Research Note 2014-02

Development of the 'TARGET' Training Effectiveness Tool and Underlying Algorithms Specifying Training Method – Performance Outcome Relationships

Shaun D. Hutchins
Thomas F. Carolan
Beth M. Plott
Patty L. McDermott
Alion Science and Technology

Karin A. Orvis
U.S. Army Research Institute

May 2014

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS, Ph.D.
Director

Research accomplished under contract
for the Department of the Army by:

Alion Science and Technology, Incorporated

Technical review by:

Foundational Science Research Unit, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Research Note has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: DAPE-ARI-ZXM, 6000 6th Street (Bldg 1464 / Mail Stop: 5610), Fort Belvoir, Virginia 22060-5610.

FINAL DISPOSITION: Destroy this Research Note when it is no longer needed. Do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy): May 2014	2. REPORT TYPE: Interim	3. DATES COVERED (from. . . to) September 2009 to December 2013

4. TITLE AND SUBTITLE
Development of the ‘TARGET’ Training Effectiveness Tool and Underlying Algorithms Specifying Training Method – Performance Outcome Relationships

5a. CONTRACT OR GRANT NUMBER
W91WAW-07-C-0081

5b. PROGRAM ELEMENT NUMBER
611102

6. AUTHOR(S)
Shaun D. Hutchins, Thomas F. Carolan,
Beth M. Plott, Patty L. McDermott ;
Karin A. Orvis

5c. PROJECT NUMBER
B74F

5d. TASK NUMBER

5e. WORK UNIT NUMBER
2902

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Alion Science and Technology, Inc.
4949 Pearl East Circle
Boulder, CO 80301

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Institute
for the Behavioral & Social Sciences
6000 6^{TH} Street (Bldg. 1464 / Mail Stop 5610)
Fort Belvoir, VA 22060-5610

10. MONITOR ACRONYM
ARI

11. MONITOR REPORT NUMBER
Research Note 2014-02

12. DISTRIBUTION/AVAILABILITY STATEMENT: Distribution Statement A: Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
Contracting Officer’s Representative and Subject Matter POC: Dr. Gerald F. Goodwin

14. ABSTRACT: A four-year research effort was conducted to collect empirical evidence on the effectiveness of different training methods for acquiring and transferring complex cognitive skills (see Plott et al., 2014). To accomplish this goal, a series of meta-analyses were conducted examining six training methods (training wheels, scaffolding, part-task training, increasing difficulty, learner control, and exploratory learning). Algorithms were then developed to quantify the relationships between the training methods, performance, and various moderating factors. These algorithms can be used to perform tradeoff analyses to determine the effectiveness of different combinations of training method(s), task/skill type(s) being trained (e.g., perceptual, psychomotor), trainee characteristics (e.g., experience, aptitude), and type(s) of training performance outcomes (e.g., learning, transfer). Finally, to ensure these research findings and algorithms would be easily consumable by training developers and researchers, a training effectiveness tool was developed, called TARGET (which stands for Training Aide: Research and Guidance for Effective Training). This tool can aid training developers and researchers in making decisions concerning the most appropriate training method(s) to use depending on their particular training context. This report focuses on the algorithm development completed as a part of this larger research effort, as well as the algorithm incorporation into TARGET.

15. SUBJECT TERMS
Training effectiveness, Transfer of training, Training , Algorithms, Meta-analysis

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unclassified Unlimited	48	Gerald F. Goodwin (703) 545-2410

Research Note 2014-02

Development of the ‘TARGET’ Training Effectiveness Tool and Underlying Algorithms Specifying Training Method – Performance Outcome Relationships

Shaun D. Hutchins
Thomas F. Carolan
Beth M. Plott
Patty L. McDermott
Alion Science and Technology

Karin A. Orvis
U.S. Army Research Institute

Foundational Science Research Unit
Gerald F. Goodwin, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
6000 6th Street, Building
Fort Belvoir, VA 22060

May 2014

Approved for public release; distribution is unlimited

DEVELOPMENT OF THE ‘TARGET’ TRAINING EFFECTIVENESS TOOL AND UNDERLYING ALGORITHMS SPECIFYING TRAINING METHOD – PERFORMANCE OUTCOME RELATIONSHIPS

Research Requirement

The present research was conducted as part of a broader four-year effort to develop evidence-based guidelines for the relative effectiveness of six different training methods for acquiring and transferring cognitive skills involved in complex task domains (see Plott et al., 2014). The six training methods investigated were part-task training, training wheels, scaffolding, increasing difficulty, learner control, and exploratory learning. Several efforts were employed to develop these evidence-based guidelines. First, a comprehensive research database was created via a broad literature search, six comprehensive meta-analyses, and several supplemental research experiments in order to generate effect size estimates of the relative effectiveness of the six training methods and how various factors (e.g., task/skill type, trainee characteristics, performance outcomes) moderate training effectiveness. Second, we developed algorithms to empirically synthesize the relationships between the six training methods and various moderating factors. Moreover, to ensure these research findings and algorithms would be easily consumable by various users, we developed a *user-friendly graphical user interface tool, called TARGET* (which stands for Training Aide: Research and Guidance for Effective Training). This tool integrated the comprehensive research database with the developed algorithms, and contains several visualization tools for identifying training conditions under which particular training methods are more/less effective. The purpose of this report was to provide information on the algorithm development completed as a part of this larger research effort, as well as their use in TARGET.

Procedure

First, we developed of a coding scheme of research study attributes, which was necessary **i**n order to generate the meta-analyses and algorithms underlying TARGET. The coding scheme was used to code, classify, and extract necessary data from all of the primary research studies in order to compute the meta-analytic effect size estimates. We also reviewed the literature on meta-analysis methods for documentation on the sets of computations required for computing effect size statistics, as well as the procedures combining each set of computations under different conditions. Next, meta-analytic effect size statistics were calculated by converting raw individual-level research study data, and these were then used to develop the algorithms. Finally, we devised an extrapolation process for estimating effect size statistics for particular moderator/attribute combinations in which there was no extant research available.

Findings

We developed TARGET, a web-based training tool, which is publicly accessible at **http://bldr-webtest.alionscience.com/Target/**. This tool summarizes the cognitive skill training research and identifies the conditions under which a particular training method is more or less effective. TARGET’s key capabilities were accomplished via the underlying innovative algorithms. First, the algorithms provide real-time computational capability so that users of TARGET can select a number of combinations of moderating attributes and get effect size estimates for a particular training method. The algorithms supported the required capability of computing effect sizes for individual studies from raw study-level information, as well as the summary-level effect size statistics compiled across primary research studies. Note that there were several combinations of moderating variables for which there is no extant research available (e.g., the impact of trainee experience on the effectiveness of part-task training of perceptual skills) and thus the algorithms could not be directly employed to generate effect sizes. Accordingly, we also defined and implemented an innovative extrapolation process for estimating effect size statistics for particular attribute combinations in which there was no extant research available. Although the functionality mentioned above is certainly useful, these effect size estimates risk becoming dated unless the tool's evidential database is continuously updated. Thus, a second key capability is that algorithms provide the capability of long-term updateability of TARGET's comprehensive research database, allowing users to add research findings/effect size data from new primary research studies conducted. Although other commercial off-the-shelf software packages exist that can be used to perform meta-analytic studies, TARGET adds value by providing a built-in database that is easily accessible on the web to any potential users.

Utilization and Dissemination of Findings

This report serves as documentation of procedures and processes for developing the algorithms and their implementation in TARGET. In summary, the algorithms offer several potential benefits to the Army. As implemented in TARGET, they can be used to perform tradeoff analyses for different combinations of training methods. The algorithms make the research findings from this larger 4-year research effort available to the Army training, development, and research communities, allowing users to systematically explore training methods and design components that would be effective for a particular set of circumstances for acquiring cognitive skills. In short, they estimate the expected costs or benefits of the six training methods on performance, for various combinations of task/skill type, trainee characteristics, and performance outcomes. Beyond TARGET, the algorithms can also be readily applied in a variety of other applications. For example, they can be used as input for human performance models to analyze the impact of different training or technological approaches. As a more specific example, the algorithms could also be adapted for use in IMPRINT to determine how different training methods might impact performance.

DEVELOPMENT OF THE 'TARGET' TRAINING EFFECTIVENESS TOOL AND UNDERLYING ALGORITHMS SPECIFYING TRAINING METHOD – PERFORMANCE OUTCOME RELATIONSHIPS

CONTENTS

APPENDICES

Page

FIGURES

DEVELOPMENT OF THE 'TARGET' TRAINING EFFECTIVENESS TOOL AND UNDERLYING ALGORITHMS SPECIFYING TRAINING METHOD – PERFORMANCE OUTCOME RELATIONSHIPS

Introduction

Research Project Background

The present research was conducted as part of a broader four-year effort to develop evidence-based guidelines for the relative effectiveness of different training methods for learning and transferring cognitive skills involved in complex Army-relevant task domains (see Plott et al., 2014). The emphasis was on how to improve not only learning, but also how to improve training transfer – that is, the degree to which trainees are able to apply and use what they learned once back on-the-job or in a new context (see Kraiger, Ford, & Salas, 1993). Six training methods were investigated: part-task training, training wheels, scaffolding, increasing difficulty, learner control, and exploratory learning (See Appendix A for definitions of each training method). These methods were expected to be influential not only on learning and transfer, but also on managing trainee effort/workload and engagement during training - factors which have been demonstrated to be key predictors of several training performance outcomes (Bell & Kozlowski, 2008; Paas, Renkl, & Sweller, 2003; van Merrienboer, Kester, & Paas., 2006).

