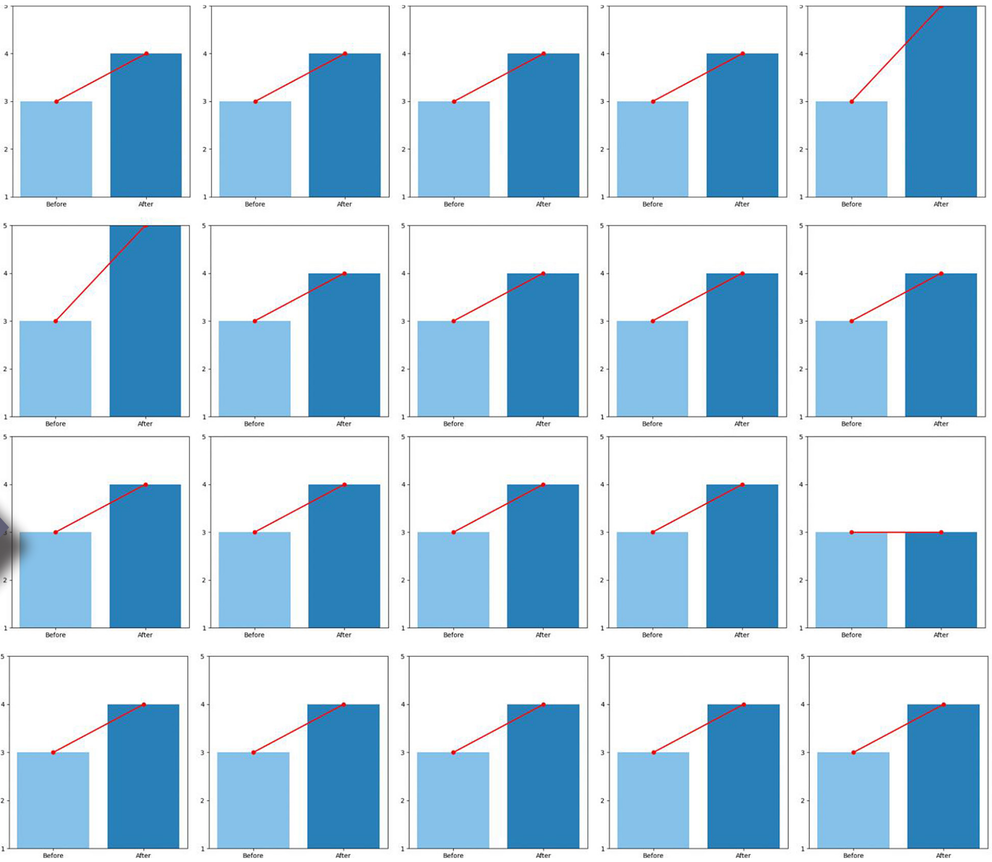
Columbia Climate School
National Center for Disaster Preparedness



ENSURING EMERGENCY MANAGEMENT TRAINING TRANSLATES INTO ACTION:

CHALLENGES AND SOLUTIONS FOR MEASURING LEVEL 3 OF THE KIRKPATRICK MODEL

NATIONAL CENTER FOR DISASTER PREPAREDNESS,
COLUMBIA CLIMATE SCHOOL, COLUMBIA UNIVERSITY

NOVEMBER 2024

RECOMMENDED CITATION

Chandler, T., Hendra, R., Huang, S., DeVincenzo, J.L., Yang, Y. (2024). Ensuring Emergency Management Training Translates Into Action: Challenges and Solutions for Measuring Level 3 of the Kirkpatrick Model. *National Center for Disaster Preparedness, Columbia Climate School, Columbia University*.

REPORT AUTHORS

Thomas Chandler, PhD

National Center for Disaster Preparedness, Columbia Climate School, Columbia University

Richard Hendra, PhD

The New School

Shuyang Huang, MS, MArch

National Center for Disaster Preparedness, Columbia Climate School, Columbia University

Josh DeVincenzo, EdD

National Center for Disaster Preparedness, Columbia Climate School, Columbia University

Yaxuan Yang, MA

National Center for Disaster Preparedness, Columbia Climate School, Columbia University

475 Riverside Drive, Suite 401
New York, NY 10115-0095
ncdp@columbia.edu
ncdp.columbia.edu

TABLE OF CONTENTS

I. INTRODUCTION

The Importance of Evaluating Training Programs

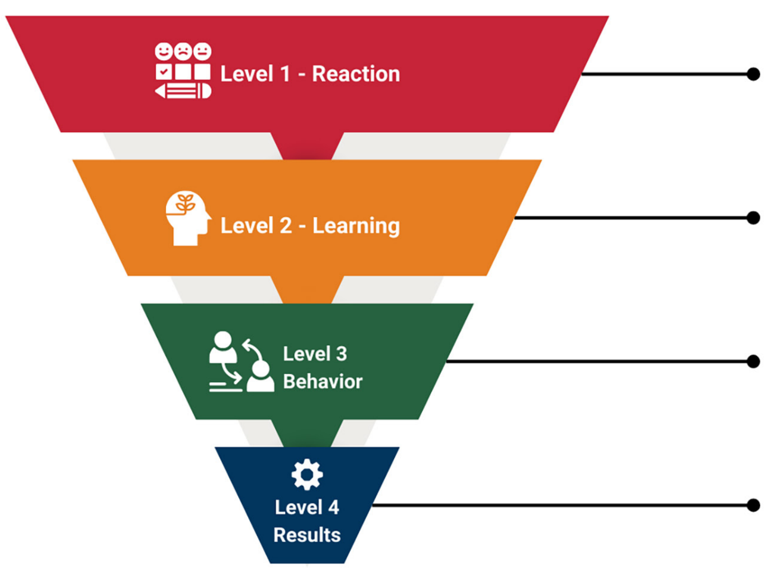
As disasters in the United States increase in frequency and intensity, preparedness is now more critical than ever. To that end, training disaster management professionals on hazard mitigation, preparation, response, and recovery is a cornerstone of new planning efforts. Annually, the U.S. government, at all levels, invests significant resources in training programs. Taking advantage of "blue sky days" to learn best practices, work through real-life scenarios, and practice using tools and templates is essential.

