DOI: 10.1002/ev.20619

ORIGINAL ARTICLE

WILEY

Leveraging the Kirkpatrick four-level model to evaluate evaluation capacity building work

Lana Rucks[1] | Lori Wingate[2] | Megan López[2] | Lyssa Wilson Becho[2] | Mike FitzGerald[1] | Kathleen Lis Dean[1]

[1]The Rucks Group, Dayton, Ohio, USA

[2]The Evaluation Center, Western Michigan University, Kalamazoo, Michigan, USA

Correspondence
Lana Rucks, The Rucks Group, Dayton, OH, USA.
Email: lrucks@therucksgroup.com

Abstract

In this article, we reflect on a decade of using the Kirkpatrick four-level model to evaluate a multifaceted evaluation capacity building (ECB) initiative. Traditionally used to assess business training efforts, the Kirkpatrick model encourages evidence to be gathered at four levels: reaction, learning, behavior, and results. We adapted these levels to fit the context and information needs of the EvaluATE project, an ECB initiative funded by the National Science Foundation. As members of the external evaluation and project teams, throughout the article we describe how each level was modified and translated into evaluation questions. Our adapted Kirkpatrick levels are implementation and reach, satisfaction, learning, application, and impact. Using these adapted Kirkpatrick levels to ground our evaluation challenged us to integrate multiple data sources to tell a comprehensive story that served the information needs of the project team and the funder. Overall, we found the Kirkpatrick model to be practical, accessible, and flexible, allowing us to capture the multidimensional aspects of the ECB initiative. However, there are opportunities to enhance the utility of the Kirkpatrick framework by integrating other evaluation approaches, such as culturally responsive and equitable evaluation and principles-focused evaluation.

INTRODUCTION

As the evaluation field takes a closer look at strategies for evaluation capacity building (ECB), we need a parallel effort to examine and share strategies for evaluating the quality and impact of that work (Bourgeois et al., 2023). Among the many potential approaches, the Kirkpatrick four-level model (Kirkpatrick & Kirkpatrick, 2006, 2016) stands out as a promising framework for navigating the evaluation of ECB initiatives. The Kirkpatrick model is commonly used in the business world to evaluate training based on four levels of outcomes (Alsalamah & Callinan, 2020). It is not well established within the program evaluation field's published (or scholarly) literature, where the focus is typically on evaluating social and educational interventions. A notable exception is Guskey's (2000) adaptation of the model for use in education settings. Over more than a decade of work in this area, we have found that the Kirkpatrick model is a simple, flexible, and useful framework for conducting a comprehensive evaluation of our ECB initiative.

In this article, we reflect on and share our experiences using the Kirkpatrick model based on our perspectives and experiences both as project staff within the ECB initiative and as its external evaluation team. First, we provide a brief review of the model, followed by an overview of the context in which we used it. Then, we describe how we used the model as a foundation for evaluating the ECB work. Finally, we conclude with reflections on the lessons learned and suggestions for factors to consider when applying the Kirkpatrick model to other ECB efforts.

OVERVIEW OF THE KIRKPATRICK MODEL

The Kirkpatrick model posits that evidence should be gathered at four levels to demonstrate the impact and value of training endeavors (Kirkpatrick & Kirkpatrick, 2016). Level 1, Reaction, calls for measuring the extent to which participants perceive the training as relevant, engaging, and worthwhile. Level 2, Learning, moves beyond reaction to understand how the training affected participants' knowledge, skills, confidence, and commitment related to the training topic. Level 3, Behavior, aims to determine the extent to which training participants apply what they learned following the training. Kirkpatrick and Kirkpatrick (2016) elaborated on this level to underscore the importance of implementers intentionally integrating support structures and strategies in the trainees' work contexts to increase the likelihood of behavior change. Here, the model calls for monitoring, reinforcing, encouraging, and rewarding the desired behaviors related to the training. It is notable that the model crosses the boundaries of what is traditionally considered the evaluation space and advises on the training intervention and follow-up activities—something that is generally avoided in other program evaluation approaches. Level 4, Results, seeks to understand the extent to which the training achieves its intended outcomes and addresses the underlying need that motivated it. Together, the Kirkpatrick model's four levels require data collection at multiple stages of the training intervention, including immediately after the intervention and through postevent follow-ups (see Figure 1).

