Journal of Organizational Behavior

Journal of Organizational Behavior, J. Organiz. Behav. (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/job.2102

Research Article

The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): two novel evaluation methods for developing optimal training programs

MATT C. HOWARD[1,2]* AND RICK R. JACOBS[2]

[1]*Department of Management, Mitchell College of Business, University of South Alabama, Mobile, A.L., U.S.A.*
[2]*Department of Psychology, Pennsylvania State University, University Park, PA, U.S.A.*

Summary Current methodologies in training evaluation studies largely employ a single method entitled random confirmatory trials, prompting several concerns. First, practitioners and researchers often analyze the effectiveness of their entire omnibus training, rather than the individual elements or identifiable components of the training program. This slows the testing of theory and development of optimal training programs. Second, a common training is typically administered to all employees within an organization or workgroup; however, certain factors may cause individualized training to be more effective. Given these concerns, the current paper presents two training evaluation methodologies to overcome these problems: the multiphase optimization strategy and sequential multiple assignment randomized trials. The multiphase optimization strategy is a method to evaluate a standard training, which emphasizes the importance of a multi-stage training evaluation process to analyze individual training elements. In contrast, sequential multiple assignment randomized trial is used to evaluate an adaptive training that varies over time and/or trainees. These methodologies jointly overcome the problems noted earlier, and they can be integrated to address several of the key challenges facing training researchers and practitioners. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: training; methodology; statistics

Despite the widespread study of organizational training, certain methodological issues systematically appear in training scholarship. Currently, the most used method to evaluate a training program is called random confirmatory trials (RCTs; Campbell, 1988; Tannenbaum & Yukl, 1992). While this method provides many benefits, RCTs also have many drawbacks. Notably, although a training program may consist of several individual elements,[1] RCTs only analyze the effectiveness of the overall training (Bass & Avolio, 1990; Burke & Day, 1986; Smith & Smith, 2007). For example, Barling, Weber, and Kelloway (1996) investigated the effectiveness of a training program to improve managers' transformational leadership. While the training consisted of four clearly identifiable elements to achieve this goal, their training evaluation methodology only analyzed the effectiveness of all the elements together. Although RCTs have provided insightful evidence on general training, little is known about the effectiveness of the individual training elements. It is possible that a single element entirely drives employee improvement, or some elements may even detract from the overall effect. Unfortunately, through only analyzing an entire regimen, the successful or unsuccessful elements cannot be identified (Isler et al., 2009; Lesch, 2008). Organizations may

*Correspondence to: Matt C. Howard, 5811 USA Drive S., Rm. 346, Mitchell College of Business, University of South Alabama, Mobile, AL, 36688, U.S.A. E-mail: mhoward@southalabama.edu.

[1]The term "training" refers to an intervention created to improve employee attributes and/or performance, with the assumption that most training programs involve several elements. The term "element" refers to an individual module or aspect of the training. The term "module" refers to a section of instructional material that provides direct information about certain knowledge, skills, or abilities (e.g., presentation, hand-out, etc.). The term "aspect" refers to an attribute of the training that is primarily meant to influence trainee motivation and/or reactions (e.g., paying trainees, method of delivery, etc.). All training programs consist of one or more modules, but aspects are optional.

Received 24 February 2015
Revised 01 February 2016, Accepted 15 February 2016

be spending excessive amounts on extraneous elements, and researchers are unable to further theory through the analysis of particular training elements.

Additionally, although many individual differences that cause individuals to experience differential training effects have been discovered (Bauer et al., 2012; Driskell et al., 1994; Martocchio & Judge, 1997), authors have noted that only occasional efforts have been made to scientifically harness and integrate these differences into adaptive training programs (Gully & Chen, 2010; Tannenbaum & Yukl, 1992). The dearth of adaptive training programs may be due to RCTs' poor ability to evaluate the interaction between training elements with individual differences and/or the cost of individualizing training. While the second issue is likely to remain, the first issue should be evaluated. Given these reoccurring issues in the training literature, the current study reports on two unique methods to better understand the effectiveness of training programs and their elements. Both of these methods are likely unfamiliar to many organizational researchers, as these methods were developed in other research areas (Engineering and Public Health).

The first is the multiphase optimization strategy (MOST; Collins et al., 2005; Collins, Murphy, & Strecher, 2007). Before the omnibus training evaluation, additional steps are taken to determine individual training elements' effectiveness, interaction effects, and optimal treatment levels. These steps are reliant on experimental designs, and the current article explores the feasibility of four experimental designs that could be used in organizational contexts. The successful implementation of MOST removes ineffective training elements, resulting in an optimized training. Also, compared with RCTs, MOST provides richer information about a training program and included elements, benefiting the investigation of theory. Many research questions can be answered with MOST that would otherwise be impossible to explore.

The second methodology is the sequential multiple assignment randomized trial (SMART; Murphy, 2005; Murphy et al., 2007). SMART enables researchers to investigate the main effects and interactions of individual time-varying adaptive training elements. This method involves the random assignment of employees to alternative training conditions within an experiment, followed by further random assignment to conditions in follow-up experiments based on tailoring variables and decision rules. From the results of these experiments, practitioners and researchers can determine the best training considering the individual characteristics of a trainee, and thus provide adaptive or customized interventions. Compared with RCTs, SMART provides more detailed information about the unfolding dynamics of individuals, training programs, and their interactions. Once again, many research questions can be answered with SMART that would otherwise be impossible to explore.

Both of these methods have been repeatedly applied in other fields to streamline the evaluation of costly interventions (Collins et al., 2005; Rivera et al.,2007). The intention of this article is to discuss when and under what conditions MOST and SMART might be most useful to organizational practitioners and researchers. The following sections (1) provide an overview of RCTs, MOST, and SMART; (2) discuss examples of organizational training situations that might benefit from using MOST and SMART; and (3) denote the implications of these methods and propose directions for future research.

Training Background

In recent decades, training research has become increasingly nuanced and theoretically driven, demonstrating that any successful training involves a careful consideration of a complex series of necessary factors (Grossman & Salas, 2011; Kirkpatrick, 1994; Rouiller & Goldstein, 1993). Despite great advancements in this area, authors have continuously noted the limited use of methodologies to evaluate a training program, which hampers the investigation of new theories and important training elements. In Campbell's (1988) review, he noted that "by far" the most popular training evaluation methodology was the comparison of a single desired training program (whether newly created or existing) to a control condition, which may be an alternative training program or none at all. This sentiment was echoed by Tannenbaum and Yukl (1992), adding that "this research type has only marginal utility for improving our understanding of training" (p. 407). They further noted the concerns with this design and the importance of new methods by stating,

DOI: 10.1002/job

'Demonstration' studies do not reveal why a particular method or combination of methods facilitates learning or how the method can be used more effectively. Even a study that pits one method against another can be inconclusive as it is likely that the relative effectiveness of different methods will depend on the purpose and objective of the training, the attributes of the trainees, and the effectiveness criteria selected. (Tannenbaum & Yukl, 1992, p. 407).

Recently, authors have made few direct statements toward this issue, but their silence speaks for itself. Aguinis and Kraiger (2009) and Salas and colleagues (2012) provide a comprehensive overview of the current state of training research. The authors note several improvements to theory in identifying important outcomes to evaluate the impact of successful training, such as advancements to Kirkpatrick's hierarchy (1994), but they did not make any comments about training evaluation methodologies. The dearth of information, by no means, is a shortcoming of the two articles, but an adequate reflection of the state of training evaluation methodologies.

It is incorrect to state that no improvements have been made whatsoever. Sackett and Mullen (1993) provide a theoretical distinction between the two most common training evaluation research questions: (i) "How much change has occurred?" and (ii) "Has a target performance level been reached?" Nevertheless, suggested methods to address these two questions can all be considered RCTs. Cascio and Aguinis (2005) present seven possible research designs to test a training program's effectiveness, each with their benefits and weaknesses. These described methods are somewhat limited, as they can all be considered variations of a RCT. Others have proposed novel methods for formative evaluations, which occur during the creation of a training program (Brown & Gerhardt, 2002); however, these evaluation methods cannot provide decisive inferences about the effectiveness of a particular training or its elements beyond an alternative, and they are often based upon qualitative inferences derived from very small sample sizes (i.e., $n = 6$, Medley-Mark & Weston, 1988; $n = 6$, 3, and 3, Saroyan, 1992). Other authors only apply RCTs for their formative evaluations (Dick & Carey, 1996; Geis, 1987). Lastly, some authors have sporadically implemented alternative designs, but systematic application of these methods has not been seen (Aguinis & Kraiger, 2009; Shadish et al., 2002; Salas et al., 2012). Together, these articles demonstrate a collective improvement in training evaluation methodologies, but most researchers still rely on designs that cannot address the need of modern training research.

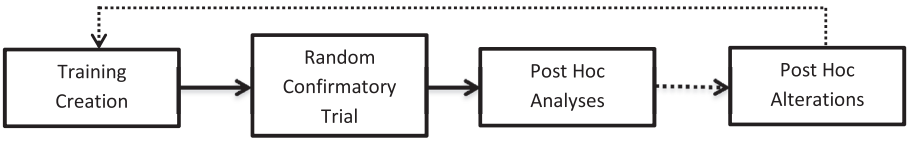
Given the apparent need, the current article clearly notes the deficiencies with current training evaluation methodologies and provides a proper label to the most popular current methodology, RCT. Then, the current article introduces two training evaluation methodologies, MOST and SMART, which explicitly overcome these shortcomings. These objectives address the call of previous authors (Campbell, 1988; Tannenbaum & Yukl, 1992) to propose new training evaluation methodologies, and further future research and practice

Random Confirmatory Trials

Current practitioners and researchers largely employ a single training evaluation method (Bass & Avolio, 1990; Burke & Day, 1986; Smith & Smith, 2007), called RCTs or the "all or nothing" method (Collins et al., 2009, p. 21). This methodology is relatively simple and provides valuable information about the overall training. An RCT poses an entire training against an alternative training program or none at all. The treatment group undergoes the training, whereas the control group receives an existing training, an unrelated training, or nothing at all. Then, a post-test is administered. Analyses can determine the training effectiveness, often decided through the group with the highest post-test score. For this method, analyses require a sample size of 128 employees to achieve power of .80 for detecting moderate effect sizes ($d = .5$; independent sample t-test with two-tail hypothesis) and 788 employees for small effect sizes ($d = .2$). This may be followed by *post hoc* alterations, which would require a repeat of this entire process to test the desired training's effectiveness. This research design is presented in Figure 1.

