7

PERSONNEL PSYCHOLOGY
1993, 46

BEYOND FORMAL EXPERIMENTAL DESIGN: TOWARDS AN EXPANDED VIEW OF THE TRAINING EVALUATION PROCESS

PAUL R. SACKETT, ELLEN J. MULLEN
Industrial Relations Center
University of Minnesota

Textbook treatments of training evaluation typically equate evaluation with the measurement of change and focus on formal experimental design as the mechanism for controlling threats to the inference that the training intervention produced whatever change was observed. This paper notes that two separate questions may be of interest: How much change has occurred? and, Has a target performance level been reached? We show that the evaluation mechanisms needed to answer the two types of questions are markedly different, and discuss circumstances under which an evaluator's interests will focus on one, the other, or both of these questions. We then discuss alternatives to formal design as mechanisms for reducing various threats to validity, and discuss trade-offs between internal validity and statistical conclusion validity.

In this paper we wish to develop a series of ideas that expand on traditional approaches to training program evaluation. First, we note that evaluating training is typically equated with the measurement of change, and evaluation designs are critiqued in terms of their adequacy for answering the question, Can the degree of change attributable to training be quantified? We will argue that while quantifying the degree of change is important in some circumstances, in a variety of applied situations the organization's primary interest is in determining whether trainees have reached some target performance level. We will show that the training evaluation methods needed to quantify the degree of change due to training are different from those needed to determine whether a target performance level has been reached. We will develop prescriptions for when it is necessary to measure change and when it is sufficient to measure level of achievement.

Second, textbook treatments often view formal experimental design as the sole mechanism for avoiding threats to internal validity in settings

The authors wish to thank Rich Arvey, Ray Noe, and Cheri Ostroff for their helpful comments on earlier drafts of this manuscript.

Correspondence and requests for reprints should be addressed to Paul R. Sackett, Industrial Relations Center, University of Minnesota, 271 19th Ave. South, Minneapolis MN 55455.

where it is important to assess change. In light of the very real constraints on organizations in terms of the feasibility of formal experimental design, we wish to explore alternative methods of evaluating the degree to which various potential threats to internal validity are serious impediments to conclusions about training effectiveness.

Third, textbook treatments give precedence, if not exclusive treatment, to internal validity over other forms of validity in evaluating training programs. Recent work (e.g., Arvey & Cole, 1989; Arvey, Cole, Hazucha, & Hartanto, 1985;) has made clear the sample size requirements needed to insure adequate statistical power. The findings are sobering: For many organizations, evaluation via formal experimental design is simply not feasible. In light of this work on statistical power, we will examine trade-offs between threats to internal validity and threats to statistical conclusion validity.

Experienced training evaluators no doubt recognize that decisions about whether to continue, modify, or eliminate training programs may need to be made on the basis of less than ideal evaluation data, as do some textbook authors (e.g., Goldstein, 1986). Our goal here is to focus attention on the trade-offs involved in deviating from formal experimental design as the mechanism for training evaluation.

Measuring Change and Measuring Level of Achievement

Consider the following set of events. A psychologist is designing an assessment center for a high-level management position. One exercise requires each candidate to deliver an oral presentation proposing a course of action that a firm should take based on the candidate's review of extensive data about the firm's operations and financial position. Based on experience developing assessment centers in a variety of settings, the psychologist believes that untrained assessors cannot provide accurate ratings of candidate performance in exercises such as this. Thus, a small group of potential assessors are identified, and receive training which familiarizes the assessors with the dimensions to be rated and with all of the potentially important firm data that will be available to the candidates. The trainees then rate a set of videotaped presentations (using role-playing candidates) which have been rated by an expert panel, thus permitting a comparison of trainee ratings and "true scores."

Only trainees whose ratings are within a specified distance from the true scores are eligible to serve as assessors.[1]

We would argue that effective training evaluation has occurred. The relevant skill has been measured after the training, and it can now be documented that assessors possess the desired level of this skill. From an applied problem solving point of view, the problem has been solved.

