A Brief Review of Prevention Program Evaluation in Institutions of Higher Education

Elissa L. Perry, Ph.D., Columbia University

Commissioned paper prepared for the National Academies of Sciences, Engineering, and Medicine
Committee on Developing Evaluation Metrics for Sexual Harassment Prevention Efforts

Sexual Harassment in Higher Education

Based on their review of empirical research conducted in higher education, the National Academies of Sciences, Engineering, and Medicine concluded that greater than 50% of female faculty and staff and 20-50% of female students experience sexual harassment in academia (National Academies of Sciences, Engineering, and Medicine 2018). The costs of sexual harassment are well established. Sexual harassment negatively affects targets' mental and physical health and impairs employees' work-related attitudes (e.g., job satisfaction, commitment) and outcomes (e.g. performance, absenteeism) as well as students' educational outcomes (e.g., lower grades; Bondestam and Lundqvist 2020; Fitzgerald and Cortina 2018; Henning et al. 2017; National Academies of Sciences, Engineering, and Medicine 2018; Shaw, Hegewisch, and Hess 2018; Wood et al. 2018). Sexual harassment can also negatively impact colleagues and peers who witness it (Berdahl and Raver 2011; Bondestam and Lundqvist 2020) and the institutions in which it occurs. Institutions can incur legal costs if there are formal charges, as well as costs from increased absences, turnover, and lower motivation and commitment among employees. Unfortunately, there is considerable evidence that formally reporting sexual harassment to their institutions is the least common response by students and faculty who experience it (Bondestam and Lundqvist 2020; National Academies of Sciences, Engineering, and Medicine 2018).

The persistence of sexual harassment, the costs it incurs, and the reluctance of targets to report it make the need for academia to focus on sexual harassment prevention efforts particularly urgent. This need is even more evident in the context in which harassment is occurring today; there is a greater social awareness of and mobilization against sexual abuse and harassment in the form of the #MeToo movement. However, despite this obvious need, a 2016 Equal Employment Opportunity Commission (EEOC) report concluded that employers' efforts to prevent sexual harassment have been woefully ineffective (Feldblum and Lipnic 2016). In response to this state of affairs, the National Academies of Sciences, Engineering, and Medicine established an Action Collaborative initiative on Preventing Sexual Harassment in Higher Education, bringing together more than 60 institutions in higher education to identify, develop, and implement evidence-based

policies and practices to address and prevent sexual harassment (National Academies of Sciences, Engineering, and Medicine 2020).

Sexual harassment interventions have been described as taking one of three approaches: primary, secondary, or tertiary (Hunt et al. 2010; McDonald, Charlesworth, and Graham 2015). *Primary* interventions attempt to address the root cause of the problem, preventing the problem from arising in the first place. This approach typically includes the development and communication of sexual harassment policies and the provision of sexual harassment awareness education and training (Hunt et al. 2010; McDonald, Charlesworth, and Graham 2015). *Secondary* interventions are directed toward how organizations respond after sexual harassment has occurred. This approach focuses on developing and ensuring an effective complaint procedure; preventing further incidents, and dealing with the immediate consequences of the sexual harassment. Finally, *tertiary* interventions focus on longer-term restorative responses (e.g., provision of counseling) that address lasting consequences of the sexual harassment and supporting victims after sexual harassment has occurred. There is a general "absence of empirical evidence examining the effectiveness of many of these strategies" (Hunt et al. 2010, p. 661), with tertiary interventions, in particular, receiving the least research attention (Hunt et al. 2010; McDonald, Charlesworth, and Graham 2015). While this typology is helpful, some intervention strategies span more than one of these approaches, and others do not fit neatly into this framework. Institutional factors including leadership, organizational structure, practices, and systems (not directly related to victim support or complaint procedures) may also play a role in preventing sexual harassment. In reality, a combination of strategies is likely to be most effective; a literature review of prevention programs addressing a variety of social issues (e.g., substance abuse, risky sexual behavior) found that comprehensive prevention strategies employing multiple interventions across multiple settings (e.g., community, peers, school) are most effective (Nation et al. 2003). A combination of strategies, with visible support from top leadership, is likely to contribute to an institutional climate that does not tolerate sexual harassment (Fitzgerald and Cortina 2018).

Prevention Evaluation

Research has not evaluated the effects of major efforts designed to prevent sexual harassment (Bondestam and Lundqvist 2020). However, to be effective, prevention programs must be continuously evaluated (Nation et al. 2003). Information learned from evaluations of interventions serve a number of purposes (Griffin, 2012). Information from evaluations can be used to justify the allocation of resources and time, assess whether the intervention is having the intended impact, and consequently whether elements or the whole of the intervention need to be revised or discontinued. Research in the context of training interventions finds that the frequency of evaluations and what is evaluated (e.g., reactions, behavior, knowledge, results) has implications for whether the training intervention effectively transfers beyond the training program. For example, transfer of training beyond the training context is more likely when organizations evaluate their training more frequently, and when they assess the impact of training on trainees' behaviors and department and organizational performance rather than on trainees' reactions to the training and how much they learned in the training (Saks and Burke 2012). Despite all of the good reasons to evaluate, research on program evaluation and particularly training interventions indicates that few organizations do it. Research has identified obstacles that limit the extent to which program evaluation occurs including a lack of resources, assumption that the training works, a lack of agreement on what should be evaluated, and a lack of knowledge about how to conduct evaluations. Institutions may also be reluctant to evaluate their sexual prevention programs because of concerns that learning that they are potentially ineffective may put them at risk. Additionally, when evaluation is undertaken, there is often a disconnect between what is measured and the outcomes training is intended to impact (Medeiros and Griffith 2019). For example, when evaluation of training is undertaken it tends to focus on the affective reactions of trainees (e.g., satisfaction with the training), which may be weak predictors of training effectiveness (Griffin 2012; Wang and Wilcox 2006).