But how do we know if a given emergency management training has actually been effective? Are participants really incorporating what they have learned into their everyday work practices or even as part of an actual disaster response? This paper addresses the challenges and opportunities of incorporating the Kirkpatrick Model's Level 3 evaluation into determining whether a training program has been successful in the long term.

Background

The Kirkpatrick Model (Kirkpatrick, 1959) is one of the most established frameworks for evaluating the effectiveness of emergency management training programs. Consisting of four levels: Reaction, Learning, Behavior, and Results, each level builds on the previous one, offering a deeper understanding of the training's impact.

The Kirkpatrick Model



Level 1 - Reaction

Level 1 (Reaction) evaluates how participants responded to the training (e.g., did participants enjoy their time in the classroom learning the material?). This level is the easiest to assess and can be measured during the training itself.

Level 2 - Learning

Level 2 (Learning) determines if the participants actually learned the material presented (e.g., did participants gain new knowledge on a post-test aligned with learning objectives in the class?).

Level 3 - Behavior

Level 3 (Behavior) considers whether participants have applied what they learned on the job (e.g., have they actually implemented what they learned?).

Level 4 - Results

Level 4 (Results) evaluates if the training has positively impacted key outcomes (e.g., did the training affect disaster management outcomes for an entire organization?).

I. INTRODUCTION

When considering the implications of evaluation findings, these levels are not of equal importance nor of equal difficulty methodologically. As mentioned, measuring Levels 1 and 2 is relatively straightforward. Measuring Level 3, on the other hand, is more difficult and, at the same time, most vital because it focuses on what people do with what they have learned rather than just how they "feel" about the training or what they learned as defined by a given test. Finally, Level 4, while a laudable goal, is the most difficult to measure as it is extremely challenging to disentangle what contribution a given training has made to enhancing an entire organization's effectiveness.

Given that Level 3 is of such crucial importance, it has become a mainstay of program evaluation among U.S. federal agencies such as the Department of Defense (DoD) and the Federal Emergency Management Agency (FEMA), particularly because it can help program leadership justify previous training expenditures as well as future budget allocations. Also known as the "Learning Transfer" level, the concept was first introduced by Donald Kirkpatrick in his book *Evaluating Training Programs* (Kirkpatrick, 1959). Yet, while Kirkpatrick Level 3 offers valuable benefits, such as assessing the practical application of learning and providing actionable feedback, it also presents many challenges. These include difficulties in isolating the effects of training from other influences and the potential complexity of gathering enough reliable data. Understanding both the advantages and the limitations of using the Kirkpatrick Level 3 framework is essential for organizations seeking to maximize the effectiveness of their training initiatives.

II. Data Collection Challenges

Although the Kirkpatrick Level 3 evaluation is widely used and highly regarded in practice (Phillips, 1991), researchers have noted several roadblocks to enabling it to bear fruitful results, particularly because much of the data are retrospective and self-reported via surveys or interviews (Reio et al., 2017). This approach is subject to a number of biases and limitations, which can, in turn, affect the accuracy and reliability of the findings, thereby providing organizations with false impressions about the impact of their training programs on actual job performance. Given that Kirkpatrick Level 3 evaluations are predominantly disseminated via post-training surveys, associated data collection and analysis issues will be addressed in the next section in relation to social desirability bias, recall bias, low survey response rates, and threats to the validity of causal inferences.

Social Desirability Bias

People who respond to surveys are motivated by a range of factors. Yet, there is a tendency to divorce survey responses from human factors. Tourangeauet et al. 2000, in "The Psychology of Survey Response," reminds us that this form of response is a social act. People want their responses to add value; in addition, they wish to show empathy.

One major concern with regard to survey responses is social desirability bias, whereby participants tend to respond to surveys by providing the response that they think will be viewed most favorably (Fisher, 1993). This is a pervasive problem in survey research and has been documented widely in the literature. As an example, in drug-use studies, it is well established that environmental or biological samples indicate much higher use than is documented in surveys (Colón et al., 2001). Similarly, Yudkowsky et al. (2019) asserted that self-reported data are subject to response bias in training environments, frequently occurring when learners do not answer questions truthfully or accurately. This is often due to the same phenomenon - individuals seek to provide socially acceptable answers (in this case, about the role of training in their job performance). Learner input in relation to on-the-job decision-making and perceptions of personal competence is especially difficult to document or analyze, given that there is frequently a desire to deny personal responsibility or even culpability in stressful, complex situations, such as disaster response, in which there are legal ramifications or even lives at stake (Sui & Humphreys, 2015; Symons & Johnson, 1997).

