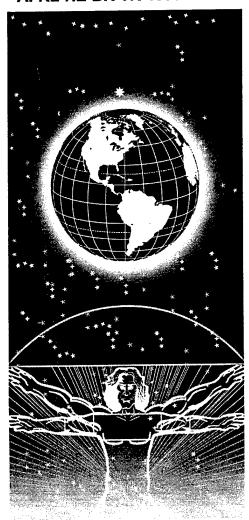AFRL-HE-BR-TR-1998-0130

UNITED STATES AIR FORCE
RESEARCH LABORATORY

A META-ANALYSIS OF THE RELATIONS AMONG TRAINING CRITERIA

George M. Alliger
State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222

Scott I. Tannenbaum
Executive Consulting Group
409 Vesper Ct.
Slingerlands, NY 12159

Winston Bennett, Jr.
HUMAN EFFECTIVENESS DIRECTORATE
MISSION CRITICAL SKILLS DIVISION
7909 Lindbergh Drive
Brooks AFB, TX   78235-5352

Holly Traver
Allison Shotland
State Univeristy of New York at Albany
1400 Washington Avenue
Albany, NY 12222

19991004 068

May 1998

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
HUMAN EFFECTIVENESS DIRECTORATE
7909 Lindbergh Drive
Brooks Air Force Base, TX   78235-5352

DTIC QUALITY INSPECTED 4

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 1999	Interim Report - Septmber 1996-December 1997

4. TITLE AND SUBTITLE

A Meta-Analysis of the Relations among Training Criteria

5. FUNDING NUMBERS

C-   F41624-93-C-5011
PE-   62202F
PR-   1123
TA-   A2
WU   23

6. AUTHOR(S)

George M. Alliger   Holly Traver
Scott I. Tamnenbaum   Allison Shotland
Winston Bennett, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Executive Consulting Group
409 Vesper Court
Slingerlands, NY 12159

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Mission Critical Skills Division
7909 Lindbergh Drive
Brooks AFB, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFRL-HE-BR-TR-1998-0130

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Technical Monitor: Winston Bennett (480) 988-6561, DSN 474-6297

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

An augmented framework for training criteria based on Kirkpatricks' (1959a, 1959b, 1960a, 1960b) model divides training reactions into affective an utility reactions; and learning into posttraining measures of learning, retention, and behavior/skill demonstration. Meta-analysis results among criteria using this framework include the finding of substantial reliabilities across training criteria and reasonable convergence among subdivisions of criteria within a larger level. Utility-type reaction measures were more strongly related to learning or on-the-job performance (transfer) than affective-type reaction measures. Moreover, utility-type reaction measures were stronger correlates of transfer than were measures of immediate or retained learning. These latter findings support recent concurrent thinking regarding use of reactions in training (e.g., Warr & Bunce, 1995). Implications for choosing and developing training criteria are discussed.

14. SUBJECT TERMS		15. NUMBER OF PAGES
Meta-Analysis Training Evaluation Training Criteria Training Effectiveness		23
		16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION ABSTRACT
UNCLASSIED	UNCLASSIED	UNCLASSIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std Z-39-18
298-102 COMPUTER GENERATED

Table of Contents

PREFACE

SUMMARY

An augmented framework for training criteria based on Kirkpatrick's (1959a, 1959b, 1960a, 1960b) model divides training reactions into affective and utility reactions; and learning into posttraining measures of learning, retention, and behavior/skill demonstration. Meta-analytic results among criteria using this framework include the finding of substantial reliabilities across training criteria and reasonable convergence among subdivisions of criteria within a larger level. Utility-type reaction measures were more strongly related to learning or on-the-job performance (transfer) than affective-type reaction measures. Moreover, utility-type reaction measures were stronger correlates of transfer than were measures of immediate or retained learning. These latter findings support recent concurrent thinking regarding use of reactions in training (e.g., Warr & Bunce, 1995). Implications for choosing and developing training criteria are discussed.

A META-ANALYSIS ON THE RELATIONS AMONG TRAINING CRITERIA

INTRODUCTION

The Current State of Training Evaluation

Training researchers agree on the importance of evaluating training effectiveness (e.g., Goldstein, 1993). There is equally strong agreement among training practitioners on the difficulty of doing so (Carnevale & Schulz, 1990). For any training evaluation to be valuable, however, training criteria must be psychometrically sound, meaningful to decision-makers, and must be able to be collected within typical organizational constraints (Tannenbaum & Woods, 1992). Research has revealed that by far the most commonly collected training criteria are trainee reactions (Saari, Johnson, McLaughlin, & Zimmerle, 1988) which, although easy to collect, may or may not be related to other, often more meaningful indicators of training effectiveness.

Perhaps unsurprisingly, Kirkpatrick's four-level model (1959a, 1959b, 1960a, 1960b) continues to be the most prevalent framework for categorizing training criteria. Annual data from the American Society for Training and Development (1997) show that from 1994 to 1996, all companies evaluate on Level 1, with percentages falling regularly as one increases levels: 80 to 90% of companies use Level 2, about 60 to 80% use Level 3 and about 30 to 45% report using Level 4 evaluation. These same ASTD data show that the percentage of courses evaluated at a given level, is quite different: while over 90% of courses are evaluated at Level 1, only about a third are evaluated at Level 2, about 10% at Level 3 and an almost vanishing percent at Level 4.