Several efforts were employed to develop the evidence-based guidelines for these six training methods. First, a *comprehensive research database* needed to be developed, summarizing the extant training research on these training methods, as well as the factors/attributes that moderate the effectiveness of each method. Such moderating factors/attributes include the task/skill type being trained (e.g., psychomotor, problem solving), types of trainees that would participate in the intended training (e.g., their average experience level, ability level), and type of training performance outcome(s) of interest (e.g., procedural knowledge, transfer). To create this database, we conducted a broad literature search, six comprehensive meta-analyses, and several supplemental research experiments in order to generate estimates of the relative effectiveness of the six specific training methods and how the moderating factors impact their effectiveness. For details on the literature review, meta-analyses and supplemental experiments see publications by Carolan, Wickens, Hutchins and Cumming (in press), Hutchins, Wickens, Carolan, and Cumming (2013), and Wickens, Hutchins, Carolan, and Cumming (2013). Other information can also be gleaned from ARI technical reports (McDermott, Carolan, Fisher, Gronowski, & Gacy, 2013; Plott et al., 2014) and conference proceedings (Carolan, McDermott, Hutchins, Wickens, & Belanich, 2011; McDermott, Carolan, & Gronowski, 2012; McDermott, Carolan, & Wickens, 2012; McDermott, Gronowski, Carolan, & Fisher, 2013).

Second, we developed *algorithms* to empirically synthesize/describe the relationships between the six training methods and various moderating factors. Moreover, to ensure these research findings and algorithms would be easily consumable by training developers and researchers, we developed a *user-friendly graphical user interface tool, called TARGET* (which stands for Training Aide: Research and Guidance for Effective Training). This tool integrated the comprehensive research database with the developed algorithms, and contains several visualization tools, such that in-depth statistical knowledge is not required to benefit from this tool. TARGET enables training developers and researchers to identify training conditions under which particular training methods are more/less effective, which helps maximize effective learning and transfer.

Purpose of Report

The purpose of this report is to provide information on the algorithm development and their use in TARGET. Guided by anticipated use cases, we sought the following key functionality from the algorithms embedded in TARGET. First, the algorithms needed to provide real-time computational capability so that users of TARGET could select a number of combinations of moderating factors/attributes (e.g., trainee characteristics, task/skill types) and get effect size estimates for a particular training method. An *effect size* is a statistical concept that measures the strength of the relationship between two variables (Preacher & Kelly, 2012). For example, a user may be interested in the whether or not the effectiveness (i.e., the effect size) for the ‘part-task training’ method differs based on the experience level of the group of trainees. As such, the algorithm must not only be able to compute effect sizes from raw study-level information (e.g., means, standard deviations), but also summarize multiple effect sizes across all relevant primary studies as indicated by the selected combination of moderating factors/attributes. Note that there were several combinations of moderating factors for which there is no extant research available (e.g., the impact of trainee experience on the effectiveness of part-task training of perceptual skills) and thus the aforementioned algorithms could not be directly employed to generate effect sizes. Accordingly, we also needed to develop an innovative extrapolation process for estimating effect size statistics for particular moderator combinations in which there was no extant research available. Although the functionality mentioned above is certainly useful, these effect size estimates risk becoming dated unless the tool’s evidential database is continuously updated. Thus, a second key functionality is that algorithms were also needed to provide long-term updateability of TARGET’s comprehensive research database, allowing users to add research findings from new primary research studies conducted.

Note that while the algorithms were implemented in TARGET, and thus we focus this report on TARGET as the context, they can also be readily applied in a variety of other applications. For example, they can be used as input for human performance models to analyze the impact of different training or technological approaches. As a more specific example, the algorithms could also be adapted for use in IMPRINT to determine how different training methods might impact performance.

Organization of Report

The remainder of this report unfolds as follows. First, we briefly describe the four main components of TARGET to provide greater context for how the underlying algorithms were generated and used in this training tool. Secondly, we describe our development of a coding scheme of research study attributes, which was necessary **i**n order to generate the meta-analyses and algorithms underlying TARGET. The coding scheme was used to code, classify, and extract necessary data from all of the primary research studies in order to compute the meta-analytic effect size estimates. Third, we describe how the meta-analytic effect size statistics were calculated by converting raw individual-level research study data, as well as how these statistics informed the algorithm development. Finally, we describe our devised extrapolation process for estimating effect size statistics for particular moderator/attribute combinations in which there was no extant research available.

Overview of TARGET

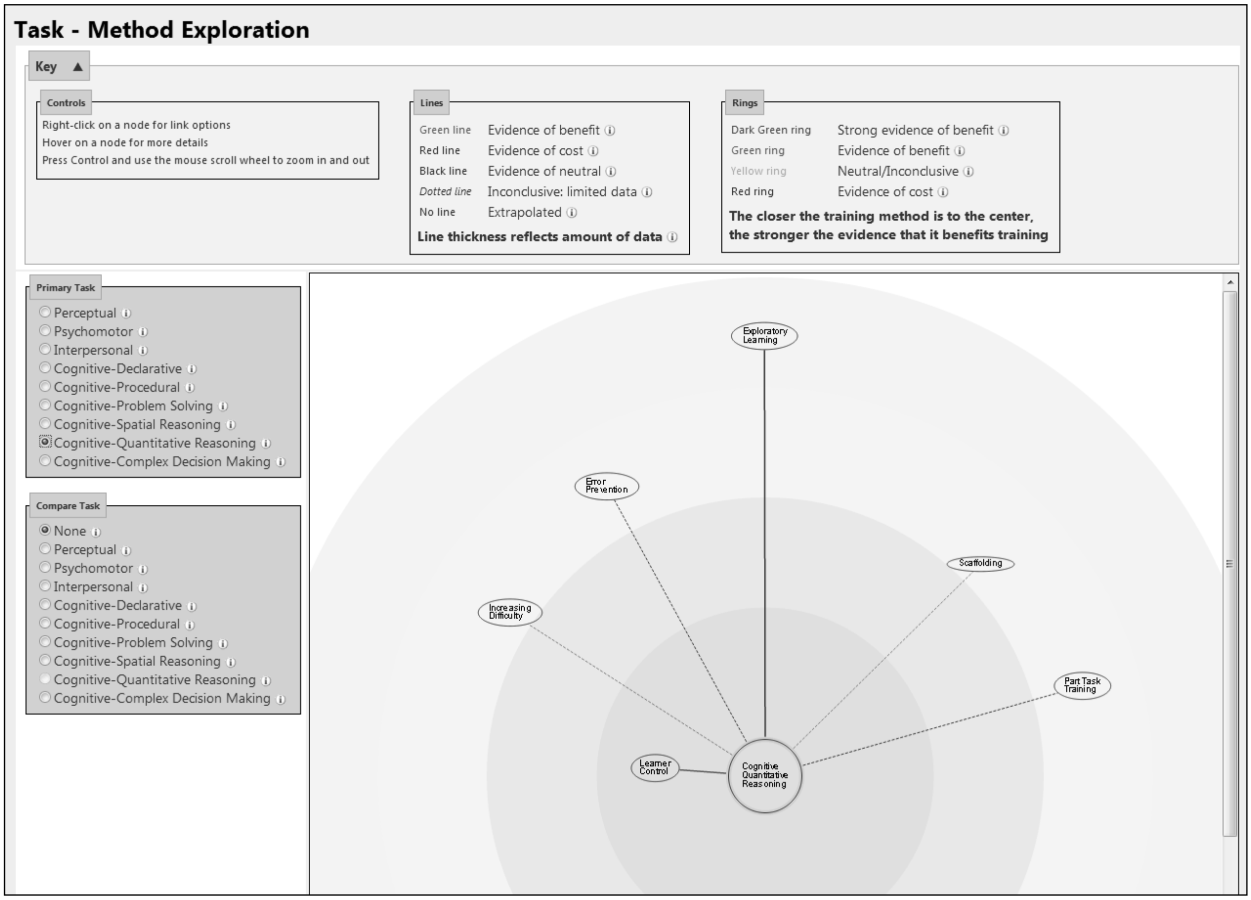
As aforementioned, a key goal of the current project was to assist training developers and researchers in better understanding the relative effectiveness of different training methods for acquiring/transferring various cognitive skills. As such, a comprehensive research database was generated that included qualitative summaries and meta-analyses of prior empirical research concerning the training effectiveness of six training methods (i.e., part-task training, training wheels, scaffolding, increasing difficulty, learner control, and exploratory learning). Findings concerning the factors that moderate the effectiveness of each method were also included, such as task/skill type being trained (e.g., psychomotor, problem solving), trainee characteristics (e.g., experience level, ability level), and type of training performance outcome (e.g., procedural knowledge, transfer).

To ensure the valuable findings from this research database would be easily consumable by training developers and researchers, the training effectiveness tool, TARGET, was developed. TARGET was designed to assist users in making evidence-based decisions concerning the most suitable training method(s) to use depending on the task/skill type(s) being trained, types of trainees that would participate in the intended training, and/or the performance outcomes sought. It provides query-based searches of the research database as well as numerous visualizations, textual summaries, and data summaries of the relationships between different training methods and performance outcomes (via the underlying computational algorithms). It is also updateable as additional training research is generated so that the database/tool stays current with state-of-the-art research developments.