For these reasons, it is important to consider how social desirability bias can undermine Kirkpatrick's Level 3 measurement. Training participants typically understand the learning objectives of a particular training, why a survey is being administered, and what the most socially acceptable answer is. In the context of Kirkpatrick Level 3, the more socially desirable response is the one that indicates that they are using the material they were trained on while performing their job tasks. For instance, consider the following example from a disaster housing recovery training: A participant receives a survey after the training and reads a question about starting a disaster housing task force. The participant remembers that this was an important theme during the training, yet is well aware that the agency has not made any real movement towards starting a task force. The participant understands that the survey is being used to judge the quality of the training and doesn't want the agency to be perceived negatively. Consequently, they select a "yes" response, as this is generally considered the most desirable response. As a result, the trainer's instructional performance is checked as being satisfactory, and the participant's agency looks more aligned with best practices.

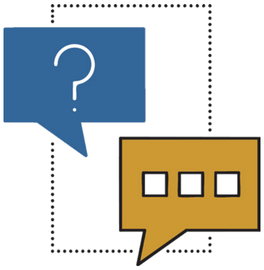
II. Data Collection Challenges

There are numerous ways to rationalize such a response. First, it's commonly assumed that follow-up tracking of responses will be minimal. Further, although participants haven't really started a task force, they have talked about it, and the trainee has some notes on the topics that were discussed during the training. Can't that be considered a "start"? This simple scenario is a clear illustration of social desirability bias. Put simply, survey respondents often report what they think interviewers want to hear (Chandon et al., 2005).

A common means of addressing social desirability bias is to assure respondents of anonymity (Paulhus, 1993). However, while this likely provides some protection, it is unlikely sufficient to address the issue completely. More elaborate means of addressing social desirability bias include list randomization (Brownback & Novotny, 2018), whereby survey respondents are randomly assigned to receive lists of statements of different lengths. For example, those responding to List A might be asked if they agree with three noncontroversial statements about the training, such as "I attended the training program,"; "I discussed it with my supervisor,"; and "I have shared materials with colleagues." List B would include these same statements, along with more specific Kirkpatrick Level 3 statements: "I have applied the training content to my everyday work." Respondents would indicate the number of statements they agree with rather than specifying which statement they agree with. The logic behind list randomization is that it is possible to isolate the percentage responding to the more controversial question by comparing affirmative responses across these two lists without ever asking the question directly.

Another means of addressing social desirability bias is to collect objective observations (using some of the performance tracking data noted above, for example). Other approaches are discussed in Podsakoff et al., 2012. We will discuss the latter further in the section on Interrupted Time Series Designs.

Recall Bias



A related problem is recall bias. Vaidya et al. (2023) found that learners can be subject to memory bias, sometimes occurring when they have difficulty recalling past events or experiences accurately. This can lead to errors in self-reporting, particularly when trying to measure long-term outcomes or impacts. Self-reported data may also not accurately reflect the true nature of events, given that the data are based on respondents' perceptions rather than objective measurements. Recently, several lines of research within the fields of cognitive neuroscience and psychology have focused on how recollection of value-relevant information serves learning and decision-making (Bakkour et al., 2018; Bornstein & Norman, 2017). Their findings suggest that the relationship between training outcomes and on-the-job performance is often murkier than assumed.

Recall Bias *(continued)*

It is not difficult to imagine how this might play out in a real-world setting. Consider a situation in which a training participant is working at an agency that is implementing a new housing planning process. The participant is an emergency manager who is not usually a part of agency decision-making. After completing the training, the participant gained familiarity with additional planning processes, layered on top of various changes to standard operating procedures (SOPs). Then, six months after the training, the participant receives a survey asking: "Since the training, has your agency started a new housing planning process?" When getting to this question, the manager quickly checks off "yes". It's been a hectic year, and they know they started a plan, and they know they started training. They don't really remember which came first, so they just check off that "yes" they started the planning process after the training.

The example provides a realistic portrait of the dynamic of survey response and speaks to the reality of recall bias. For example, it is difficult for respondents to remember sequences, particularly as more time has passed, leading to a tendency called "telescoping," in which respondents overestimate the number of events that have happened in a given time period. One way to circumvent this type of bias is through "anchoring," in which surveyors attempt to anchor memory to landmark events that might have happened since the training (including milestones or major disasters that have recently occurred).

Low Survey Response Rates

A low survey response rate can also significantly undermine the effectiveness of a Kirkpatrick Level 3 evaluation, which relies heavily on collecting feedback to assess behavioral changes in the workplace. When only a small portion of participants respond, the data gathered may not accurately represent the entire group, leading to biased or incomplete insights. It is plausible, for example, that those who had a positive training experience because they were able to implement training concepts are more likely to respond, resulting in an inaccurate assessment of whether the training has been effectively applied on the job. Additionally, a low response rate can hinder the ability to identify trends or common challenges faced by employees in implementing the new skills, ultimately limiting the evaluation's reliability and the organization's capacity to make informed decisions about future training programs (Keeter et al., 2017). Although some studies have argued that the relationship between survey response rates and survey data quality is weak (Hendra & Hill, 2019), this issue continues to be the most challenging for organizations to overcome. In Section IV, we provide guidelines and best practices when implementing web and phone-based surveys to improve response rates, noting that learning management system technologies can help in the data collection process.