Although RCTs are the most common training evaluation method (Campbell, 1988; Tannenbaum & Yukl, 1992), numerous issues have been noted. First, aside from possible pilot studies, the training is largely finalized before any

DOI: 10.1002/job



Note: Solid lines are necessary paths that must be taken to achieve final results.
Dotted lines are optional and dependent upon the researcher's discretion.

Figure 1. Visual representation of the "All or Nothing" method for training evaluations

testing. This approach is not conducive to optimizing a training program, as RCTs do not guide researchers to insufficient training elements. Second, although many training programs are composed of several elements, RCTs cannot analyze the impact of individual elements. Third, some analyses that follow RCTs are "not based on random assignment" (Collins et al., 2007, p. S113). In addition to testing for overall training effectiveness, many practitioners and researchers are interested in the effect of attending only a portion of the training. Through the "all or nothing" method, the only employees that do not complete the entire training (and therefore only attended a portion of the training) are those who withdrew for extenuating circumstances. Comparing these employees against those that completed the entire training may provide biased results, as individual differences of those that withdrew may cause any observed effects.

Fourth, RCTs are very limited in their ability to test and further theory. A multitude of authors have proposed theoretical models of training effectiveness (Alvarez et al., 2004; Cannon-Bowers, Salas, Tannenbaum, and Mathieu 1995; Grossman & Salas, 2011; Kraiger et al., 1993). These models note several factors that impact learning and development, and they are often applied to create particular training programs. When using RCTs alone, separate experiments must be performed to test the effect of each individual element — a costly and timely process. Further, these models propose several moderators and mediators (Aguinis & Kraiger, 2009; Blume et al., 2010; Kraiger et al., 1993), and RCTs are extremely limited in the evaluation of these moderators with individual training elements. RCTs can only identify moderators and mediators of the entire training, such as the effect of climate on the relationship between an entire training with an outcome (Rouiller & Goldstein, 1993; Tracey & Tews, 2005). It is impossible to discover any moderated or mediated relationship of elements through RCTs. Lastly, researchers using RCTs struggle to assess adaptive elements and unfolding training dynamics, as RCTs usually do not include measurement occasions during the training. Even if included, RCTs can only detect the effect of the all elements together, providing limited information on developments during the training. Therefore, despite the increasing nuance of theories — both those applied to understand the learning and transfer process as well as those applied to develop particular training programs — the current dominant evaluation method, RCTs, cannot test the complexity of these theories.

Although we have many concerns about RCTs, they are still appropriate to use in certain situations. Practitioners and researchers are often limited in their resources and/or have very simple goals, which include demonstrating a difference between trained and untrained employees. Further, an evaluation may only need to test the impact of a single additional element in a training program. In these cases, RCTs would be ideal. For any practitioners or researchers' purpose, it may be confusing to determine when RCTs are appropriate. For this reason, a flowchart of when to use RCTs, MOST, and SMARTs is presented in Figure 2. Each method's relative benefits are presented in Table 1. Both are further explained below when explicating MOST and SMART.

A description of RCTs is provided, previously, but certain aspects of RCTs may still be unclear. For this reason, an example RCT is presented to further clarify proper procedures.

Random Confirmatory Trial Example — Study 1

The current article presents illustrative examples, using actual participants, of each training evaluation method to highlight their strengths and weaknesses. In each example, the methods are used as a formative evaluation, during training creation, but there is no reason why the methods cannot be used as a summative evaluation, after the training

DOI: 10.1002/job

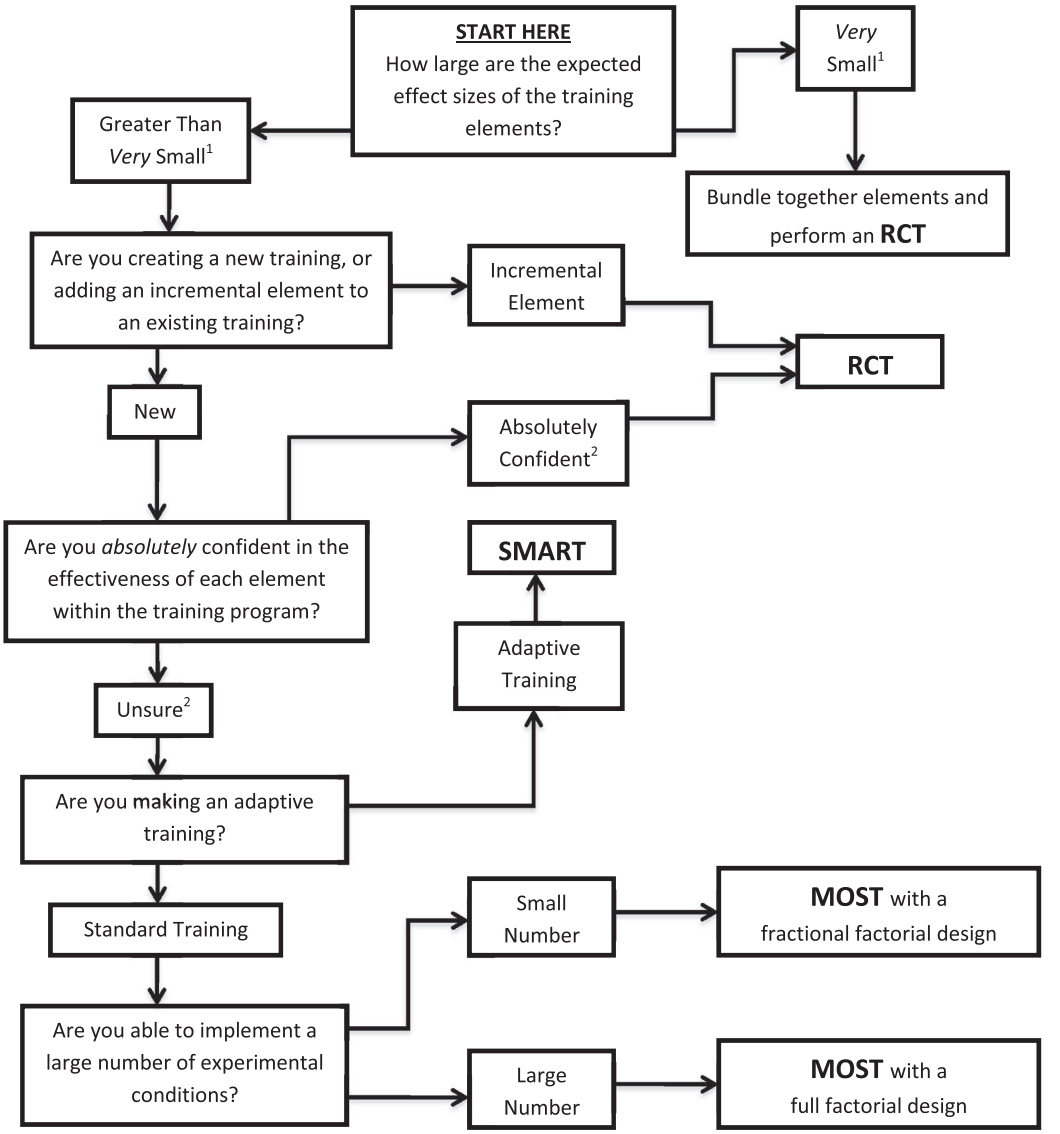


Figure 2. Flowchart of when to use random confirmatory trials (RCTs), multiphase optimization strategy (MOST), and sequential multiple assignment randomized trial (SMART)

DOI: 10.1002/job

Table 1. Comparisons of RCTs, MOST, and SMARTs.

	Number of conditions	Sample size	Application of multiple measure designs	Test of overall training	Test of individual training elements	Test of training elements interactions	Test of time-varying elements
(1.) RCT							
(a.) Any form	2	$2n$	√	√			
(2.) MOST							
(a.) Screening phase only							
(i.) Multiple two-condition experiments	$2k$	$k * 2n$	√		√	√	
(ii.) Comparative treatment	$k+1$	$(k+1) * n$	√		√	√	
(iii.) Full factorial	2^k	$2n$	√		√	√	
(iv.) Fractional factorial	2^{k-x}	$2n$	√		√	√	
(b.) Screening and confirming							
(i.) Multiple two-condition experiments	$2k+2$	$(k*2n)+2n$	√	√	√	√	
(ii.) Comparative treatment	$(k+1)+2$	$[(k+1)*n]+2n$	√	√	√	√	
(iii.) Full factorial	$(2^k)+2$	$2n+2n$	√	√	√	√	
(iv.) Fractional factorial	$2^{k-x}+2$	$2n+2n$	√	√	√	√	
(3.) SMART							
(a.) Standard	Variable	$2n$	√				√

Note: k = number of training elements; n = sample size; x = chosen exponential reduction; RCT = random confirmatory trial; MOST = multiphase optimization strategy; SMART = sequential multiple assignment randomized trial.

DOI: 10.1002/job

creation. For the RCT example, two self-guided online safety training programs were created. These training programs educated trainees on safety regulations and best practices for blue-collar workplaces, and information was taken from the Occupational Safety and Health Administration (OSHA) website (www.osha.gov) or websites linked through the OSHA website. The training programs educated trainees on eight separate areas of safety regulations (i.e. falling risks, electrical risks, etc.).

Random confirmatory trial method

Two self-guided online safety training programs were created. The first represented an existing training with minimal features, and it presented textual information to trainees on a webpage. In total, eight separate topics were taught on eight separate pages, resulting in the training consisting of eight mandatory modules.

