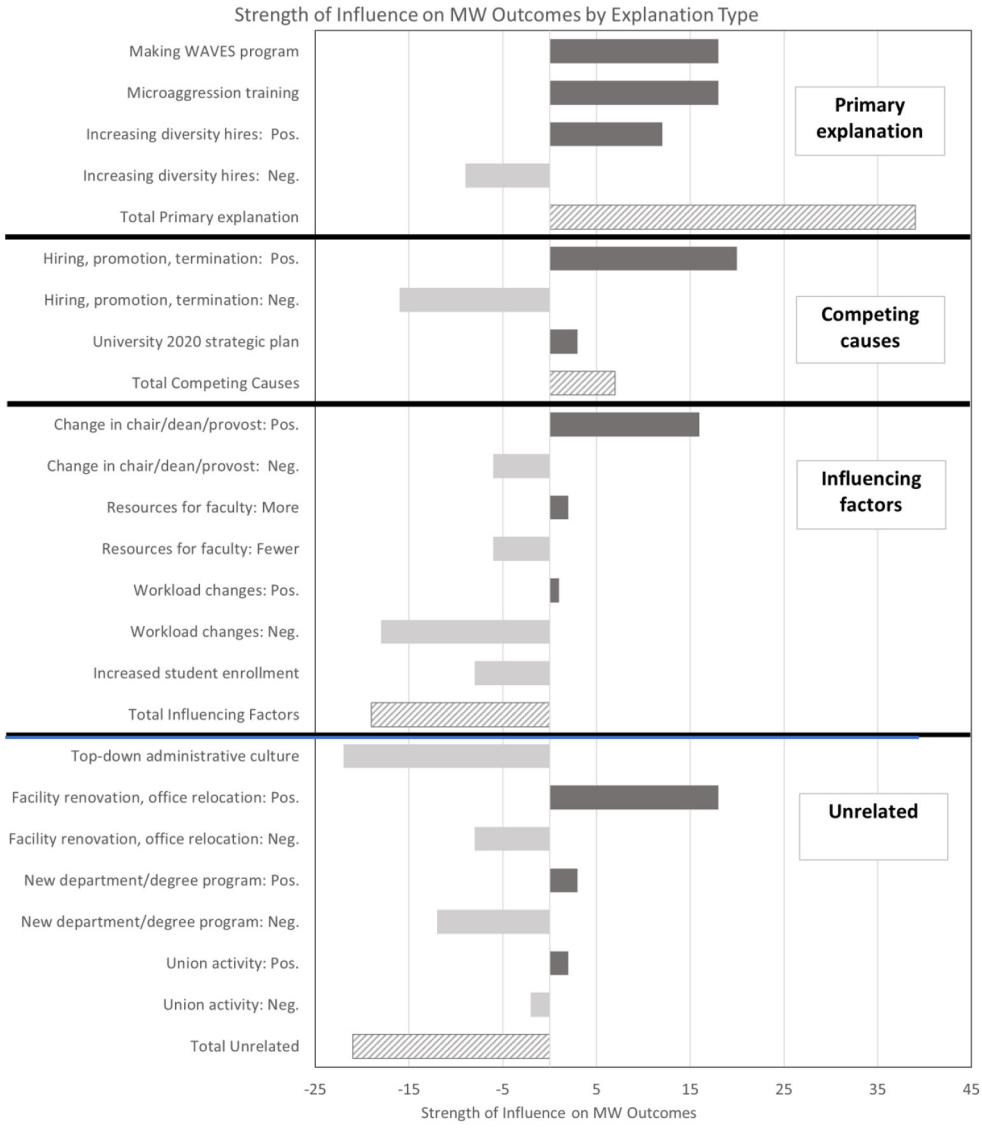

Figure 4. Strength of influence by explanation type on planned Making Waves (MW) DEI program outcomes.

that these events have exerted an opposing force on the measured outcomes of the MW interventions. A similar magnitude and pattern of influence of negative events were apparent in the unrelated category (-21).