Yet from the point of view of many textbooks on the topics of training and chapters on training in human resource management texts, the process above is woefully inadequate as an evaluation. There is no pretest, and there is no control group at all, much less random assignment to experimental and control groups. Many texts would exhort students to strive for a more rigorous evaluation than that described above. Some would go so far as to denounce the above process as useless: "We would suggest that no outcome evaluation at all is preferable to an evaluation without pretesting or some form of control group. Since evaluations without such precautions are without validity the company might as well save the money" (Camp, Blanchard, & Huszczo, 1986, p. 167). "If no control group is available and directors cannot obtain a pretraining measure of performance, a post-training evaluation becomes fairly meaningless" (Grove & Ostroff, 1991, p. 5214).

Both sets of authors quoted above explicitly equate "training evaluation" with "measuring change" (e.g., either obtaining an effect size measure to quantify the extent of change attributable to the training program or simply documenting that statistically significant change occurred). An examination of textbooks devoted solely to training (Camp et al, 1986; Goldstein, 1986; Wexley & Latham, 1991), and of a variety of human resource management textbooks reveals that this perspective is common. Some explicitly equated evaluation with change measurement (e.g., "the measurement of training and development outcomes is a special case of the much broader problem of change measurement," Cascio, 1991, p. 407), while others did not make an explicit statement, but did so implicitly by presenting formal experimental design as the mechanism by which training is to be evaluated.

---

[1] Reviewers expressed concerns that our example was unrealistic in failing to use a pretest to determine whether some individuals were already competent, and thus the organization could save the expense of training them. Our response is twofold. First, we certainly agree that there are many situations in which such a pretest makes sense. We view this as a judgment call on the part of the evaluator, based on the evaluator's assessment of the likelihood that candidates are already competent and the costs of a pretest. In our example we noted that prior experience led the evaluator to the judgment that untrained candidates could not perform the assessment task without training. Second, contrary to reviewer claims that such a pretest would be routine in organizational settings, we have yet to encounter an assessment center in which a pretest is used as the basis for waiving assessor training for new assessors.

We believe that it is useful to expand training evaluation to incorporate two different classes of questions. While in some situations the organization may want an answer to the question of precisely how much change occurred, in others organizations may be asking, Has a specific level of the skill, knowledge, or performance attribute of interest been achieved? Applying behavior science measurement principles to document that a desired end state has been achieved is, we believe, a form of evaluation which may be very appropriate for answering the organization's question. In the case of our assessment center example above, given the absence of a pretest or a control group, the psychologist cannot answer with any confidence the question of how much change occurred. But the psychologist can document that assessors have achieved the desired level of rating accuracy. Documenting this answers the applied question of interest to the organization, namely, Can we be confident that our assessors are capable of providing accurate ratings? It is also responsive to the *Guidelines and Ethical Considerations for Assessment Center Operations* (Byham, 1991), which call for some measurement of assessor performance prior to actual service as an assessor to insure that they are sufficiently trained to function as assessors. It is critical to insure that achievement is measured reliably. Our assessment example incorporated the rating of multiple videotaped presentations as the basis for certifying achievement, rather than the rating of a single presentation, as a means of increasing reliability.

Thus, we suggest a broader framework for training evaluation. We believe that evaluation encompasses two different classes of questions, How much change has occurred? and, Has a specific level of performance been achieved? The two questions are qualitatively different. The standard outcomes of a study using formal experimental design are a measure of statistical significance and an effect size measure (e.g., the standardized difference between the experimental and control group posttest means). Reporting that a training program produced, on average, a statistically significant one-half standard deviation improvement in some outcome variable answers the question, How much change has occurred? but not the question, Has a specific level of performance been achieved? Left undetermined is how many employees perform up to the standard of competence desired by the organization. Conversely, a posttest only no control group evaluation design can answer the question, Has a specific level of performance been achieved? without indicating quantitatively, How much change has occurred? Thus, it is important to ascertain whether the requirements of a given situation call for the measurement of change, the measurement of achievement, or both.