OVERVIEW OF THE ECB CONTEXT

In this article, we reflect on our use of the Kirkpatrick model to frame the evaluation of an ECB initiative called EvaluATE. Housed at The Evaluation Center at Western Michigan University, EvaluATE is the evaluation hub for the National Science Foundation's (NSF's) Advanced Technological Education (ATE) program. This NSF program aims to "support the education of technicians for the high-technology fields" by disseminating roughly $75 million in funding annually (NSF, 2021, n.p.). Most grants are awarded to community colleges engaged in developing academic programs, courses, or educational materials for technician education (Marshall et al., 2021). Each of the approximately 400 ATE projects that

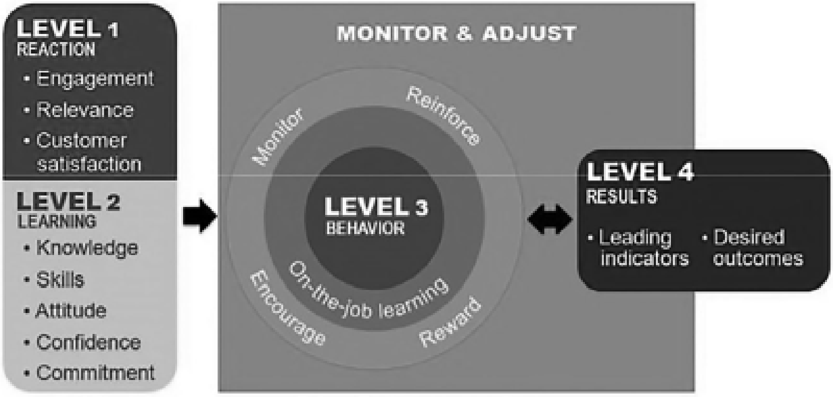

FIGURE 1 The Kirkpatrick model levels and influencing factors. *Note*. From Kirkpatrick and Kirkpatrick (2016, p. 35). Copyright 2016 by Kirkpatrick and Kirkpatrick. Reprinted with permission; license ID 1508699-1.

receive grant funding is required to evaluate its work. NSF funds EvaluATE to support these efforts by developing the evaluation capacity of evaluators and evaluation users associated with the program.

The individuals who comprise EvaluATE's intended audience include project-level evaluators, project leaders and staff, grants professionals, and others involved in projects funded by the ATE program. That is, the audience includes both evaluators and evaluation users and encompasses a wide range of levels of familiarity with evaluation. Audience members also have different needs depending on where they are in their project's lifecycle. EvaluATE supports individuals in, for example, writing evaluation plans for proposals, hiring evaluators, working with evaluators, designing and conducting evaluations, reporting, and using evaluation findings. Projects are typically funded for 3 to 5 years, which means that people are constantly entering and leaving the program, which changes the makeup of EvaluATE's audience. These factors create a continual need for ECB.

EvaluATE has engaged in a range of activities to build evaluation capacity among evaluators and evaluation users. Since its initial funding in 2008, a core component of EvaluATE's ECB strategy has been providing training through online webinars, in-person workshops, and conference presentations. EvaluATE has also developed a robust library of resources to assist users in carrying out evaluation-related tasks. Since 2019, efforts have expanded to community-building efforts that emphasize one-on-one coaching, peer-to-peer learning, and networking through virtual discussions, message boards, and peer gatherings. EvaluATE conducts research on evaluation practices to better understand and provide guidance on topics such as using evaluations for project improvement and integrating diversity, equity, and inclusion into evaluations. In addition to a core set of activities that serve as a foundation for EvaluATE's theory of change, the team has the flexibility to respond to emergent needs through different types of events and materials. Additional information about the EvaluATE project, including its evaluation reports and logic models, is available at www.evalu-ate.org.

How we used the Kirkpatrick model for EvaluATE's evaluation

We have used the Kirkpatrick model as a framework for evaluating EvaluATE since 2012. The model offered an intuitive, straightforward framework for organizing our evaluation questions and data. It is also highly compatible with EvaluATE's logic model, which depicts a progression of outcomes that align with the Kirkpatrick model levels. Like the theoretical assumptions underpinning a logic model, Kirkpatrick and Kirkpatrick (2006) pointed to the need for a chain of evidence about results across the four levels. While the Kirkpatrick model was designed specifically for evaluating training in business contexts, we believe it is amenable to evaluating any intervention designed to influence what people know or do.

The Rucks Group has served as EvaluATE's external evaluator since 2012. The Rucks Group and the EvaluATE team collaborated to lead various aspects of the evaluation, with the EvaluATE team focusing more on the lower levels of the Kirkpatrick model and The Rucks Group focusing on the higher levels. We use multiple types of data to inform the evaluation of EvaluATE's work across the levels defined in the model, including attendance records, website analytics, biennial evaluation surveys, social network analyses, and two interrupted time series studies.

The project uses evaluation findings for accountability purposes, to measure project impact, and to inform project improvements and evolutions (e.g., changes in direction and content of subsequent funding proposals). These findings are presented in two comprehensive evaluation reports, each based on 5 years' worth of data across all four levels (FitzGerald et al., 2024; Rucks et al., 2018).