Next, we briefly describe the four main components of TARGET (*Explore Tasks*, *Explore Methods*, *Explore Documents,* and *Add New Document*) to provide greater context for how the underlying algorithms were generated and are used in this tool. Readers interested in additional details regarding TARGET are encouraged to access and use the tool via its publically available website (http://bldr-webtest.alionscience.com/Target/), as well as the accompanying TARGET User Guide (Plott & Hutchins, 2013).

Component 1: Explore Tasks

The *Explore Tasks* component of TARGET provides users with a range of interactive features for viewing accumulated research evidence by task/skill type (e.g., cognitive-quantitative reasoning skills, perceptual skills, psychomotor skills) in order to identify which training methods are likely to be useful in training a particular task/skill type(s). The assumption is that a training developer or researcher will have information about the type of task/skill(s) that need to be trained. This component serves as a starting point for those users to visualize which training methods have been successfully used to train the task/skill type of interest. The visualization shows the task/skill type at the center of the screen surrounded by the six training methods. The closer a given training method is to the task/skill at the center of the screen, the more *evidence* there is that this training method benefits performance for this task/skill (see Figure 1 as an example). Users can examine up to two task/skill types simultaneously to determine which training methods are likely to be most effective for their training needs.



Task - Method Exploration

Key ▲

Controls

Right-click on a node for link options
Hover on a node for more details
Press Control and use the mouse scroll wheel to zoom in and out

Lines

Green line	Evidence of benefit ⓘ
Red line	Evidence of cost ⓘ
Black line	Evidence of neutral ⓘ
Dotted line	Inconclusive: limited data ⓘ
No line	Extrapolated ⓘ

Line thickness reflects amount of data ⓘ

Rings

Dark Green ring	Strong evidence of benefit ⓘ
Green ring	Evidence of benefit ⓘ
Yellow ring	Neutral/Inconclusive ⓘ
Red ring	Evidence of cost ⓘ

The closer the training method is to the center, the stronger the evidence that it benefits training

Primary Task

○ Perceptual ⓘ
○ Psychomotor ⓘ
○ Interpersonal ⓘ
○ Cognitive-Declarative ⓘ
○ Cognitive-Procedural ⓘ
○ Cognitive-Problem Solving ⓘ
○ Cognitive-Spatial Reasoning ⓘ
● Cognitive-Quantitative Reasoning ⓘ
○ Cognitive-Complex Decision Making ⓘ

Compare Task

● None ⓘ
○ Perceptual ⓘ
○ Psychomotor ⓘ
○ Interpersonal ⓘ
○ Cognitive-Declarative ⓘ
○ Cognitive-Procedural ⓘ
○ Cognitive-Problem Solving ⓘ
○ Cognitive-Spatial Reasoning ⓘ
○ Cognitive-Quantitative Reasoning ⓘ
○ Cognitive-Complex Decision Making ⓘ

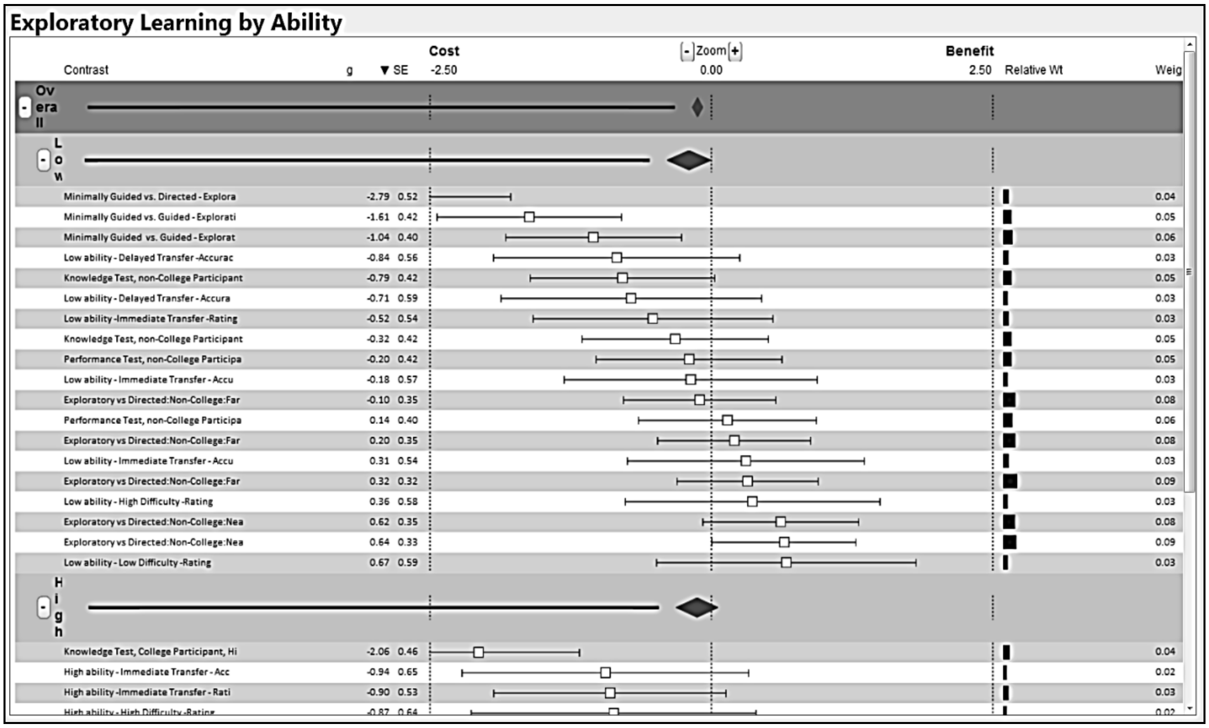
Figure 1. **Example 'Explore Tasks' Graphical Visualization**

Component 2: Explore Methods

The *Explore Methods* component graphically displays results of the meta-analyses of prior training research. The visualization allows users to view interactive forest plots of effect size statistics for a range of studies under one of the six training methods. For example, if the users select exploratory learning, a forest plot is graphically displayed with effect size data based on all experimental studies in the research database that used this training method. The effect size would reflect the strength of the relationship between this training method and training effectiveness.

Users can further select moderator variables for which they want to see subsets in the training method forest plot. The moderator variable selected could be a task/skill type of interest, a trainee characteristic (e.g., trainee ability level), training method-specific moderators, or what specific training performance outcome is of interest (e.g., speed, accuracy, knowledge, near transfer). For example, if trainee ability was selected as the moderator, the user would now see the exploratory learning effect size data divided into two groupings – research studies that used this training method with a group of lower ability trainees and another grouping of studies that used this method with higher ability trainees. This enables the user to visually see whether or not the effectiveness of the 'exploratory learning' training method *depends* on the ability level of the group of trainees (in addition to seeing the actual effect size values, which are also reported). Figure 2 shows the exploratory learning research studies categorized by the individual factor of ability. As shown, the effect size data is summarized across the studies via the three diamonds (one for overall, low ability, and high ability, respectively). These overall effect size estimates are further broken down into the primary studies from which information was collected.

Note that users have a variety of other advanced options to choose from, such as whether they want to use a fixed effects or random effects model to calculate the effect sizes (described in Appendix B), or if they want to include 'extrapolated effects.' As will be explained in greater detail in the *Effect Size Extrapolation Approach* section of this report, using extrapolated effects enables an effect size estimate for a subset of moderator variables, even when no data was available in the research database (i.e., from the existing training research). Instead, these estimates are based on extrapolating effects from a nearest 'parent' set of moderator variables.



Exploratory Learning by Ability

Contrast	g	▼ SE	Weig
Overall			
Low			
Minimally Guided vs. Directed - Explora	-2.79	0.52	0.04
Minimally Guided vs. Guided - Explorati	-1.61	0.42	0.05
Minimally Guided vs. Guided - Explorat	-1.04	0.40	0.06
Low ability - Delayed Transfer -Accurac	-0.84	0.56	0.03
Knowledge Test, non-College Participant	-0.79	0.42	0.05
Low ability - Delayed Transfer - Accura	-0.71	0.59	0.03
Low ability -Immediate Transfer -Rating	-0.52	0.54	0.03
Knowledge Test, non-College Participant	-0.32	0.42	0.05
Performance Test, non-College Participa	-0.20	0.42	0.05
Low ability - Immediate Transfer - Accu	-0.18	0.57	0.03
Exploratory vs Directed:Non-College:Far	-0.10	0.35	0.08
Performance Test, non-College Participa	0.14	0.40	0.06
Exploratory vs Directed:Non-College:Far	0.20	0.35	0.08
Low ability - Immediate Transfer - Accu	0.31	0.54	0.03
Exploratory vs Directed:Non-College:Far	0.32	0.32	0.09
Low ability - High Difficulty -Rating	0.36	0.58	0.03
Exploratory vs Directed:Non-College:Nea	0.62	0.35	0.08
Exploratory vs Directed:Non-College:Nea	0.64	0.33	0.09
Low ability - Low Difficulty -Rating	0.67	0.59	0.03
High			
Knowledge Test, College Participant, Hi	-2.06	0.46	0.04
High ability - Immediate Transfer - Acc	-0.94	0.65	0.02
High ability -Immediate Transfer - Rati	-0.90	0.53	0.03
High ability - High Difficulty -Rating	-0.87	0.64	0.02