The second training represented a desired training that was meant to improve upon the existing, minimal training. This second training presented identical textual information to trainees, but six elements were added to the training. These elements were chosen due to their similarity to common training additions. These six elements were as follows: (1) the presentation of pre-training material, which provided workplace safety statistics given at the onset; (2) a short pre-test, which gave several example post-test items given at the onset; (3) the notification of a chance to win $25 if a certain percentage was scored on the post-test; (4) the opportunity to write notes throughout the presentation of material but removed before the post-test; (5) the inclusion of material that provided information about general workplace safety given at the onset; and (6) a short safety video presented midway through the material. Elements one and five represents additional modules to the training, but they were not tested upon in the training post-test (described later). All other elements are considered training aspects. In sum, it was assumed that these six additional elements would result in a more effective training.

To test the RCT methodology, participants were assigned to one of these two groups. After completing the training, they were given a brief questionnaire to gauge reactions followed by a post-test. Finally, they were debriefed on the current study, and all participants were entered into the $25 raffle regardless of their assignment to training condition.

Random confirmatory trial participants

To test RCTs, 105 participants were recruited from an undergraduate student participant pool in return for extra course credit. Fifty participants were placed in the minimal training, and 45 completed the entire training and post-test. Fifty-five were placed in the enhanced training, and 48 completed the entire training and post-test. While this sample is too small to detect small effects, it is suitable for detecting medium and large effects. Because of confidentiality and anonymity purposes, demographic information was not recorded. Based upon previous data collection from this student participant pool, it is expected that these participants are approximately 19 years of age, mostly Caucasian, and with a majority being female.

Random confirmatory trial measures

Total time on training

As the training was administered online, participants were able to progress through material at their own pace. This may enhance learning for some participants as they are able to reread difficult material and process complicated ideas. Other participants, however, may quickly progress through material to complete it quickly. For this reason, the amount of time each participant spent on each page was digitally recorded. Then, the time spent on each of the required pages, included within the minimal and enhanced training, was averaged together. For the current study, z-scores were calculated to detect outliers in time spent on training. Extreme outliers on this variable may represent individuals that started the training but did not close the training program while they attended to other matters.

DOI: 10.1002/job

Within the dataset, two individuals' z-scores were above 3.5. These individuals' responses were rescaled to the next highest individual's average time on training score, which has a z-score of 2.9.

Post-test

To gauge learning, 28 multiple-choice items were created. These items focused on declarative knowledge as the topic, safety, involves an ample amount of required declarative knowledge. An example question is, "In general industry workplaces, at what elevation must fall protection be provided," and the correct answer is "Four feet." The Cronbach's alpha of this measure was .80 within the current example.

Random confirmatory trial results

To test for the differences between the minimal and enhanced training, an independent samples *t*-test was performed. For the total time on training variable, Levene's test for the equality of variances was significant ($F = 5.866$, $p < .05$), so equal variances were not assumed for this analysis. Those within the enhanced group spent more time on the training than those within the minimal group ($t(72.388) = -3.492$, $p < .01$). Those within the minimal training spent, on average, 56 seconds per training topic page (Std. Dev. = 66). Those within the enhanced training spent, on average, 127 seconds per training topic page (Std. Dev. = 129). Also, differences between participant's post-test scores were analyzed. Those within the enhanced group received higher scores upon the post-test than those within the minimal group ($t(93) = -4.331$, $p < .001$). Those within the minimal training, on average, correctly answer 48 percent of the post-test questions (Std. Dev. = 18 percent). Those within the enhanced training, on average, correctly answered 63 percent of the post-test questions (Std. Dev = 16 percent). Together, these results indicate that the enhanced training is more effective than the minimal training. Further, these results are also analyzed through regression (Table 2). These results are provided because the MOST and SMART examples are analyzed *via* regression. Providing consistent analyses for all example allows for a clearer understanding of RCTs, MOST, and SMART.

Random confirmatory trial discussion

As evident from the example, the RCT demonstrates the enhanced training is superior to the minimal program, but RCTs are unable to provide much in the way of additional information. A single element may drive the entire impact of the enhanced training's effectiveness, and all other elements of the training may simply waste time and organizational resources. When using an RCT, an organization would remain unaware of the ineffective elements, and practitioners would be unable to exactly gauge which elements prompt the positive results. Also, researchers would endure similar difficulties, which prevents a honed investigation of all theory — those used to understand the learning and transfer process as well as those used to develop certain training programs. For this reason, an alternative training evaluation method, MOST, is described.

Table 2. Regression analysis results of random confirmatory trial example.

	Post-test score				Time spent on training			
	B	Std. error	*β*	*t*	*B*	Std. error	*β*	*t*
Constant	.559	.017		32.094***	91.369	10.183		8.972***
Group	.075	.017	.410	4.331***	35.770	10.183	.333	3.513**
R^2				.168				.111

Note: The condition is effect coded (−1 is minimal training, 1 is enhanced training).
*$p < .05$, **$p < .01$, ***$p < .001$

DOI: 10.1002/job

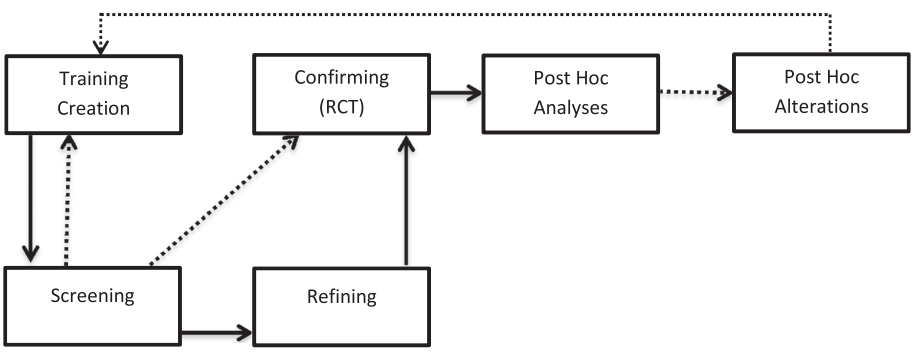
Multiphase Optimization Strategy

An alternative method to the traditional RCT is the MOST. MOST was explicitly created to overcome the shortcomings of RCTs and was founded within the field of engineering (Collins et al., 2005; Collins et al., 2007). This research design incorporates the existing prevalent methodology, RCTs, but adds several prior steps to determine which and/or how much of each element should be included. The goal of these prior steps is to identify a cheaper, more effective, and optimized training, which can afterwards be tested within an RCT. The three steps within MOST are screening, refining, and confirming; the confirming phase is identical to an RCT. An illustration of MOST is presented in Figure 3.

Like the "all or nothing" method, MOST begins with creating the initial training, along with any necessary pilot testing to ensure feasibility and practicality. Then, the screening phase begins. The purpose of the screening phase is to determine which individual training elements are effective through randomized experiments, with the belief that removing ineffective elements can make the training quicker, cheaper, and more effective. The randomized experiments in the screening phase can calculate effect sizes for each individual element, and interactions between elements can be discovered. *A priori* alpha levels, effect sizes, or cost–benefit ratios can aid in choices to retain or discard elements. No hard criteria exist for element selection decisions. At the conclusion of the screening phase, an identified set of effective training elements should be obtained, which need to be further analyzed. If no training elements are deemed effective, then the researcher should restart the MOST process with an entirely new set of training elements. Further, it should be noted that randomized experiments refer to methods that randomly assign individual participants to training programs (experimental design) as well as methods that randomly assign groups of participants to training programs (quasi-experimental design).

Assuming some training elements are retained, the next step is the refining phase. During the refining phase, the retained elements are further scrutinized. Particularly, considerations over the optimal amount of individual elements are made, and the refining phase can also determine whether certain elements have differential effects on certain participant populations. Much like the screening phase, retention decisions are made from *a priori* alpha levels, effect sizes, and/or cost–benefit ratios. After the refining phase, an optimized version of the training is identified. It should be noted, however, that the refining phase is not mandatory. If time or participants need to be conserved, the screening and refining phase can be combined. Particularly, if certain elements must be included, it is unnecessary to test their effectiveness in the screening phase. Instead, it is possible to test different levels of the elements during the screening phase, such as 8-week/12-week or minor/intensive. Also, if the evaluator is simply not interested in various levels of elements, then the refining phase is unnecessary.

After the refining phase is the confirming phase. In the confirming phase, an RCT is performed to determine whether the overall training is more effective than an alternative, followed by the same analyses and alterations



Note: Solid lines are necessary paths that must be taken to achieve final results.
Dotted lines are optional and dependent upon the researcher's discretion.

Figure 3. Visual Representation of multiphase optimization strategy

DOI: 10.1002/job

as before. With this final step, the MOST procedure is concluded, and the organization should reach an optimal training. To clarify MOST, a visual representation is provided in Figure 3.

Randomized experiments to test individual elements are applied in both the screening and refining phase of MOST, causing careful consideration of experimental designs to be extremely important. For this reason, the next section expands upon the benefits of several experimental designs that lend themselves to implementation in MOST.

Experimental Designs

During the screening and refining phases, each of the training elements must be evaluated through a randomized experiment. Many methods exist to achieve this goal. The first is to construct separate two-condition experiments, each with a control and intervention group, and only one element present in each intervention group. Many problems exist with this method (Collins et al., 2011; Dziak et al., 2012). Notably, each experiment requires an adequate sample size to ensure suitable power. For each experimental condition, a sample size of 64 is needed to obtain power above .8 for detecting a moderate effect size ($d = .5$; independent group t-test[2]), resulting in a total sample size of 128 for each experiment. Given this, 768 participants are needed to test the effectiveness of six training elements, which is largely unattainable for most researchers and practitioners. Further, this method can only test main effects of elements, and all element interaction effects are unknown. These shortcomings make this design unsatisfactory.