Program Evaluation 1.0

Much has been written about best practices for evaluating programs in general and training programs more specifically. A discussion of important considerations in the evaluation process follows next.

Step 1: Conduct Needs Assessments

Few would disagree with the importance of implementing prevention efforts in higher education institutions, therefore many institutions currently offer a wide variety of interventions. Although it seems logical to suggest that interventions are likely to be most effective when they address the specific needs of a given institution, it is not always clear that the interventions that are developed and implemented are based on formal assessments of the needs of the institution and the people affiliated with it. A needs assessment involves diagnosing what needs to be trained, for whom, and when (Salas et al. 2012). Implementing a sexual harassment prevention program without first conducting a needs assessment is akin to giving a patient (in this case an individual or organization) medicine without fully diagnosing their illness and understanding how it is affecting them specifically.

Different stakeholder groups within higher educational institutions (e.g., faculty, students, staff) likely have different needs that require different interventions. A person analysis can help identify individuals who are most likely to benefit from an intervention and the type of intervention that is most appropriate for them. In some cases a climate survey can serve as a diagnostic tool if it provides insight into the prevalence of sexual harassment, the institution's perceived tolerance of it, and knowledge about reporting procedures. For example, a climate survey may help identify which stakeholder groups on a given campus experience the highest levels of sexual harassment and therefore should be targeted for immediate action. Supplementary data may be required to determine which target groups have a good understanding of information related to sexual harassment (e.g., reporting channels, institutional policies) and which do not, and therefore which groups should be targeted to receive this information. An intervention designed to provide individuals with information about reporting channels is most effective when it targets only those who require this information.

Additionally, an organizational assessment can help identify intervention priorities based on the institution's goals and determine whether the institution has the resources and will to support interventions that are developed (Salas et al. 2006). For example, if an important goal of an institution is to promote female leaders into positions of authority, an intervention designed to target this specific outcome should be prioritized. It is also important to consider the institution's

organizational climate. Ideally, interventions that align with the institution's strategic priorities and those that foster an environment that supports the intervention should be given top priority. Annual climate surveys may provide useful information about the climate in which interventions are or will be implemented, and thus help identify potential obstacles to the success of these interventions. It is also important to assess the institution's resources and will; interventions will only be successful when the institution provides financial support, proper equipment, and necessary materials to develop and implement interventions.

An important outcome of a needs assessment is an understanding of the specific needs and whose needs the intervention is designed to address (e.g., greater awareness, changed behavior, improved climate) and thus greater clarity around the outcomes that an intervention should be designed to impact. The needs assessment should be tied directly to an evaluation plan (National Academies of Sciences, Engineering, and Medicine 2018). For example, some institutions may identify and prioritize changing a particular group's (e.g., members of fraternities and sororities) attitudes related to sexual harassment, while others may focus on addressing the institution's need to limit legal vulnerabilities by reducing sexual harassment.

Step 2: Choose Appropriate Program Outcome Measures

Training and program evaluation fall into one of two categories: formative and summative. *Formative* evaluations focus on improving the quality of the program, including its delivery and design. *Summative* evaluations are used to assess whether the program achieved its intended goals and outcomes (Griffin 2012; Wang and Wilcox 2006). Further, summative evaluations can focus on short-term outcomes such as reactions of and learning by participants or longer-term outcomes such as behavior change and institutional impact (Wang and Wilcox 2006). Historically, training evaluation has been based on outcomes from four hierarchical levels: reactions, learning, behavior, and results (Salas et al. 2012), with the majority of organizations relying on the first two levels. In other words, evaluation typically focuses on short-term outcomes that occur immediately after an intervention while ignoring its longer-term impact. *Reactions* might refer to whether the participant liked the program and thought it and the person delivering the program were effective. In the context of training interventions, scholars have suggested that questions about the usefulness of the intervention may be more strongly

associated with learning than questions about whether participants liked the intervention (Aguinis and Kraiger 2009). It may also be useful to ask people how interested and motivated they are to participate in the intervention because these perceptions may have implications for its success (Wang and Wilcox 2006). *Learning* outcomes indicate how well participants learned program-provided facts and information (e.g., understand the institution's policy, know what different types of sexual harassment are, understand the institution's reporting procedures). *Behavioral* evaluation pertains to the extent to which the program leads to behavioral change outside of the program. This requires direct observations of behavior outside of the program, which is often difficult, time consuming, and costly (Tannenbaum and Woods 1992). This might include observations that program participants are more willing to report inappropriate behavior when they are a third party. Finally, *results* assess whether the program impacts institutional outcomes such as a reduction in formal complaints of sexual harassment or turnover related to sexual harassment. These distinctions are important, and care must be taken to collect the data that are most consistent with program goals. For example, while participant reactions can be used to understand interest, attention, and motivation to engage with a program, they say little about whether the person has successfully acquired new skills or knowledge. However, learning evaluations that measure knowledge and skills can be used together with reaction evaluations to improve a program.