Within the primary explanation type themes, participants expressed awareness for all three making waves intervention types (microagression training, mentoring, and policies to promote equity) across the three EMA workshops. One theme in this category, "increasing diversity hires (negative)" captured faculty perceptions about overzealous enforcement of diversity hiring policies (see Table 3, intervention row 3). Although these participants saw a high negative impact on the work environment, the strength score (-9) represented only a modest counter force in relation to the overwhelming perceptions of positive work environment impact in this category.

The Competing Causes themes scored the lowest overall strength of influence (+7). All themes in this category were assessed as relevant to all MW outcomes (Table 4, column B). The themes with the highest strength of influence were faculty hiring, promotion, tenure, and termination, where the strength of the positive events (+20) was largely counterbalanced by the strength of the negative events (-16). The event, "University 2020 strategic plan" offered a competing cause for all MW outcomes (A, B, and C) (strength +3). This event was rated as high impact to work environment but was mentioned by only one participant.

The Influencing Factors type themes (i.e., factors unrelated to MW that could amplify or dampen observed MW outcomes) collectively scored a strength of influence of -19, indicating a pattern of more events judged as negative (and more strongly negative) than positive. Perceived positive changes in administrative leadership (chair/dean/provost) were the most salient MW amplifying forces within this category (strength +16). We interpreted from the data that administrative changes could have positively influenced the MW goals of disrupting microaggressions and promoting equitable policies and procedures (Table 4, column B). Workload concerns were the most salient identified damping force (strength -18); increases in workload could have interfered with participation in MW mentoring opportunities (e.g., organizing a 50/50 lecture or IDEA group). Similarly, fewer resources for faculty (strength -6) and increased student enrollment (strength -8), both of which are related to workload, were also theorized as possible obstacles to participating in MW mentoring or other activities. Taken together, the overall prevailing direction and magnitude of negative work environment impact scored in this category could indicate possible interference of these contextual factors in allowing Making WAVES interventions to achieve their maximum effect size.

Themes in the Unrelated type accounted for an overall strength of influence of -21, slightly higher than the influencing factors category. In this category, we again see a net balance of negatively perceived events. Although we were not able to theorize any specific relationship between these events and Making WAVES outcomes, the preponderance of negative experiences could nonetheless interfere with the overall MW goal of a supportive work environment.

Case Conclusion. The goal of the Making WAVES initiative was to increase gender equity and promote a supportive academic environment for female STEM faculty in a public higher education institution. The most salient influencing factors of the institutional context that were identified for their potential *amplifying* effects on MW outcomes were: positive experiences of faculty hiring, promotion, and termination; positive changes in administrative leadership; and improvements in office or laboratory facilities. On the other hand, the most salient threats or counterforces to MW outcomes were negative experiences of faculty hiring and promotion, top-down management climate, and workload changes (increases) arising from a combination of budget cuts and increases in class size. The enhanced EMA generated a vivid picture of the dynamic institutional setting during a relatively short time period and facilitated a nuanced evaluation of the MW program against this background.

Discussion

This case demonstration of the EMA method illustrated its use as a supplementary evaluation activity for an institutional change intervention, in this case, one that sought to benefit STEM faculty DEI in higher education. Whereas the EMA had been used previously to evaluate contextual factors relevant to smaller scale interventions, we found the method also generates useful contextual data relevant for an institution-level intervention. As summarized in the case conclusion, EMA workshops provided specific data about possible competing causes and influencing factors (i.e., intervention "modifiers") in the program setting that enabled us to address questions of attribution. These were primarily amplifying and dampening forces (i.e., Influencing factors) and other events in the work environment

(Unrelated). The fact that faculty identified such a high number of so-called "unrelated" events signaled that these contributed meaningfully to the overall faculty experience and therefore should be taken into account, even though they were not specifically linked by respondents to DEI outcomes. By characterizing the possible mechanisms of influence for these events, we also gained insight into the situations where MW might or might not be effective. Themes identified were not duplicative of, but instead complemented other MW program evaluation data collected (M. Haynes-Baratz, personal communication, June 9, 2022) to facilitate more accurate judgments about attribution and effect sizes of the MW program.