III. Threats to Validity of Causal Inferences

Even if good data are available, there are numerous threats to the validity of causal inferences related to measuring the value-added of training content. As mentioned, Kirkpatrick Level 3 is focused on whether a given training has led to behavior change. One challenge with this approach is that while activities that align with training learning objectives may be in place, they may have happened anyway - without the training - a problem known in research methodology as the *counterfactual*. In order to know the true effect of the training on behavior, therefore, it is essential to assess what would have happened if the training had not occurred.

It is often assumed that if participants implement specific steps after the training, it is evidence that the training is being implemented. However, the counterfactual problem is more subtle and difficult to address than one might suppose. To understand why this is so, we introduce a pernicious problem in evaluation research known as *selection bias*.

It is easy to see how selection bias can threaten causal inferences in this context. First, consider the motivation for attending a training course. Going back to the previous housing program training example, consider that a new manager is overseeing a project and wants to bolster the disaster housing mission at the agency. The manager might implement a change management agenda in which staff are asked to work on a new disaster housing plan. They might also be asked to attend training on the subject with the goal that the team can then improve the plan with the concepts learned. A few months later, the training team responded to a survey asking whether several activities captured in the training were being implemented. One of the questions asked whether the team had started on a disaster housing plan. The team accurately responds "yes" to the fact that they have begun this process. For many, this can be considered a Level 3 Kirkpatrick "win." However, the problem here is that this team had already started the plan before the training, so it is not accurate to attribute the fact that they are implementing a housing plan to the training. In fact, causality has run in the reverse direction. It was the housing plan itself that revealed the need for the training.

Constructing a Counterfactual

In principle, it might be possible to statistically control for factors such as selection bias and possibly even some of the response bias factors mentioned above. That is, it is possible to show that - in theory - if one can create a perfect statistical model laying out all the differences between responders and non-responders, selection bias can be corrected. However, it is well known that constructing such a model is very difficult, if not impossible. While statistical methods (such as the classic "Heckman correction") exist, data on all of the various psychological and behavioral factors that might distinguish respondents from non-respondents are rarely available. However, we can address omitted variable bias through a number of different research designs.

III. Threats to Validity of Causal Inferences

Addressing Omitted Variable Bias

Randomized Controlled Trial: Gold Standard but Difficult to Implement.

The most rigorous means of addressing selection bias is to conduct a randomized controlled trial (RCT). In an RCT, trainees are randomized to either participate or not participate in a given training. Because the only difference between participants and non-participants is the random assignment, researchers can be assured that, in expectation, there will be no difference between these groups. Hence, any differences that emerge would be due to the training intervention rather than selection bias. While this design has great appeal, it is very difficult to implement in practice. For example, many disaster management training programs struggle to enroll enough participants. Randomization would reduce attendance even more due to the need to assign participants to a control group. Further, some organizations would be reluctant to merely sign up only for the chance to attend a training.

Crossover Designs: A Means of Making Randomization More Feasible.

Crossover designs address some of the feasibility concerns presented by randomized controlled trials. In this approach, everyone receives the training eventually, but in different time-released cohorts. Consider, for example, a state emergency management agency wants to train all of its regions in disaster housing preparation. It may be feasible to randomize the regions and send a first cohort to the training, followed by a second cohort several months later. Randomization is best done in this scenario at the office level to guard against spillover effects (in which training content could spill over by changing practices). This design has been used to evaluate training in the past and is more feasible than a traditional RCT. However, it requires a fair amount of coordination, and agencies might not have the bandwidth to train in cohorts this way.

Interrupted Time Series Design.

A natural way to measure the effect of a training program on employee behavior (Kirkpatrick's Level 3) is to measure employee behavior over time. A long-standing technique for measuring impact over time is the interrupted time series design (Cook et al., 2002). With this design, everyone has the opportunity to experience a given training, but it represents a step down in terms of validity. Using this design, outcomes are measured for several time periods before the intervention so that a time trend can be developed (see Bloom, 1999 for statistical power considerations). Then, the technique measures whether the intervention (i.e., the training) affects the slope of the time trend in the outcome measured. The difference between the projected slope (which provides the counterfactual) and the actual slope is the impact.

While this approach has promise, the key challenge is collecting enough data on the participant behaviors that are implicated in the training. These data might not exist. Fortunately, important strides have been made in performance management, including frameworks such as key performance indicators (Parmenter, 2005) and objectives and key results (Doerr, 2018), that are increasingly stored in management information systems (MIS). If the training logic model maps to these same outcomes, then a time series analysis is feasible.

III. Threats to Validity of Causal Inferences

Interrupted Time Series Design *(continued)*

One of the main threats to the validity of a time series design is called the "history threat," which refers to the fact that other factors that are contemporaneous with the training can affect employee behaviors. Suppose, for example, that an employee took a training class during the COVID-19 pandemic. While they were in the training, the workforce was dramatically restructured to a work-at-home model, which would very likely have effects on employee behavior outcomes that could be larger than the training effect.

There are various other threats, such as maturation and regression to the mean, as discussed in Cook et al. (2002). One way to try to reinforce the validity of the interrupted time series design against these threats is to follow another group of staff who were not in the training (perhaps from a different office) to see how their outcomes change. This design is called a *comparative interrupted time series (CITS)*. Continuing with the COVID-19 example, this would provide some ability to control for the contextual effects of the pandemic because both the training and the comparison employees were experiencing the same events. The CITS design would, therefore, control for this particular threat to validity. Such a design is easier to implement than randomization but is considered less rigorous.