Addressing whether a program is effective requires a clear idea of the purpose of the program and what the program is meant to impact and is likely to require multiple measures of different types of outcomes (e.g., reactions, learning, behavior) (Salas et al., 2012). Greater clarity of purpose can be achieved by a formal needs assessment, which can help identify who is in need (e.g., undergraduates do not have a good understanding of the institution's reporting channels) and what is needed (e.g., provision of information about reporting channels). An intervention program designed to provide this information could then be evaluated on the basis of the extent to which undergraduate students who participate in the program understand the various reporting channels available to them (e.g., a learning outcome perhaps measured by a quiz). Collecting data on a given outcome from multiple sources where appropriate (e.g., program participants, peers) may provide evidence of convergence and thus greater confidence in results (Griffin 2012).

It is important to consider how quickly evaluation information is needed and how long it will take for the impact of the program to impact the outcomes of interest (McLinden 1995; Tannenbaum and Woods 1992). For example, increasing program participants' knowledge will likely take less time to impact and could therefore be assessed more quickly than a program designed to change behaviors, which will likely require a longer window. The choice of outcome measures should be based on the results of a needs assessment as well as the purpose for which the measures are being used. For example, measures needed for compliance reporting may differ from those that may be helpful for making improvements to the prevention program. Unfortunately, institutions tend to use data that are easily measured or available rather than obtain data that align with program objectives (Funnell 2000). Consistent with this, in their review of research on sexual harassment and assault training programs, Medeiros and Griffith (2019, 10) observed that there is a general "misalignment between the constructs being measured and the intended outcomes of training." Once appropriate outcomes have been identified, attention must be paid to finding or developing valid measures.

Step 3: Choose Appropriate Evaluation Designs

Before a particular evaluation design is adopted, the level of evidence required to show program impact must be decided. Typically, greater effort (i.e., more rigorous and sophisticated designs) is required to establish greater evidence that the program is effective. More rigorous designs, which include use of control groups and pretests, help reduce counter-arguments that the program was effective. More rigorous evaluation designs are also more costly and require more skill to employ. However, these designs are not always necessary depending on the purpose of the evaluation (McLinden 1995). For example, less rigor is required to establish how open participants are to a new program, which can be assessed by collecting reaction data at the end of a program. However, if the program is large (i.e., impacting a large number of people), ongoing, and perceived as important by the institution, a more rigorous design may be appropriate. Therefore, it is important to establish the level of evidence required to establish program effectiveness up front.

Many people have suggested that the most powerful experimental designs (i.e., randomized controlled trials) are difficult, if not impossible, to implement in real-world settings, particularly

when they involve human program participants as is the case in sexual harassment prevention programs (Chatterji 2007; Kraiger, McLinden, and Caspar 2004). However, some simpler quasi-experimental designs may be effective. For example, where possible, it can be useful to make comparisons between those who have been exposed to the program and a control group of individuals who have not. Logically, if both groups are impacted by the same extraneous factors (e.g., college communications), these factors are less likely to explain differences between the two groups. It may also be possible to use program participants who will but have not yet participated in the program as a comparison group for those who have. Where comparison groups are not available or possible, using trend lines may be a possibility: tracking the outcome of interest (e.g., level of satisfaction with reporting sexual harassment procedures as measured in an annual campus climate survey) over time (Shadish, Cook, and Campbell 2002). Outcome measures (e.g., level of satisfaction) collected at multiple times prior to the program can be used to forecast future trends; support for program effectiveness occurs when the post-program data are higher than the forecasted trend line.

Evidence of impact can also be demonstrated by showing that the outcome targeted by the program (e.g., knowledge of the institution's reporting procedures) is higher following the program compared to before (pre-test/post-test design). An even stronger version of this design would be to compare changes in outcomes that are expected to change as a function of the program (planned) relative to those that are not (unplanned) (Haccoun and Hamtiaux 1994). If, for example, a program is designed to improve participants' ability to identify different types of sexually harassing behavior, responses to items assessing this knowledge should improve by a greater amount than items assessing knowledge of the institution's sexual harassment reporting procedures (Haccoun and Hamtiaux 1994; Kraiger, McLinden, and Casper 2004). Finally, in some cases, assessing impact only after the intervention may be appropriate if the program developers are more interested in whether a particular level of the outcome (e.g., skill, knowledge, performance) has been achieved rather than the amount of change in that outcome (Sackett and Mullen 1993). For example, a program designed to impart information about the institution's resources for sexual harassment can establish effectiveness by demonstrating that program participants are familiar with at least 80% of the campus resources related to filing a complaint of sexual harassment. In this case, it may be less important to establish how many

more resources participants are familiar with following the program (i.e., pre-post program change), than that they are familiar with 80% of them (i.e., achieved the benchmark). It is typically more difficult to make the case for program impact when no-control group designs are used. However, when these designs are used, they should include a careful investigation (e.g., using interviews or questionnaires) into whether other extraneous factors (e.g., course curricula, other campus initiatives) may have played a role in what appears to be a program effect (Sackett and Mullen 1993).