Our results provided valuable insights about contextual factors relevant for gender equity interventions such as Making WAVES. Other scholars of gender equity for STEM faculty have underscored the importance of assessing the "social, cultural, normative, economic, and organizational settings" of the intervention host institution and the lack of studies that adequately assess these factors (Kalpazidou Schmidt & Cacace, 2017, p. 103). In their evaluation framework, Kalpazidou Schmidt and Graversen (2020) emphasized organizational context (working conditions, organizational climate) and the need for a complex approach acknowledging how "multiple interacting influences contribute to a particular outcome" (p.79). The EMA method can be helpful in future evaluations of STEM gender equality interventions and others that are similarly sensitive to having a supportive environment for the intervention to have the desired effects.

This study extends the previously published EMA method (Edwards & Winkel, 2018) by including a new analytical step to help with interpretation of data generated from the workshops. In addition to the already-established protocol for assessing perceived direction and size of influence of effect-modifying variables, we developed a modified version of the "Relevant Explanation Finder" framework (Lemire et al., 2012) to classify themes according to the type (mechanism) of influence on intervention outcomes. Other causal classification systems have been reported in Realist Evaluation (Cartwright et al., 2020; Roodbari et al., 2022) and Implementation Science (Lewis et al., 2018) studies. The REF typology was particularly useful here for differentiating the salient work environment conditions that could have independently explained or influenced the planned outcomes of the Making WAVES program.

A few scholars have posited ways to assess competing causes in program impact evaluation (Cartwright et al., 2020; Dybdal et al., 2010; Lemire et al., 2012; Renmans et al., 2020; Roodbari et al., 2022). These suggested approaches to context assessment include reviewing available literature, which may be limited by not being grounded in the specific study setting; forming "middle range theories" about mechanisms and the context in which they may produce observed outcomes; and theorizing in the evaluation design phase about possible threats to the assumptions underlying the program theory of change, which might overlook factors unrelated to the program's design and implementation methods. Some of these authors also pointed to data analysis procedures useful for making inferences; for example, Cartwright and co-authors (2020) state that Realist Evaluation is "method neutral." However, no specific procedures have been described for collecting the needed data on background conditions relevant to intended program outcomes. The EMA uses a concrete, structured protocol specifically designed to elicit responses from program beneficiaries about the organizational environment. This approach provides an opportunity to learn about salient factors that might not have been anticipated if the data collection instruments were designed only by starting from the program logic model.

A strength of the EMA method is that it generates useful data from the perspective of the intended program beneficiaries rather than from the program implementers. Facilitating the EMA workshops in a neutral context, temporally and physically separated from the intervention activities, helps to avoid contribution bias, a possible threat to validity (Budhwani & McDavid, 2017). In an EMA workshop, evaluators begin by posing broad questions to prompt participants to think about the most

significant aspects of their environment, regardless of their relevancy to the program being evaluated. In this study, the hierarchical structure of the EMA workshop questions first captured broad changes in the work environment; then changes in work climate; and finally, changes related to faculty gender equality. Incrementally narrowing the scope of each question allowed characterization of the events or changes that could have impacted university culture while MW was being implemented, without pointing participant attention to the specific program being evaluated until the final step.

Study Strengths and Limitations

This study engaged STEM faculty representing a broad spectrum of disciplines, ranks, job tenure, and gender, and who were the intended beneficiaries of the Making WAVES program. Although we conducted only three EMA focus groups and the sample represented a small segment (5%) of the eligible population, participants were equally divided between college units and diverse by gender and rank. The diversity of the participants surfaced a wide range of insights about changes in the work environment during the MW program. The composition of faculty characteristics (including departmental representation) in each focus group was interdisciplinary, which might have offered greater psychological safety (and richer data from more open discussion) when discussing work environment concerns than what could be achieved if participants were worried about consequences of sharing opinions directly faculty from the same department. The value of the EMA results relies on recruiting a diverse mix of intended beneficiaries to participate. There is always a risk of nonrepresentativeness or selection bias in such a sample, which should be actively countered by measures such as using multiple dissemination avenues and screening potential participants during enrollment to obtain a diverse study sample.