There are various permutations to the basic strategy of looking at change over time. The weakest time series design involves looking at just one post-training time point. This model is weakest because there is no evidence to suggest if the behaviors were preexisting or due to the training itself. An improvement (but still low rigor) approach would be to change over time using a pre/post design. A further improvement would involve including a pre/post measurement for a comparison group of trainees. This design, commonly referred to as a difference-in-differences design, is heavily used in evaluation research (Angrist & Pischke, 2014). Including several observations (using a time series or panel design) would be better than just two observations. The strongest feasible design in this family of approaches is CITS.

But what if you don't have a time series? While performance measurement has improved in recent years, many organizations lack good measures of performance that they can track over time to determine whether training has translated into behavior change. Thus, point-in-time observations or surveys may be the only way to collect such data. In this case, there is no "history" as would be necessary in the time series designs. A pilot project could examine common performance management metrics tracked by emergency response agencies to see whether it is feasible to use MIS or other administrative data to measure behavior change.

Matching.

While time series designs can include comparison groups (such as the CITS), these comparison groups are not as *reliable as control groups in RCTs*. One way to improve these comparison groups is through matching.

If a comparison group can be identified, then matching can be used to tailor a better comparison group. To some extent, this can address concerns such as selection bias, particularly if methods such as propensity score matching are used. However, matching methods are still not as rigorous as a randomized controlled trial. One way to create a matched comparison group is to pick a set of variables that might be considered likely to correlate with training participation, such as job experience, level, and job title, and match those characteristics. A more sophisticated approach would be to build a statistical model (such as a logistic regression model) to predict the likelihood of participation in the given training. Such a model could be created using the full potential matching pool and the treatment group. The model would then enable the researcher to know the probability of each participant to train based on their characteristics. This probability, in turn, could be used to match each participant to their "nearest neighbor" (the individual with the closest probability) in the comparison group. Finally, the time series methods mentioned above could be employed. That said, while this would help further reduce selection bias (beyond simply a time series model), there would still be threats to validity. Most notably, the matching models would likely lack some of the key factors associated with participant outcomes and, therefore, not fully address selection bias.

Other research designs exist for making causal inferences such as regression discontinuity designs and synthetic control designs (Abadie, 2021). However, they are less relevant to the kinds of training programs being described in this paper. For regression discontinuity designs, one would need to use a continuous variable to determine eligibility for training (such as income or age) and collect data for a large sample of both eligible and ineligible participants. The synthetic control design has become popular in recent years, but it requires a great deal of data collection that is beyond the scope of most training organizations.

IV. Learning Management System Evaluation Capabilities

As noted by Cahapay (2021), the limitations to Kirkpatrick's Level 3 evaluation described in this paper sometimes present roadblocks, but there is still a feasible path forward for implementing successful Level 3 evaluations. For example, self-reported data on the transfer of knowledge from the classroom to the job can still be ascertained and successfully analyzed in certain situations, especially when combined with the latest technological best practices and evaluation methods, particularly through the incorporation of learning management systems (LMS). The latter often have automated capabilities to distribute thousands of surveys at specific points in time, such as six months after the completion of a training.

V. Re-Envisioning the Future of Level 3

Considering the previously identified limitations of Kirkpatrick's Level 3 evaluation, we argue that Kirkpatrick's model is most useful when considered as more of a framework than a prescription. When viewed through this lens, it can present valuable insights for the evaluation of training programs. However, the less-than-optimal response rate remains a pressing concern for ensuing investigations, especially for asynchronous web-based trainings. Consequently, one strategy is not only to use the Level 3 analysis but also to actively seek and integrate potential enhancements that could refine the Level 3 analysis concept and increase the response rate. We have categorized enhancement of Level 3 evaluations across behavioral, design, extension, and technological approaches below.

1. Behavioral Approach

The use of incentives may help to yield more survey responses for level 3 evaluation. Incentives can provide "nudges" to engage in a variety of forms. Some options may include:

- **Monetary incentives**: Providing monetary incentives serves as a sign of appreciation for their time and commitment, encouraging more people to participate. Singer and Ye (2013) illustrated this perspective through a systematic review of survey incentives, analyzing both major journal articles and unpublished papers from 2002 to 2013. Their findings consistently highlighted the impact of monetary rewards in enhancing response rates. Furthermore, they concluded that lotteries or prize draws are less effective in web surveys than in other survey formats (Singer & Ye, 2013).
- **Certificate Expiration and Re-Validation:** Learning management systems such as Catalog/Canvas utilized by many higher education institutions, provide an option to assign an expiration date to certificates. In this condition, we can set a specific date for the certificates to expire and re-validate them once the participant has completed the Level 3 survey. This approach potentially motivates individuals to participate in the survey, thereby boosting the response rate for a Level 3 evaluation.

2. Survey Design Approach

Survey design guidelines hold implications for the user experience and overall data collection.