An alternative method is the comparative treatment design, which is similar to multiple two-condition experiments but the same control group is used for all experimental group comparisons. Whereas 12 experimental groups were needed for six two-condition experiments, a comparative treatment design would only need seven. Although this is an improvement, it still suffers from the same problems as before. A sample of 64 is still needed for each experimental condition, resulting in a total sample of 448. Once again, all interaction effects between elements are left unstudied. Also, when using the same control group for all comparisons, an unusual sampling for this one group could bias all inferences taken from analyses. These shortcomings have caused researchers to seek other methods, which require fewer participants, test for interaction effects, and avoid major methodological concerns.

A method to achieve this goal is the factorial analysis of variance (ANOVA) design. Typically, when using an ANOVA design, each element is fully crossed. That is, a condition exists for each combination of elements, known as a complete factorial design, and each element is concurrently analyzed using the same sample. Whereas other experimental methods require new samples to test each element, a complete factorial design only needs one. This reduces the total sample needed to the size of one experiment, which is only 128 individuals ($d = .5$; dummy coded regression). Also, this method allows researchers to probe for element interactions. In fact, all interactions can be tested, including interactions between all training elements; however, while a complete factorial design is an improvement in these two aspects, it still has drawbacks. A complete factorial design is difficult for experiments with a large number of elements. In the example given, the training consists of six elements that have two levels (element/no element), which would require 64 conditions (2^6) in a complete factorial design. This design is likely problematic for a practitioner or researcher. Alternative factorial designs have been created to investigate the impact of individual elements while utilizing fewer resources.

An alternative to the complete factorial design is the fractional factorial ANOVA design. In this research design, the required total sample size is the same as complete factorial design, but the required number of conditions is reduced based on careful considerations. Most often, main effects or lower-order interactions (two-way) demonstrate significant effects, and higher-order interactions (three-way and higher) are inconsequential. Given this, certain conditions that are required to test these higher-order interactions can be removed, and resources can be allocated toward

[2]For this example, the only measurement occasion is after the intervention. The required sample size is reduced if pre-training measures are administered and analyzed through mixed between-within subjects ANOVAs, but the number of measurement occasions and type of statistical analyses do not change the inferences of this passage.

DOI: 10.1002/job

testing more important effects. A fractional factorial design can reduce a complete factorial design of 64 conditions to 32, 16, or even 8, based on the needs of the researcher.

Although fractional factorial designs allow elements to be studied using fewer resources, a drawback is a reduction of information. That is, not all interactions are studied, and significant effects may remain unknown. Further, the effects of unmodeled interactions are subsumed into lower-order effects (Collins et al., 2005; Collins et al., 2007). This is not a concern when interactions are non-significant and inconsequential, as their effects are null and zero. If the interactions are significant and/or consequential, however, the observed lower-order effects are misleading. In practice, the element retention decision will likely be unaffected. A practitioner would otherwise consider the impact of the main effects and interactions, together, when all interactions are modeled. When unmodeled, the effects are grouped together in lower-order effects, and the perceived impact of elements would remain unchanged. In research, the elements would be misunderstood, thereby obfuscating the true nature of training. For these reasons, it is very important to carefully consider which effects to analyze before deciding the conditions to remove, causing existing theory to be important in fractional factorial designs (Gunst & Mason, 2009). Most notably, if a researcher has a plausible reason to suspect that a higher-order interaction between elements is consequential, then they should use a research design which tests the higher-order interaction.

Furthermore, a balanced factorial design is still possible when using fractional factorial designs. In balanced designs, all elements are present in an equal number of conditions relative to each other element, providing several statistical benefits. First, main effects and interactions are estimated without bias, assuming that all significant effects are modeled. Second, the sample size and statistical power for testing each main effect is the same as a single experiment testing one element alone. Therefore, balanced fractional factorial designs require fewer conditions than complete factorial designs through eliminating conditions based on previous theory, with only a modest loss of information, while retaining statistical power of single condition experiments.

Additionally, a point should be made about measurement when using a factorial design. Cascio and Aguinis (2005) presented several alternative RCT designs that focused upon the use of a comparative treatment group and number of measurement occasions; however, there is no reason that their measurement designs cannot be applied to any other experimental design. Post-test only, pre-test and post-test, multiple pre-tests and post-tests, and several others could surely be applied to fractional factorial designs. Given the resources, greater power can be achieved with more measurement occasions, and these additional measurement occasions should be included.

Together, the application of fractional factorial designs allows all training elements of interest to be concurrently investigated with reasonable power. When incorporated into MOST, these designs can provide beneficial inferences about training programs. To further support this notion, an example is presented to better understand the logic and method of MOST.

Multiphase Optimization Strategy Example — Study 2

Multiphase optimization strategy method

For the MOST example, self-guided online safety training programs were analyzed, which included identical material and elements to the RCT training programs. They were as follows: (1) the presentation of pre-training material, which provided workplace safety statistics given at the onset; (2) a short pre-test, which gave several example post-test items given at the onset; (3) the notification of a chance to win $25 if a certain percentage was scored on the post-test; (4) the inclusion of material that provided information about general workplace safety given at the onset; (5) and the opportunity to write notes throughout the presentation of material but removed before the post-test (6) a short safety video presented midway through the material.

The first phase of MOST is the screening phase. To test the effectiveness of these elements during the screening phase, a fractional factorial design was chosen. Three-way and higher interactions were not of interest for this

DOI: 10.1002/job

example, and only main effects and two-way interactions were of interest. The PROC FACTEX program in SAS was used to determine which experimental groups should be included in the research design to only test for main effects and two-way interactions.[3] Through only entering the number of variables and statistical effects of interest, this program notes which experimental conditions may be eliminated while producing unbiased results. The results produced 32 conditions, a large reduction from 64, that could test for the effects of interest. Thus, for the screening phase, 32 conditions were created, and the included elements within each condition were mandated through the PROC FACTEX output.

The second phase of MOST is the refining phase. For the reasons noted later on, the refining phase was not applied within the current example. The final phase of MOST is the confirming phase. To compare the optimized training discovered through the screening phase, a RCT was performed, which compared the optimized training to the existing minimal training. The results of this phase provide strong evidence about the improvement, if any, of the optimized training beyond the existing training, and conclude the MOST process.

Multiphase optimization strategy participants

For the screening phase, 103 participants were recruited from an undergraduate student participant pool in return for extra course credit, and 98 completed the entire training. Of these 98 participants, three were placed within each of the 32 conditions, and the extra two participants were randomly assigned. This final sample size of 98 is comparable with the RCT example sample size of 93. Also, because of the inherent proprieties of factorial designs, the sample size of the MOST screening phase provides identical power to detect main effects and interactions as the RCT to detect the difference of two groups (Collins et al., 2009; Dziak et al., 2012), and the sample of the current study is considered appropriate.

For the confirming phase of MOST, 74 participants were recruited from an undergraduate student participant pool in return for extra course credit, and 69 completed the entire training. Of these 69 participants, 34 were placed within the minimal training condition, and 35 were placed within the optimized training condition. The final sample size of 69 is less than the RCT example sample size of 93 because it was assumed that the optimized training would need less power to detect differences beyond the minimal training because of increased effect sizes from removal of detrimental components during the screening phase.

Multiphase optimization strategy measures

The measures administered during the MOST example were the same as those used in the RCT example: total time on training and the post-test. Those who had a z-score above 3.5 on total time on training were rescaled to the next highest value.

Results

The results of this regression analysis is presented in Table 3. Only one element demonstrated a significant impact on post-test scores and was marginally significant on time spent on training. This element was the opportunity to write notes that were not provided during the post-test. No two-way interactions were significant. For this reason, only the element of providing the opportunity to write notes was included after the screening phase, and this training with the retained element is henceforth considered the optimized training.

Within the current example, no refining phase was performed. This is because different levels of the retained element, providing the opportunity to write notes, could not be created. The confirming phase was immediately conducted after the screening phase, and it is meant to ensure that the optimized training from the screening and refining phases is actually an improvement beyond an existing training. Two groups were created for the confirming phase.

[3]The orthoplan procedure in SPSS, the FrF2 package in R, and the support. CEs package in R can provide similar information as the PROC FACTEX program in SAS.

DOI: 10.1002/job

Table 3. Regression analysis results of multiphase optimization strategy.

	Post-test score				Time spent on training			
	B	Std. error	*β*	*t*	*B*	Std. error	*β*	*t*
Constant	.530	.017		30.405***	127.542	23.835		5.351***
Element 1	.005	.017	.027	.262	36.747	23.812	.155	1.543
Element 2	.011	.017	.064	.614	-6.742	23.806	-.028	-2.83
Element 3	.013	.017	.081	.774	7.249	23.810	.031	.304
Element 4	-.009	.017	-.054	-.519	19.178	23.835	.081	.805
Element 5	.042	.017	.255	2.437*	40.718	23.832	.172	1.709†
Element 6	-.023	.017	-.138	-1.317	-37.284	23.835	-.157	-1.564
1 × 2 interaction	-.002	.017	-.015	-.141	-8.181	23.773	-.035	-.344
1 × 3 interaction	-.013	.017	-.078	-.749	24.186	23.785	.102	1.017
1 × 4 interaction	-.014	.017	-.086	-.822	17.814	23.812	.075	.748
1 × 5 interaction	.001	.017	.009	.085	20.784	23.801	.088	.873
1 × 6 interaction	-.002	.017	-.011	-.109	-29.104	23.812	-.123	-1.222
2 × 3 interaction	.002	.017	.012	.115	-28.991	23.801	-.122	-1.218
2 × 4 interaction	-.010	.017	-.058	-.555	22.744	23.806	.096	.955
2 × 5 interaction	.007	.017	.044	.416	26.619	23.785	112	1.119
2 × 6 interaction	.003	.017	.017	.158	19.611	23.806	.083	.824
3 × 4 interaction	.021	.017	.128	1.227	-3.074	23.810	-.013	-.129
3 × 5 interaction	-.006	.017	-.036	-.342	-.757	23.773	-.003	-.032
3 × 6 interaction	.010	.017	.060	.574	15.657	23.810	.066	.658
4 × 5 interaction	.009	.017	.052	.499	30.715	23.832	.130	1.289
4 × 6 interaction	018	.017	-.108	-1.035	.068	23.835	.000	.003
5 × 6 interaction	.020	.017	.121	1.152	-6.273	23.832	-.026	-.263
R^2				.169				.196