These results are interesting in part because ASTD could complete a survey where the respondents had no trouble categorizing their training evaluation efforts in terms of Kirkpatrick's model, even if they had not done so previously. This underscores why this taxonomy of training criteria became very popular in business and academia in the first place -- because it addressed a need to understand training evaluation simply yet systematically (Shelton & Alliger, 1993). A first critical point about training evaluation today is then:

1. Kirkpatrick's model of training criteria is the overwhelmingly popular approach to discussing training evaluation.

Of course, this is not all we know, about the state of training evaluation. At little risk of error, we also suggest that:

2. It is generally recommended that training evaluation should be considered an integral part of the ISD process,

3. Trainee reactions (Level 1) are *by far* the most commonly used form of training criteria,

4. It is easier to gather trainee reactions (Level 1) than any other form of evaluation data,

5. Sometimes decisions makers need more than reaction (Level 1) data, and

1

6. There is interest in being able to infer whether training "works" (i.e., whether learning, behavior, or results follow training) from Level 1 data.

Bassi, Cheney, and Van Buren (1997) provide a further discussion of and data supporting most of these points.

Problems with Kirkpatrick's model

The simplicity of Kirkpatrick's four-level model is appealing but, as revealed in more recent work, this simplicity is also a liability. Alliger and Janak (1989) conducted a meta-analytic review of the literature based on Kirkpatrick's model. They concluded that:

> [Kirkpatrick's model] provides a vocabulary and rough taxonomy for criteria. At the same time, Kirkpatrick's model, through its easily adopted vocabulary and a number of (often implicit) assumptions, can tend to misunderstandings and overgeneralizations (pp. 331-332).

While there are problems with Kirkpatrick's model, just how best to think about training criteria is not clear. Perhaps Kirkpatrick's taxonomy requires revision. Some researchers have gone further and presented compelling arguments that entirely different and better models of training evaluation are needed (Holton, 1996; Kraiger, Ford, & Salas, 1993). New approaches are undoubtedly called for, and a thorough model of training effectiveness must include much more than is addressed by any taxonomy of training criteria (cf. Holton, 1996). Nonetheless, the Kirkpatrick typology remains by far the most influential and prevalent approach among practitioners, and, to a certain extent, researchers. For this reason, it can still serve as a point of departure for communicating understandings about training criteria.

This article builds in part upon the meta-analytic work of Alliger and Janak (1989) in examining the relationship among training criteria. It extends their work in the following ways. First, we augmented Kirkpatrick's typology to be less coarse, while maintaining a broad enough framework to facilitate a generalized understanding of training evaluation results. Second, we conducted a meta-analysis of the relationships among training criteria based on this augmented model, using over twice as many studies and four times as many correlations as Alliger and Janak had available in 1989.

Six Key Questions

The goal of this meta-analysis was to address a number of questions which, given the current state of training evaluation, seem to be critical to advancing the field. These questions are:

1. Are typical training evaluation measures reliable?
2. Are trainee reactions related to other measures of training effectiveness?
3. Are all trainee reactions equal in meaning and import?
4. Is trainee learning related to subsequent on the job behavior?
5. Are all trainee learning measures equal in meaning and import?
6. Overall, what measures of training effectiveness are most recommended?

2

AN AUGMENTED TAXONOMY

Below we briefly introduce a scheme for augmenting Kirkpatrick's taxonomy into one that is still simple but captures some additional important distinctions among criteria. Table 1 outlines the differences between the original and modified frameworks. The augmented framework guides the meta-analyses that follow. Please note that we do not propose the new framework as a comprehensive replacement for Kirkpatrick's original. Rather, the goal only was to provide an approach to coding for the current meta-analysis that was at least somewhat more sensitive to reasonable distinctions among training criteria. Table 1

TABLE 1: Training Criteria Taxonomies

Kirkpatrick's Taxonomy	Augmented Framework

Reactions	Reactions
	    Affective Reactions
	    Utility Judgments

Learning	Learning
	    Immediate Knowledge
	    Knowledge Retention
	    Behavior/Skill Demonstration

Behavior	Transfer
Results	Results

Level 1. Reactions

Definition. Originally, to assess "reactions" was to ask trainees how they liked and felt about training. That is, reactions were emotionally-based opinions. Indeed, the term "reactions" seems to imply an immediate, more or less unthinking, response. However, the boundary between feeling and a more considered opinion is fuzzy at best, and trainers have asked a wide variety of "reaction" questions. Several researchers have suggested that reaction measures that directly ask trainees about the transferability or utility of the training should be more closely related to other criteria than would reactions measures that ask about "liking" (e.g., Alliger & Janak, 1989; Tannenbaum & Yukl, 1992). Attitude theorists acknowledge the difference between affective and more behaviorally evaluative responses (Eagly & Chaiken, 1992). Therefore, we have broken reactions into two basic components, affective and utility reactions. For the sake of the meta-analyses, we also needed a third sub-category that was a combination of the first two, since in many cases researchers combined both types of questions in one scale.

Interestingly, and independent of this research, Warr and Bunce (1995) suggested a tripartite division of reaction measures: enjoyment of training, usefulness of training, and

3

difficulty of training. Warr and Bunce's (1995) first two concepts are our affect and utility reactions; the third concept, difficulty, is not captured in our scheme. In any case, training difficulty seems to be rarely asked of trainees.

Level 1a. Reactions as affect. Liking of training is easily the most common form of assessment of training. Usually such reactions are obtained via easy-to-administer post-training questionnaires. Reactions of trainees are extremely important for several reasons. First, trainees may be considered one of the "customers" of training. As such, assessment of their satisfaction with training seems entirely in keeping with most current models of provision of organizational services. Second, whether training is liked could have substantial influence on such distant variables as later training attendance, "word of mouth" advertising, subsequent training funding, and so forth.