Finally, where quasi-experimental and experimental designs are not feasible (e.g., when only correlational data obtained from a survey or from a large database are available), a confirmatory evaluation approach may help strengthen support for a causal relationship between program participation and targeted outcomes (Chatterji 2016; Reynolds 2005). This approach is based on a clear theory of how the program is expected to impact the outcomes of interest and looks for and interprets patterns of relationships based on the theory. The following are examples of the types of evidence that can provide greater support for conclusions that the program impacted targeted outcomes (Chatterji 2016; Reynolds 2005):

1. Temporality. When outcomes are measured *after* program participation.
2. Gradient. When greater program intensity (e.g., number of participation hours, number of sessions attended) is accompanied by greater outcomes.
3. Size. When size of the relationship between the program and outcomes is sufficiently large.
4. Specificity. When a specific type of intervention consistently influences a specific type of outcome.
5. Consistency. When the intervention impacts the same outcomes similarly across contexts and groups of people.
6. Coherence. When the findings taken as a whole provide a convincing story for the effects of the program on outcomes.

Program Evaluation 2.0

In practice, institutions may implement sexual harassment prevention programs that are complex because they involve multiple intervention strategies, target multiple stakeholders (i.e., faculty,

staff, and students), are delivered by human agents, and are housed in institutions that are themselves complex. Some scholars have suggested that evaluating complex social programs requires a different approach than evaluating programs that are less complex (Chatterji 2007, 2016). Complex social programs often do not lend themselves to the most rigorous experimental designs; they are difficult to standardize due to human delivery agents and to randomly assign and manipulate, and often more than one program can influence outcomes of interest. As a result, it is difficult to isolate the net effects of the entire program on targeted and measurable outcomes over and above the effects of other contextual factors. This reality runs counter to assumptions of a traditional evaluation approach, in which a standardized treatment can be experimentally manipulated and randomly assigned to a clearly defined target group, resulting in an indisputable causal impact on targeted outcomes (Chatterji 2016).

Two approaches have been suggested to evaluate the effects of complex programs on targeted outcomes. First, it may be helpful for multiple stakeholders in the institution, including those involved in the development and implementation of the program, to develop a "logic model" or "program theory" that depicts the causal pathways by which the program is expected to lead to outcomes in target populations before it is evaluated (Chen and Rossi 1983). The logic model makes the underlying assumptions about how the program is expected to work explicit (Reynolds 2005). This model can then be used to guide the evaluation process (Chatterji 2004). The model can be based on social science theory, different stakeholder beliefs, and local information. The model should identify program-relevant constructs (e.g., different aspects of the content of the program and how it is delivered), map relationships between these and program outcomes, and identify aspects of the larger context (e.g., campus population, institutional leadership, climate) that may influence the program and outcomes (Chatterji 2004, 2016).

Additionally, this model should outline the relationship between the program and a sequence of hierarchical outcomes: immediate (e.g., intended targets participate in the program), intermediate (e.g., participants learned something), and long-term impacts (e.g., there is greater intervention on the part of bystanders on campus) (Funnell 2000). For each level, attention should be given to what the outcome of interest is, what success looks like, what program factors and non-program factors may impact success, and what information should be collected, from what sources, and

using what methods (Funnell 2000). The collection of sequential outcomes over time provides information about intervening processes that affect program effectiveness and allows course correction along the way (Chen and Rossi 1983). Different questions about impact are likely to arise during different phases of the program (e.g., early compared to later stages) and these different questions will require different methodological approaches. For example, earlier in the evaluation process, the model can be used to assess whether the program is actually being implemented in a way that is consistent with how it was conceived. As data are collected and changes to the program are made, the model may evolve over time and incorporate feedback loops (e.g., showing how outcomes impact contextual factors). See Figure 1 for a depiction of a sample logic model that may guide the implementation of a sexual prevention program.

Some scholars have also advocated for multi-phased mixed methods designs that rely on a logic model in addition to other strategies. These designs employ qualitative and quantitative (e.g., correlational, quasi-experimental, and experimental) research methods at different points in the life of the program to assess whether it had the intended effects (Chatterji 2004, 2016; McLinden 1995). This approach suggests adopting a more exploratory perspective in earlier phases where the goal is to learn about the organizational context, refine and stabilize the content and delivery of the program, and better understand the context in which it is implemented. Only when a program is implemented successfully (and the logic model may help establish this) can it be determined whether it has had its intended impact (Chen and Rossi 1983; McLinden 1995). The next, middle, stage may involve preliminary examination of relationships between the program and its intended outcomes. Finally, guided by an underlying logic model or program theory, later stages would take a more confirmatory approach and more formally assess the impact of the program on intended outcomes using the comparative experimental or quasi-experimental designs described earlier (Chatterji 2004, 2007, 2016). This approach suggests that more rigorous evaluation designs should occur in later stages when the program is better defined, and potentially confounding environmental and organizational variables are more clearly identified and can be observed and analyzed. Across these phases, multiple qualitative (e.g., interviews) and quantitative (e.g., surveys) methods may be used for process and outcome assessments, and may complement one another to provide a more comprehensive understanding of the program (Chatterji 2016).