After selecting the Kirkpatrick model to structure EvaluATE's evaluation, we articulated evaluation questions specific to EvaluATE aligned to each level of the model. In doing so, we realized that some adaptations to the levels were needed to better suit our needs. This adaptation included adding a level to precede the Kirkpatrick model's initial level to capture the reach of our initiative and create a focal point for assessing the project's implementation. We also renamed some of the levels to better match our context. In the following subsections, we describe how we applied the Kirkpatrick model in EvaluATE's evaluation, including the rationale for including and naming an additional level, details about indicators we used at each level, and examples of evaluation findings and how they informed EvaluATE's work. We provide these example findings not to dwell on the details of EvaluATE's, work but to provide a blueprint of the types of statements that result from using the Kirkpatrick model in the evaluation of an ECB initiative (see Figure 2).

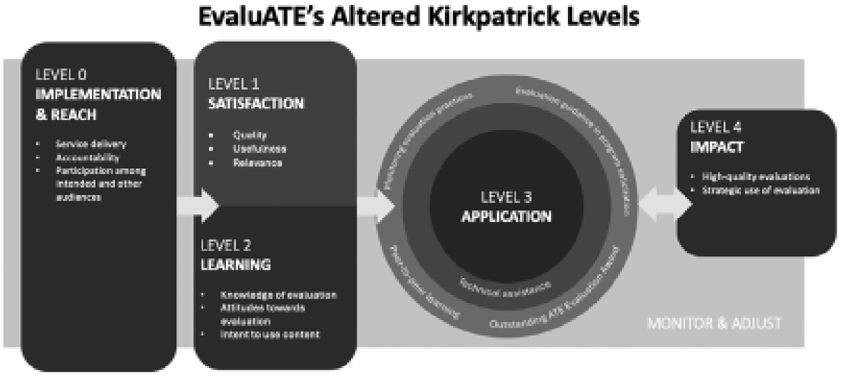

FIGURE 2 EvaluATE's adapted Kirkpatrick levels and influencing factors.

Level 0: Implementation and reach

Referring to EvaluATE's ECB strategy as only training would be an oversimplification. It is a multifaceted initiative involving traditional training through webinars and workshops, as well as resources, one-on-one support, community building, research, and program monitoring. This complexity and the need for basic accountability evidence led us to add a level that focuses on assessing EvaluATE's implementation and reach prior to Kirkpatrick's Level 1, Reaction. The premise behind this level is that to build evaluation capacity for a group of NSF grantees, EvaluATE should understand the extent to which it is reaching its intended audience. NSF does not require ATE program grantees—or their evaluators—to engage in the opportunities or resources that EvaluATE provides. This absence of a requirement differs from other contexts that typically use the Kirkpatrick model, where employees are required or incentivized to attend the training that is being assessed.

Implementation

We have observed that evaluators tend to think about implementation in a dichotomous way—that is, project activities were conducted as intended or not. Instead, EvaluATE's evaluation focuses on *how* it develops and institutes its activities. By focusing on implementation in a directed way, the project team seeks to make data-informed decisions about EvaluATE activities for continual improvement purposes. The evaluation of EvaluATE's implementation is particularly important, considering the project team often innovates new approaches to build ECB for their intended audience. The resulting evaluation questions on implementation are*: To what extent has EvaluATE effectively implemented its activities? What opportunities/successes are being leveraged? What challenges/barriers have emerged? How have they been addressed?*

Reach

Evaluating EvaluATE's reach is important to determine how well the ECB initiative engages its intended audience. At this level, we also measure EvaluATE's reach beyond its primary audience of those funded by the NSF ATE program. While EvaluATE is not accountable for serving entities outside of this group, attracting unaffiliated individuals is a marker of the quality and relevance of its ECB work. The open-access nature of EvaluATE's resources and training events can make it difficult to understand the entirety of who is reached in detail.

To measure reach, the evaluation question asks*: To what extent has EvaluATE reached its intended audience and other audiences?* The main data sources for assessing engagement are attendance records from EvaluATE's various events and website analytics, including pageviews and downloads. Additionally, respondents to EvaluATE's biennial external evaluation survey are asked if they shared any EvaluATE content with others. In 2022, 93% of respondents said they shared EvaluATE materials with their colleagues and networks.

The EvaluATE team leveraged implementation- and reach-related evaluation findings in the case of EvaluATE's evaluation coaching service. The coaching service aimed to provide free, one-on-one evaluation coaching to 100 project staff or evaluators per year. However, fewer than 30 individuals took advantage of the service over 3 years. To address this implementation-related challenge of low participation counts, EvaluATE's staff and external evaluators came together to reflect on the available data around coaching services in one of the quarterly joint meetings that the team calls "Bring Your Own Data (BYOD)."

Each BYOD meeting presents an opportunity to take a deep dive into evaluation data to address an outstanding question or specific concern. Project and evaluation team members review data from both the project's external and internal evaluation efforts to interpret their meaning and cocreate solutions to problems. The event about the coaching services utilized attendance data, information collected from service inquiry forms, and website traffic data. The evaluation team gathered additional data through surveys and interviews with intended audiences to gain insights into their perspectives on the service. Based on this information, we identified opportunities to better market the service, such as reframing the service to highlight the peer-to-peer learning aspect rather than the expert-driven teaching approach that the term "coaching" may imply. We additionally increased marketing efforts that highlighted how the service could be utilized to answer questions specific to the time of year and where people are in an ATE grant cycle. Although these modifications and others were implemented, coaching remained underutilized despite ATE participants' general interest in the service. Therefore, the project team felt confident about discontinuing the service and planning alternative ECB resources and initiatives (see also Table 1).