- **Invitation**: Web survey invitations typically use a conventional format, consisting of an email, with a link to the survey. The design of the email can further influence its effectiveness, including:
  - **Personalization:** Using a personalized approach by addressing participants with their first name, as in "Dear [First Name]," can boost their engagement and interest. (Dillman et al., 2014, p.329).
  - **Legitimacy**: Displaying logos or affiliations of the sponsoring institution or organization on the invitation can enhance the trust between the institution and the participants. Moreover, it can foster a sense of contribution among participants that they are making positive impacts to society. (Groves et al., 1992, p.477).
  - **Effort/Time Indication**: It's crucial to manage participant's expectations. By providing a clear indication of the time commitment required, like "the survey will take about 10 minutes to complete", respondents can have a better understanding of the effort they need to finish the survey. (Kaplowitz et al., 2011, p. 346)

- **Invitation** (continued)
  - **Location of the URL:** The positioning of the survey link within the email also matters. Kaplowitz et al. (2011) found that placing the survey URL closer to the bottom of the email invitation led to better engagement and more responses.

Human Dialogue

Integrating human dialogue into web surveys has significantly improved the accuracy and reliability of participant responses, as demonstrated by Conrad et al. (2007). The authors conducted several experiments that included a version of questions that was concise and another version that incorporated longer questions and clarification dialogs. Their findings suggest that the system with embedded clarification in the questions enhanced the precision of the responses, particularly when the clarification was tailored to individual characteristics, such as age (Conrad et al., 2007). Therefore, for a Level 3 survey, it might be beneficial to offer participants the option for clarification. For instance, a feature, such as an “about” button, which will show more information upon clicking it, could potentially provide a participant-friendly and more precise survey experience.

Optimal Time to Send the Survey

Industry research has shown that survey response rates can vary based on the day of the week and the time when the surveys are sent out. For example, in 2015, CheckMarket, an online survey platform, analyzed a sample of 1,500 surveys to determine the optimal times for sending B2B (Business-to-Business) and B2C (Business-to-Customer) surveys. (CheckMarket, n.d.) For our purposes, when considering the Level 3 model, the findings from the B2C segment are most relevant.

The study revealed that sending surveys between 18:00 and 20:59 resulted in both higher click and completion rates. As for the day of the week, Tuesday emerged as the ideal day for shorter B2C surveys (those taking under 15 minutes to complete). Meanwhile, longer surveys (taking more than 15 minutes) saw optimal response rates on Wednesdays and Fridays. (CheckMarket, n.d.)

However, it's crucial to note that these results may differ based on regional, contextual, or cultural differences. For a Level 3 analysis, implementing an A/B test is a prudent approach to ascertain the most favorable time for administering surveys.

3. Extended Methods Approach

If Level 3 does not use a survey but rather builds upon other quantitative and qualitative evaluation methods, other options are available.

As noted, online surveys present several limitations, including sampling bias, relatively low validity, and a low response rate. Using a phone interview format rather than a web-based survey may address some of these concerns.

One of the primary challenges associated with online surveys is the non-random nature of the respondents. While a good response rate may occur in some circumstances, the reality is that surveys with low response rates mostly reflect individuals who are inclined to provide feedback, also known as respondent-driven, through online surveys, thereby introducing bias into the sample characteristics. This bias potentially excludes a significant portion of the population that might offer their opinions if given the opportunity to provide feedback in another format.

For instance, Roster et al. (2004) conducted a comparative analysis between web and phone interviews. Their findings indicated that older participants, who are often underrepresented in online surveys, showed a higher propensity to engage in phone interviews. Additionally, the study found that phone interviews had a higher representation of respondents with college degrees or those engaged in post-graduate endeavors.

Beyond creating a more diversified response pool, phone interviews offer a deeper understanding of participants. That is, the dynamic nature of a conversation allows researchers to ask follow-up questions based on the respondent's answers. Furthermore, the personalized approach of a phone interview often results in higher response rates. People tend to feel more valued when directly engaging in one-on-one dialogue, and this can make them more willing to participate and share their insights.

4. LMS as a Research Instrument Approach

LMS programs offer another avenue for data collection that spans the duration and the automation of Level 3 evaluation administration.

NCDP is leveraging the Canvas learning management system (LMS); the following examples assume a Canvas LMS instance or similar. Canvas is a web-based LMS that provides an online place for students and educators to access and manage course materials and communicate about the learning process (Canvas Community, n.d.)The remainder of this section will present the pros and cons of the LMS features for data collection as well as introduce new strategies employed by NCDP for learning analytics.

- **Integrate Assessment Tools into the LMS**: Some features and capabilities of Canvas include the incorporation of assessment tools such as surveys, self-assessments, and performance evaluations.
    - **Canvas Survey**: Most universities and organizations use surveys to get feedback, especially through the “Quizzes” feature on Canvas. At the beginning of setting up a quiz, Canvas will ask the instructor to choose between “Classic Quizzes” and “New Quizzes.” Both methods are used to conduct the survey, but there are differences between them. The classic quizzes will no longer get updates with new functionality, while the new quizzes will receive updated functionality over time (Cornell University, 2023). Instructors are able to choose the tool they prefer to use. However, as “New Quizzes” doesn’t support surveys, we will not consider it in the future. There are many guidelines about making survey through “quizzes,” including how to set up a survey in Canvas and how to view the results once learners have completed the survey. Examples:
        1. University of Nebraska-Lincoln: https://teaching.unl.edu/images/Workshops/Teaching
        2. University of Oregon: https://teaching.uoregon.edu/sites/teaching2.uoregon.edu/files/2020-12/how-to-create-a-student-survey-in-canvas.pdf
        3. UCSB: https://help.lsit.ucsb.edu/hc/en-us/articles/7652336119067-Surveys-in-Canvas