Table 1 summarizes what the approach proposed by Chatterji (2004, 2016) might look like when applied to a hypothetical sexual harassment prevention program comprising multiple moving components in a higher education institution. For example, a program like this might include education and policy training offered to students and employees at all levels; new institutional offices and processes for reporting and addressing incidents of sexual harassment; and maintenance of longitudinal survey data and records on sexual harassment incidents.

Examples of Evaluation in Practice

A 2017 Association of American Universities report summarized the results of a survey of its 62 member institutions and provided examples of the interventions these institutions employed to prevent and respond to campus sexual assault and misconduct (Association of American Universities 2017). Eighty-four percent of the institutions that responded indicated they were developing new or improved ways of measuring effectiveness of policies, programs, and interventions. Evaluations were based on student opinion and feedback, trends in surveys conducted over time and pre-post evaluations of interventions. The report found a particular focus on assessing students' knowledge about and use of campus policies and resources related to sexual assault and misconduct. Additionally, greater than 50% of institutions indicated they assessed faculty and staff knowledge. The report further indicated that climate surveys were frequently used by institutions (e.g., Massachusetts Institute of Technology, Rutgers University—New Brunswick, Indiana University, Yale University), and in some cases paired with focus groups, to assess prevalence of incidents, students' knowledge and attitudes related to policies and resources, and student satisfaction after using particular campus resources (Association of American Universities 2017). These surveys identified problems (e.g., frequency of sexual misconduct, lack of knowledge of college reporting procedures) and thus the needs a prevention program could be developed to address. Similarly, the National Academies of Sciences, Engineering, and Medicine's Year 1 Annual Report of the Action Collaborative on Sexual Harassment Prevention in Higher Education found that a majority of the institutions participating in the collaborative reported relying on quantitative climate survey data or formal reports of sexual harassment to understand harassment (National Academies of Sciences, Engineering, and Medicine, 2020). Institutions that indicated that they were working on evaluation most often described their efforts as revising climate surveys, expanding the

population surveyed, or conducting new surveys. These reports leave little doubt that institutions of higher education are collecting a large amount of data. What is less clear is whether these data are being collected and used in the systematic fashion recommended by evaluation researchers discussed above.

Cornell University employed an evaluation approach considered to be among the most rigorous to assess its Intervene program. The program includes a 20-minute online video and a 60-minute facilitated workshop. The video, a product of a joint collaboration between Cornell's health center and its theater ensemble, depicts a number of scenarios (including sexual harassment) and shows how students can make a difference in each of them (health.cornell.edu/intervene). The university conducted a randomized controlled trial to evaluate the standalone effectiveness of the video among graduate and undergraduate students. Students were randomly assigned to either watch the video or were assigned to a comparison group that did not watch the video. Results of their study indicated that students who watched the video reported a higher likelihood to intervene for most situations compared to the comparison group that was not shown the video (Association of American Universities 2017).

Argonne National Laboratory took a non-experimental approach to assess the impact of its Core Values Shout-Outs Program. The objective of the program was to create a welcoming and inclusive environment by recognizing colleagues who demonstrate core values through their behaviors. Three months after the program concluded, a series of employee climate pulse surveys was conducted. Respondents were asked about positive activities that influenced their attitudes and behavior, including the laboratory's focus on Core Values generally and programs such as Shout-Outs more specifically (Sexual Harassment Collaborative Repository, The National Academies of Sciences, Engineering and Medicine, https://www.nationalacademies.org/our-work/action-collaborative-on-preventing-sexual-harassment-in-higher-education/repository, accessed April 9, 2021).

The Department of Cell and Developmental Biology at Vanderbilt University intends to take a non-experimental (post-test only) approach to evaluate the impact of its revised admissions process. The objective of the revised process was to address advisor-student conflicts and

turnover among graduate students who entered the program through direct admission to a lab. This form of admission was revised to address the hierarchical and dependent relationship this system created between students and advisors. The Department intends to use three criteria to assess the success of the revised admissions process. First, it will use qualitative data to assess the experience of direct-admit students; whether they experienced conflict with their advisor, their academic success, and their intention to stay in the doctoral program. Second, data will be collected to assess how many direct-admit students had worked with their advisor prior to joining a lab. Revisions to the program were designed to decrease the possibility that direct-admit students would not have worked with their lab advisor prior to joining the lab. Third, data will be collected on the overall number of direct-admit graduate students. Lower levels will be considered evidence of the program's success. The small numbers prevent the use of a more sophisticated assessment, but the use of multiple types of data from different sources is a strength. However, this design could be further strengthened by considering the impact that other extraneous factors may have on the chosen metrics (Sexual Harassment Collaborative Repository, National Academies of Sciences, Engineering, and Medicine, https://www.nationalacademies.org/our-work/action-collaborative-on-preventing-sexual-harassment-in-higher-education/repository, accessed April 9, 2021).