TABLE 1 EvaluATE's indicators for Level 0: Implementation and reach.

Kirkpatrick model	EvaluATE indicators
	Implementation
N/A	• Activities delivered (e.g., webinars, workshops, conference sessions, webchats, networking events, coaching sessions, resources, white papers, newsletters)
	Reach
	• Event registration and attendance • Resource downloads • Coaching recipients • Web page views • Social media engagement • Newsletter and direct email open rates • Users' reports of sharing information from EvaluATE with others

Level 1: Satisfaction

We renamed Kirkpatrick's reaction level as "satisfaction," because we think satisfaction is a more direct and intuitive descriptor, and it is more common in program evaluation. We operationalized this level as users' perceptions of the quality, relevance, and utility of EvaluATE's activities and materials, including its webinars, workshops, webchats, videos, resource materials, and coaching.

EvaluATE's evaluation question about satisfaction is, *To what extent are EvaluATE's users satisfied with EvaluATE's activities and resources?* This question is addressed with data from participant surveys after each learning event and a biennial external evaluation survey. Similar to the preceding 5-year report (Rucks et al., 2018), findings from EvaluATE's most recent 5-year evaluation report (FitzGerald et al., 2024) concluded that the vast majority of respondents were highly or extremely satisfied with the quality of EvaluATE's webinars and resources.

Because satisfaction ratings have been consistently high over time, there have been moments when the utility of gathering data about satisfaction has been questioned by EvaluATE staff, external evaluators, and advisory board members. That is, we wondered whether, after so many years, we were not learning anything new and whether the findings influenced EvaluATE's work. Ultimately, everyone agreed that EvaluATE should continue to evaluate at this level for three primary reasons. First, the qualitative component of satisfaction measurement (i.e., an open-ended survey item) provided useful information. For example, analysis of qualitative responses from webinar feedback surveys revealed that presenter characteristics and the perceived utility of content were most associated with higher perceptions of webinar quality. In contrast, satisfaction with the technology and design characteristics of webinars most often contributed to lower perceptions of webinar quality.

Second, since EvaluATE often implements new and novel activities and training opportunities, it is important to continue to assess satisfaction with these activities.

Finally, satisfaction measures are useful to consider in combination with other levels of evaluation. In the prior example of the coaching service, it was important to consider whether its underutilization was related to its implementation or, perhaps, poor satisfaction with the service. In this case, those who did receive the service were satisfied with it, suggesting that other factors were at play. Biennial evaluation survey respondents are also asked to rate the utility of each of EvaluATE's activities and materials on a 4-point scale from *not at all useful* to *very useful*. These data are more actionable, with the *very useful* ratings ranging from 13% to 65% across resource types, helping inform EvaluATE's decisions about where to invest (and where not to invest) its resources for maximum benefit (see Table 2).

TABLE 2 EvaluATE's indicators for Level 1: Satisfaction.

Kirkpatrick model	EvaluATE indicators
Reaction	*Satisfaction*
• *Customer satisfaction* • Engagement of participants in learning experiences • *Relevance of training to participant job*	• Rating and description of event quality • Rating and description of what was particularly effective about training events • Rating of usefulness of training events and resources • Satisfaction with coaching services received

Level 2: Learning

Kirkpatrick's Level 2, *Learning*, aligns with EvaluATE's intended short-term outcome to improve evaluation knowledge and skills among evaluators and evaluation users. The evaluation question at this level is, *To what extent has EvaluATE's work led to improvements in users' knowledge of evaluation?* To evaluate learning, we gather data about how EvaluATE users' knowledge changes based on their engagement with EvaluATE. For event-specific evaluations, such as webinars, this is straightforward. We use retrospective pre–post survey questions that ask respondents to rate their knowledge of the webinar topic before and after the webinar on a 7-point scale, from *no knowledge* to *advanced knowledge*.

We chose a retrospective pre–post design to compensate for respondents' possibly overestimating their knowledge at the beginning of a training and, therefore, showing no gain in knowledge using a traditional pre–post design (Little et al., 2020). This format pairs two items: The first asks respondents to think back to before the training session and rate their knowledge of the topic. The second item asks respondents to rate their knowledge after completing the training session. For a broader assessment of learning, we ask biennial evaluation survey respondents, "How much has your understanding of each of the following aspects of your evaluation(s) improved due to information you obtained from EvaluATE?" We ask respondents to report the degree of change they attribute to EvaluATE specifically because it is highly likely they receive information about evaluation from other sources. This is an imperfect yet practical way of dealing with the attribution versus contribution challenge in evaluations that do not employ a control or comparison mechanism.