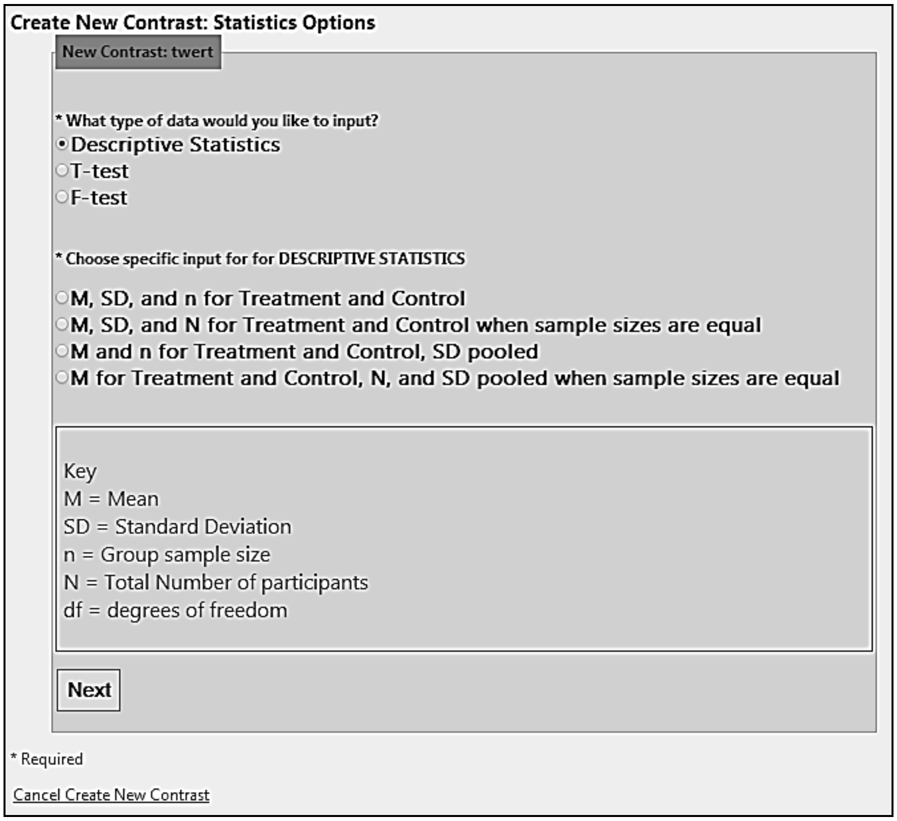
Figure 2. **Example 'Explore Methods' Forest Plot Visualization.**

Component 3: Explore Documents

This component of TARGET enables the user the search the research database and review the extant training literature with queries built around a set of attributes. For example, a user may want to search by the attribute of 'training method' or 'task/skill type' (or both simultaneously). As another example, searching for the term 'exploratory learning' will pull up a list of all the training research in the database that has employed this training method. Then, detailed summaries of each study can be viewed, as well as any relevant effect size statistics pulled from the given studies. Additional attribute examples in which a user can use to search for prior literature are listed in Appendix A. Note that these attributes will also be discussed in greater detail in the *Coding Scheme for Research Studies* section of this report (see pp. 7-8).

Component 4: Add New Document

The *Add New Document* component provides users with the capability to add a new research study to the database and keep TARGET up-to-date with the latest training research. A wizard tool walks the user through how to enter a new study, including how to code its attributes, provide a qualitative summary of the study's purpose, methodology and results, as well as how to enter raw study data (e.g., means, standard deviations, sample sizes) so that effects size data (e.g., Hedge's *g*, standard error of *g*) can be automatically calculated by TARGET (see Figure 3 as an example of how this wizard assists a user step-by-step in entering raw study data).



Create New Contrast: Statistics Options

New Contrast: twert

* What type of data would you like to input?

● Descriptive Statistics
○ T-test
○ F-test

* Choose specific input for for DESCRIPTIVE STATISTICS

○ M, SD, and n for Treatment and Control
○ M, SD, and N for Treatment and Control when sample sizes are equal
○ M and n for Treatment and Control, SD pooled
○ M for Treatment and Control, N, and SD pooled when sample sizes are equal

Key
M = Mean
SD = Standard Deviation
n = Group sample size
N = Total Number of participants
df = degrees of freedom

* Required

Figure 3. **Example Enter Effects Data Fields**

Coding Scheme for Research Studies

In order to generate the meta-analyses and algorithms underlying TARGET concerning the training effectiveness of the six training methods, the first required step was the development of a coding scheme. The coding scheme was used to code, classify, and extract necessary data from all of the primary research studies in order to compute the meta-analytic effect size estimates. For example, each study was coded for which training method was employed (i.e., part-task training, training wheels, scaffolding, increasing difficulty, learner control, and exploratory learning), as well as a variety of moderating factors. These factors included: *trainee characteristics* (i.e., experience, ability), *task characteristics* (i.e., task/skill type, task difficulty), *performance outcomes* (e.g., near transfer, far transfer), and training method-specific factors. By *training method-specific factors*, we mean training design features that varied *within* a given training method. For example, for the training method part-task training, a training method-specific moderator was whether the different component tasks were trained concurrently or sequentially. As another example, for the scaffolding training method, one important moderator was whether the scaffolds were removed according to the trainee's progress/performance during the training (i.e., adaptive scaffolding) or on a set schedule regardless of trainee performance (i.e., fixed scaffolding). Each training method had approximately 2-3 training method-specific moderating factors. Additional examples of the coding scheme attributes (and their definitions) are provided in Appendix A. For the full set of attributes (and definitions) used within TARGET and the underlying algorithms, please see Appendix A in the TARGET User Guide (Plott & Hutchins, 2013).

Note that for a given research study (referred to as a document in TARGET), data was collected (e.g., study means and standard deviations) and could be coded for *multiple* effects sizes. As aforementioned, the effect size is a statistical concept that measures the strength of the relationship between two variables (Preacher & Kelly, 2012). We focused on Hedges' *g* (Hedges & Olkin, 1985) as a standardized measure of effect size between the treatment group (e.g., experimental group receiving the training method) and the control group (e.g., experimental group receiving no training or a lesser degree of the given training method). A research study may have multiple effect sizes when there are a number of different training treatment groups examined and/or collected data on a number of moderating factors. For example, consider a research study that examined the training method of scaffolding compared to no scaffolding (i.e., the control group), but also investigated two different types of scaffolding administration (i.e., Fixed and Adaptive). In this case, there would be two effect sizes to code: Control vs. Fixed Scaffolding, and Control vs. Adaptive Scaffolding. Further, if the research study also examined the task characteristic of task difficulty (i.e., high vs. low task difficulty) as a moderator, then the total number of effect sizes to be coded would double to four.

Utility for Subset Analysis within TARGET

This coding scheme supported the capability in TARGET, via the underlying algorithms, for a user to filter, analyze, and display subsets of information. For example, within the *Explore Methods* component, a user can filter, analyze, and display subsets of effect size data via the forest plots. The example displayed in Figure 2 represents such a subset analysis. Specifically, the effectiveness of the ‘exploratory learning’ training method is divided into two subsets – effect size data for trainees with lower ability and data for higher ability trainees. Further, if a user wanted to know whether the effectiveness of exploratory learning depended on both trainee ability *and* whether the transfer distance (far transfer or near transfer), a subset analysis could break down the effect size data into four subsets: lower ability trainees/far transfer, lower ability trainees/near transfer, higher ability trainees/near transfer, and higher ability trainees/far transfer. This allows the user to specific under what training conditions they want to view the training effectiveness information for a given training method.

Secondly, within the *Explore Documents* component, TARGET can pool and display document summaries for all research studies using a particular training method or ones which train a particular type of cognitive skill. Moreover, if a user was interested in only examining *part-task training* for training *psychomotor skills*, with *far transfer* being the focus outcome of interest, then the search capabilities within the component would enable only this subset of summaries to be displayed to the user. Finally, users can also receive information concerning how much of each attribute is represented in the overall TARGET research database (e.g., 30% of the research studies examining methods in which to train psychomotor skills). In sum, this capability enables users to view and examine summary information for only the studies that pertain to their immediate interests. The subset of studies can be displayed both as a list of individual study summaries and as frequency counts of each specified attribute.

Computing Meta-Analytic Effect Size Statistics

Computing meta-analyses for the six training methods played a key role in synthesizing the extant training literature. Further, these meta-analytic effect sizes became the basis for the algorithms and weightings that underlie the TARGET tool. The key benefit of meta-analyses is that they can empirically summarize the collective wisdom on a topic. They also provide a way to systematically evaluate the impact of specific moderators (i.e., the attributes described in the prior section) on given relationships of interest. For example, the impact of trainee experience on the relationship between the part-task training and far transfer. In other words, meta-analysis does not just provide an overall rating of whether part-task training “benefits” or “costs/hinders” transfer, but rather provides insight into the specific conditions under which this training method may amplify or diminish its influence on performance.

Hedges’ *g* (Hedges & Olkin, 1985) was used as the meta-analytic effect size statistic. Hedges’ *g* is a well-accepted effect size metric for comparing standardized mean differences between a treatment group (e.g., experimental group receiving the training method) and a control group (e.g., experimental group receiving no training or a lesser degree of the given training method) (Rosenthal, 1991). This effect size statistic was used for three primary reasons. First, as an index from the family of standardized mean differences, the standardized scale of Hedges’ *g* allows comparison of the magnitude of the difference between groups across studies. Second, Hedges’ *g* is an effect size index commonly associated with analysis of variance designs, thus capturing the overlap between distributions of experimental groups as the standardized mean difference. The majority of the research designs from the extant training literature utilize experimental or quasi-experimental research designs. Third, Cohen’s *d’* index, another common index of standardized mean differences, tends to be an upwardly biased estimate of effect size for small samples. Hedges' *g* is a sample adjusted correction to Cohen's *d'* to address the upward bias of *d'* (Borenstein et al., 2009; Hedges & Olkin, 1985; Sheskin, 2007).