The University of California (UC) Berkeley illustrates how institutions may evaluate more complex programs. UC Berkeley developed a Toolkit, "Preventing Sexual Harassment in Your Academic Department," designed to help decision-makers in academic departments implement a plan to prevent sexual harassment in their academic community. This toolkit entails multiple components including incorporating diversity and inclusion values into hiring, promotion, and the admissions process, fostering prosocial behaviors on the part of community members, and developing leadership skills to reinforce community values and address harmful behavior. A working group of multiple stakeholders within the College of Engineering used the toolkit. The plan to evaluate the impact of their work includes assessing both processes and results, using different types and sources of data. Process evaluation includes assessing whether prevention recommendations were generated and implemented. Results evaluation includes the use of climate surveys to track changes in departmental climate over time and assessing reactions of the working group members to their participation. For example, members of the working group were

asked to indicate their own understanding of how to prevent sexual harassment in the College, which improved from limited or no understanding to good or excellent understanding. The collection of different data from different data sources is a strength, but it would be useful to consider the role of extraneous environmental factors on outcomes (Sexual Harassment Collaborative Repository, National Academies of Sciences, Engineering and Medicine, https://www.nationalacademies.org/our-work/action-collaborative-on-preventing-sexual-harassment-in-higher-education/repository, accessed April 9, 2021).

Conclusion

It is clear that institutions of higher education are engaged in many different types of sexual harassment prevention efforts aimed at different target groups. However, it is unclear how much attention has been given to the evidence bases for the programs that are developed and the approaches that are taken. It also is not apparent whether institutions have undertaken person, group, or organizational needs assessments prior to the development and implementation of their prevention programs. Needs assessments are typically undertaken prior to an intervention to understand what needs to be addressed, for whom, and when (Salas et al. 2012). Programs can be most effective when they are developed based on measured needs rather than the presumed needs they are intended to address. The results of these assessments can help inform program development (e.g., a person assessment may suggest that different types of programs are required for different types of people), identify where problems lie (e.g., identify labs or groups that have higher rates of sexual harassment), and help set program priorities (e.g., who most requires an intervention and when). Additionally, institutions of higher education are collecting an extraordinary amount of data related to sexual harassment generally and the impact of prevention interventions more specifically. It is difficult to divine from available public reports whether these data collection efforts are guided by a clear theory of how the program should work and thus employ a systematic approach to the evaluation process. Given the complexity of some of the programs that institutions offer, this would seem to be essential to establishing the effectiveness of the prevention efforts that are undertaken.