Retrospective pre–post webinar survey data indicate that most participants report at least a one-step gain in knowledge on a 7-point scale following participation. Considering this gain alongside the previously mentioned findings about webinar quality, we concluded that there is strong evidence for the practical effectiveness of ECB delivered through short learning opportunities (FitzGerald et al., 2024).

EvaluATE uses evaluation findings at this level in two primary ways. First, the findings help the team identify what is going well or needs to be improved, nearly in real time. After every event, EvaluATE team members engage in a postevent reflection that includes reviewing responses to feedback surveys. Events that have less than a one-step gain in knowledge on average are flagged for further investigation. Second, responses to questions on the biennial external evaluation survey about changes in understanding are used to identify topics for future resources or training events (see Table 3).

TABLE 3 EvaluATE's indicators for Level 2: Learning.

Kirkpatrick model	EvaluATE indicators
Learning	*Learning*
• Knowledge and skill • Attitude • Confidence • Commitment	• Rating of change in knowledge (retrospective pre–post) • Rating of change in attitudes toward evaluation • Rating of improvement in understanding evaluation topics • Rating and description of intent to apply information from training events

Level 3: Application

Level 3 of the Kirkpatrick model is called Behavior. We renamed this level Application because our primary focus is on the extent to which EvaluATE's users apply what they learn from our activities, rather than on evaluation behaviors in general. The evaluation question related to the application is, *To what extent has EvaluATE's work prompted users to modify their evaluation practices?* EvaluATE expects its users will adopt the practices recommended in its activities and materials.

Immediately following a webinar, we ask participants an open-ended question related to application—specifically, "If you plan to use something you learned in this webinar, please describe." We found this open-ended question more illuminating than a closed-ended rating about intent to use. In addition to providing insights for this level of the evaluation, this question sheds light on what users found most relevant to their work, whether or not the participants would actually go on to incorporate what they learned into their work. Respondents typically identify a specific tool or strategy from the webinar that they aim to use in their work (e.g., logic modeling, data collection matrix; FitzGerald et al., 2024).

To further measure applicability, the biennial evaluation survey presents respondents with various actions associated with evaluation practice and asks them to indicate the extent to which information they obtained from EvaluATE influenced them to take those actions. The majority of respondents noted a significant influence of EvaluATE's materials on their evaluation-related decisions, such as, in the cases of evaluation users integrating evaluation more fully into their projects (FitzGerald et al., 2024).

As noted previously, the Kirkpatrick model includes guidance for facilitating participants' adoption of behaviors promoted by the training activities through monitoring, reinforcing, encouraging, and rewarding. They call these system levers the "missing link in moving from learning to results" (Kirkpatrick & Kirkpatrick, 2016, p. 59). The EvaluATE team has taken steps to promote the application of learning through multiple avenues that align with these facilitators, as we described next.

Monitoring

EvaluATE conducts a program monitoring survey that more than 90% of the leaders of ATE projects complete. Respondents are asked about their project-level evaluations, among other aspects of their work. Specifically, they are asked if they have an evaluator, how frequently they meet with them, what types of decisions their evaluation findings informed, and with whom they shared their evaluation results. These questions allow EvaluATE to monitor evaluation activity in the program. However, they also send a tacit message to the project leaders that evaluation should be an integral part of their projects.

Reinforcing

A substantial proportion of EvaluATE's work is dedicated to reinforcing learning by providing its users with an array of job aids, including checklists, templates, and quick-reference guides. EvaluATE's monthly newsletter highlights resources that its audience may find especially helpful at different phases of their projects' life cycles.

A pivotal opportunity to reinforce the content of EvaluATE's training activities and materials was to recommend language for the ATE program's request for proposals document (called a program solicitation by the NSF). In 2018, NSF program officers revised the guidance for evaluation in the program solicitation based on suggestions made by EvaluATE (National Science Foundation [NSF], 2018, 2021). The revised guidance aligned more closely with EvaluATE's recommendations in its webinars and resource materials.

Encouraging

According to the Kirkpatrick model, key mechanisms for encouragement are mentoring and coaching, typically delivered by managers and supervisors (Kirkpatrick & Kirkpatrick, 2016). EvaluATE's role as an evaluation hub is unlike that of a manager or supervisor, so its options for encouragement are limited. Furthermore, as explained previously, a key lesson learned from EvaluATE's implementation evaluation was that its audience was not inclined to engage in one-on-one coaching. Kirkpatrick and Kirkpatrick (2016) reported that they had observed a shift away from a supervisor- and manager-led mentoring toward peer mentoring. This shift aligns with EvaluATE's work to create a community of evaluators and evaluation users who can provide feedback, support, and guidance among themselves. From 2019 to 2021, EvaluATE's evaluators conducted a social network analysis of evaluators working with ATE projects to track community ties. Most evaluators reported interacting with other evaluators by exchanging guidance, information, or resources (FitzGerald & Siwierka, 2022). We also saw the frequency of interaction between evaluators increase between 2019 and 2021. Despite the continual turnover of the people who evaluate ATE projects and the mass disruptions caused by the COVID-19 pandemic, the network maintained a general level of connectedness—and a core group of individuals who were highly connected—throughout these years (FitzGerald & Siwierka, 2022).