It is unknown whether conducting a greater number of focus groups would have produced different results. Our workshops achieved near saturation (i.e., very few new topics emerged) by the third session, indicating additional workshops may not have produced new information (Savin-Baden & Major, 2013). Given the timing of this study, which coincided with the onset of the global pandemic and the end of the academic year, we were not able to run more workshops to test whether more workshops would have continued producing data different from earlier sessions. Future research could investigate optimal EMA sampling strategies (number and composition of groups) for maximizing the likelihood that important contextual factors and competing causes are surfaced for evaluation.

It is possible that our estimates of "strength of influence" do not represent the true magnitudes of influence on MW outcomes. For example, the theme, "University 2020 strategic plan for diversity" was identified by only one faculty member (i.e., identified in a single event-note) who explained that diversity was a key institutional goal for 2020, as established 10 years earlier. Although this event was scored as high impact on work environment, its strength of influence score (-3) demonstrates that this was a minority opinion among workshop participants, even though the plan was an institution-wide effort. Our data do not permit us to resolve whether or not the strategic plan was influencing developments on campus during the MW program. Overall, the EMA provides a qualitative assessment of potential contextual forces that could have influenced intervention outcomes, but the strength of influence should not be equated with effect size measurement. The strength of influence metric does seem useful, however, for indicating the relative importance of specific events or conditions in the program environment that could influence or explain observed program outcomes. This metric may also be useful in future studies for targeting events or themes with extremely high scores for additional inquiry.

Future research involving the EMA should assess its use in a range of programs and explore how to integrate the EMA data collection and analysis with other evaluation efforts. Questions remain

about the optimal timing of EMA data collection in sequence with other program evaluation efforts, and how to interpret EMA results in concert with other evaluation data.

There are a number of practical considerations when choosing the EMA method. Evaluators using the EMA should have skills in qualitative research methods, knowledge of the program setting and intended beneficiaries, and access to the program implementers throughout all phases of the data collection and analysis. Establishing collaborative contacts internal to the organization is especially important if evaluators are external to the program setting. Last, the workshops are perceived as a positive experience by respondents as they are able to share experiences about their work experiences with each other.

Conclusion

The EMA method provides a concrete, novel protocol to assess the contextual factors that influence a program or intervention's mechanisms and outcomes. It uses an exploratory approach, centered on participants' own experiences in the program setting. The case demonstration documents the first application of the EMA method for evaluating an institutional-level change effort. At the same time, the method remains flexible enough to be used for many different programs or intervention types and levels. Evidence from this and prior studies show that the EMA method can help address a weakness in theory-based evaluation studies—that is, a lack of data collection protocols for assessing competing causes and contextual influencers. Given the resource challenges of complex program evaluations, evaluators might hesitate to spend limited resources to assess competing causes for observed program effects. However, overlooking this kind of assessment runs the risk of making inappropriate claims of program attribution and possibly drawing faulty conclusions. The EMA method is valuable for uncovering potential alternative explanations for any apparent program results, whether very strong or even null effects. In this way, the EMA method can help evaluators generate a more complete set of evidence to answer questions about attribution during program impact evaluation studies.

Acknowledgments

The author(s) sincerely thank our program partners at the University of Massachusetts Lowell Center for Women and Work, Meg Bond, PhD, Michelle Haynes-Baratz, PhD, and Brita Dean, PhD, for permitting us to pilot this version of the EMA to evaluate the Making WAVES program. The authors thank James Hughes for his assistance with data analysis and the faculty at the University of Massachusetts Lowell for their participation in this research. The facilitator guide for the EMA protocol is available as supplementary material in the Nobrega et al. (2021) reference of this paper. This content is solely the responsibility of the authors and does not represent the official views of NIOSH.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institute for Occupational Safety and Health (grant numbers U19 OH008857, U19 OH012299).

ORCID iD

Suzanne Nobrega https://orcid.org/0000-0002-2354-0996

References

APA Dictionary of Psychology, Retrieved on June 6, 2022 from https://dictionary.apa.org/

Befani, B., & Mayne, J. (2014). Process tracing and contribution analysis: A combined approach to generative causal inference for impact evaluation. *IDS Bulletin*, *45*(6), 17–36.