Level 1b. Reactions as utility judgments. A second kind of reaction variable is operationalized by asking such questions as, "To what degree will this training influence your ability later to perform your job?", "Was this training job relevant?", and "Was the training of practical value?" These questions attempt to ascertain the perceived utility value, or usefulness, of training for subsequent job performance. An interesting question is whether answers to these utility questions correlate more or less strongly with later on-the-job application of trained skills or knowledge than do answers to affect-type reaction measures.

Level 1c. Combined reactions. In some cases correlations for reaction scales were available, but separate items for affective or utility reactions were reported only as a combined score. In this case, a "combined reactions" category was coded for. It is interesting to note that to the extent that reaction scores are reported only in combined form, practitioners and researchers are apparently acting on the assumption that all reactions tap a single underlying construct.

Level 2. Learning

Definition. Usually, learning as a training criterion is indexed by results of traditional tests of declarative knowledge. Many forms of knowledge assessment, however, could fit under this label. The assessment of mental models, for example, is a type of knowledge assessment, albeit relatively newer and much less common than typical measures of knowledge (Kraiger et al., 1993). Procedural knowledge, or performance of trained tasks immediately after training, also demonstrates learning. Hence, we include three sub-categories of learning: knowledge that is assessed immediately after training (most common), knowledge that is assessed at a later time, and behavior/skill demonstration assessed immediately after training.

Level 2a. Immediate posttraining knowledge. Immediate posttraining assessment of learning is fairly common in the training literature. Usually knowledge is assessed by multiple choice test responses, answers to open-ended questions, listings of facts and so forth. That is, trainees are asked to indicate, in one of several ways, how much they know about the training topic(s). From our literature search we found overwhelming indication that traditional tests (e.g., multiple choice tests) are by far the most common, while newer methods of eliciting knowledge structure or associations are, as yet, virtually unused.

4

Level 2b. Knowledge retention. Sometimes training evaluators assess knowledge at a later time rather than (or in addition to) immediately after training. We coded such studies as "retention" measures of learning. That is, Level 2b measures are equivalent to Level 2a measures except that they are administered at some point later than just after training.

Level 2c. Behavior/Skill Demonstration. Kirkpatrick actually used the term "behavior" to refer to any behavioral changes that occur as a result of training. However, he did not make a clear distinction between behavior demonstrated in the training context and behavior demonstrated on the job. That is, his Level 3 may include both results of behavioral skill tests administered at the conclusion of training (i.e., indications of "can do") as well as on-the-job performance (i.e., indications of "does do"). It seems in keeping with Kirpatrick's original intent, however, to retain Level 3 as representative of transfer of training to the job environment. Therefore, we include in Level 2c any indicators of behavioral proficiency when these are measured within the training, rather than the work environment. Thus, in addition to simulations, such immediate posttraining measures as behavioral role plays, behavioral reproduction, scores/grades in a performance-centered class, and ratings of training performance are typical of the kinds of measures that were classified here.

Level 3. Transfer.

Definition. While skilled performance that was assessed at the conclusion of training was coded as Level 2c, Level 3 we term "transfer," to emphasize the on-the-job nature of criteria in this category. A measure was classified as indicating on-the-job performance whenever it appeared that the measure was not only taken some time after training, but that it was in fact some measurable aspect of job performance. Sometimes ratings were used to indicate on-the-job performance; work samples, and work outputs and outcomes were also reported.

Note that while retained knowledge falls under Level 2, behavior that is retained and applied to the workplace is considered transfer, Level 3. This seems most in keeping with an important distinction: it is application to the job that, in most cases, defines training success (Alliger, Tannenbaum, & Bennett, 1995). Simple indication of retained knowledge may not.

Level 4. Results

Definition. "Results" criteria are meant to be those where organizational impact is indexed. Examples of results criteria include productivity gains, customer satisfaction, cost-savings, employee morale (for manager training) and profitability. In some ways, organizational results criteria represent, for training evaluation, the "ultimate" criteria (Brogden & Taylor, 1950). That is, they are at once the most distal from training, and often perceived as the most fundamental to judging training success. These qualities make results Level 4 training criteria seem highly desirable -- and, indeed, when available, it may make good sense to collect them. However, organizational constraints greatly limit the opportunities for gathering Level 4 data (Tannenbaum & Woods, 1992; Shelton & Alliger, 1993), and most training efforts are incapable of directly affecting results level criteria. Unfortunately, the expectations of sponsors of training in regard to results-level outcomes may be unrealistic. Those charged with evaluation efforts may need to manage unrealistic sponsor expectations early in the evaluation process (Tannenbaum, 1996).

Another problem with Kirkpatrick's framework is that it is fairly vague about results level criteria. Some indicators such as employee attendance or scrappage rates could be categorized equally well as behavioral (Level 3) or results (Level 4) criteria. In any case, only three studies provided correlations that might be categorized as being based on Level 4 criteria, so we have not focused on this level in our current meta-analysis.

Content Overlap: Moderator Analysis

Fundamental differences between the different levels of criteria may explain the modest intercorrelations found in Alliger and Janak (1989) (e.g., learning tests are different from behavioral indicators). However, another explanation could be that different measures focused on different training content. For example, a learning measure might have focused on very specific aspects of handling hazardous materials as addressed in training, while a performance measure from the same study may have been a more global rating of job performance. In this case, a low correlation between the criteria may be attributable to low content overlap rather than a fundamental difference between learning and behavior. Therefore, in an attempt to better understand the relationship between criteria, we decided to conduct a moderator analysis based on the closeness of criteria to training content (that is, on the content validity of the measures). Each measure was rated as either being close to training content (i.e., highly content valid) or divergent from training content (less content valid). The prediction here is that if both measures are rated close to training content, then they would correlate more highly than if one or both measures were not constructed in such a way as to closely reflect training content.

METHOD

Procedure.