- **Canvas Poll:** Some organizations are also using social polls to get feedback. The social poll allows embedding polls or surveys in Canvas pages (LX at UTS, n.d.). "Social poll" will appear in the shortcut menu in the Apps list if it is used most frequently. Otherwise, instructors would select "View All" and then type "Social Poll" in the search bar to get it (LX at UTS, n.d.). Different from quizzes, polls show learners how other learners in the course responded immediately after they responded (LX at UTS, n.d.). There are four steps to making a poll:

    1. Embed a poll
    2. Choose a question type
    3. Configure the question settings
    4. Edit, view, and take a poll

- **Pros And Cons Of Each Feature**
    - **Pros Of Classic Quizzes**
        - Supports survey
        - Has a rich content editor
        - Have question banks per course/account
    - **Cons Of Classic Quizzes**
        - Does not support tagging question banks with metadata
        - Searching for huge item banks is cumbersome
    - **Use Classic Quizzes If Instructors Want To:**
        - Use question banks to create quizzes
        - Use graded or ungraded surveys
        - Use the Student Analysis Report and/or the CSV download
        - Download the quiz submissions in bulk/all at one time
        - Customize content in the quiz instructions by using design tools
    - **Pros Of Social Poll:**
        - Offers three question types: single response, multiple response, ranked choice
        - Allows configuring more settings for the questions by selecting "Show Advanced Setting"
        - Provides optional feedback to a specific answer, which will show up after learners submit their response
        - Displays immediate results after learners' submission – percent, number, or both (LX at UTS, n.d.)

V. Re-Envisioning the Future of Level 3

- **Use Social Poll If Instructors Want To:**
  - Mainly use single response, multiple response, or ranked choice in the survey
  - Configure the question settings
  - Add optional feedback to a specific answer for each question
  - Display the result immediately after learners submit their answers (LX at UTS, n.d.)

■ **Available Statistics And Analysis Methods**

- **Survey through Quizzes:** Instructors can download CSV files for a survey. The Student Analysis Report is available and must be downloaded as a CSV file, while Item Analysis is unavailable for surveys. Moreover, instructors can choose to enable the anonymous option before or after the survey has received submissions. Instructors will be able to view the results of the survey after at least one learner has taken the survey. Particularly for graded surveys, common data such as average score, high score, low score, standard deviation, and average time are available as statistics. The CSV download also provides the following calculations and counts:
  - Question ID
  - Question Title
  - Answered student count
  - Quiz question count (total number of quiz questions)
- **Survey Through "Social Poll"**
- **Analysis Methods:** Based on our experience using the LMS as a database, we have found the following analysis strategies conducive with the data available:
  - Statistics
  - NLP/ML
  - Visualization
  - Sentiment Analysis – has been helpful in analyzing a summary of how learners feel about a course. Common sentiment labels like positive, neutral, and negative can be used to filter the results.

Overall, we note the importance of exploring other data collection methods that span the duration of a Level 3 evaluation and ways to automate Level 3 evaluation administration as drivers for the success of implementing the model. Throughout the process, collecting feedback from participants on the data collection experience is important to make continuous improvements. We recommend integrating assessment tools into the LMS and choosing the "assess" feature based on what function can be achieved, what type of data is desired as a result, and what analysis method is available under such a feature.

VI. Conclusion

This paper has described some of the challenges of measuring Level 3 of the Kirkpatrick model while also presenting numerous solutions. One way forward is to use carefully developed surveys with designs such as list randomization and techniques such as "anchoring" to collect more reliable data and to disseminate them through an automated learning management system (LMS) during specific time periods. A further improvement would be to use one of the designs discussed above to address selection bias. Perhaps the most feasible approach would be to use a crossover design that randomizes the timing of trainings and measures outcomes relative to the date of training. Data collection and research design often overlap. Consequently, when implementing a design using one of the comparison group models (like randomization or matching), there is often less need for historical data, thus alleviating some of the burden for the analysis. Combining a crossover design with the survey administrative benefits of an LMS would significantly mitigate (though not eliminate entirely) the threats to validity mentioned earlier in this paper. Additional areas of promise involve aligning training measurements with performance measurement systems already in place at many organizations. Undoubtedly, this would require some accommodation within performance management systems. Since training is supposed to be focused on best practices, encoding key learning objectives into performance management systems should be a standard practice.

In reconsidering Kirkpatrick's Level 3 evaluation model, this paper also advocates for viewing the model as a flexible framework rather than a rigid prescription to better evaluate training programs. Noting persistent challenges, particularly low response rates in asynchronous web-based training, the authors outline potential strategies to enhance Level 3 assessments. Behavioral strategies, such as using incentives, could improve response rates, while refined survey design—personalizing invitations, optimizing timing, and embedding clarifications—may enhance user engagement. Alternative methods, including phone interviews, are proposed to mitigate online survey biases, expand demographic representation, and gather more nuanced feedback. Additionally, leveraging Learning Management Systems (LMS) for automated data collection and assessment tools, including quizzes and social polls, offers a path for integrating analytics and diversifying feedback collection, such as through sentiment analysis. By broadening the Level 3 model through these strategies, the paper emphasizes a continuous improvement approach, ensuring the model's responsiveness to evolving educational and technological contexts.