Rewarding

The ultimate reward for excellence in evaluation within the ATE program is for evaluations to deliver actionable information that projects can use to improve their work and demonstrate impact. Strong evaluative evidence can help projects obtain additional grant dollars and garner political support for their efforts. However, these types of rewards are beyond EvaluATE's control. In place of those leverage points, EvaluATE created an award program to recognize and promote excellence in evaluation in the ATE program. The winners are announced publicly at the program's annual conference. EvaluATE posts the reports associated with the winning evaluations on its website and highlights them in newsletter and webinars. Thus, the award program not only rewards strong evaluations, but the winners serve as exemplars for other evaluators in the program, which, in turn, potentially also serves the purposes of encouragement and reinforcement (see Table 4).

TABLE 4 EvaluATE's indicators for Level 3: Application.

Kirkpatrick model	EvaluATE indicators
Behavior	*Application*
• Critical behaviors (i.e., "the few, specific actions, which, if performed consistently on the job, will have the biggest impact") • Required drivers (i.e., "processes and systems that reinforce, monitor, encourage and reward performance of critical behaviors") • On-the-job learning	• Rating and description of EvaluATE's influence on evaluation practice • Evaluation practices maintained by projects • Changes in the number, types, and frequency of interaction among evaluators of Advanced Technological Education projects • Applications to and recipients of the Outstanding Advanced Technological Education Evaluation Award

Level 4: Impact

In the Kirkpatrick model, Level 4 is called Results. In the business world, this might refer to increased revenue or improved productivity as a result of training. However, in the context of training on evaluation, this was a potentially confusing term, given that evaluators often refer to "results" as findings from the evaluation. Therefore, we renamed this level Impact to underscore that the focus is on long-term outcomes and the deep impact that EvaluATE's ECB initiative aims to have.

Kirkpatrick and Kirkpatrick (2016) defined this level as "the reason that training is performed" (p. 60). EvaluATE's primary reason for existence is to improve the quality of evaluations in the ATE program so that evaluation plays an ongoing, strategic role in advancing the program's goals. Therefore, EvaluATE operationalized Level 4 as the extent to which users improve aspects of their evaluations.

In the biennial evaluation survey, we ask respondents to estimate how much aspects of their evaluations have improved because of information they obtained from EvaluATE. We specifically ask about evaluation plans, logic models, instruments, data collection, analysis or interpretation, data visualization, and reports. Similar to the questions about learning, these questions ask respondents to report the degree of improvement on a 4-point scale from *no change* to *greatly improved*. Across multiple years, most respondents indicated that information had greatly or moderately improved their evaluations.

Over time, we increasingly felt the need to identify ways to attribute changes in evaluations to EvaluATE's ECB work without relying on self-reported data. In 2021, we conducted an interrupted time series study examining evaluation plans from 169 ATE proposals funded between 2004 and 2017 (Wingate et al., 2022a, 2022b). The "interruption" or marker between pre- and post-EvaluATE occurred in 2010, when EvaluATE was first identified as a source of information about evaluation in the ATE program solicitation (i.e., the first broad announcement about the evaluation hub to all program grant seekers). The study corroborated what we had found in the self-reported data. Specifically, we found that the level of detail related to four of the six elements of evaluation plans that we examined correlated with the proposal's award year. That is, the evaluation plans became more detailed over time and more reflective of what EvaluATE had recommended to its users via webinars and resource materials. While these findings are consistent with previous findings, we acknowledge that this study was not designed to provide unequivocal evidence of EvaluATE's role, or lack thereof, in these changes (see Table 5).

TABLE 5 EvaluATE's indicators for Level 4: Impact.

Kirkpatrick model	**EvaluATE indicators**
Results	*Impact*
• High-level organizational mission • Leading indicators "help to bridge the gap between individual initiatives and efforts, and organizational results" (e.g., customer satisfaction, employee engagement, sales volume, cost containment, quality, market share)	• Change in the quality of evaluation plans in proposals for the Advanced Technician Education program • Rating and description of change in the quality of evaluation attributable to EvaluATE

DISCUSSION

In this article, we describe our experience using the Kirkpatrick model as a framework for evaluating ECB. We discuss the adaptations we made to the levels and how they were operationalized to fit the design and intentions of an ECB intervention. In this concluding section, we share our reflections and lessons learned for the benefit of others interested in applying this evaluation approach to their ECB work.