Note: Effect coded variables were created to indicate the presence of each element. Values were coded −1 if the element was not present and 1 if the element was present. Two-way interaction terms were created through the product of the two relevant elements.
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

The first group underwent a minimal training with only the core elements of the training present, identical to the minimal training within Study 1. The second group underwent the optimized training with the significant element retained, providing the opportunity to write notes. To test the differences between the two groups, an independent samples *t*-test was performed. Levene's test for the equality of variances for the *t*-test comparing the total time spent on training was significant ($F = 4.817$, $p < .05$), indicating that equal variances should not be assumed for this analysis. Individuals within the optimized training spent more time on the required training components compared with the existing training ($t(59.458) = 3.696$, $p < .001$). Those within the optimized training condition spent 175 seconds on each topic page on average (Std. Dev = 150 seconds), whereas those within the minimal training spent 66 seconds on each topic page on average (Std. Dev. = 96 seconds). Also, the optimized training group received higher post-test scores than the existing training group ($t(66) = 2.272$, $p < .05$). The average post-test score of the optimized training group was 57 percent (Std. Dev. = 17 percent), whereas the minimal training group received an average score of 48 percent (Std. Dev. = 16 percent). Together, the optimized training is a significant improvement beyond the minimal training with only a single element added, and MOST is adept in evaluating and discovering an optimized training.

Multiphase Optimization Strategy Discussion

When comparing the MOST example to the RCT example, several aspects should be noted. MOST is more involved than performing an RCT. More steps are required to optimize a training program than simply comparing two training

DOI: 10.1002/job

programs, but this extra effort provides several benefits. In the RCT example, all elements were included in the final training, as an RCT is unable to compare individual elements. In MOST, the screening phase demonstrated that most of these components did not have an impact on any training outcomes, and some demonstrated a negative relationship with the outcomes. In the RCT, these null and problematic components remained in the final training, but they were removed during MOST. Removing these elements in MOST created a more effective and cheaper training with less trainee and trainer time required, and this optimized training is an improvement over the existing training.

Additionally, MOST can provide greater theoretical inferences than RCTs. As mentioned, several authors have proposed several training effectiveness models, each suggesting various factors and interactions that impact learning and development (Alvarez, Salas, and Garofano 2004; Cannon-Bowers et al., 1995; Grossman & Salas, 2011; Salas et al., 2012). Unlike RCTs, MOST can test the effects of many elements within a single experiment. For instance, in the example provided, six separate RCTs would be required to test each individual element, whereas a single MOST achieved the same outcome — and more. MOST also provides direct insights about the moderated and mediated effects of individual elements, an impossible feat with RCTs. Thus, in addition to practical advancements, MOST has several implications for theory development.

While MOST can surely benefit organizational scholars and practitioners, another training evaluation method can provide even more benefits. This method is entitled SMART.

Sequential Multiple Assignment Randomized Trial

Individuals vary in their personal characteristics and perceptions, and these differences engender differential training reactions and outcomes (Bauer et al., 2012; Collins, Murphy, and Bierman, 2004, Driskell, et al. 1994; Martocchio & Judge, 1997). Despite interest in such topics, little empirical research has investigated the impact of certain training elements on these trainee characteristics, also called attribute–treatment interactions (ATIs; Gully & Chen, 2010; Tannenbaum and Yukl, 1992). As Gully and Chen (2010) note, "the most surprising thing about previous research on ATIs is how little work has been done" (p. 38). Further yet, even less research has created adaptive training programs that systematically present certain elements, when appropriate, to harness trainee characteristics to prompt trainee success. A possible reason is the difficulty to study and apply.

First, individual differences that cause differential training effects are not random; instead, they are systematic. Many research designs demand random error to analyze certain phenomenon and cannot be applied to this scenario. Second, the methodologies required to evaluate adaptive training programs are often more complicated than traditional training program evaluations — RCTs. Given these difficulties, training programs, which harness the individual differences between employees, are more uncommon than general, broad training programs.

Fortunately, a particular experimental research design, SMART, can systematically analyze elements that are catered to individual trainees, with the purpose of discovering which elements counteract (accentuate) the individual differences that prompt negative (positive) outcomes. In other words, SMART analyzes the main effects of time-varying adaptive training elements as well as the elements' interactions with each other and with trainee characteristics. Before this design is discussed, the type of training needed to cater to individual employees should be explained. This type of training is analogous to adaptive interventions (Rivera et al., 2007), and the current article labels it adaptive training.

Adaptive Training

An adaptive training is a "multistage process that adapts to the dynamics of the 'system' of interest (e.g., individuals, couples, families, or organizations) *via* a sequence of decision rules that recommend when and how the [training] should be modified in order to maximize long-term primary outcomes" (Nahum-Shani et al., 2012, p. 457). Adaptive

DOI: 10.1002/job

training programs are individualized to the needs of participants, and they are adjusted over time based on these needs. Many authors have commented on the need for their inclusion in organizational research and practice (Entin & Serfaty, 1999; Gully & Chen, 2010). In fact, as long ago as 1969 have researchers been calling for adaptive training (Kelly, 1969).

Several conceptual advantages arise from adaptive training programs. Most training models propose elements that produce effects only after certain sequencing and/or only impact certain trainees (Alvarez et al., 2004; Cannon-Bowers et al., 1995; Grossman & Salas, 2011). In traditional training programs, practitioners underutilize these relationships. Traditional programs often administer a uniform training to each trainee, and any element sequencing is theoretically assumed (but not empirically justified) to produce the greatest effects. More effective outcomes can be obtained through adaptive training programs that have the sequencing of events empirically justified. Likewise, in traditional trainings, elements that are assumed to only impact certain individuals are often applied to all individuals. An adaptive training that only presents elements specifically believed to be effective for the individual trainee would be more effective and cost reducing. Therefore, the need for adaptive training is present, as they can produce many solutions in organizational research and practice.

As mentioned, adaptive training programs change over time. These changes are determined by decision rules. Decision rules are planned choices to guide elements, based upon participants' "tailoring variables" that moderate training effects. Tailoring variables can include demographics (e.g., gender), personality (e.g., conscientiousness), motivation (e.g., high/low), outcomes (e.g., initial performance), self-perceptions (e.g., self-efficacy), and many others. The proper use of a tailoring variable ensures that participants are categorized into the most beneficial training. For example, if previous studies have shown that highly motivated employees benefit from a longer training whereas poorly motivated employees benefit from a shorter training, then employees should be directed to a training program based on their motivation.

Once particular tailoring variables and decision rules have been chosen, logic steps should be clearly defined using IF and THEN statements. An IF statement frames the outcomes of a tailoring variable, whereas THEN statements direct the follow-up actions to be taken based on outcomes. To clarify, take the following example. For the organization, they decided that all employees should get the standard, three-part engineering training. Then, after a period of time, trainees are evaluated on their training progress, which serves as the tailoring variable. A decision rule then decides which follow-up actions should be taken. If the evaluation is positive, employees continue the standard training. If the evaluation is negative, employees are placed into an intensive training with an additional element. The logic step for this procedure is:

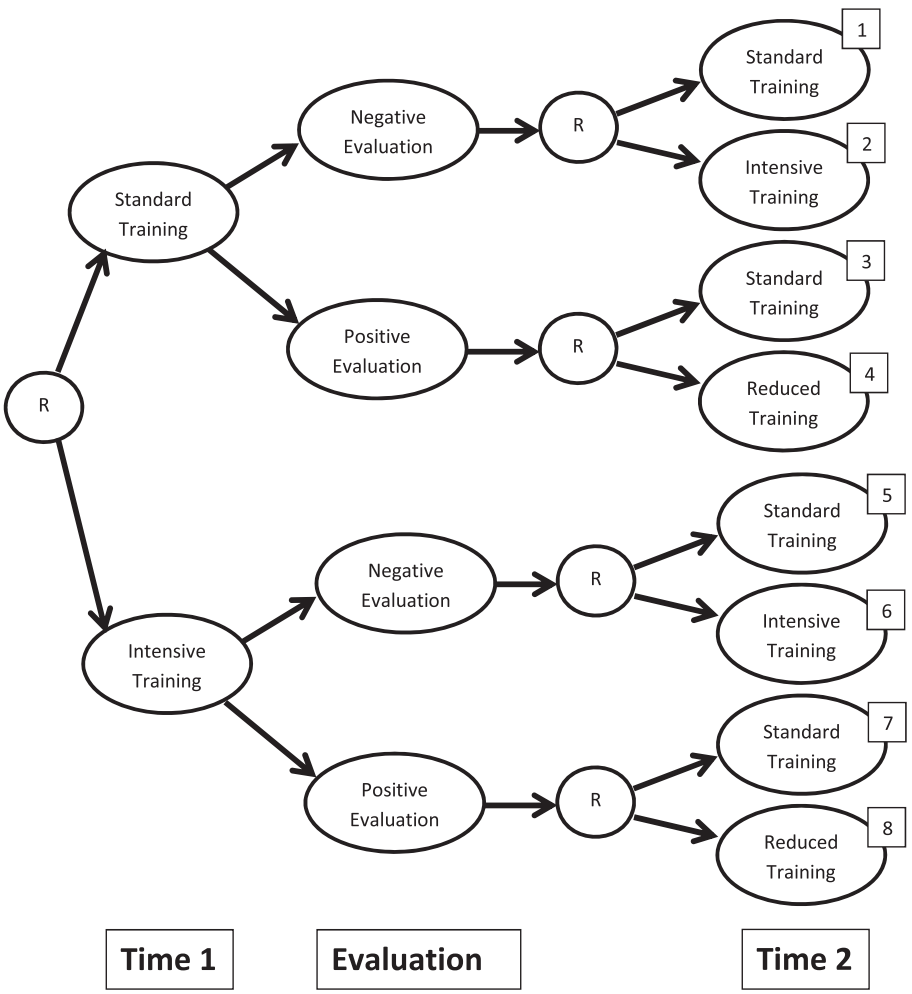
```
First-stage training = {Standard Training}
    IF evaluation = {positive}
        THEN at second stage = {Continue Standard Training}
    ELSE IF evaluation = {negative}
        THEN at second stage = {Administer Intensive Training}
```

Sequential Multiple Assignment Randomized Trial Overview

Sequential multiple assignment randomized trial (SMART) can efficiently evaluate an adaptive training (Murphy, 2005; Murphy et al., 2007), and the method was created as a direct improvement over RCTs. Like MOST, SMART was created in the field of engineering and quickly adapted to evaluate efficient large-scale adaptive interventions (Collins et al., 2007; Rivera et al., 2007). The goal of this section is to demonstrate how SMART can evaluate efficient and powerful adaptive training programs.