We see three situations which call for change measurement. The first involves situations in which the evaluator wishes to estimate the utility

of the training program. Regardless of the complexity of one's utility model, one component of all contemporary versions of such models is an effect size measure *d*: the standardized difference between trained and untrained groups. The evaluator may wish to estimate utility for a variety of reasons, including (a) an external request for evidence of training effectiveness, (b) a self-imposed need to know how effective a program was, or (c) the marketing value of documented past successes, among others.

The second involves situations in which the evaluator wishes to compare the efficacy of two different training programs. Attempting to answer the question, Is approach A or approach B more effective? demands greater methodological sophistication than the simple post-training evaluation needed to answer the question, Has a specified performance level been achieved? Answering this comparative question requires training some individuals using each approach, and a fair comparison of the approaches demands an evaluation design that permits unambiguous interpretation of study outcomes (e.g., assurances that the superior post-training performance of one group reflects the superiority of the training method rather than assignment of better performers to that group).

We suggest that one factor that will influence taking the comparative question, Is approach A or approach B more effective? seriously is the number of times a training program will be repeated. When a program will be offered only once, as in the case of training accompanying the introduction of some new equipment, training program designers take their best shot at estimating what approach will be most effective under the given cost and time constraints, but there is neither the opportunity to compare training approaches, nor strong reason to do so. (If you somehow found out after the fact that another approach would have been more effective, what could you do but shrug?) On the other hand, when a program will be offered repeatedly, as in the case of a program given to a group of new hires every month, the long-term consequences make comparative research potentially worthwhile.

The third type of situation where quantifying the degree of change produced is of interest occurs when the interests of the training evaluator go beyond applied problem solving for the benefit of the organization in question, and extend toward a broader contribution to the research base regarding various training approaches. While a simple posttest may suffice to answer the organization's question, Has a specified performance level been achieved? the evaluator with an eye toward contributing to a better understanding of training processes may wish to introduce a formal experimental design if possible. We suggest that the evaluator ought to make clear to the organization that the purpose of the formal design,

namely, a contribution to scientific knowledge, may go beyond the immediate needs of the organization. At the same time, though, we certainly hope evaluators are persuasive in convincing organizations to think beyond the short term and to support the high-quality evaluation research needed to extend our knowledge about the training process.

We now move to a description of situations in which a measure of a specified level of performance is needed. We see two critical features of such situations, both of which must be present for a performance level-oriented evaluation to be viable. The first is that a clear target performance level exists. Often this is externally imposed, as in the case where a firm commits itself to providing a specific level of performance. For example, in a manufacturing setting, a firm may bid on a government contract to produce parts within specified tolerances. Sometimes the target is internally imposed, as in our assessment center example earlier. The decision as to what constitutes an acceptable degree of similarity between trainees' ratings and true scores reflects a somewhat arbitrary judgment that the firm can modify at will.

The second is that there is an interest in documenting the performance of individual trainees. Given the specification of the target performance level, assessment of whether each trainee meets this target is used as the basis for some decision about the trainee. For example, successful trainees may be certified as qualified in the domain covered by the training program, while unsuccessful trainees may repeat the program, receive some other remedial treatment, be assigned other duties, be terminated from the training program, or, at the extreme, be terminated from the organization. We reiterate here again the need for reliable measurement of training outcomes, given that these outcomes will be the basis for decisions about individuals.

Thus, a key distinction between change-oriented evaluation and performance level-oriented evaluation is that the former is focused on making decisions about whether the training program is functioning as intended, while the latter involves evaluating both the program and the individuals participating in the program. With change-oriented evaluation, a common goal is inferring the degree of change that can be expected with future repetitions of the program. Subsequent repetitions of the program are typically not subjected to the same formal evaluation; only with some substantial change, such as a change in curriculum or in the trainee population, is re-evaluation deemed necessary. In contrast, given the need to document the level of individual performance for decision making purposes, performance level-oriented evaluation is typically ongoing: The level of performance achieved by each class of trainees is assessed.