All evaluation approaches (whether they are called approaches, models, frameworks, lenses, or something else) serve as heuristics to aid evaluators in determining how to focus and carry out their evaluations. Evaluation approaches may also inform things like what to prioritize or emphasize in an evaluation, what to measure, how to frame evaluation questions, and whom to involve, among other things (Montrosse-Moorhead et al., 2024). Generally speaking, current evaluation approaches are not expected to be used as rigid, prescriptive recipes for conducting evaluations (Fitzpatrick et al., 2023).

With this view of evaluation approaches in mind, we found that grounding EvaluATE's evaluation in the Kirkpatrick model allowed for context-specific adaptations that reflect the multidimensional nature of ECB, inspired a varied and robust data collection strategy, and generated insights that fulfilled the information needs of the project team and our funders. We expand on these themes below.

- **Captured the multidimensional nature of ECB efforts.** The effectiveness of an ECB initiative cannot be measured by a single construct, especially when looking at a multifaceted initiative that seeks change on both individual and community levels. This complexity is, of course, acknowledged in current models of the ECB (Labin, 2014; Preskill & Boyle, 2008); however, it also needs to be reflected in the evaluation of the ECB intervention. The adapted Kirkpatrick levels provided such a multidimensional approach to ECB evaluation.

- **Provided motivation and a structure to integrate multiple data sources**. The adapted Kirkpatrick levels challenged us to consider what kinds of evidence best spoke to each dimension. This led us to branch out from a traditional survey in which respondents self-reported behaviors to an interrupted time series study of evaluation artifacts and social network analysis. It also provided a framework to weave a variety of data sources together to tell a comprehensive story about EvaluATE's impact on the evaluation capacity of the ATE program.
- **Ensured evaluation findings met the needs of both the project team and the funder**. Our adapted Kirkpatrick levels (Implementation and Reach, Satisfaction, Learning, Application, and Impact) span questions of process and outcomes. The model provided a framework to zoom in and out as needed. For example, the EvaluATE team could focus on individual feedback forms or a set of questions to build a nuanced understanding of how to improve training activities. We could also communicate our ECB efforts' quality, effectiveness, and overall impact on the ATE community to our funder, the NSF. As the funder, they are most interested in hearing about the impact of our activities as a result of their investment.

Although this framework is highly flexible and adaptable, we have found some areas related to the evaluation of our ECB work that we would like to explore further. As with any approach that provides direction and guidance on what to focus on, it inevitably means that other concepts are not given attention. Two areas in particular stand out, as we describe next.

While EvaluATE has examined how the concepts of diversity, equity, and inclusion are being measured in ATE evaluations (Boyce et al., 2022), we have yet to turn that critical lens on evaluating our ECB work. Our adapted Kirkpatrick levels currently do not remind us to hold space for questions about diversity, equity, or inclusion. Nor does this approach speak to what it means to conduct an evaluation of an ECB initiative in a culturally responsive or equitable way. Further research and experience integrating culturally responsive and equitable evaluation (CREE) would benefit both ECB practitioners and participants.

Similarly, our adapted Kirkpatrick levels do not currently call for questions about our project's values or principles and how they have been enacted. After the EvaluATE team engaged in a strategic process to revisit the project's mission and vision, we articulated guiding values. These discussions highlighted the need to hold the project accountable for carrying out these values in EvaluATE's work and demonstrating their importance and impact. The field would benefit from the future integration of a principles-focused evaluation approach to ECB initiatives.

The Kirkpatrick model has served as a practical, accessible, and adaptable framework for EvaluATE's evaluation efforts for over a decade. With some adaptation to the original levels, the structure fits quite naturally and intuitively with the ECB initiative. As EvaluATE's activities continue to grow, there is an opportunity for its evaluation to grow and evolve alongside the project. Evaluation approaches are often woven together to fit the context and needs of the intervention (Bledsoe & Graham, 2005; Montrosse-Moorhead et al., 2024). We look forward to seeing the woven tapestry the evaluation of EvaluATE becomes as we integrate perspectives from other approaches and continue to innovate our methods.

ORCID

Lana Rucks https://orcid.org/0000-0002-8952-2686
Lori Wingate https://orcid.org/0000-0003-4630-9604
Megan López https://orcid.org/0000-0002-5723-7143
Lyssa Wilson Becho https://orcid.org/0000-0002-3562-7933
Mike FitzGerald https://orcid.org/0000-0002-8546-5506
Kathleen Lis Dean https://orcid.org/0009-0009-9366-495X

REFERENCES

Alsalamah, A., & Callinan, C. (2020). The Kirkpatrick model for training evaluation: Bibliometric analysis after 60 years (1959–2020). *Industrial and Commercial Training*, *54*(1), 36–63. https://doi.org/10.1108/ICT-12-2020-0115

Bledsoe, K. L., & Graham, J. A. (2005). The use of multiple evaluation approaches in program evaluation. *American Journal of Evaluation*, *26*(3), 302–319.