DOI: 10.1002/job

The goal of SMART is to determine which training elements and sequences produce the best outcomes. To do this, initial (Time 1) training programs (i.e., standard and intensive) and follow-up (Time 2) actions (i.e., administer standard, administer intensive, and administer reduced) are created. Then, tailoring variables are chosen. These tailoring variables are used to categorize individuals after the initial (Time 1) training; however, when performing SMART, random assignment guides participants to follow-up (Time 2) actions, instead of decision rules. This random assignment allows for inferences about element effectiveness and sequencing. If decision rules were used to categorize participants into follow-up actions during SMART, then a confounding effect would be introduced based on the tailoring variable and comparisons could not be made. For these reasons, SMART cannot be considered an adaptive training and can only be used to evaluate a potential adaptive training. A visual representation of SMARTs is presented in Figure 4. Finally, after an adaptive training has been chosen, an RCT can compare the adaptive training with an alternative. Once again, the following example clarifies SMART.



Note: R represents random assignment to groups.

Figure 4. Visual Representation of a sequential multiple assignment randomized trial (SMART) design

DOI: 10.1002/job

Sequential Multiple Assignment Randomized Trial Example — Study 3

For the SMART example, self-guided online safety training programs were analyzed. The material was identical to the RCT and MOST examples, but the SMART example was not interested in all six elements. Instead, SMART evaluates an adaptive training. For this reason, participants were initially randomly assigned into a standard or intensive training. In the standard training, no extra elements were added. The intensive training included the following: (1) the presentation of pre-training material, which provided workplace safety statistics given at the onset; (2) the notification of a chance to win $25 if a certain percentage was scored on the post-test; (3) the inclusion of material, which provided information about general workplace safety given at the onset; and (4) a short safety video presented midway through. After the introductory section, before any informational material was presented, a tailoring variable was added. This variable questioned participants about whether they wanted to learn more about the post-test and was meant to represent performance motivation. Regardless of their answer, trainees were assigned to one-of-two conditions. The first was a reduced training, which did not include any extra elements. The second was an enhanced training that included (1) a short pre-test, which gave several example post-test items, was given after the tailoring variable and (2) the opportunity to write notes throughout the presentation of material but removed before the post-test. Together, this research design analyzes the efficacy of the first training conditions, standard or intensive, as well as the efficacy of the second training conditions, reduced or enhanced. More importantly, the training also determines the effect of sequences and tailoring variables.

Sequential multiple assignment randomized trial participants

For SMART, 103 participants were recruited from an undergraduate participant pool in return for extra course credit. Of these 103 participants, 55 were placed within the initial standard training, and 48 were placed within the initial intensive training. Then, 52 participants were randomly assigned to the reduced training, and 51 participants were randomly assigned to the enhanced training. In total, 28 participants attended the initial standard training and then the reduced training, 27 participants attended the initial standard training and then the enhanced training, 24 participants attended the initial intensive training and then the reduced training, and 24 participants attended the initial intensive training and then the enhanced training.

Sequential multiple assignment randomized trial measures

The measures administered during the SMART example were the same as those used in the RCT and MOST example: total time on training and the post-test. Those who had a z-score above 3.5 on total time on training were rescaled to the next highest value.

Tailoring variable

In addition to outcome variables, the SMART example included a tailoring variable. The tailoring variable was a single question that read, “In the current study, you’ll read a section about safety behaviors in businesses, then take a test. Would you like to know more about the test?” A response of “Yes” was treated as indicative of higher motivation. Those who answered “Yes” were coded as “1,” and those who answered no were coded as “0.” Regardless of their answer, participants were randomly assigned to a second training condition.

DOI: 10.1002/job

Sequential multiple assignment randomized trial results

First, the efficacy of the initial training assignments, to the standard or intensive training, as well as the second training assignments, to the reduced or enhanced training, was tested using a regression analysis (Table 4). In regards to the average time spent upon each section, the results demonstrated that a main effect of the initial training assignment was not significant ($\beta = .073$, $t(101) = .751$, $p > .05$), the main effect of the second training assignment was significant ($\beta = .245$, $t(101) = 2.502$, $p < .05$), and their interaction was not significant ($\beta = -.011$, $t(101) = -.108$, $p > .05$). With regards to the only significant effect, those who were assigned to the enhanced training (mean seconds = 108.95, Std. Dev. = 15.279) spent a significantly longer time on each section than those assigned to the reduced training (mean seconds = 55.426, Std. Dev. = 14.979). Alternatively, with regards to the total correct post-test questions, the main effect of the initial raining assignment was not significant ($\beta = .053$, $t(101) = .529$, $p > .05$), the main effect of the second training assignment was not significant ($\beta = .100$, $t(101) = .994$, $p > .05$), and their interaction was not significant ($\beta = .077$, $t(101) = .763$, $p > .05$). Overall, the second assignment was the only significant predictor of any variable, which is further discussed in the succeeding text.

Second, the differential effect of the training upon certain individuals was tested. In the current example, the tailoring variable was whether participants wanted to learn more about the post-test, which 42 replied "Yes" and 58 replied "No." To test this, regression analysis was conducted, which included the initial training assignment, the second training assignment, participants' response to the tailoring variable, and their interactions. In this analysis, the interaction between the tailoring variable and the second assignment denotes whether the second training assignment demonstrated a significant differential impact upon individuals depending upon their tailoring variable response. This interaction was not statistically significant (time spent on training, $\beta = -.076$, $t(101) = -.743$, $p > .05$; post-test score, $\beta = -.037$, $t(101) = -.360$, $p > .05$). Also, the three-way

Table 4. Regression analysis results of SMART example with and without tailoring variables.

	Time spent on training				Post-test score			
	B	Std. error	*β*	*t*	*B*	Std. error	*β*	*t*
SMART example without tailoring variables								
Constant	82.188	10.698		7.682***	.498	.019		26.618***
FA	8.031	10.698	.073	.751	.010	.019	.053	.529
SA	26.762	10.698	.245	2.502*	.019	.019	.100	.994
Interaction	-1.154	10.698	-.011	-.108	.014	.019	.077	.763
R^2				.066				.017
SMART example with tailoring variables								
Constant	81.592	11.268		7.241***	.508	.019		26.167***
FA	6.478	11.268	.059	.575	.007	.019	.039	.373
SA	24.163	11.268	.219	2.144*	.013	.019	.072	.696
TV	5.445	11.268	.049	.483	.043	.019	.228	2.228*
FA and SA interaction	-2.051	11.268	-.019	-.182	.023	.019	.121	1.166
FA and TV interaction	-10.346	11.268	-.094	-.918	.005	.019	.025	.238
SA and TV interaction	-8.373	11.268	-.076	-.743	-.007	.019	-.037	-.360
FA, SA, and TV interaction	-12.688	11.268	-.114	-1.126	.007	.019	.040	.386
R^2				.095				.069

Note: Each condition was dummy coded. In the first assignment variable, participants were assigned a 0 if they attended the standard training, and they were assigned a 1 if they attended the intensive training. In the second assignment variable, participants were assigned a 0 if they attended the reduced training, and they were assigned a 1 if they attended the enhanced training. FA = first assignment; SA = second assignment; TV = tailoring value; SMART = sequential multiple assignment randomized trial.
$^*p < .05$, $^{**}p < .01$, $^{***}p < .001$

DOI: 10.1002/job

interaction tests the cumulative differential impact of the initial training assignment, the second training assignment, and trainees' response to the tailoring variable. The three-way interaction was not statistically significant (time spent on training, $\beta = -.114$, $t(101) = -1.126$, $p > .05$; post-test score, $\beta = .040$, $t(101) = .386$, $p > .05$). Therefore, participants' success within the training did not rely upon differential training assignments based upon their tailoring variable. These results are included within Table 4.

The SMART revealed that the only significant training impact was the second training assignment that did not demonstrate a differential impact based upon participants tailoring variable. From these results, the training creator would most likely opt for the initial standard training followed by the second enhanced training for all participants. This would result in the following logic step for the created training:

First-stage training = {Standard Training}
THEN at second stage = {Administer Enhanced Training}

This finding demonstrates a benefit of SMART. Although SMART was created to produce adaptive training programs, the results do not force an optimized training to be adaptive.

Lastly, in a follow-up RCT, the training program identified through SMART was compared with the existing, minimal training program. Levene's test for the equality of variances for the *t*-test comparing the total time spent on training was significant ($F = 5.263$, $p < .05$). Individuals in the optimized training program did not spend more time on the required training components compared with the existing training program ($t(59.313) = -.276$, $p > .05$). On the other hand, the optimized training group received higher post-test scores than the existing training group ($t(110) = 2.789$, $p < .01$). The average post-test score of the optimized training group was 55 percent (Std. Dev. = 20 percent), whereas the minimal training group received an average score of 45 percent (Std. Dev. = 19 percent). Together, the optimized training program is a significant improvement beyond the minimal training program, and SMART is adept in evaluating and discovering an optimized training program — even if it is not adaptive.