Better Evaluation. Approaches page. Retrieved on November 22, 2021 from https://www.betterevaluation.org/en/approaches

Bond, M. A., & Haynes-Baratz, M. C. (2022). Mobilizing bystanders to address microaggressions in the workplace: The case for a systems-change approach to getting A (collective) GRIP. *American Journal of Community Psychology*, *69*, 221–238. https://doi.org/10.1002/ajcp.12557

Budhwani, S., & McDavid, J. C. (2017). Contribution analysis: Theoretical and practical challenges and prospects for evaluators. *Canadian Journal of Program Evaluation*, *32*(1), 1–24. https://doi.org/10.3138/cjpe.31121

Cartwright, N., Charlton, L., Juden, M., Munslow, T., & Williams, R. B. (2020). Making predictions of programme success more reliable. CEDIL Methods Working Paper 1. Oxford: Centre of Excellence for Development Impact and Learning (CEDIL). https://cedilprogramme.org/publications/making-predictions-of-programme-success-more-reliable/

Copestake, J. (2014). Credible impact evaluation in complex contexts: Confirmatory and exploratory approaches. *Evaluation*, *20*(4), 412–427.

Davidson, E. J. (2000). Ascertaining causality in theory-based evaluation. *New Directions for Evaluation*, *87*, 17–26. https://doi.org/10.1002/ev.1178

Delahais, T., & Toulemonde, J. (2012). Applying contribution analysis: Lessons from five years of practice. *Evaluation*, *18*(3), 281–293. https://doi.org/10.1177/1356389012450810

Dybdal, L., Bohni Nielsen, S., & Lemire, S. (2010). Contribution analysis applied: Reflections on scope and methodology. *Canadian Journal of Program Evaluation*, *25*(2), 29–57.

Edwards, K., Cooper, R. G., Vedsmand, T., & Nardelli, G. (2020). Evaluating the agile-stage-gate hybrid model: Experiences from three SME manufacturing firms. *International Journal of Innovation and Technology Management*, *16*(8), 1950048. https://doi.org/10.1142/S0219877019500482

Edwards, K., Prætorius, T., & Paarup Nielsen, A. (2020). A model of cascading change: Orchestrating planned and emergent change to ensure employee participation. *Journal of Change Management*, *20*(4), 342–368. https://doi.org/10.1080/14697017.2020.1755341

Edwards, K., & Winkel, J. (2018). A method for effect modifier assessment (EMA) in ergonomic intervention research. *Applied Ergonomics*, *72*, 113–120. https://doi.org/10.1016/j.apergo.2018.05.007

Guerin, R. J., Harden, S. M., Rabin, B. A., Rohlman, D. S., Cunningham, T. R., TePoel, M. R., Parish, M., & Glasgow, R. E. (2021). Dissemination and implementation science approaches for occupational safety and health research: Implications for advancing total worker health. *International Journal of Environmental Research and Public Health*, *18*(21), 11050.

Haynes-Baratz, M. C., Bond, M. A., Allen, C. T., Li, Y. L., & Metinyurt, T. (2021). Challenging gendered microaggressions in the academy: A social–ecological analysis of bystander action among faculty. *Journal of Diversity in Higher Education*, *15*(4), 521–535. https://doi.org/10.1037/dhe0000315

Haynes-Baratz, M. C., Metinyurt, T., Li, Y. L., Gonzales, J., & Bond, M. A. (2021). Bystander training for faculty: A promising approach to tackling microaggressions in the academy. *New Ideas in Psychology*, *63*, 100882. https://doi.org/10.1016/j.newideapsych.2021.100882

Jørgensen, R., Edwards, K., Scarso, E., & Ipsen, C. (2021). Improving public sector knowledge sharing through communities of practice. *Vine Journal of Information and Knowledge Management Systems*, *51*(2), 318–332. https://doi.org/10.1108/vjikms-08-2019-0115

Jørgensen, R., Scarso, E., Edwards, K., & Ipsen, C. (2019). Communities of practice in healthcare: A framework for managing knowledge sharing in operations. *Knowledge and Process Management*, *26*(2), 152–162. https://doi.org/10.1002/kpm.159