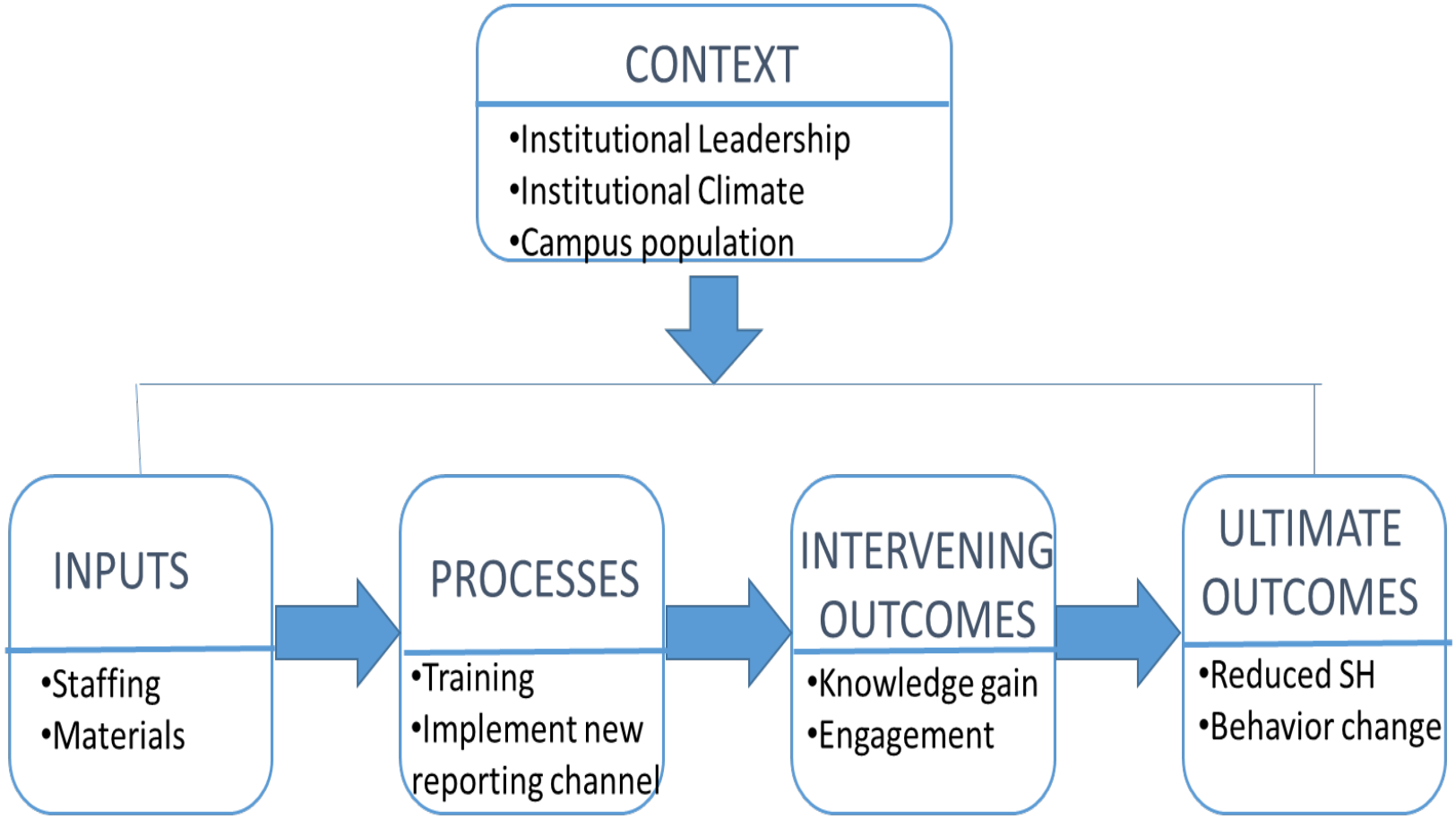

FIGURE 1 Sample logic model.

NOTE: SH = sexual harassment.

TABLE 1 Evaluation of Complex Sexual Harassment Prevention Programs: Multi-Phased Mixed Methods Approach

PROGRAM STAGE	EVALUTION APPROACH
EARLY STAGE	**Primary Type of Evaluation**: Formative, exploratory **Methodology in Use**: Multi-method (primarily non-experimental, qualitative, descriptive) **Purpose of Evaluation Includes**: • Program development (e.g., needs assessment to identify where and what type of programming is most needed) • Program implementation and management (e.g., monitoring whether a new reporting channel is operating as intended) • Program improvement and tightening (e.g., adding or revising a training program's curriculum where deficiencies are identified) • Program theory assessment (e.g., assessing whether the "program theory" or "logic model" requires revision to better understand program impact)
MIDDLE STAGE	**Primary Type of Evaluation**: Formative and preliminary outcomes monitoring **Methodology in Use**: Multi-method (qualitative and quantitative including correlational and quasi-experimental) **Purpose of Evaluation Includes**: • Assess preliminary associations between the program and outcomes, considering the role of contextual factors

LATER STAGE	**Primary Type of Evaluation**: Summative, confirmatory **Methodology in Use**: Multi-method (more rigorous experimental designs supplemented with qualitative and descriptive data on context factors, guided by a refined "program theory" or "logic model") **Purpose of Evaluation Includes**: • Assess program impact accounting for alternative causes identified in prior stages • Determine whether the main causal pathway between the program and targeted outcomes was confirmed. On whom or where did the programs yield the best effects, and under what conditions? • Make summative decisions (e.g., program effectiveness, continuation, expansion)

References

Aguinis, H., and K. Kraiger. 2009. "Benefits of Training and Development for Individuals and Teams, Organizations, and Society." *Annual Review of Psychology* 60: 451−74. https://doi.org/10.1146/annurev.psych.60.110707.163505.

Association of American Universities. 2017. "AAU Campus Activities Report: Combating Sexual Assault and Misconduct." Accessed March 19, 2020. https://www.aau.edu/key-issues/aau-campus-activities-report-introduction.

Berdahl, J.L., and J.S. Raver. 2011. "Sexual harassment". In *APA Handbook of Industrial and Organizational Psychology, vol. 3,* edited by Sheldon Zedeck, 641–69. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12171-018.

Bondestam, F., and M. Lundqvist. 2020. "Sexual Harassment in Higher Education—A Systematic Review." *European Journal of Higher Education*. https://doi.org/10.1080/21568235.2020.1729833.

Chatterji, M. 2004. "Evidence on 'What Works': An Argument for Extended-Term Mixed-Method (ETMM) Evaluation Designs." *Educational Researcher* 33 (9): 3–13.

———. 2007. "Grades of Evidence Variability in Quality of Findings in Effectiveness Studies of Complex Field Interventions." *American Journal of Evaluation* 28 (3): 239–55. https://doi.org/10.1177/1098214007304884.

———. 2016. "Causal Inferences on the Effectiveness of Complex Social Programs: Navigating Assumptions, Sources of Complexity and Evaluation Design Challenges." *Evaluation and Program Planning* 59: 128–40. https://doi.org/10.1016/j.evalprogplan.2016.05.009.

Chen, H., and P.H. Rossi. 1983. "Evaluating with Sense: The Theory-Driven Approach Methodology." *Evaluation Review* 7 (3): 283–302.

Feldblum, CR, and VA Lipnic. 2016. *Select Task Force on the Study of Harassment in the Workplace: Report of Co-Chairs Chai R. Feldbum & Victoria A. Lipnic—Executive Summary and Recommendations*. U.S. Equal Employment Opportunity Commission, Washington, DC. Accessed April 20, 2021. https://www.eeoc.gov/eeoc/task_force/harassment/report_summary.cfm.

Fitzgerald, L.F., and L.M. Cortina. 2018. "Sexual Harassment in Work Organizations: A View from the Twenty-First Century." In *Handbook on the Psychology of Women*, edited by

Jacquelyn W. White and Cheryl B. Travis. Washington, DC: American Psychological Association.