When conducting the meta-analyses, our first step was to convert raw data from an individual research study (e.g., data such as means, standard deviations, sample size, test statistics, and p values) into the effect size statistic *d'* (see Appendix B for detailed formulas concerning these conversion). Next, to address the upward bias of *d'*, the *d'* is used to calculate the individual study-level Hedges' *g* effect size statistics (e.g., Hedges' *g*, variance of *g*, standard error of *g*, and 95% confidence interval around *g*) within a *single* study (see Appendix B for formulas). Finally, we calculated summary-level Hedges' *g* effect size statistics, which represents the mean, or average, effect size statistics *across* multiple research studies. For example, say we had 15 research studies that all examined the effectiveness of the 'exploratory learning' training method. For each study, we would calculate in the individual study-level *d'* and Hedges' *g* effect size statistics. Then, we would compute the summary-level Hedges' *g* effect size statistics across all 15 studies. This latter set of statistics provides the better estimate of the true effectiveness of this training method, as it takes into account the findings from all 15 prior research studies.

Algorithm Development

Using the meta-analytic effect size statistics, algorithms were then developed to quantify the relationships between the training methods, performance outcomes, and the various moderating attributes. In particular, within the context of TARGET, this tool's key capabilities were accomplished via the underlying innovative algorithms. First, algorithms were developed to provide real-time computational capability so that users of TARGET could select a number of combinations of moderating attributes and get effect size estimates for a particular training method. Both individual-level and summary-level Hedges' *g* effect size statistics can be generated real-time, using either a fixed effects model or a random effects model. In short, these algorithms allow users to systematically explore the meta-analytic evidence base to understand the expected costs or benefits of the six training methods on performance, for various combinations of task/skill types, trainee characteristics, and performance outcomes. For example, the algorithms help answer questions such as: How do worked examples impact performance on a psychomotor task? Does this effect depend on the experience or skill of the trainee?

Secondly, algorithms were developed to allow real-time filtering of the qualitative summaries of all the research studies in the research database, based on the interests of the user. Filtering is accomplished with the attribute coding scheme that allows users to view study summaries by selecting key attributes of interest such the training method, task/skill type, or outcome measure.

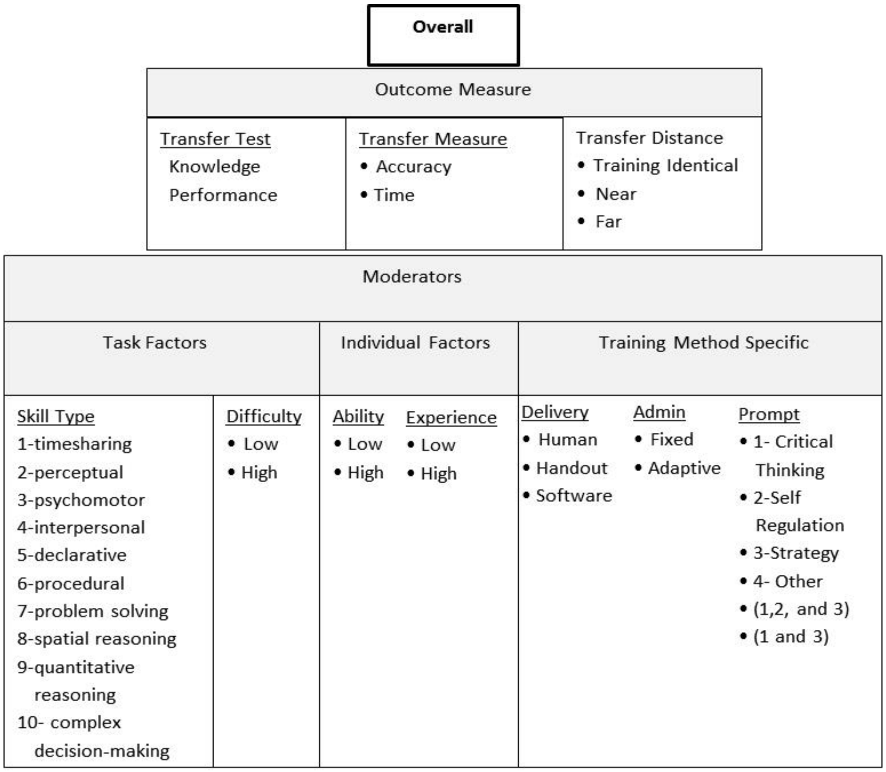
Finally, the developed algorithms also provide the capability of long-term updateability of TARGET's comprehensive research database, allowing users to add research findings/effect size data from new primary research studies conducted. These algorithms can transform a number of different types of raw study-level data (e.g., descriptive statistics, *t*-test statistics, and

F-test statistics) into standardized individual-level effect size statistics; these are then used to update in the summary-level effect size statistics in real-time across all relevant research studies.

Effect Size Extrapolation Approach

Description of the Moderator Analysis Problem

As depicted in Figure 4, effect size estimates within TARGET are organized hierarchically. The highest level in the hierarchy is the least specific - the *overall estimated effectiveness* of a given training method. This estimate is calculated by averaging effects across all studies despite any differences those studies may have. Consider, for example, two studies examining the effectiveness of the Scaffolding method for transfer of training. Study A may conceptualize transfer as a knowledge test that occurs immediately after training. Study B may conceptualize transfer as a performance on a skills test that occurs some time after the training occurred. For the purposes of the overall Scaffolding effectiveness estimate, these differences are unimportant and are thus ignored. As such, the results of Study A and Study B are combined to arrive at an overall estimate of Scaffolding's effectiveness. In some instances, however, the TARGET user may be interested in the differences between Studies A and B, which can be examined by moving down levels of the hierarchy to arrive at more specific comparisons.



Overall

Outcome Measure

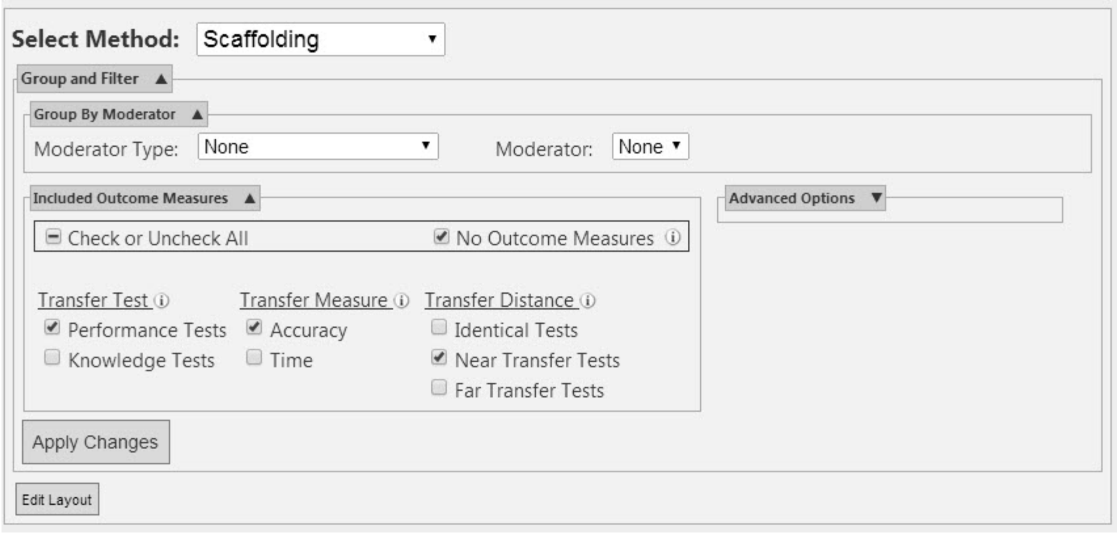
Transfer Test	Transfer Measure	Transfer Distance
Knowledge Performance	• Accuracy • Time	• Training Identical • Near • Far

Moderators

Task Factors		Individual Factors		Training Method Specific		
Skill Type	Difficulty	Ability	Experience	Delivery	Admin	Prompt
1-timesharing 2-perceptual 3-psychomotor 4-interpersonal 5-declarative 6-procedural 7-problem solving 8-spatial reasoning 9-quantitative reasoning 10- complex decision-making	• Low • High	• Low • High	• Low • High	• Human • Handout • Software	• Fixed • Adaptive	• 1- Critical Thinking • 2-Self Regulation • 3-Strategy • 4- Other • (1,2, and 3) • (1 and 3)

***Figure 4*. Scaffolding Moderators and Moderator Variables**

Immediately beneath the overall estimate level of the effect size hierarchy is the *outcome measure* level of specificity. This level allows users to decompose the overall effect estimate into different ways of conceptualizing performance outcomes. Continuing with the example above, a TARGET user may be interested in whether different transfer of training conceptualizations would affect the conclusions about Scaffolding's effectiveness. In other words, is the effectiveness of Scaffolding different when transfer is conceptualized as a knowledge test (Study A) rather than a skills performance test (Study B)? Thus, the overall estimate of Scaffolding effectiveness is further broken down into the effectiveness for knowledge and performance transfer tests. Within the outcome measure level, users can also decompose the overall effect size estimates into differences across transfer measures and transfer distance. For instance, a user may wish to view effect size estimates for skill performance tests where accuracy was the transfer measure and the transfer task was similar to the training task (i.e., near transfer). The user would specify such a search using the options depicted in Figure 5.