The above discussion has outlined situations which would lead to a change-oriented evaluation or a performance level-oriented evaluation. Note that the two are certainly not mutually exclusive. A firm can certainly be simultaneously interested in documenting both the degree of change produced by a training program and that individual trainees do or do not meet the target performance standard. Whichever is the case, the evaluation should be undertaken with the greatest degree of rigor possible, as will be discussed later in the paper.

Finally, we consider situations in which there is no concern for either change-oriented or performance level-oriented evaluation. We see four reasons why evaluation may be undertaken: (a) to make decisions about the future use of a training program or technique (e.g., continue, modify, or eliminate), (b) to make decisions about individual trainees (e.g., certify as competent, provide additional training, etc.), (c) to contribute to a scientific understanding of the training process, or (d) for political or public relations purposes (e.g., documenting success may increase the training function's credibility and visibility within an organization). Absent concern for the one of the first two (i.e., a program will not be repeated and no personnel decisions will be based on trainee performance), evaluation is unlikely unless the evaluator has a strong enough interest in the third or fourth to expend the resources needed for evaluation. Professional or personal development programs offered as an employee benefit and positioning the organization as a progressive place to work may be examples of programs not likely to undergo rigorous outcome evaluation.

Pre-Experimental Designs and Threats to Internal Validity

In this section we focus on evaluation methodology in situations in which measurement of change is seen as important. Treatments of measuring change in training texts and HR texts focus on the use of experimental and quasi-experimental designs to reduce threats to internal validity. Such treatments commonly make three points: (a) one should strive for the most rigorous evaluation strategy possible, (b) rigorous evaluation is often very difficult in applied settings, and (c) some compromise may be necessary. When interested in measuring change, we concur with Goldstein (1986) that "the job of the training analyst is to choose the most rigorous design possible and to be aware of its limitations" (p. 144).

The question is whether there are limits to the degree one is willing to deviate from the ideal of formal experimental design with random assignment to treatment and control groups. There is little dispute with some forms of quasi-experimental designs, such as a pretest-posttest

nonequivalent control group design, and others that Cook and Campbell (1976) label as "generally interpretable designs." However, various authors draw the line at three designs labelled "pre-experimental designs" by Campbell and Stanley (1963) and labelled "generally uninterpretable designs" by Cook and Campbell (1976). These are the posttest-only no control group design ("... is completely worthless and should not be used to evaluate training," Fisher, Schoenfeldt, & Shaw, 1990, p. 351), the pretest-posttest no control group design ("should never be used to measure training outcomes," Cascio, 1991, p. 396), and the posttest-only nonequivalent control group design.

Our concern about textbook treatments of training evaluation is that they focus heavily on design issues as the mechanism for controlling threats to validity. While one should generally take advantage of the opportunity to make use of a design that neatly controls for various threats, situations will often arise where the best one can do is to implement one of the three pre-experimental designs labelled above as generally uninterpretable. Given the realities of the training world, we'd like to encourage consideration of methods other than formal design in order to address threats to the inference that training has produced a change of a given magnitude.

Campbell and Stanley's (1963) famous table which summarized with "+," "–," and "?" the degree to which various designs controlled the various threats to internal validity contained a footnote that tends to disappear when the table is reprinted or modified: "It is with extreme reluctance that these summary tables are presented because they are apt to be 'too helpful.'... it is against the spirit of this presentation to create uncomprehended fears of, or confidence in, specific designs" (p. 8). As Cook and Campbell (1979) assert, "estimating the internal validity of a relationship is a deductive process in which the investigator has to systematically think through how each of the internal validity threats may have influenced the data" (p. 55).