Bourgeois, I., Lemire, S. T., Fierro, L. A., Castleman, A. M., & Cho, M. (2023). Laying a solid foundation for the next generation of evaluation capacity building: Findings from an integrative review. *American Journal of Evaluation*, *44*(1), 29–49. https://doi.org/10.1177/10982140221106991

Boyce, A. S., Tovey, T. L., Onwuka, O., Moller, J. R., Clark, T., & Smith, A. (2022). Exploring NSF-funded evaluators' and principal investigators' definitions and measurement of diversity, equity, and inclusion. *American Journal of Evaluation*, *44*(1), 50–73. https://doi.org/10.1177/10982140221108662

FitzGerald, M., Becho, L. W., López, M., Dean, K., Wingate, L. A., & Rucks, L. J. (2024). *Evaluation of EvaluATE: 2018–23*. The Evaluation Center, Western Michigan University. http://www.evalu-ate.org/about/evaluation/

FitzGerald, M., & Siwierka, J. (2022). *ATE evaluation community social network analysis results for year 3*. The Rucks Group. https://evalu-ate.org/about/evaluation/

Fitzpatrick, J. L., Sanders, J. R., Worthen, B. R., & Wingate, L. A. (2023). *Program evaluation: Alternative approaches and practical guidelines* (5th ed.). Pearson.

Guskey, T. R. (2000). *Evaluating professional development*. Corwin.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). Berrett-Koehler Publishers.

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. ATD Press.

Labin, S. N. (2014). Developing common measures in evaluation capacity building: An iterative science and practice process. *American Journal of Evaluation*, *35*(1), 107–115. https://doi.org/10.1177/1098214013499965

Little, T. D., Chang, R., Gorrall, B. K., Waggenspack, L., Fukuda, E., Allen, P. J., & Noam, G. G. (2020). The retrospective pretest–posttest design redux: On its validity as an alternative to traditional pretest–posttest measurement. *International Journal of Behavioral Development*, *44*(2), 175–183. https://doi.org/10.1177/0165025419877973

Marshall, V. A., Sturgis, E., Becho, L. W., Wingate, L. A., & Gullickson, A. (2021). *ATE Survey: 2021 report*. The Evaluation Center, Western Michigan University. https://atesurvey.evalu-ate.org/

Montrosse-Moorhead, B., Schröter, D., & Becho, L. W. (2024). The garden of evaluation approaches. *American Journal of Evaluation*, *45*(2), 166–185. https://doi.org/10.1177/10982140231216667

National Science Foundation. (2018). *Advanced technological education: Program solicitation*. (NSF 18–571). National Science Foundation. https://www.nsf.gov/pubs/2018/nsf18571/nsf18571.htm

National Science Foundation. (2021). *Advanced technological education: Program solicitation*. (NSF 21–598). National Science Foundation. https://new.nsf.gov/funding/opportunities/advanced-technological-education-ate/nsf21-598/solicitation

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation*, *29*(4), 443–459. https://doi.org/10.1177/1098214008324182

Rucks, L. J., Wingate, L. A., FitzGerald, M., Schwob, J., Becho, L., & Perk, E. (2018). *Evaluation of EvaluATE: 2012–17*. The Evaluation Center, Western Michigan University. http://www.evalu-ate.org/about/evaluation/

Wingate, L. A., Robertson, K., FitzGerald, M., Rucks, L. J., Tsuzaki, T., Clasen, C., & Schwob, J. (2022a). *Characteristics of evaluation plans in ATE proposals over time*. The Evaluation Center, Western Michigan University. https://evalu-ate.org/wp-content/uploads/2021/07/proposal_one-pager_final-1.pdf

Wingate, L. A., Robertson, K., FitzGerald, M., Rucks, L. J., Tsuzaki, T., Clasen, C., & Schwob, J. (2022b). Thinking outside the self-report: Using evaluation plans to assess evaluation capacity building. *American Journal of Evaluation*, *43*(4), 515–538. https://doi.org/10.1177/10982140211062884

How to cite this article: Rucks, L., Wingate, L., López, M., Wilson Becho, L., FitzGerald, M., & Lis Dean, K. (2024). Leveraging the Kirkpatrick four-level model to evaluate evaluation capacity building work. *New Directions for Evaluation, 2024*, 81–94. https://doi.org/10.1002/ev.20619

AUTHOR BIOGRAPHIES

Lana Rucks is the founder and principal consultant at The Rucks Group, LLC.

Lori Wingate is the executive director of The Evaluation Center at Western Michigan University.

Megan López is a senior research associate at The Evaluation Center at Western Michigan University.

Lyssa Wilson Becho is a principal research associate at The Evaluation Center at Western Michigan University.

Mike FitzGerald is a senior research and evaluation associate at The Rucks Group, LLC.

Kathleen Dean is a senior research and evaluation associate at The Rucks Group, LLC.