Sequential multiple assignment randomized trial discussion

Sequential multiple assignment randomized trial (SMART) analyzes the effectiveness of adaptive training programs, whereby each combination of all possible adaptive training programs is tested. In the current example, this was seen through analyzing the main effects and interactions of the initial and second training programs. RCTs are unable to efficiently perform this analysis and can only compare a single adaptive training program against an alternative. Also, SMART analyzes the differential impact of training programs on participants through tailoring variables. In the example, the adaptive training programs did not demonstrate differential impacts, indicating that a standard training was suitable. Although an adaptive training was not produced, the example demonstrates the ability of SMART to identify the most suitable training, even if it is not adaptive.

Additionally, the theoretical benefits of SMART should be emphasized. The current example did not observe any sequencing effects between elements, and the tailoring variable did not demonstrate any significant effects. While not statistically significant, SMART still provided information about unfolding dynamics during training, and future authors should likewise apply smart to uncover emergent training effects. To start, training self-efficacy relates to training outcomes and reactions (Esfandagheh, Harris, and Oreyzi, 2012; Howard, 2014; Simosi, 2012; Sitzmann & Weinhardt, 2015). In fact, Cannon-Bowers and colleagues (1995) directly call for the use of training self-efficacy as a tailoring variable, as trainees with low training self-efficacy often receive poor outcomes from training. From their perspective, trainees should have their training self-efficacy gauged, and those with low self-efficacy should be assigned to a pre-training self-efficacy development program. Since this suggestion, authors have used less-than-ideal methods, such as RCTs, to test the effect of training self-efficacy on training

DOI: 10.1002/job

elements and sequencing, but SMART can provide greater depth to training theory investigation. Once the unfolding dynamics of training self-efficacy have been understood, authors could approach other variables in a similar manner.

Overall Discussion

Random confirmatory trials (RCTs) are limited in their ability to evaluate organizational training programs, resulting in drawbacks to theory and practice. The primary limitation of RCTs is their inability to analyze particular training elements. Therefore, other training evaluation methods should be considered.

Multiphase optimization strategy (MOST) and SMART are able to evaluate individual training elements, among many other benefits when compared with RCTs. In MOST, the screening and refining stages are added before the RCT to optimize a training program, and these stages can provide information about the effectiveness of each element. Then, once the effective elements have been identified, an RCT can analyze the successful elements together. Through this follow-up RCT, the effectiveness of an optimized training can be demonstrated. Additionally, SMART likewise incorporates a multiphase stage that discovers an optimal adaptive training before it is compared in a RCT. In SMART, participants are randomly assigned to conditions to determine the efficacy of certain training combinations. This random assignment allows unbiased analyses of all training combinations of interest. Also, with the inclusion of tailoring variables, authors can determine the differential impact of various training combinations upon individuals with particular responses to the tailoring variables. The SMART can identify the adaptive training most effective for these individuals, providing an individualized training for trainee populations.

Despite these strengths, RCTs may still be the preferred evaluation in some situations. First, it is entirely possible that a training program or environment is so complex that MOST or SMART are simply too difficult to implement, and the traditional "all or nothing" method is more appropriate. Second, if a new training is only an incremental improvement of an existing training, RCTs may be more effective than MOST or SMART. Third, some training programs may consist of several elements that individually demonstrate small effect sizes but yield a more sizeable impact when analyzed together. In this case, an RCT with all the elements grouped together may be the most effective method. Given that some situations will call for certain training evaluation methods, Figure 2 presents a flowchart that aids in determining the ideal situation to use RCTs, MOST, and SMART. Table 1 can also aid in researcher methodological decisions, as it labels some of the strengths and weaknesses of these research designs.

Lastly, MOST and SMART provide opportunities for the advancement of theory and practice. Several authors have proposed factors that impact the effectiveness of a training program, such as a climate for learning (Kraiger et al., 1993; Lim & Morris, 2006; Rouiller & Goldstein, 1993). While these factors may impact an overall training, it is unclear which elements they impact. It is possible that these factors only impact elements with particular characteristics, and MOST and SMART may uncover these occurrences. Also, MOST and SMART allows the simultaneous investigation of multiple research questions that may hasten the understanding of theory. Particularly, authors can test multiple theoretically justified training elements within a single MOST or SMART, which would otherwise take multiple RCTs. For example, Liu and colleagues (2014) performed a meta-analysis investigating the impact of multiple job search intervention elements, which they called components. Their meta-analysis combined the results of multiple RCTs to observe overall direct effects, but similar conclusions could be obtained through a single application of MOST or SMART while providing inferences about interactions. Similarly, SMART can further explain individual differences in training outcomes and reactions. In performing these tests, it should also be reiterated that the sample size required for MOST and SMART are the same as RCTs, although these methods investigate a greater number of effects. Therefore, MOST and SMART provides several benefits without requiring greater sample sizes.

DOI: 10.1002/job

Limitations

Certain limitations of the current article should be noted. The examples used student samples and did not represent a workplace training evaluation. Many issues may arise when MOST or SMART are used in organizational settings. For example, MOST involves alternative conditions, causing employees to possibly perceive unequal treatment (Cook et al., 1979; Shadish, Cook and Campbell 2002). Also, some organizations may be logistically unable to assign employees to the number of conditions that MOST and SMART require. Practitioners and researchers could benefit from a study investigating the practicality of MOST and SMART in organizational settings. This would provide substantial support for the evaluation methods, or could uncover problems that may prompt an adaptation of these methods for organizational settings. Such work would likely uncover hybrid methods that might not be as encompassing as the originals but still an improvement over RCTs. If such new methods are desired, authors should incorporate Sackett and Mullen's (1993) discussion of threats to validity in training evaluations and ensure that these hybrid methods are still able to accurately draw conclusions about training programs.

Additionally, in RCT, MOST, and SMART, researchers and practitioners must make decisions about *a priori* effect size cutoffs, and several authors have proposed methods to choose these cutoffs (Collins et al., 2007; Collins et al., 2011). The current article focused on statistical significance testing, as readers are likely familiar with these analyses; however, statistical significance testing poses several concerns (Cohen, 1995; Nickerson, 2000), and other cutoffs may be more ideal. Particularly, cost–benefit ratios may prove advantageous. When using cost–benefit ratios, the cost of each element is compared with the utility of its outcome, resulting in a dollar amount that the element benefits the organization. The element is discarded if it costs more than it achieves, even if it is statistically significant. Future research should test the potential of cost–benefit ratios and other cutoffs.

Lastly, the current article proposed the basic form of MOST and SMARTs. Recent authors have proposed variations to these two methods for specific purposes (Collins et al., 2007; Danaher & Seeley, 2009). Future researchers should investigate instances which variations to MOST and SMARTs are appropriate and answer relevant training research questions. Again, this may lead to yet more evaluation models, some that are more workable for given situations.

Future Research

An important first step in furthering MOST and SMART is an analysis of where such evaluations might be most effectively implemented. Clearly, these approaches require more trainees, experimental conditions, and organizational planning. These should not be reasons to dismiss complex approaches, as MOST and SMART can have tremendous payoffs. A hard look is needed at organizations that train vast amounts of employees and/or organizations that provide training almost continuously. Examples are abound. Public transportation organizations must train thousands of bus operators each year, and New York City alone trains between 800 and 1,000 each year (Jacobs et al., 1996). If the training is viewed from the perspective of a national imperative, an organization like the American Public Transit Association could provide the leadership and necessary sample size, and alternative approaches to identify the best training program would benefit the entire industry. Relatedly, military and government settings also include sufficient sample sizes and organizational control to apply complex training evaluations.

We could also look to those organizations that hire vast amounts of individuals because of high rates of employee turnover. Call centers and hotels have been well described as examples of high turnover, and high levels of new hires resulting in massive training efforts (Aksin et al., 2007; Lam et al., 2002). Similarly, there

DOI: 10.1002/job

are high hiring rates for large retail stores during the holiday season. Looking to these organizations and planning a study that might unfold over several years of seasonal hiring could create fertile ground for MOST and SMART.

Also, smaller organizations could immediately apply MOST and SMART to evaluate training programs that are easy to modify. The current article applied MOST and SMART to an online training, which is an immediate possibility for others. Online training programs, while requiring resources to develop, require little additional effort to include or exclude particular elements. Likewise, this has been seen in Public Health research, as online interventions were the first application of MOST and SMART (Collins et al., 2005; Murphy, 2005).

Several additional avenues of future research should also be explored. MOST is ideal for many theoretical questions. For example, transformational leadership consists of four primary dimensions (Bass & Avolio, 1990; Bass, 1991), and authors have created training programs that are composed of four modules to target each dimension. MOST can determine which modules are most effective for overall trainee development. More importantly, MOST can also determine whether the training modules actually enhance the dimensions they are meant to improve, or whether a single module improves each dimension of transformational leadership. Findings such as these could prompt a deeper understanding of training and content of interest, leadership.

Similarly, SMART can determine the best sequencing of modules, as certain modules may be most effective when building upon a previously trained skill. In a transformational leadership training, for example, a module training idealized influence may be most effective when following an individualized consideration module, as the trainee may need to understand methods to devote personal attention before they can become an ideal role model. In addition to creating an ideal adaptive training, SMART benefits theory development through discovering the unfolding and longitudinal dynamics of training elements. Further, the extant research upon ATIs may provide guidance for the construction of adaptive training programs through SMART (Gully & Chen, 2010; Tannenbaum and Yukl, 1992). Therefore, module order may be important for a host of other constructs and models, which can be easily analyzed through SMART.

Additionally, sample size considerations with non-nested data were noted, but companies often consist of nested units. Sample size considerations for factorial experiments using nested subjects should be investigated further. Only one substantial discussion of the topic has been published. Dziak and colleagues (2012) only considered group sizes of 20 and 100 in their power analyses, whereas they explored the use of 25, 30, 40, and 50 groups. Many researchers use group sizes of much smaller than 20, and the effect of varying group size is largely unknown. Similarly, researchers and practitioners are often unable to assign individual employees to different training programs, and the current article noted that RCTs, MOST, and SMART can all use quasi-experimental designs. The current article did not, however, go in-depth about the threats to validity or sample size considerations when using such designs with MOST and SMART — a possible avenue for future discussion (Cook et al., 1979; Shadish et al., 2002).