Thus, our purpose here is to call for careful consideration of the risk of various threats in a given situation, and also to suggest that in organizational settings a variety of strategies other than formal design exist for minimizing threats to internal validity. Cook and Campbell (1979) refer to "contextual knowledge" and "intelligent presumptions" as means of inferring that the treatment, rather than some other event, caused the change in question. In other words, we can often use rational judgment and previous knowledge about employees to substitute for the benefits of a pretest and a control group. This resembles Scriven's (1976) "modus operandi" approach to evaluation, modeled after a police detective searching for clues, to which Cook and Campbell (1979) also refer: "The researcher can sometimes function as a detective, noting the

level of different variables and using this information to rule out some threats to both internal and construct validity" (p. 97).

For example, history effects as an alternative explanation for change can be examined through direct inquiry, via interview or questionnaire, as to whether trainees have undergone other experiences concurrently with the training program which could affect posttest measures. Maturation effects may be ruled out on logical grounds in many cases simply due to the short time duration of the training program. Instrumentation is typically not an issue when measurement of training outcomes is standardized. Access to performance ratings, selection test scores, and other file data may provide insight into the risk of statistical regression or selection as plausible threats to validity, as they may offer insight into the atypicality of a group receiving training or into similarity between training and control groups not created randomly. Similar data may help address mortality issues, as the status on these file variables of individuals not completing all of the measurements in the evaluation study can be examined. While not an exhaustive list of threats to internal validity, this set of examples is intended to give a flavor for the kind of "detective work" that can be used to assess the risk of various threats.

We must acknowledge, though, that there are two possible results of an investigation into the likelihood that a specific threat (say, history) is a viable alternative explanation for the observed change. The first is that our investigation leads to the conclusion that history is not a viable threat: We can detect no other activity occurring concurrently with training that could plausibly produce the observed change, thus contributing to our hypothesis that training was effective. The second is that the investigation leads to the conclusion that history is a viable threat: Other events relevant to the variable of interest did occur concurrently with the training. In this case our investigation has not reduced ambiguity about the cause of the observed change; rather, it has bolstered the case for an alternative explanation. Thus, while the strategy of careful investigation into the plausibility of various threats can yield useful information, it is not clear in advance whether that information will or will not result in increased confidence that observed effects are due to training. Formal experimental design, in contrast, is intended as a mechanism allowing an inference that observed effects are due to training whether or not various factors such as history are operating. Thus, formal design is generally the preferred strategy; when it is not possible we still advocate attempting an evaluation, even if all that is possible is a pre-experimental design. We argue that a pre-experimental design, paired with careful investigation into the plausibility of various threats, is still better than no evaluation at all, given that organizations must make decisions about future training efforts with or without evaluation data.

Some comments on specific pre-experimental designs are in order. A pretest-posttest no control design or a posttest-only nonequivalent control group design does at least permit the computation of a measure of change; the problem with these designs revolves around attributing the change to training or to some other factor(s). The posttest-only no control group design, though, is the most frequently criticized of the pre-experimental designs (recall our earlier quotations labeling this as "completely worthless," "fairly meaningless," and "the company might as well save the money"). This design highlights the distinction we draw between performance level-oriented evaluation and change-oriented evaluation in that it is generally a perfectly adequate design for answering the question, Has a target level of performance been achieved? but generally is inadequate to answer the question, Has changed occurred, and, if so, how much? The critical feature of this design is that it produces only one measure: trainees' standing on the variable of interest after training. What is lacking is a measure of their standing prior to training, and thus a measure of change cannot be obtained (except in situations where one can safely infer a pre-training score of zero, as in the case of a totally novel technology or area of knowledge).

The underlying message here is that there are alternatives to true experimental design which can, under some circumstances, be of value and may be better than no evaluation at all. Heneman, Schwab, Fossum, and Dyer (1989) would agree, for they assert, "It is often worthwhile to evaluate training programs with a less sophisticated design than not to evaluate them at all" (p. 448). As we have described, the potential threats to internal validity can be adequately overcome in many instances.