An updated manual search was conducted for each of the journals used in Alliger and Janak (1989) (*Journal of Applied Psychology, Academy of Management Journal, Academy of Management Review, Journal of Applied Behavioral Science, Personnel Psychology, Personnel,*

and Training and Development Journal). Additionally, a manual search was conducted examining other journals reporting training research (*Group and Organizational Studies, Human Factors, Human Relations, International Journal of Man-Machine Studies, International Journal of Psychology, Journal of Management, Journal of Organizational Behavior, Public Personnel Management Journal, and Training and Education*).

A computer search was conducted using Psyclit and ERIC, databases that contain abstracts of psychological and educational research, as well as Dissertation Abstracts International. Direct communications were also made to researchers conducting training research in order to obtain unpublished studies or studies in press. Finally, references in review papers were also searched.

Alliger and Janak's (1989) meta-analysis produced 12 studies and 26 correlations. The current meta-analysis almost tripled the number of studies, providing a total of 34, and more than quadrupled the number of correlations, yielding 111.

The method of meta-analysis used was that developed by Hedges and Olkin (1985). Like other approaches, this method of research synthesis permits the generation of mean effect sizes and related confidence intervals, and tests of statistically significant differences between mean effect sizes. The effect size of interest of course for this article is the correlation coefficient.

RESULTS

The results of the meta-analysis are presented in Table 2. An appendix including the citation for each study, the format of the measure, measure reliability, timing of measurement, class under each major level (e.g., "affective" under "reactions"), and coding of measure closeness to course content, including a description and/or example of each study measure is also available from the authors.

Reliabilities

Table 2 presents, on the diagonal, the results of a meta-analysis of criterion reliabilities. Average reliabilities for all categories of measures were above .57, with six out of seven above .75.

Correlations involving reactions

Levels 1a and 1b. Affective and utility reactions tended to correlate positively, although only three studies provided such correlations. The range of correlations was from moderate (.19) to large (.64).

TABLE 2: Mean Sample-size Weighted Correlations Among Training Criteria

	Reactions						Learning						Transfer	
	Affective		Utility		Combined		Immediate		Retained		Behavior/ Skill			
	r	n	r	n	r	n	r	n	r	n	r	n	r	n
Reactions														
Affective	.82 (.81)	12	.34 (.28)	3			.02 (.01)	11			.03 (.01)	9	.07 (.03)	6
Utility			.86 (.85)	5			.26 (.20)	6			.03 (-.08)	3	.18 (.12)	3
Combined					.82 (.80)	5	.14 (.09)	6			.12 (.07)	8	.21 (.16)	9
Overall							.08 (.06)	23			.05 (.03)	20	.13 (.10)	18
Learning														
Immediate							.77 (.75)	14	.35 (.29)	2	.18 (.16)	13	.11 (.08)	16
Retained									.58 (.53)	2	.14 (.05)	2	.08 (.03)	4
Behavior/Skill Demonstration											.85 (.84)	9	.18 (.11)	7
Overall													.11 (.09)	27
Transfer													.86 (.85)	13

Note: Values in parentheses show lower 95% confidence bound for mean correlation; n is number of studies combined in calculating each mean correlation. Empty cells indicate that one or fewer correlations were available. Reliabilities are on the diagonal.

Levels 1 and 2a. The overall average correlation between reactions of any type and immediate learning was only .08. This result corroborates Alliger and Janak's (1989) findings -- training reactions should not be used blindly as a surrogate for the assessment of learning of training content. Affective reactions alone correlated on average just about zero with immediate learning (.02). However, there were indications that utility reactions, considered alone, did correlate somewhat with immediate learning (.26). As would also be predicted, those reactions which appeared to have both affective and utility characteristics correlated less than utility reactions alone, but more than affective reactions alone (.14).

Levels 1 and 2b. The few reaction - retention correlations available provided an average index of relationship that again was very small (.04).

Levels 1 and 2c. One advance of this study over Alliger & Janak (1989) is the separate examination of immediate post-training behavior/skill demonstration. On average, again, reactions did not correlate highly with this index of training effectiveness (.05).

Levels 1 and 3. An interesting result of this meta-analysis is that utility and combined reactions, as in the case of immediate and retained learning, again correlated on average somewhat with transfer (.18, .21). Affective reactions, however, correlated less strongly with transfer (.07).

Correlations involving learning (other than with reactions)

Levels 2a and 2b. Only two studies reported correlations between immediate and delayed learning measures. As reasonably expected, these correlations (which are a kind of index of reliability of learning over time) average moderately positive (.35).

Levels 2 and 2c. Both immediate learning (2a) and retention (2b) correlated positively with immediate behavior/skill demonstration measures (.18 and .14, respectively). The weighted average correlation between learning and immediate behavior/skill demonstration was found to be .18. Thus, the moderate positive linkage between knowledge and behavior/skill demonstration found by Alliger and Janak (1989) was confirmed.

Levels 2 and 3. The average correlation between learning and on-the-job performance was somewhat smaller (.11), with immediate learning and job performance correlating at .11 and retention and job performance correlating at .08. Note that these correlations are actually less than those found for utility reactions. On average, immediate posttraining behavior/skill demonstration and on-the-job performance correlated at .18. It is interesting to note that this correlation is similar to that for utility or combined reactions and on-the-job performance (.18, .21). Thus, given these results, behavioral learning predicts "transfer" no better than utility or combined reactions.

Correlations involving results

Immediate learning correlated substantially with results on average (.52). There were, however, only two studies reporting such correlations. Only one study reported a correlation between performance and results. The correlation was .48.