Funnell, S. C. 2000. "Developing and Using a Program Theory Matrix for Program Evaluation and Performance Monitoring." *New Directions for Evaluation* 87(Fall).

Griffin, R. 2012. "A Practitioner Friendly and Scientifically Robust Training Evaluation Approach." *Journal of Workplace Learning* 24 (6): 393–402. https://doi.org/10.1108/13665621211250298.

Haccoun, R.R., and T. Hamtiaux. 1994. "Optimizing Knowledge Tests for Inferring Learning Acquisition Levels in Single Group Training Evaluation Designs: The Internal Referencing Strategy." *Personnel Psychology* 47: 593–604.

Henning, M. A., C. Zhou, P. Adams, F. Moir, J. Hobson, C. Hallett, and C.S. Webster. 2017. "Workplace Harassment among Staff in Higher Education: A Systematic Review." *Asia Pacific Education Review* 18 (4): 521–39. https://doi.org/10.1007/s12564-017-9499-0.

Hunt, C.M., M.J. Davidson, S.L. Fielden, and H. Hoel. 2010. "Reviewing Sexual Harassment in the Workplace—an Intervention Model." *Personnel Review* 39 (5): 655–73. https://doi.org/10.1108/00483481011064190.

Kraiger, K., D. McLinden, and W.J. Casper. 2004. "Collaborative Planning for Training Impact." *Human Resource Management* 43 (4): 337–51. https://doi.org/10.1002/hrm.20028.

McDonald, P., S. Charlesworth, and T. Graham. 2015. "Developing a Framework of Effective Prevention and Response Strategies in Workplace Sexual Harassment." *Asia Pacific Journal of Human Resources* 53 (1): 41–58. https://doi.org/10.1111/1744-7941.12046.

McLinden, D. J. 1995. "Proof, Evidence, and Complexity: Understanding the Impact of Training and Development in Business." *Performance Improvement Quarterly* 8 (3), 3–18.

Medeiros, K., and J. Griffith. 2019. "#Ustoo: How I-O Psychologists Can Extend the Conversation on Sexual Harassment and Sexual Assault through Workplace Training." *Industrial and Organizational Psychology* 12 (01), 1–19. https://doi.org/10.1017/iop.2018.155.

Nation, M., C. Crusto, A. Wandersman, K.L. Kumpfer, D. Seybolt, E. Morrissey-Kane, and K. Davino. 2003. "What Works in Prevention: Principles of Effective Prevention Programs." *American Psychologist* 58 (6/7): 449–56. https://doi.org/10.1037/0003-066X.58.6-7.449.

National Academies of Sciences, Engineering, and Medicine. 2020. *Action Collaborative on*

Preventing Sexual Harassment in Higher Education: Year One Annual Report of Member Activities. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. 2018. *Sexual Harassment of Women: Climate, Culture, and Consequences in Academic Sciences, Engineering, and Medicine.* Washington, DC: The National Academies Press. https://doi.org/10.17226/24994.

Reynolds, A. J. 2005. "Confirmatory Program Evaluation: Applications to Early Childhood Interventions." *Teachers College Record* 107 (10): 2401.

Sackett, P.R., and E.J. Mullen. 1993. "Beyond Formal Experimental Design: Towards an Expanded View of the Training Evaluation Process." *Personnel Psychology* 46: 613−27.

Saks, A.M., and L.A. Burke. 2012. "An Investigation into the Relationship between Training Evaluation and the Transfer of Training." *International Journal of Training and Development* 16 (2): 118–127. https://doi.org/10.1111/j.1468-2419.2011.00397.x.

Salas, E., S.I. Tannenbaum, K. Kraiger, and K.A. Smith-Jentsch. 2012. "The Science of Training and Development in Organizations." *Psychological Science in the Public Interest* 13 (2): 74–101. https://doi.org/10.1177/1529100612436661.

Salas, E., K.A. Wilson, H.A. Priest, and J. Guthrie. 2006. "Training in Organizations: The Design, Delivery and Evaluation of Training Systems." In *Handbook of Human Factors and Ergonomics, 3rd ed.,* edited by Gavriel Salvendy, 472–512. Hoboken, NJ: John Wiley.

Shadish, W. R., T.D. Cook, and D.T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Shaw, E., A. Hegewisch, and C. Hess. 2018 (October). "Sexual Harassment and Assault at Work: Understanding the Costs." Report No. IWPR #B376. Institute for Women's Policy Research.

Tannenbaum, S.I., and S.B. Woods. 1992. "Determining a Strategy for Evaluating Training: Operating within Organizational Constraints." *Human Resource Planning* 15 (2): 63−81.

Wang, G.G., and D. Wilcox. 2006. "Training Evaluation: Knowing More Than Is Practiced." *Advances in Developing Human Resources* 8 (4): 528–539. https://doi.org/10.1177/1523422306293007.

Wood, L., S. Hoefer, M. Kammer-Kerwick, J.R. Parra-Cardona, and N. Busch-Armendariz. 2018. "Sexual Harassment at Institutions of Higher Education: Prevalence, Risk, and Extent." *Journal of Interpersonal Violence*. https://doi.org/10.1177/0886260518791228.