One danger that we do wish to preclude is the tendency to focus solely on the evaluation's design. As Cascio (1991) argues, "First of all, exclusive emphasis on the design aspects of measuring training outcomes is rather narrow scope" (p. 399). Certainly the purpose of the evaluation, the content and objectives of the training course, and the characteristics of the employees and the work context all deserve first consideration, and formative evaluation (Scriven, 1967) merits emphasis as an adjunct to the summative evaluation emphasized in this paper. The evaluation can then be designed to fit the purpose, within the bounds of the situation. That is, we can use the most rigorous evaluation design that is feasible, practical, and logical given the training content and its context and the characteristics of the trainees.

Trade-Offs Between Internal Validity and Statistical Conclusion Validity

Recent writing by Arvey and coworkers (Arvey & Cole, 1989; Arvey, Cole, Hazucha, & Hartanto, 1985) has examined the statistical power of

various evaluation designs. With small samples, power is often quite low. This creates a dilemma for the individual reading textbook treatments of training evaluation: On the one hand the reader is told that training should always be evaluated, and on the other hand one is told that there is little to be gained from evaluation designs with inadequate power. In some situations, the evaluator has the flexibility to simply increase the sample size for an evaluation study in order to insure adequate power; in others, sample size is fixed. For example, consider a new plant start-up where plans are to hire a class of 20 every 2 months until the plant is fully staffed. The firm wishes to try out a training program on the first class and decide whether to continue to use it with subsequent classes. In yet other situations, sample size is limited by constraints on resources such as time and money.

Cook and Campbell (1976, 1979) discuss trade-offs between the various forms of validity (i.e., internal, external, construct, and statistical conclusion). They note that basic and applied researchers may have different priorities, and they speculate as to the relative importance of the various forms of validity for different types of research problems. Regardless of the setting, though, statistical conclusion validity (i.e., the determination of whether groups are different on the variable of interest, regardless of cause) is judged by Cook and Campbell to be of lesser importance than internal validity.

We would like to offer the proposition that statistical conclusion validity needs to take first priority in applied training evaluation research. The question, Is there a difference between trained and untrained groups? needs to be answered before addressing, Can the difference be attributed to the training intervention? What follows from this proposition is that it may be reasonable in some settings to trade off internal validity for statistical conclusion validity. Consider the following example. A firm wishes to evaluate a training program in a pilot plant with 20 employees, intending to train additional cohorts of new hires if the program is successful. The firm is willing to use a true experimental design, with pretest, posttest, and random assignment to treatment and control groups. Based on Arvey et al. (1985), we calculate the statistical power of this design to detect what Cohen (1988) labels a moderate effect size of .5 standard deviations, to be .46 (analyzed using the pretest score as a covariate, and assuming a population pretest-posttest correlation of .50). On the other hand, if one does not establish a control group, but simply administers a pretest, trains all 20 employees, and administers a posttest, the power of the test of the pretest-posttest difference is .85. With this pre-experimental design, power is higher since the analysis is based on an N of 20 pretest and 20 posttest values, in contrast with the

ANCOVA design based on 10 treatment groups values and 10 control group values.

Clearly, the interpretation of the no-control group design is much more ambiguous than the true experimental design, given the threats to internal validity inherent in the no-control group design. But in terms of power, the true experimental design is strikingly inadequate. In terms of the choice between threats to internal validity and threats to statistical conclusion validity, note that the previous section of this paper offered alternatives to formal design as mechanisms for assessing the risk of various threats to internal validity. In contrast, we see no ready mechanism for combatting the low statistical power of the true experimental design in this setting where N is constrained. Thus, we suggest that there may be settings in which pre-experimental designs should be chosen over true experimental designs. Constrained N settings where a decision about future training activities must be made may be exceptions to the general rule that one ought to use the most rigorous design possible. Even if the organization will permit random assignment to treatment and control groups, the evaluator may be better off saying no.