The Kirkpatrick model has stood the test of time because it is a common-sense way of thinking about how to translate training into actual skills and practice. We have provided a series of ideas for how Level 3 of the model can be measured in future applications. It would be beneficial for emergency management training organizations to consider developing more robust mechanisms to capture, analyze, and act upon Level 3 data so that the Kirkpatrick model can achieve the aims for which it was originally designed.

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2), 391-425.

Alliger, G., & Janak, E . Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology*, 42(2), 331-341. https://doi.org/10.1111/j.1744- 6570.1989.tb00661.x

Angrist, J. D., & Pischke, J. S. (2014). Mastering metrics: *The path from cause to effect.* Princeton University Press.

Bakkour, A., Zylberberg, A., Shadlen, M. N., & Shohamy, D. (2018). Value-based decisions involve sequential sampling from memory. *BioRxiv*. https://doi.org/10.1101/269290

Bloom, H. S. (1999). *Estimating program impacts on student achievement using "short" interrupted time series. MDRC.*

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*. https://doi.org/10.1038/nn.4573

Brownback, A., & Novotny, A. (2018). Social desirability bias and polling errors in the 2016 presidential election. J*ournal of Behavioral and Experimental Economics*, 74, 38-56.

Cahapay, M. (2021). Kirkpatrick model: Its limitations as used in higher education evaluation. *International Journal of Assessment Tools in Education*, 8(1), 135-144.

Canvas Community. (n.d.). *What is Canvas?* https://community.canvaslms.com/t5/Canvas-Basics-Guide/What-is-Canvas/ta-p/45

Chandon, P., Morwitz, V. G., & Reinartz, W. J. (2005). Do intentions really predict behavior? Self-generated validity effects in survey research. *Journal of Marketin*g, 69(2), 1-14. https://doi.org/10.1509/jmkg.69.2.1.60755

CheckMarket. (n.d.). *What's the best time to send a survey?* Retrieved from https://www.checkmarket.com/blog/survey-invitations-best-time-send/

Colón, H. M., Robles, R. R., & Sahai H. (2001). The validity of drug use responses in a household survey in Puerto Rico: Comparison of survey responses of cocaine and heroin use with hair tests. *International Journal of Epidemiology*, 30(5), 1042-1049. doi: 10.1093/ije/30.5.10

Conrad, F. G., Schober, M. F., & Coiner, T. (2007). Bringing features of human dialogue to web surveys. Applied Cognitive Psychology: *The Official Journal of the Society for Applied Research in Memory and Cognition*, 21(2), 165-187.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (pp. 103-134). Houghton Mifflin.

Cornell University. (2023). *Comparison of quiz tools: Canvas quizzes and new quizzes*. https://learn.canvas.cornell.edu/comparison-of-quiz-tools-canvas-quizzes-and-new-quizzes/#:~:text=The%20New%20Quizzes%20tool%20and,receive%20updated%20functionality%20over%20time.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: *The tailored design method*. John Wiley & Sons.

Doerr, J. (2018). Measure what matters: *How Google, Bono, and the Gates Foundation rock the world with OKRs*. Penguin.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2), 303-315.

References

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56(4), 475-495.

Hendra, R., & Hill, A. (2019). Rethinking response rates: New evidence of little relationship between survey response rates and nonresponse bias. *Evaluation Review*, 43(5), 307-330.

Kaplowitz, M. D., Lupi, F., Couper, M. P., & Thorp, L. (2012). The effect of invitation design on web survey response rates. *Social Science Computer Review*, 30(3), 339-349.

Keeter, S., Hatley, N., Kennedy, C., & Lau, A. (2017). What low response rates mean for telephone surveys. *Pew Research Center*, 15(1), 1-39.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training Directors*, 13(11), 3-9.

LX at UTS. (n.d.). Social poll in Canvas. https://lx.uts.edu.au/collections/communicating-in-canvas/resources/social-poll-in-canvas

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, 15(3), 263-280.

Parmenter, D. (2015). *Key performance indicators: Developing, implementing, and using winning KPIs*. John Wiley & Sons.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). Academic Press.

Phillips, J. (1991). *Handbook of training evaluation and measurement methods* (2nd ed.). Gulf Publishing Company.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539-569.

Rachidi, A. *Is your training and technical assistance (T/TA) effective? Considerations for measuring performance* (No. 43d6c2336da54d02a514b1ae543d9379). Mathematica Policy Research.

Reio, T. J., Rocco, T. S., Smith, D. H., & Chang, E. (2017). A critique of Kirkpatrick's evaluation model. *New Horizons in Adult Education & Human Resource Development*, 29(2), 35-53.

Roster, C. A., Rogers, R. D., Albaum, G., & Klein, D. (2004). A comparison of response characteristics from web and telephone surveys. *International Journal of Market Research*, 46(3), 359-373.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 112-141.

Sui, J., & Humphreys, G. W. (2015). The integrative self: How self-reference integrates perception and memory. *Trends Cogn Sci*, 19, 719-728.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychol Bull*, 121, 371-394.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Vaidya, A., Castillo, J., Torres, A., & Badre, D. (2023). *Influences of familiarity and recollection on value-based decision-making*.

Yudkowsky, R., Park, Y. S., & Downing, S. (2019). *Introduction to Assessment in the Health Professions, 2nd edition*. Routledge. eBook ISBN9781138054394

Columbia Climate School
National Center for Disaster Preparedness

475 Riverside Drive, Suite 401
New York, NY 10115-0095
ncdp@columbia.edu
https://ncdp.columbia.edu