Moderator analysis: Closeness of criteria to training content

A moderator analysis that examined closeness of the criteria to training content (content validity of measures) was carried out. Each measure was rated as either being close to training content (i.e., highly content valid) or divergent from training content (less content valid). The hypothesis was that if both measures are rated close to training content, then they would correlate more highly than if one or both measures were not constructed in such a way as to closely reflect training content. In general, this finding was supported. Very few correlations were coded divergent-divergent. But, in accordance with prediction, the mean correlations between measures that were coded close-close was higher (mean weighted r = .16, based on 45 correlations) than those coded close-divergent (mean weighted r = .04, based on 42 correlations). The lower 95% confidence intervals for these two mean correlations do not include zero, but the difference between them is statistically significant (p < .001). It should be noted that to most of

the measures coded divergent were at the reaction level, hence this analysis is to some extent redundant with the inter-level comparisons.

DISCUSSION

One hundred and eleven correlations were identified for inclusion in this meta-analysis. Although four times the amount uncovered in 1988, the number seems small given the expansiveness of the search. However, we have learned again what Alliger and Janak (1989) noted, namely, that even when studies clearly measured criteria on several levels, the intercorrelations among levels are often not noted in publications. Fortunately, this state of affairs seems to have improved somewhat, with more recent studies reporting such correlations. It may well be that in the past, the emphasis of training research was almost entirely on training effectiveness as indicated by mean change or mean group differences in criteria. Recently, however, there appears to be increasing awareness about understanding training effectiveness more broadly (e.g., Kraiger, et al., 1993), including the relationships among criteria.

The results reveal that at most, there are modest correlations between the various types of training criteria. Not surprisingly, the strongest correlations were exhibited between different criteria from within the same level, indicating convergent validity. Affective and utility reactions were correlated more strongly with each other (.34) than with other measures, and immediate and retained learning measures were correlated more strongly with each other (.35) than with other measures as well.

Because reaction measures are the easiest to collect, it would be ideal if they could be used as surrogate measures of learning and transfer. Unfortunately, most of the correlations between reactions and other indicators of effectiveness were quite small. Nonetheless, some intriguing differences were exhibited. While overall reactions generally failed to correlate with either immediate or delayed learning or with behavior/skill demonstration, they did correlate somewhat with on-the-job behavior (.13). Interestingly, the magnitude of the relationship between training satisfaction and job performance is about the same as that typically exhibited between job satisfaction and job performance (Iaffaldano & Muchinsky, 1985).

As expected, utility and combined reactions correlated with on-the-job performance more highly (.18, .21) than did affective reactions (.07). This may be attributable to the more specific nature of the utility measures. The more behaviorally specific attitudes are, the more likely they are to predict behavior. Ajzen & Fishbein (1973) write, "While measures of attitude toward an object, such as obtained by the Thurstone, Likert, or semantic differential techniques, may perhaps be related to a person's general behavioral tendency with respect to the object, it appears that for the prediction of a given act, attitudinal variables more specific to the act in question will have to be considered" (p. 56). Weigel, Vernon and Tognacci (1974) also found that the more specific the content of the attitude measure is to the behavioral criterion, the higher the relationship between the two. "Attitude measures should be expected to predict only behaviors that are appropriate to or specified by the attitude under consideration" (p. 728). Ajzen and Fishbein (1977) propose that low correlations may stem from the inconsistency between the levels of specificity of attitude and behavior variables. They reviewed a number of studies examining the attitude-behavior relationship in terms of the specificity or the congruence

10

between target and action elements. The results tend to suggest "the relations between attitudes and behaviors tend to increase in magnitude as the attitudinal and behavioral entities come to correspond more closely in terms of their target and action elements" (p. 911). A recent meta-analysis by Kraus (1995) also found higher correlations between attitude and behavior when the two variables were measured at corresponding levels of specificity. Thus, utility reactions may in fact be more useful, for known theoretical reasons, for predicting on-the-job behavior than are typical affective measures.

While the utility reactions versus affective reactions results were anticipated, the utility versus learning results were not. Surprisingly, utility reactions exhibited higher correlations with on-the-job performance than did either immediate or retained measures of learning (.18 versus .11 and .08). One possible explanation assumes that trainees' utility reactions are influenced by their knowledge of the work environment to which they will return. Transfer (or "on-the-job performance") is a function of both knowledge acquisition and the favorability of the work environment for allowing new skills to be applied (Baldwin & Ford, 1988; Tracey, Tannenbaum, & Kavanaugh, 1995). Objective measures of learning indicate whether trainees have acquired knowledge but do not capture whether they will be able to use it on the job. When asked to provide utility reactions, trainees may implicitly consider the situational constraints they will face when they return to the job (Peters, O'Connor, Eulberg, & Watson, 1988), as well as how much they believed they learned during training.

Regardless of their relationship with other measures, from a pragmatic perspective, trainee reactions are important. While positive reactions do not guarantee organizational support, negative reactions can often have an adverse effect on the training department. However, overall the results of this meta-analysis support Alliger and Janak's findings that reaction measures cannot be used as surrogates of other measures. In particular, affective reactions are unrelated to other indicators -- liking does not equate to learning or to performing. If the purpose for collecting reaction measures is to predict or indicate transfer it would be best to ask utility-oriented questions. If both utility and affective questions are asked, it appears that these should be treated separately, with the utility questions being used to provide the better estimate of potential transfer.

With regard to reliability, three points can be made. First, training criteria, overall, seem to have reassuringly high reliabilities. A second point is that immediate measures have slightly higher reliability than more delayed measures of the same criterion. This makes sense if the training intervention is at its maximum general impact in the most immediate case (Bennett, 1995). In other words, as time passes the intervention has differential impact on trainees, with some staying at the same level of proficiency or knowledge as immediately after training, while others gain or decrease in this regard. Third and finally, learning measure reliabilities are somewhat lower than either reaction or performance reliabilities. This may reflect the tendency for these measures to be more heterogeneous in content (i.e., they may cover a broader conceptual range) than the assessment of either reactions or performance. In any case, since one important criterion characteristic is reliability, it is reassuring to know that most training research criteria do possess that characteristic.