Select Method: Scaffolding

Moderator Type: None

Moderator: None

⊟ Check or Uncheck All

☑ No Outcome Measures ⓘ

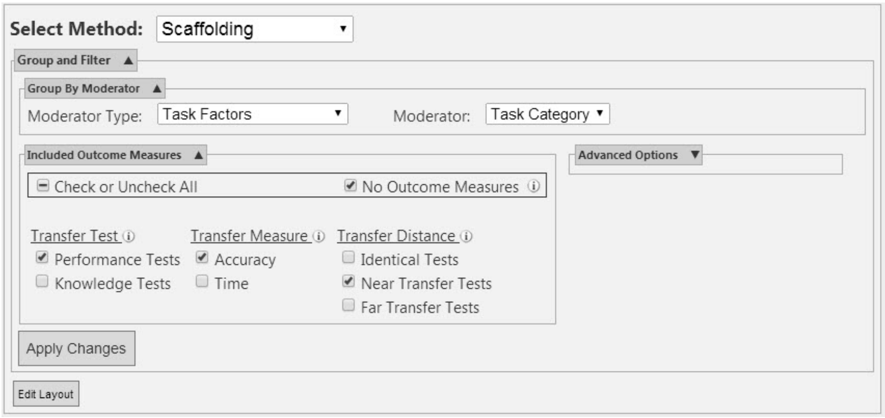
Transfer Test ⓘ
☑ Performance Tests
□ Knowledge Tests

Transfer Measure ⓘ
☑ Accuracy
□ Time

Transfer Distance ⓘ
□ Identical Tests
☑ Near Transfer Tests
□ Far Transfer Tests

Figure 5. **Selecting Outcome Measure Subsets**

The final level of the hierarchy is the other *moderator variables* level. This level is the most specific and includes task (task/skill type, task difficulty), individual (ability, experience), and training method-specific factors (scaffolding delivery, scaffolding administration, prompt type). This allows users to further decompose outcome measure estimates according to even more specific moderators. For example, the user may further decompose Scaffolding effect size estimates for accuracy-based evaluations of near transfer skill performance into even more detailed task categories (e.g., the task/skill types), as shown in Figure 6. Thus, each level of specification down in the hierarchy returns estimates of Scaffolding's effectiveness for an increasingly specific set of circumstances. However, the likelihood of *multiple* primary studies examining the same set of circumstances decreases as the user moves down in the hierarchy because the search criteria become increasingly specific. For example, it is unlikely that enough studies have examined the effectiveness of Scaffolding for knowledge-based, near transfer evaluations of accuracy for psychomotor skills to make a meaningful comparison. Nevertheless, users may be interested in precisely this level of specificity. For this reason, we developed an innovative extrapolation procedure that uses the information that is available to derive effectiveness estimates that are as close as possible to the user’s desired search criteria.


Select Method: Scaffolding

Group and Filter

Group By Moderator

Moderator Type: Task Factors

Moderator: Task Category

Included Outcome Measures

⊟ Check or Uncheck All

☑ No Outcome Measures

Transfer Test
- ☑ Performance Tests
- ☐ Knowledge Tests

Transfer Measure
- ☑ Accuracy
- ☐ Time

Transfer Distance
- ☐ Identical Tests
- ☑ Near Transfer Tests
- ☐ Far Transfer Tests

***Figure 6.* Selecting Outcome Measure and Moderator Subsets**

Extrapolation Overview

Extrapolation has been discussed in the context of generalizations from meta-analyses (e.g., Matt & Cook, 2009; Pigott, 2010) using a framework outlined by Shadish, Cook, and Campbell (2002). In this context, Shadish and colleagues define extrapolation as generalizing beyond the sampled data set. TARGET uses an effect size extrapolation process to estimate effect sizes for two types of moderator combination subsets.

Type 1 stays within the *outcome measure* level of specificity and involves estimating an effect size for performance outcome variable subsets (e.g., the effect size estimate for (a) performance tests where (b) accuracy was the transfer measure and (c) the transfer task reflect near transfer; see Figure 5) that has not been examined by prior research studies. Type 2 also involves the other *moderator variables* level, by including one of the other moderators (task/skill type, task difficulty, trainee ability, trainee experience, or training method-specific moderators, e.g. Figure 6). Note that for each of these types, extrapolation can only be conducted *within* a single training method (e.g., scaffolding)

The extrapolation strategy for TARGET is designed to maintain effect size relationships between moderator variables while minimizing algorithm complexity. For example, with Type 1, the objective is to estimate an effect size for a subset of outcome measure types without any primary study data available, based on the difference between the effect size values of the each of the moderator subsets (e.g., performance test, accuracy measure, and near transfer) and the overall moderator (e.g., Transfer Test, Transfer Measure, Transfer Distance). The differences between individual moderator effect size values and the overall moderator effect size values are assumed to be additive and the total effect size (gain or loss) is computed to be the extrapolated effect size estimate of these three outcome measures.

Extrapolation Procedure

The extrapolation procedure for these two types of moderator combination subsets are described next and illustrated with examples from the TARGET scaffolding data. Table 1 shows the scaffolding meta-analysis effect size data for 'overall' (i.e., across all research studies), each of the three outcome measure moderators (i.e., Transfer Test, Transfer Measure, Transfer Distance), each of the potential outcome measure combination subsets (e.g., performance, accuracy, near transfer), the task/skill type moderator, and the task by outcome measure combination subsets. The gray rows are the moderators or moderator subsets without any primary study data available. The white rows contain the effect size data from the research data and the black rows with white text are the subset rows that are populated with extrapolated data. The outcome measure combination subsets are labeled with their first initials (e.g., **PAN** represents the following combination of outcomes: **P**erformance, **A**ccuracy, **N**ear Transfer). Since there is no contrast data for the 'Time' outcome measure, or the 'Far' transfer task distance (and therefore nowhere to 'extrapolate from'), 8 of the 12 outcome measure combinations cannot be populated.

Let's use the **PAN** 3-outcome moderator combination as an example of how the Type 1 extrapolation process works. In this case, each individual outcome moderator variable is represented in the meta-analytic data set; however, the 3-outcome combination is not represented. Thus, the objective is to estimate an effect size for PAN based on the additive differences between the effect sizes for P, A, and N, and the overall moderator effect sizes for Transfer Test, Transfer Measure and Transfer Distance, respectively. The following 5-step procedure is applied to calculate the extrapolated effect size.

1. Get the single moderator summary-level effect size (i.e., Hedges' *g*) for each of the three outcome variables (i.e., performance, accuracy, near transfer). These represent the average effect size based on the prior studies that include each moderator variable.
    - Example: P = .97, A = .46, N = .66.

2. Get the differences between these values and the overall effect size for the parent moderator (e.g., Transfer Test, Transfer Measure, Transfer Distance). This represents the benefit or cost relative to the overall parent moderator.
    - Example: P = .97 - .46 = .51 , A = .46 - .46 = 0, N = .66-.46 = .20

3. Add the effect size from Step 2 to get a 'total difference estimate' (gain or loss) due to the 3-outcome moderator combination.
    - Example: .51 + 0 + .20 = .71

4. Add the 'total difference estimate' to the 'overall' effect size (across all studies) to get the extrapolated overall effect size for 3-outcome moderator combination (e.g., PAN).
    - Example: .46 + .71 = 1.17.

Table 1.

Scaffolding: Outcome and Task/Skill Type Moderator Subset Observed and Extrapolated Effect Size Values

Moderator Combination		# Data Points	k	*g*	SE	Lower 95% CI	Upper 95% CI
Overall		7	21	.46	.07	.33	.59
Transfer Test - Overall		7	21	.46	.07	.33	.59
Knowledge		5	17	.29	.08	.13	.44
Performance		2	4	.97	.13	.38	1.48
Transfer Measure - Overall		7	21	.46	.07	.33	.59
Accuracy		7	21	.46	.07	.33	.59
Time		0	0	0	0	0	0
Transfer Task Distance-Overall		7	21	.46	.07	.33	.59
Training **I**dentical		6	18	.44	.07	.31	.58
Near		1	3	.66	.23	.20	1.12
Far		0	0	0	0	0	0
Transfer Outcome Moderator Variable Combinations - Overall							
KAI		4	14	.24	.08	.08	.40
KAN		1	3	.66	.234	0.20	1.12
KAF			0				
KTI			0				
KTN			0				
KTF			0				
PAI		2	4	.97	.132	.71	1.23
PAN		0	0	***1.17***	***.366***	***.25***	***1.69***
PAF			0				
PTI			0				
PTN			0				
PTF			0				
Task/Skill Type		7	21	.46	.07	.33	.59
5 – Declarative - Overall		4	14	.24	.08	.08	.40
KAI			14	.24	.08	.08	.40
KAN			0	***.44***	***.32***	***-.18***	***1.06***
PAI			0	NA	NA	NA	NA
PAN			0	NA	NA	NA	NA
7-Problem Solving - Overall		3	7	.89	.12	.67	1.12
KAI			0	***.77***	***.20***	***.38***	***1.16***
KAN			3	.66	.23	.20	1.12
PAI			4	.97	.13	.71	1.23
PAN			0	***1.60***	***.48***	***.66***	***2.54***
9- Quantitative - Overall		1	2	.47	.20	.08	.87
KAI			0	***.24***	***.29***	***-.32***	***.80***
KAN			0	***.67***	***.44***	***-.19***	***1.53***
PAI			2	.47	.20	.08	.87
PAN			0	***1.18***	***.57***	***.06***	***2.30***

Note. KAI = **K**nowledge, **A**ccuracy, **I**dentical Training; KAN = **K**nowledge, **A**ccuracy, **N**ear Transfer; KAF = **K**nowledge, **A**ccuracy, **F**ar Transfer; KTI = **K**nowledge, **T**ime, **I**dentical Training; KTN = **K**nowledge, **T**ime, **N**ear Transfer; KTF = **K**nowledge, **T**ime, **Far** Transfer; PAI = **P**erformance, **A**ccuracy, **I**dentical Training; PAN = **P**erformance, **A**ccuracy, **N**ear Transfer; PAF = **P**erformance, **A**ccuracy, **F**ar Transfer; PTI = **P**erformance, **T**ime, **I**dentical Training; PTN = **P**erformance, **T**ime, **N**ear Transfer; PTF = **P**erformance, **T**ime, **F**ar Transfer.