This concept is presented more systematically in Table 1, which documents the statistical power of various evaluation designs as a function of the effect size one wishes to detect, the sample size available, and the population pretest-posttest correlation. This table is modeled on Arvey and Cole (1989), who compared the power of three data analytic approaches that can be used with a true experimental design: (a) a posttest-only analysis, which simply compares the posttest scores of a training group and a control group, (b) a gain score analysis, which computes a pretest-posttest difference score and compares the difference scores of a training group and a control group, and (c) a covariance analysis, which compares posttest scores of a training group and a control group, treating pretest score as a covariate. Table 1 includes these three data analytic approaches, and also includes the pre-experimental design discussed above, namely, the pretest-posttest no control group design. The table includes power values for effect sizes of .2, .5, and .8 standard deviations, labelled small, medium, and large effect sizes respectively by Cohen (1988). Arvey and Cole tabled power values for pretest-posttest correlations of .1, .3, .5, .7, and .9; in the interest of space, we report power only for the intermediate values of .3, .5, and .7. The table includes power values for total N of 20, 40, 60, 80, and 100. It is critical to interpret this value correctly: An N of 40 means that 40 individuals are available for the study. In the case of a true experimental design, this means 20 are assigned to a treatment group and 20 to a control group. In the case of the pretest-posttest no control group design, pretraining and posttraining measures are obtained for all 40.

TABLE 1 - FOR ENTRY

Statistical Power of Various Evaluation Designs as a Function of Effect Size, Sample Size, and Pretest-Posttest Correlation

	PO		Gain			ANCOVA			Pre-post/ no control		
n		r_{xy}=	.3	.5	.7	.3	.5	.7	.3	.5	.7
Effect size = 0.2											
20	.11		.10	.11	.14	.13	.15	.20	.19	.22	.31
40	.15		.13	.15	.20	.19	.22	.31	.28	.35	.49
60	.19		.16	.19	.26	.24	.29	.41	.37	.46	.63
80	.22		.19	.22	.31	.28	.35	.49	.45	.56	.74
100	.26		.22	.26	.36	.33	.41	.56	.52	.63	.81
Effect size = 0.5											
20	.29		.24	.29	.40	.36	.46	.63	.74	.85	.96
40	.46		.37	.46	.64	.59	.71	.87	.84	.93	.98
60	.61		.49	.61	.79	.74	.85	.96	.95	.98	.99
80	.72		.59	.72	.88	.84	.93	.98	.98	.99	.99
100	.80		.67	.80	.94	.91	.97	.99	.99	.99	.99
Effect size = 0.8											
20	.53		.43	.53	.72	.66	.79	.94	.90	.96	.99
40	.80		.67	.80	.94	.90	.96	.99	.99	.99	.99
60	.92		.82	.92	.99	.97	.99	.99	.99	.99	.99
80	.97		.91	.97	.99	.99	.99	.99	.99	.99	.99
100	.99		.96	.99	.99	.99	.99	.99	.99	.99	.99

Note: N refers to the total number of individuals in the evaluation study. PO = posttest-only, with control group; Gain = pretest-posttest control group design, analyzed by testing difference between gain scores; ANCOVA = pretest-posttest control group design, treating pretest as a covariate.

Examination of Table 1 shows the markedly higher power obtained through the use of the pre-experimental design. For example, with a total available N of 40 and a .5 pretest-posttest correlation, the power to detect a moderate effect size (i.e., .5) is .71 with the most powerful true experimental design available, but reaches .93 with the no-control group design. (Values in this table will not correspond to Arvey and Cole. They erroneously state that "N" in their table refers to total N, when in fact it refers to the sample size in each group.)

Thus, in constrained N situations in which true experimental designs have inadequate power one choice is to use a true experimental design and consequently risk failure to detect an effective training intervention. Another choice is to use the pretest-posttest no control group design which will be more powerful, and thus more likely to detect change. However, the change cannot be unambiguously attributed to the training program. Which is the more costly error? To wrongly abandon a useful training program or to persist in using an ineffective one? Our sense is that the latter error is more amenable to correction, as larger scale evaluation may become feasible in the future if sample size constraints are eased or as more data is accumulated over time. In addition,

the risk of wrongly attributing change to a training program which is, in fact, ineffective drops to the extent that threats to validity can be examined and ruled out on logical or empirical grounds, as discussed in the previous section of this paper. Both of these features lead us to conclude that there are likely to be situations in which it makes sense to trade off internal validity for statistical conclusion validity.