11

The results of the moderator analysis suggest that if training criteria do not overlap in content, convergence between or among them should not be expected. A lack of convergence, then, among criteria could simply be a sign that the criteria were not, in fact, designed to cover the same aspects of training. Thus, depending on the goals of the evaluation, convergence might or might not be expected or even desirable. In some cases, it might be best to use multiple criteria with minimal overlap to get a more complete picture of the effect of training. The important point is for the researcher/evaluator to pursue a conscious criterion-development strategy vis-à-vis content validity; in this way, regardless of the relationships among criteria, the chance for misinterpreting those relationships is minimized.

Limitations and Future Research

One limitation of the current meta-analysis is the number of studies that fall within the domain. Although the total has grown significantly since 1988, there are still a number of cells for which there are few or no correlations. In particular, there are very few published studies based on Level 4 criteria. We believe this reflects the inherent difficulty of conducting Level 4 evaluations rather than a reporting bias. However, it does limit us from providing any useful information about the relationship of Level 4 criteria with other training criteria.

A second limitation is that the meta-analysis is based upon Kirkpatrick's model. As we noted earlier, this model has several shortcomings. While the augmented framework may be an improvement, it cannot address all concerns. New taxonomic models capture recent developments from areas like cognitive psychology that are not addressed in Kirkpatrick's framework. Currently, Kirkpatrick's model remains the most prevalent, but as new taxonomic models, for example that of Kraiger et al. (1993), become more common, then the work from this study will need to be revisited.

We focused a great deal of attention on reactions because they are by far the most common criterion of training effectiveness. One area in which future research may prove fruitful is the examination of alternative methods of gathering reaction data. Reactions are typically gathered from training participants at the conclusion of training. However, trainees are not the only customers of training, nor is the conclusion of training necessarily the optimal time to collect reaction data. Trainees may not always be the most important or best judges of training effectiveness. Future research should examine the value of gathering utility reaction-type data from supervisors of training participants and other business leaders. Do their observations confirm the value or utility of the training? This is consistent with developments in the area of 360-degree feedback.

Affective reaction data is appropriately and easily gathered at the conclusion of training. Trainees should be able to offer immediate judgments of whether they liked the training. However, utility reactions require trainees to speculate about the future usefulness of training. By gathering reaction data one, three, or six months after training, trainees will have experienced whether the training was in fact useful, and should be in a better position to judge the utility of the training. Future research should examine when best to collect reaction data.

12

CONCLUSIONS

In this section of the report we answer the six questions we posed in the Introduction. The answers are based on the meta-analytic results, filtered through the lens of our own knowledge and interpretation of the field of training and training evaluation. In this sense, we have tried to generate practical answers to the questions; answers which might help U.S. Air Force researchers, trainers and others within and outside the U.S. Air Force.

1. Are typical training evaluation measures reliable?

Yes. Most training evaluation measures demonstrate sufficient reliability. Reliability in and of itself is not an indication of accuracy, of course, or predictive power. It is an index of temporal stability, which is a prerequisite for accuracy and prediction.

2. Are trainee reactions related to other measures of training effectiveness?

The answer here is a mixed "yes" and "no". In general, affective reactions are unrelated to measures of learning or transfer. However, utility reactions (and those measures that contain utility type questions) demonstrate a modest relationship with immediate learning and subsequent transfer. Utility reactions are thus not a substitute for learning and transfer measures, but they do provide a general sense of direction.

3. Are all trainee reactions equal in meaning and import?

No. All types of trainee reactions can have value because strong negative reactions may suggest a need for change, and strong positive reactions can influence how others perceive training. However, utility reactions demonstrate a relationship with other criteria while affective reactions do not. "Liking" a course does not appear to influence whether people acquired knowledge or whether they subsequently applied what they learned during training. In contrast, it does appear that utility reaction measures can provide some insight about whether trainees have learned something and whether they will be able to use what they have learned when they return to the job.

4. Is trainee learning related to subsequent on the job behavior?

Only modestly. Measures of immediate and retained learning demonstrated only samll relationships with transfer; in fact, simple utility reaction measures showed a stronger relationship with transfer. This is surprising, but on possible explanation is that when trainees are asked about the practical value of training or the extent to which they will be able to use what was covered in training, they implicitly consider a) what they learned during training and b) the work conditions they will face when they return. A great deal of empirical research (and experience) has shown that the work environment plays an enormous role in transfer. Learning measures can show what someone learned but not whether they can use it in actual work conditions. Utility reactions may capture perceptions of both what was learned and of the work environment.

13

5.   Are all trainee learning measures equal in meaning and import?

No. Behavior and/or skill demonstrations such as simulations, behavioral role plays, in-class performance ratings, observations checklists, and other behavioral indicators of learning during training appear to be more closely related to transfer than knowledge testing per se.

If a researcher is seeking to diagnose where a breakdown between learning and transfer may be occurring, this meta-analysis suggests that it is important that the learning measure and the behavior measure contain overlapping content. Otherwise, a low correlation between the two may not be due to situational constraints or poor supervisory support (e.g., Tracey, Tannenbaum, & Kavanagh, 1995), but instead may simply be because the learning and behavior measures are assessing different parts of the criterion space. In other cases, training researchers should intentionally select non-overlapping criterion measures in order to get a more complete picture of the factors that influence training effectiveness.