5. Compute heuristic confidence intervals. While there is no possibility of significance testing for an extrapolated effect size, conservative confidence intervals can be heuristically estimated based on adding the standard error of each of the following two effect sizes and computing 95% confidence interval estimates around the extrapolated effect size. The two effect sizes represented are: 1) the 'extracted from' combination effect size (e.g., overall effect size for individual Performance, Accuracy, and Near Transfer moderators) and 2) the 'extracted to' combination effect size (e.g., PAN).
    - Overall Performance CI = 1.23-0.71 = 0.52
    - Overall Accuracy = 0.59-0.33 = 0.26
    - Overall Near Transfer = 1.12-0.20 = 0.92
    - Add CIs 0.52+0.26+0.92 = 1.70
    - Extrapolated *g* from step 4 = 1.17
    - Lower 95% CI = *g* - (.5*1.70) = 1.17 – 0.85 = 0.32
    - Upper 95% CI = *g* + (.5*1.70) = 1.17 + 0.85 = 2.02

Now, let's consider a second example to illustrate how the Type 2 extrapolation process works. This type involves a subset of outcome measures (as in Type 1 above) with one other, non-outcome moderators (e.g., task/skill type). It can be used for all moderator subgroups that have at least one study effect size in the training method data set (e.g., Declarative, Problem Solving, Quantitative Reasoning), but the combination with three outcome measures is not represented. The objective is to estimate an effect size for the three outcome measure subset in the case where when there is no observed data point for the other moderator variable.

Using the scaffolding example, this extrapolation process assumes that without data to provide contrary evidence, the 3-outcome combination (e.g., knowledge, accuracy, near transfer - **KAN**) will have the same additive moderating effect on a non-outcome moderator variable (e.g., the 'declarative skill' task/skill type), as it has on the variable's 'parent' moderator (e.g., task/skill type collapsed across types). As shown in Table 1, the scaffolding effects data for the task/skill type moderator is limited to three task/skill types (declarative, problem solving, and quantitative reasoning). Each task/skill type can be combined with one of the four potential outcome subsets represented in the meta-analytic data. The exception is declarative knowledge; since only knowledge test are used, the two performance tests are not applicable (NA). The subset of task/skill type by 3-outcome moderator combinations yields 10 moderator combinations, of which 4 have observed effect sizes (i.e., primary data is available). For each of the remaining combinations, the extrapolation procedure is the same: extrapolation from the 3-outcome moderator combination to the combinations of the three outcome moderators plus a fourth moderator variable (e.g., the 'declarative skill' task/skill type). The following 5-step procedure is applied to calculate the extrapolated effect size.

1. For the target 4-variable combination without an effect size (e.g., declarative skill-knowledge-accuracy-near transfer), get the effect size for the target 3-outcome measure combination only (e.g., KAN).
    - The KAN outcome combination effect size value is .66.

2. Get the effect size for the overall non-outcome moderator value (e.g., overall task/skill type).
    - Example: The overall task/skill type effect size is .46.

3. Subtract the overall task/skill type effect size from target 3-outcome measure combination effect size to get the 'total difference estimate' (gain or loss) due to the 3-outcome combination moderator (e.g., KAN).
    - Example: .66 – .46 = .20

4. Get the effect size for the task/skill type moderator variable (e.g., Declarative).
    - Example: Declarative = .24

5. Add the 'total difference estimate' from Step 3 to the moderator variable effect size (e.g., Declarative) from Step 4 to get the expected change in the task/skill type effect size when the extrapolated effect of the outcome combination is added.
    - Example: .24 + .20 = .44

6. Compute heuristic confidence interval estimates based on adding the standard error of each of the two effect sizes, 'extracted from' (e.g., overall 3-outcome variable effect size) and 'extracted to' (e.g., Declarative), and computing CI estimates around the extrapolated moderator effect size.
    - Overall KAN CI = 1.12-.20 = 0.92
    - Overall Declarative CI = 0.40 – 0.08 = 0.32
    - Add CIs 0.92 + 0.32 = 1.24
    - Extrapolated g = 0.44
    - Lower 95% CI = g-(.5*1.24) = 0.44 – 0.62 = -0.18
    - Upper 95% CI = g+(.5*1.24)=0.44+.62 = 1.06

Summary

In summary, the present research was conducted as part of a broader four-year effort to develop evidence-based guidelines for the relative effectiveness of six different training methods for acquiring and transferring cognitive skills involved in complex task domains (see Plott et al., 2014). Several efforts were employed to develop these evidence-based guidelines. First, a comprehensive research database was created via a broad literature search, six comprehensive meta-analyses, and several supplemental research experiments in order to generate effect size estimates of the relative effectiveness of the six training methods and how various factors (e.g., task/skill type, trainee characteristics) moderate training effectiveness. Second, we developed algorithms to empirically synthesize the relationships between the six training methods and various moderating factors. Moreover, to ensure these research findings and algorithms would be easily consumable by various users, we developed TARGET, a user-friendly graphical user interface tool. TARGET contains several visualization tools, such that in-depth statistical knowledge is not required to benefit from this tool. As such, users (with varying levels of expertise) can easily use TARGET's evidence-based recommendations to identify the most effective training method given a set of desired attributes/conditions. The purpose of this report was to provide information on the algorithm development completed as a part of this larger research effort, as well as their use in TARGET.

Algorithm Capabilities within TARGET and Beyond

TARGET's key capabilities were accomplished via the underlying innovative algorithms. First, the algorithms provide real-time computational capability so that users of TARGET can select a number of combinations of moderating attributes and get effect size estimates for a particular training method. Both individual-level Hedges' *g* effect size statistics from raw study-level information, as well as the summary-level effect size statistics across primary research studies can be generated, using either a fixed effects model or a random effects model. Note that there were several combinations of moderating variables for which there is no extant research available (e.g., the impact of trainee experience on the effectiveness of part-task training of perceptual skills) and thus the algorithms could not be directly employed to generate effect sizes. Accordingly, we also defined and implemented an innovative extrapolation process for estimating effect size statistics for particular attribute combinations in which there was no extant research available. Such effect size estimates were based on extrapolating from a subset of moderator variables that did include the target variables – that is, from the nearest "parent."

Secondly, the algorithms also allow real-time filtering of the qualitative summaries of all the research studies in the research database, based on the interests of the user. Filtering is accomplished with the attribute coding scheme that allows users to view study summaries by selecting key attributes of interest such the training method, task/skill type, or outcome measure. Finally, the developed algorithms also provide the capability of long-term updateability of TARGET's comprehensive research database, allowing users to add research findings/effect size data from new primary research studies conducted.

In summary, the algorithms offer several potential benefits to the Army. As implemented in TARGET, they can be used to perform tradeoff analyses for different combinations of training methods. The algorithms make the research findings from this larger 4-year research effort available to the Army training, development, and research communities, allowing users to systematically explore training methods and design components that would be effective for a particular set of circumstances for acquiring cognitive skills. In short, they estimate the expected costs or benefits of the six training methods on performance, for various combinations of task/skill type, trainee characteristics, and performance outcomes.

Beyond TARGET, the algorithms could also be adapted for use in other human performance models, such as Improved Performance Research Integration Tool (IMPRINT). IMPRINT was developed by the U.S. Army Research Laboratory, Human Research and Engineering Directorate (ARL-HRED) to support Manpower and Personnel Integration (MANPRINT) and HSI analyses. IMPRINT is a modeling tool designed to help assess the interaction of Soldier and system performance. With IMPRINT, users can gain useful information about processes that might be too expensive or time-consuming to test in the real world. Adding the training effects algorithms from this research into IMPRINT would allow users to determine how different training methods might impact performance and predict which training methods (or combination of methods) will result in the most effective performance. Either of these uses can assist program managers with training design. The algorithms can provide the basis for cost benefit analyses of different training methods and may enable program managers to make decisions concerning the amount of training that system operators and maintainers should receive, as well as what basic types of training methods should be developed to support this training.

References

Bell, B. S., & Kozlowski, S. W. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology, 93*, 296-316.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to meta-analysis.* West Sussex, UK: John Wiley & Sons, Ltd.

Carolan, T., McDermott, P.L., Hutchins, S., Wickens, C.D., & Belanich, J. (2011). Investigating the impact of training on performance. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference.* Arlington, VA: National Defense Industrial Association.

Carolan, T.F., McDermott, P.L., Wickens, C.D., Fisher, A., & Gronowski, M. (under review). Guidance and exploratory learning: Application to transfer and retention in a procedural learning task. *Cognitive Engineering and Decision Making Journal*.

Carolan, T.F., Wickens, C.D., Hutchins, S. & Cumming, J. (in press). Costs and benefits of more learner freedom: Meta-analyses of exploratory and learner control training methods. *Human Factors*.

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* London, UK: Academic Press.