Lastly, the current article largely focused on individual learning. RCTs, MOST, and SMART can all be applied to evaluate outcomes beyond learning, such as reactions, transfer, and results (Kirkpatrick, 1994) or cognitive, skill-based, and affective outcomes (Kraiger et al., 1993). Also, all these methods can be applied to evaluate team training programs (Salas et al., 2008). Thus, while the current article focused on these methods in traditional applications, these evaluation methods can also be used to investigate more innovative research questions.

Conclusion

Research on organizational training largely applies a single method, RCTs, which includes several methodological disadvantages. Particularly, RCTs can only test an entire training rather than individual training elements, causing the creation of an optimized training to be an ineffective process. Also, the analysis of adaptive training programs is difficult through RCTs. For these reasons, the current article proposed two innovative methods, MOST and SMART, which may prove useful for the evaluation of training, the optimization of training, and

DOI: 10.1002/job

the advancement of theory. Through three examples, the current article also demonstrated the proper method to perform RCTs, MOST, and SMART, as well as the relevant statistical analyses. Together, the current article may prompt a newfound interest into the methodologies applied within training research and quicken the advancement of theory and practice.

Acknowledgements

A special thanks to James LeBreton for his comments on a previous version of this manuscript.

Author biographies

Matt C. Howard — Matt C. Howard is currently a lecturer at the University of South Alabama and Pennsylvania State University, and he will be an assistant professor of Management in the Mitchell College of Business at the University of South Alabama starting Fall 2016. His research interests include workplace technologies, measurement/personnel selection, and employee development.
Rick Jacobs — Rick Jacobs is a professor of psychology at Pennsylvania State University and is CEO of EB Jacobs, a consulting firm specializing in selection and assessment. His research interests include the assessment of individuals for jobs, assessment centers and issues surrounding the measurement and reduction of adverse impact in employment decision making.

References

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology*, *60*, 451–474.
Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, *16*(6), 665–688.
Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review*, *3*(4), 385–416.
Barling, J., Weber, T., & Kelloway, E. K. (1996). Effects of transformational leadership training on attitudinal and financial outcomes: A field experiment. *Journal of Applied Psychology*, *81*(6), 827.
Bass, B. M., & Avolio, B. J. (1990). Developing transformational leadership: 1992 and beyond. *Journal of European Industrial Training*, *14*(5).
Bass, B. M. (1991). From transactional to transformational leadership: Learning to share the vision. *Organizational Dynamics*, *18* (3), 19–31.
Bauer, K. N., Brusso, R. C., & Orvis, K. A. (2012). Using adaptive difficulty to optimize videogame-based training performance: The moderating role of personality. *Military Psychology*, *24*(2), 148–165. http://dx.doi.org/10.1080/08995605.2012.672908.
Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*(4), 1065–1105.
Brown, K. G., & Gerhardt, M. W. (2002). Formative evaluation: An integrative practice model and case study. *Personnel Psychology*, *55*(4), 951–983.
Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology*, *71*(2), 232–245. http://dx.doi.org/10.1037/0021-9010.71.2.232.
Campbell, J. P. (1988). Training design for performance improvement. In J. P. Campbell, & R. J. Campbell (Eds.)and Associates (Ed.), *Productivity in organizations* (pp. 177–216)pp. San Francisco: Jossey-Bass.

DOI: 10.1002/job

Cannon-Bowers, J. A., & Salas, E., Tannenbaum, S. I., & Mathieu, J. E.(1995). Toward theoretically based principles of training effectiveness: A model and initial empirical investigation. *Military Psychology*, *7*(3), 141.

Cascio, W. F., & Aguinis, H. (2005). Applied psychology in human resource management. Pearson/Prentice Hall.

Cohen, J. (1995). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Collins, L. M., Baker, T. B., Mermelstein, R. J., Piper, M. E., Jorenby, D. E., Smith, S. S., et al. (2011). The multiphase optimization strategy for engineering effective tobacco use interventions. *Annals Of Behavioral Medicine*, *41*(2), 208–226. http://dx.doi.org/10.1007/s12160-010-9253-x.

Collins, L. M., Murphy, S. A., & Bierman, K. L. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science*, *5*(3), 185–196.

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods*, *14*(3), 202–224.

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals Of Behavioral Medicine*, *30*(1), 65–73. http://dx.doi.org/10.1207/s15324796abm3001_8.

Collins, L. M., Murphy, S. A., & Strecher, V. J. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trail (SMART): New methods for more potent eHealth interventions. *American Journal of Preventative Medicine*, *32*(5S), S112–S118.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. *351*,). Boston: Houghton Mifflin.

Danaher, B. G., & Seeley, J. R. (2009). Methodological issues in research on web-based behavioral interventions. *Annals of Behavioral Medicine*, *38*(1), 28–39.

Dick, W., & Carey, L. (1996). *The systematic design of instruction* (4th ed.,,). Reading, MA: Addison-Wesley.

Driskell, J. E., Hogan, J., Salas, E., & Hoskin, B. (1994). Cognitive and personality predictors of training performance. *Military Psychology*, *6*(1), 31–46. http://dx.doi.org/10.1207/s15327876mp0601_2.

Dziak, J., Nahum-Shani, I. R., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods*Advance online publication. http://dx.doi.org/10.1037/a0026972PMCID:PMC3351535.

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors*, *41*(2), 312–325.

Esfandagheh, F. B., Harris, R., & Oreyzi, H. R. (2012). The impact of extraversion and pre-training self-efficacy on levels of training outcomes. *Human Resource Development International*, *15*(2), 175–191.

Geis, G. L. (1987). Formative evaluation: Developmental testing and expert review. *Performance and Instruction*, *26*, 1–8.

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development*, *15*(2), 103–120.

Gully, S., & Chen, G. (2010). Individual differences, attribute-treatment interactions, and training outcomes. *Learning, training, and development in organizations*, 3–64.

Gunst, R. F., & Mason, R. L. (2009). Fractional factorial design. *WIREs Comp Stat*, *1*, 234–244. http://dx.doi.org/10.1002/wics.27.

Howard, M.C. (2014). Creation of a computer self-efficacy measure: Analysis of internal consistency, psychometric properties, and validity. *CyberPsychology, Behavior, and Social Networking*, *17*(10), 429–433.

Isler, R. B., Starkey, N. J., & Williamson, A. R. (2009). Video-based road commentary training improves hazard perception of young drivers in a dual task. *Accident Analysis and Prevention*, *41*, 445–452.

Jacobs, R. R., Conte, J. M., Day, D. V., Silva, J. M., & Harris, R. (1996). Selecting bus drivers: Multiple predictors, multiple perspectives on validity, and multiple estimates of utility. *Human Performance*, *9*(3), 199–217.

Kelly, C. R. (1969). What is adaptive training? *Human Factors*, *11*(6), 547–556.

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels*. San Francisco, CA: Berrett-Koehler(Original work published 1959).

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*(2), 311.

Lam, T., Lo, A., & Chan, J. (2002). New employees' turnover intentions and organizational commitment in the Hong Kong hotel industry. *Journal of Hospitality & Tourism Research*, *26*(3), 217–234.

Lesch, M. F. (2008). Warning symbols as reminders of hazards: Impact of training. *Accident Analysis and Prevention*, *40*, 1005–1012.

Lim, D. H., & Morris, M. L. (2006). Influence of trainee characteristics, instructional satisfaction, and organizational climate on perceived learning and training transfer. *Human Resource Development Quarterly*, *17*(1), 85–115.

Martocchio, J. J., & Judge, T. A. (1997). Relationship between conscientiousness and learning in employee training: Mediating influences of self-deception and self-efficacy. *Journal of Applied Psychology*, *82*(5), 764–773. http://dx.doi.org/10.1037/0021-9010.82.5.764.

Medley-Mark, V., & Weston, C. B. (1988). A comparison of student feedback obtained from three methods of formative evaluation of instructional materials. *Instructional Science*, *17*(1), 3–27.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, *24*, 1455–1481.

DOI: 10.1002/job

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., & TenHave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence*, *88*(S2), S24–S30.
Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., … & Murphy, S. A. (2012). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods*, *17*(4), 457.
Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241.
Rivera, D. E., Pew, M. D., & Collins, L. M. (2007). Using engineering control principles to inform the design of adaptive interventions: A conceptual introduction. *Drug and Alcohol Dependence*, *88*(Suppl 2), S31–S40.
Rouiller, J. Z., & Goldstein, I. L. (1993). The relationship between organizational transfer climate and positive transfer of training. *Human Resource Development Quarterly*, *4*(4), 377–390.
Sackett, P. R., & Mullen, E. J. (1993). Beyond formal experimental design: Towards an expanded view of the training evaluation process. *Personnel Psychology*, *46*(3), 613–627.
Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., et al. (2008). Does team training improve team performance? A meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(6), 903–933.
Salas, E., Tannenbaum, S. I., Kraiger, K., & Smith-Jentsch, K. A. (2012). The science of training and development in organizations: What matters in practice. *Psychological science in the public interest*, *13*(2), 74–101.
Shadish W. R, Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning.
Saroyan, A. (1992). Differences in expert practice: A case from formative evaluation. *Instructional Science*, *21*(6), 451–472.
Simosi, M., (2012). The moderating role of self-efficacy in the organizational culture-training transfer relationship. *International Journal of Training and Development*, *16*(2), 92–106.
Sitzmann, T., & Weinhardt, J. M.(2015). Training Engagement Theory A Multilevel Perspective on the Effectiveness of Work-Related Training. *Journal of Management*, 0149206315574596.
Smith, A., & Smith, E. (2007). The role of training in the development of human resource management in Australian organizations. *Human Resource Development International*, *10*, 263–279.
Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology*, *43*, 399–441.
Tracey, J. B., & Tews, M. J. (2005). Construct validity of a general training climate scale. *Organizational Research Methods*, *8*(4), 353–374.

DOI: 10.1002/job