6.   Overall, what measures of training effectiveness are most recommended?

The answer here is: "it depends". This study provided *some* insight into the relationship among several typical training criteria. However, it is our belief that the selection of training criteria must first be driven by the purpose of the evaluation, the expectations of our customers, and the objectives of the training. Only at that point might the results of this study add any practical insight.

If the purpose of evaluation is to demonstrate "value-add", then we should determine (and manage) our customers' expectations for the training. These may be related to the course objectives, but not always. Criteria should be selected that address their expectations. If one knows the customer's answer to the question: "If this training is successful, then what should happen?", then one can develop, adapt or adopt criteria that will be most likely to answer that question.

If the purpose of an evaluation is to identify possible course improvements, then we should examine course objectives. Criteria should be selected that provide information about the attainment of the objectives and that provide insights into possible course improvements.

STUDIES PROVIDING CORRELATIONS FOR META-ANALYSES

1. Alliger, G. M. & Horowitz, H. M. (1989). IBM takes the guessing out of testing. *Training and Development Journal, 43(4)*, 69-73.

2. Ammar, S. (1994). *The influence of individual and organizational characteristics on training motivation and effectiveness.* Unpublished doctoral dissertation, University at Albany, Albany, New York.

3. Baldwin, T. T. (1992). Effects of alternative modeling strategies on outcomes of interpersonal skills training. *Journal of Applied Psychology, 77*, 147-154.

4. Bolman, L. (1971). Some effects of trainers on their T-groups. *Journal of Applied Behavioral Science, 7*, 309-326.

5. Bretz, R. D. Jr. & Thompsett, R. E. (1992). Comparing traditional and integrative learning methods in organizational training programs. *Journal of Applied Psychology, 77*, 941-951.

6. Clement, R. W. (1982). Testing the hierarchy theory of training evaluation: An expanded role for trainee reaction. *Public Personnel Management Journal, 11*, 176-184.

7. Eden, D. & Shani, A. B. (1982). Pygmalion goes to boot camp: Expectancy, leadership, and trainee performance. *Journal of Applied Psychology, 67*, 194-199.

8. Faerman, S. R. & Ban, C. (1992). Trainee satisfaction and training impact: Issues in training evaluation. *Public Productivity & Management Review, 16*, 299-314.

9. Frayne, C. A. & Latham, G. P. (1987). Application of social learning theory to employee self-management of attendance. *Journal of Applied Psychology, 72*, 387-392.

10. Gist, M. E. (1989). The influence of training method on self-efficacy and idea generation among managers. *Personnel Psychology, 42*, 787-805.

11. Gist, M. E., Schwoerer, C. & Rosen, B. (1989). Effects of alternative training methods on self-efficacy and performance in computer software training. *Journal of Applied Psychology, 74*, 884-891.

12. Harrison, J. K. (1992). Individual and combined effects of behavior modeling and the cultural assimilator in cross-cultural management training. *Journal of Applied Psychology, 77*, 952-962.

13. Martineau, J. W. (1996). *A contextual examination of the effectiveness of a supervisory skills training program.* Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

15

14. Martocchio, J. J. & Webster, J. (1992). Effects of feedback and cognitive playfulness on performance in microcomputer software training. *Personnel Psychology, 45*, 125-147.

15. Mathieu, J. E., Martineau, J. W. & Tannenbaum, S. I. (1993). Individual and situational influences on the development of self-efficacy: Implications for training effectiveness. *Personnel Psychology, 46*, 125-147.

16. Mathieu, J. E., Tannenbaum, S. I. & Salas, E. (1992). Influences of individual and situational characteristics on measures of training effectiveness. *Academy of Management Journal, 35*, 828-847.

17. Miles, M. B. (1965). Changes during and following laboratory training: A clinical experimental study. *Journal of Applied Behavioral Science, 1*, 215-242.

18. Noe, R. A. & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology, 39*, 497-523.

19. Quinones, M. A. (1995). Pretraining context effects: Training assignment as feedback. *Journal of Applied Psychology, 80*, 226-238.

20. Reeves, E. T. & Jensen, J. M. (1972). Effectiveness of program evaluation. *Training and Development Journal, 26*, 36-41.

21. Severin, D. (1952). The predictability of various kinds of criteria. *Personnel Psychology, 5*, 93-104.

22. Smith, P. E. (1976). Management modeling training to improve morale and customer satisfaction. *Personnel Psychology, 29*, 351-359.

23. Smith, K. A. & Salas, E. (1994). *Narrowing the gap between performance and potential: The effects of team climate on the transfer of assertiveness training.* Unpublished doctoral dissertation, University of South Florida, Tampa, Florida.

24. Stroud, P. (1959). Evaluating a human relations training program. *Personnel, 36*, 52-60.

25. Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cannon-Bowers, J. (1991). Meeting trainees' expectations: The influence of training fulfillment on the development of commitment, self-efficacy, and motivation. *Journal of Applied Psychology, 76*,759-769.

26. Tannenbaum, S. I. & Woods, S. B. (1992). Determining a strategy for evaluating training: Operating within organizational constraints. *Human Resource Planning, 15*, 63-81.[1]

27. Thayer, P. W., Antoinetti, J. A., & Guest, T. A. (1958). Product knowledge and performance: A study of life insurance agents. *Personnel Psychology, 11,* 411-418.

28. Tracey, J. B., Tannenbaum, S. I., & Kavanagh, M. J. (1995). Applying trained skills on the job: The importance of the work environment. *Journal of Applied Psychology, 80,* 239-252.

29. Tziner, A. & Falbe, C. M. (1993). Training-related variables, gender and training outcomes: A field investigation. *International Journal of Psychology, 28,* 203-221.