Conclusion

This paper has suggested a broader perspective on a variety of aspects of the training evaluation process. First, we suggested that while textbook treatments commonly equate training evaluation with the measurement of change, in many applied settings the organization's concern is with certifying that a target performance level has been reached, rather than quantifying change. When the concern is simply for certifying a level of performance, evaluation designs that will be inadequate in most cases for measuring change, such as a posttest-only no control group design, will be perfectly adequate.

Second, we suggested that even in settings where change measurement is needed, textbooks overemphasize formal experimental design as the mechanism for ruling out threats to internal validity. While textbooks do note the difficulties in convincing organizations of the need for control groups and random assignment, we see little treatment of mechanisms other than formal design for assessing the degree of threat that various impediments to internal validity pose in a given situation.

Third, we suggested that trade-offs between internal validity and statistical conclusion validity need to be considered. Many training programs are undertaken in settings in which small numbers of trainees are available for study, and traditional evaluation designs may have inadequate statistical power. We endorse greater attention to statistical power issues, and suggest that in some settings where N is limited the creation of a control group to achieve greater internal validity extracts too great a price in terms of threats to statistical conclusion validity.

Finally, we disagree with writers who endorse an "if you can't do it right, don't do it at all" approach to training evaluation. We prefer a pragmatic perspective. The purpose of evaluation is to help organizations make decisions about future training activities, and we call for giving students the tools needed to assess the type of evaluation possible in a given situation, to conduct the most informative evaluation possible given the constraints of the situation, and to communicate to organizational decision makers both the strengths and the limitations of whatever evaluation data is obtained.

REFERENCES

Arvey RD, Cole DA. (1989). Evaluating change due to training. In Goldstein IL and Associates (Eds.), *Training and development in organizations* (pp. 89–117). San Francisco: Jossey-Bass.

Arvey RD, Cole DA, Hazucha JF, Hartanto FM. (1985). Statistical power of training evaluation designs. PERSONNEL PSYCHOLOGY, *38*, 493–507.

Byham WC and Associates. (1991). *Guidelines and ethical considerations for assessment center operations: Task force on assessment center guidelines*. [Monograph XVI], (rev. ed.). Pittsburgh: Development Dimensions International.

Camp RR, Blanchard PN, Huszczo GE. (1986). *Toward a more organizationally effective training strategy and practice*. Englewood Cliffs, NJ: Prentice-Hall.

Campbell DT, Stanley JC. (1963). Experimental and quasi-experimental designs for research on teaching. In Gage NL (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Cascio WF. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cohen J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook TD, Campbell DT. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In Dunnette MD (Ed.), *Handbook of industrial and organizational psychology* (pp. 223–326). Chicago: Rand McNally.

Cook TD, Campbell DT. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Fisher CD, Schoenfeldt LF, Shaw JB. (1990). *Human resource management*. Boston: Houghton Mifflin Company.

Goldstein IL. (1986). *Training in organizations: Needs assessment, development, and evaluation* (2nd ed.). Monterey, CA: Brooks/Cole.

Grove DA, Ostroff C. (1991). Training program evaluation. In Wexley KN (Ed.), *Developing human resources* (pp. 5185–5220). Washington, DC: BNA Books.

Heneman HG III, Schwab DP, Fossum JA, Dyer LD. (1989). *Personnel/human resource management* (4th ed.). Homewood, IL: Irwin.

Scriven M. (1967) The methodology of evaluation. In Tyler R, Gagne R, Scriven M (Eds.), *Perspectives of curricular evaluation* (American Educational Research Association Monograph Series on Curriculum Evaluation, pp. 39–83). Chicago: Rand McNally.

Scriven, M. (1976). Maximizing the power of causal investigation: The modus operandi method. In Glass GV (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 101–118). Beverly Hills, CA: Sage.

Wexley RN, Latham GP. (1991). *Developing and training human resources in organizations*. New York: Harper Collins.