Kalpazidou Schmidt, E., & Cacace, M. (2017). Addressing gender inequality in science: The multifaceted challenge of assessing impact. *Research Evaluation*, *26*(2), 102–114. https://doi.org/10.1093/reseval/rvx003

Kalpazidou Schmidt, E., & Graversen, E. K. (2020). Developing a conceptual evaluation framework for gender equality interventions in research and innovation. *Evaluation and Program Planning*, *79*(October 2019), 101750. https://doi.org/10.1016/j.evalprogplan.2019.101750

Kane, R., Levine, C., Orians, C., & Reinelt, C. (2017). *Contribution Analysis in Policy Work: Assessing Advocacy's Influence*. Washington DC, USA: Center for Evaluation Innovation. Retrieved on 6/24/2021 from: http://www.evaluationinnovation.org/publications/contribution-analysis.

Lemire, S. T., Nielsen, S. B., & Dybdal, L. (2012). Making contribution analysis work: A practical framework for handling influencing factors and alternative explanations. *Evaluation*, *18*(3), 294–309. https://doi-org/10.1177/1356389012450654

Lewis, C. C., Klasnja, P., Powell, B. J., Lyon, A. R., Tuzzio, L., Jones, S., Walsh-Bailey, C., & Weiner, B. (2018). From Classification to Causality: Advancing Understanding of Mechanisms of Change in Implementation Science. *Frontiers in Public Health 6*(136). https://doi.org/10.3389/fpubh.2018.00136

Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. *The Canadian Journal of Program Evaluation*, *16*(1), 1–24.

Mayne, J. (2008). *Contribution analysis: An approach to exploring cause and effect. ILAC Brief 16*. The Institutional Learning and Change (ILAC) Initiative. Retrieved on 6/24/2021 from: https://www.betterevaluation.org/en/resources/guides/contribution_analysis/ilac_brief

Mayne, J. (2012). Contribution analysis: Coming of age? *Evaluation*, *18*(3), 270–280. https://doi.org/10.1177/1356389012451663

National Science Foundation. (2020, March 6). ADVANCE: Organizational Change for Gender Equity in STEM Academic Professions. Beta.Nsf.Gov. Retrieved March 24, 2022, from https://beta.nsf.gov/funding/opportunities/advance-organizational-change-gender-equity-stem-academic-professions-advance

Nobrega, S., Ghaziri, M. E., Giacobbe, L., Rice, S., Punnett, L., & Edwards, K. (2021). Feasibility of Virtual Focus Groups in Program Impact Evaluation. *International Journal of Qualitative Methods*, *20*. https://doi.org/10.1177/16094069211019896

Pawson, R., & Tilley, N. (1997). An introduction to scientific realist evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 405–418). Sage Publications, Inc. https://doi.org/10.4135/9781483348896.n29

Poulsen, S., Ipsen, C., & Gish, L. (2015). Applying the chronicle workshop as a method for evaluating participatory interventions. *International Journal of Human Factors and Ergonomics*, *3*(3/4), 271–290. https://doi.org/10.1504/IJHFE.2015.073002

QSR International Pty Ltd. (2019). NVivo (Version 12), https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Renmans, D., Holvoet, N., & Criel, B. (2020). No mechanism without context: Strengthening the analysis of context in realist evaluations using causal loop diagramming In J. Schmitt (Ed.), causal mechanisms in program evaluation. *New Directions for Evaluation*, *167*, 101–114. https://doi.org/10.1002/ev.20424

Roodbari, H., Nielsen, K., Axtell, C., Peters, S. E., & Sorensen, G. (2022). Testing middle range theories in realist evaluation: A case of a participatory organisational intervention. *International Journal of Workplace Health Management*, *15*(6), 694–710. https://doi.org/10.1108/IJWHM-12-2021-0219

Savin-Baden, M., & Major, C. H. (2013). *Qualitative research: The essential guide to theory and practice*. Milton Park, Abingdon, Oxon: Routledge.

University of Massachusetts Lowell. (n.d.). Making WAVES: Women academics valued and engaged in STEM. Retrieved October 29, 2021, from https://www.uml.edu/Research/ADVANCE/initiatives/

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, *114*, 41–55. https://doi.org/10.1002/ev.1086