30. Warr, P. & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. *Personnel Psychology, 48,* 347-375.

31. Webster, J. & Martocchio, J. J. (1993). Turning work into play: Implications for microcomputer software training. *Journal of Management, 19,* 127-146.

32. Weitzman, D. O., Fineberg, M. L., Gade, P. A., & Compton, G. L. (1979). Proficiency maintenance and assessment in an instrument flight simulator. *Human Factors, 21,* 701-710.

33. Werner, J. M., O'Leary-Kelly, A. M., Baldwin, T. T., & Wexley, K. N. (1994). Augmenting behavior-modeling training: Testing the effects of pre- and post-training interventions. *Human Resource Development Quarterly, 5,* 169-183.

34. Wexley, K. N. & Baldwin, T. T. (1986). Post training strategy for facilitating positive transfer: An empirical exploration. *Academy of Management Journal, 29,* 503-520.

[1] Correlations were obtained from: Woods, S. B. & Tannenbaum, S. I. (1990, April). *Evaluating supervisory training: A case study.* Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, Miami, FL.

REFERENCES

Ajzen, I. & Fishbein, M. (1973). Attitudinal and normative variables as predictors of specific behaviors. Journal of Personality and Social Psychology, 27, 41-57.

Ajzen, I. & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. Psychological Bulletin, 84, 888-918.

Alliger, G.M., & Janak, E.A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42, 331-342.

Alliger, G.M., Tannenbaum, S.I., & Bennett, W. (1995). Transfer of training: Comparison of paradigms. Part of the symposium, "Searching for Solutions to the 'Transfer of Training Problem': A Multi-Disciplinary Approach," Kimberly A. Smith and Eduardo Salas, Co-chairs; presented at the 10th Annual Conference of the Society for Industrial and Organizational Psychology, Miami.

American Society for Training and Development (1997). ASTD Benchmarking Forum.

Baldwin, T.T. & Ford, J.K. (1988). Transfer of training: a review and directions for future research. Personnel Psychology, 41, 63-105.

Bennett, W. (1995). A meta-analytic review of factors that influence the effectiveness of training in organizations. Unpublished doctoral dissertation. Texas A&M University, College Station, TX.

Bassi, L.J., Cheney, S., & Van Buren, M. (1997). Training industry trends 1997. American Society for Training and Development.

Brogden, H.E., & Taylor, E.K. (1950). The dollar criterion: Applying the cost accounting concept to criterion construction. Personnel Psychology, 3, 133-154.

Carnevale, A.P. & Schulz, E.R. (1990). Evaluation practices. Training and Development Journal, 44, S-23-S-29.

Eagly, A.H., & Chaiken, S. (1992). The psychology of attitudes. San Diego, CA: Harcourt Brace Janovich.

Goldstein, I. (1993). Training in organizations: Needs assessment, development, and evaluation (3rd ed.) Pacific Grove, CA: Brooks/Cole.

Holton, E.F. (1996). The flawed four-level evaluation model. Human Resources Development Quarterly, 7, 5-15.

Iaffaldano, M.T., & Muchinsky, P.M. (1985). Job satisfaction and performance. <u>Psychological Bulletin, 97</u>, 251-273.

Kirkpatrick, D.L. (1959a). Techniques for evaluating training programs. <u>Journal of ASTD, 13</u>, 3-9.

Kirkpatrick, D.L. (1959b). Techniques for evaluating training programs: Part 2 - Learning. <u>Journal of ASTD, 13</u>, 21-26.

Kirkpatrick, D.L. (1960a). Techniques for evaluating training programs: Part 3 - Behavior. <u>Journal of ASTD, 14</u>, 13-18.

Kirkpatrick, D.L. (1960b). Techniques for evaluating training programs: Part 4 - Results. <u>Journal of ASTD, 14</u>, 28-32.

Kirkpatrick, D.L. (1985). Effective training and development, Part 2: In house approaches and techniques. <u>Personnel, 62</u>, 52-56.

Kraiger, K., Ford, K.J., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. <u>Journal of Applied Psychology, 78</u>, 311-328.

Kraus, S.J. (1995). Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. <u>Personality and Social Psychology Bulletin, 21</u>, 58-75.

Peters, L.H., O'Connor, E.J., Eulberg, J.R., & Watson, T.W. (1988). An examination of situational constraints in Air Force work settings. <u>Human Performance, 1</u>, 133-144.

Saari, L.M., Johnson, T.R, McLaughlin, S.D., & Zimmerle, D.M. (1988). A survey of management training and education practices in U.S. companies. <u>Personnel Psychology, 41</u>, 731-743.

Shelton, S., & Alliger, G.M. (1993). Who's afraid of Level 4 Evaluation? A practical approach. <u>Training & Development Journal, 47</u>, 43-46.

Tannenbaum, S.I. (1996). Customer-focused training evaluation <u>ASTD Forum in Action Update</u>.

Tannenbaum, S.I., & Yukl, G. (1992). Training and development in work organizations. <u>Annual Review of Psychology, 43</u>, 399-441.

Tannenbaum, S.I., & Woods, S.B. (1992). Determining a strategy for evaluating training: Operating within organizational constraints. <u>Human Resources Planning Journal, 15</u>, 63-82.

Tracey, J.B., Tannenbaum, S.I., & Kavanagh, M.J. (1995). Applying trained skills on the job: The importance of the work environment. Journal of Applied Psychology, 80, 239-252.

Warr, P., & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. Personnel Psychology, 48, 347-376.

Weigel, R.H., Vernon, D.T.A. & Tognacci, L.N. (1974). Specificity of the attitude as a determinant of attitude-behavior congruence. Journal of Personality and Social Psychology, 30, 724-728.