Hutchins, S., McDermott, P. L., Carolan, T. F., Gronowski, M. R., Fisher, A., & DeMay, M. (2013). *Interpersonal skills summary report*. (Research Note, No. 2013-03). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hutchins, S., Wickens, C.D., Carolan, T., & Cumming, J. (2013). The influence of cognitive load on transfer with error prevention training methods: A meta-analysis. *Human Factors, 55*, 854-874.

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

Matt, G.E., & Cook, T.D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L.V. Hedges, and J.C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 537–560). New York: Russell Sage Foundation.

McDermott, P. L., Carolan, T. F., Fisher, A., Gronowski, M. R., & Gacy, M. (2013). *Transferring from the simulator to a live robotic environment: The effectiveness of part-task and whole-task training*. (Technical Report, No. 1331). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McDermott, P.L., Carolan, T., & Gronowski, M.R. (2012). Application of worked examples to unmanned vehicle route planning. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference.* Arlington, VA: National Defense Industrial Association.

McDermott, P.L., Carolan, T., & Wickens, C.D. (2012). Part-task training methods in simulated and realistic tasks. *Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society*. Boston, MA: Human Factors and Ergonomics Society.

McDermott, P. L., Gronowski, M. ., Carolan, T. F., & Fisher, A. (2013). Error training and adaptive remediation: The impact on transfer performance in a complex planning task. *Proceedings of the 57th Annual Meeting of the Human Factors and Ergonomics Society.* San Diego, CA: Human Factors and Ergonomics Society.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*, 1–4.

Pigott, T.D. (2012). A*dvances in meta-analysis: Statistics for social and behavioral sciences.* New York: Springer Science

Plott, B., & Hutchins, S. (2013). *Training aide: Research and guidance for effective training - User guide.* (Research Product 2014-02). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Plott, B., McDermott, P. L., Archer, S., Carolan, T. F., Hutchins, S., Fisher, A., Gronowski, M. R., Wickens, C.D., & Orvis, K. A. (2014). *Understanding the impact of training on performance*. (Technical Report 1341). Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Preacher, K. J., & Kelly, K. (2012). On effect size. *Psychological Methods, 17,* 137-152.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th Ed.). New York: Chapman & Hall/CRC.

Wickens, C.D., Hutchins, S., Carolan, T., & Cumming, J. (2013). The effectiveness of part-task training and increasing-difficulty training strategies: A meta-analysis approach. *Human Factors, 55*, 461-470.

van Merriënboer, J. J. G., Kester, L., & Paas, F. (2006). Teaching complex rather than simple tasks: balancing intrinsic and germane load to enhance transfer of learning. *Applied Cognitive Psychology, 20,* 343–352.

Appendix A
Example Coding Scheme Attributes

Category	Attribute Name	Attribute Definition
General Training Attributes		
Training Method	Part-Task Training	A training method that decomposes complex tasks into a series of smaller tasks, each of which is demonstrated and practiced separately before being practiced as a whole task.
	Increasing Difficulty	A training method in which parameters of the task are initially set to lower difficulty levels, to reduce the intrinsic load early in training, and then increased as training progresses, until the difficulty reaches the level of the target task. The difficulty levels can increase in either a fixed, pre-determined schedule or adaptively based on the trainee's performance.
	Learner Control	A training method that provides trainees with decision making control over specific dimensions or activities within a structured learning environment.
	Exploratory Learning	A training method in which the trainee explores a task environment on his or her own. The level or type of guidance given to the trainee can vary within this method (e.g., only providing a user manual to reference versus the provision of input by trainers in response to trainee questions).
	Training Wheels	A training method geared towards reducing the difficulty of the target task during initial learning by reducing training task errors, as well as helping trainees acquire the appropriate schema to assimilate the target task.
	Scaffolding	A training method where assistive supports are provided to trainees to ease the demands of task performance. These scaffold supports are incrementally faded out over time until the trainee is executing the whole task independently.

Category	Attribute Name	Attribute Definition
Task/Skill Type	Timesharing	This skill category involves two (or more) subtasks/parts/components of a whole task are performed concurrently in the whole task (e.g., training vertical control and lateral control in flight task, since a good pilot will time-share these two when doing a climbing turn). Another example: learning left hand and right hand on the piano, or strumming and chording on the guitar.
	Perceptual	This skill category involves detecting and interpreting sensory information in order to gain awareness and support performance. The ability to detect and use sensory cues. Includes some perceptual reasoning tasks.
	Psychomotor	This skill category involves physical skills, such as movement, coordination, manipulation, dexterity, strength, and speed; also, includes both fine motor skills and gross motor skills.
	Interpersonal	This skill category involves social interaction skills, such as communication, exchange information, persuasion, building and maintaining relationships, managing conflict, and interacting effectively.
	Cognitive-Declarative	This skill category includes all verbal knowledge categories including facts, principles, knowledge organization, and verbal or graphical mental models, concept maps, etc.
	Cognitive-Procedural	This skill category includes performance of basic procedural skills, where the procedure may include cognitive steps and basic perceptual and motor steps, such as those involved in navigating a computer interface. Constrained sequences of physical and cognitive activities performed in predictable situations.
	Cognitive-Problem Solving	This skill category involves application of principles, rules and concepts to process information and solve problems. Includes general logical reasoning skills: inductive, deductive, diagnostic, etc.

Category	Attribute Name	Attribute Definition
Task/Skill Type (Cont.)	Cognitive-Spatial Reasoning	This skill category involves visual-spatial skills, such as representing, transforming, generating, and recalling symbolic, nonlinguistic information. Includes: mental rotation; spatial perception - ability to determine spatial relationships with respect to own body orientation; spatial visualization - multi-step manipulations of spatially presented information, requiring analysis of the relationship between different spatial representations.
	Cognitive-Quantitative Reasoning	This skill category involves the application of mathematical concepts and skills to solve problems.
	Cognitive-Complex Decision making	This skill category involves situation assessment and decision making in complex, dynamic and time-sensitive environments with changing situations, attentional demands, application of strategies, multiple goals; also, may involve coordination of perceptual and motor skills.
Task Difficulty	Low	An experimental condition where task difficulty is manipulated to be low; or the training task clearly has no interacting variables making it non-complex.
	High	A condition where task difficulty is manipulated to be high; or the training task clearly has interacting variables making it more complex.
Trainee Ability	Low	Participants were assessed or tested as having low, often less than median, general or task-related ability, skill, or aptitude.
	High	Participants were assessed or tested as having high, often greater than median, general or task-related ability, skill, or aptitude.
Trainee Experience	Low	Participants were assessed or screened through sampling as having little to no prior knowledge, familiarity, or practice with the training content.

	High	Participants were assessed, screened through sampling, or trained up to having moderate to significant prior knowledge, familiarity, or practice with the training content.

Category	Attribute Name	Attribute Definition
Transfer Distance	Near Transfer - Identical	Identical near transfer tests examine the application of what was learned in training using a task or problem identical to that used during training.
	Near Transfer - Similar	Similar near transfer tests examine the application of what was learned in training using a task or problem somewhat different, but similar, to that used during training.
	Far Transfer	Far transfer tests examine the application of what was learned in training using a testing situation that is different or new from the task or problem that was used in training.
Transfer Measure	Accuracy/ Quality	Transfer tests that assess the execution of appropriate action or response (e.g., steps, multiple choice/true-false/open-ended answers, problems, recall, actions), measured as the percent correct, # correct, accumulated points, task/test scoring correct action/timing/performance. This also includes quality rating by SMEs or inverse of error performance.
	Performance/ Response Time	Transfer tests that assess the time to respond or take appropriate action, measured as time to complete the entire transfer task (i.e., time to complete), time to complete a subtask (i.e., time to respond), time to reach a criterion level of performance (i.e., time to criterion), or trials to reach a criterion level of performance (i.e., trials to criterion).
Transfer Test	Knowledge	Knowledge tests capture the ability of a trainee to recall information or problem solve from a trained skill.
	Performance/ Cognitive Skill	Performance tests capture the ability of a trainee to physically perform a trained skill.
Training	Academic	Study sample is undergraduate or high school students.

Sample	Military	Study sample is military personnel.
	Work	Study sample is people in the workforce.
Category	**Attribute Name**	**Attribute Definition**
Training Delivery Method	Classroom	The primary delivery environment is any classroom-based delivery environment led by an instructor including the use of lecture, computers, handouts, texts, etc.
	Computer-Based	The primary delivery environment is a computer-based presentation of the training material. The computer-based training environment can be online (web-based), on a local network, or on a single computer. The training methods are implemented within the computer-based environment. Includes CAI, CBT, and many ITS systems.
	Field/Real System	The primary delivery environment is a field exercise with real equipment or training using the actual operational system or device. This includes training to use computer applications, sports, real science labs, etc.
	Gaming	The primary delivery environment is an interactive simulation with competition and scoring, as well as constraints, privileges and penalties.
	Simulation-Based	The primary delivery environment is an interactive synthetic environment that approximates the real world environment or equipment in which the target task is to be performed. It can include desktop system simulations, equipment simulators, virtual environments, and networked environments. It also includes simulation gaming environments. A key difference from computer-based is that the trainee manipulates variables in the environment to change system states and outcomes.
	Web-Based	The primary delivery environment is via a computer using the Internet, enabling instant updating, distribution, and sharing of information. This environment gives learners the ability to communicate interactively with others online and access to other information and media via the web.

Category	Attribute Name	Attribute Definition
Training Method-Specific Attributes for 'Scaffolding' Training Method		
Scaffolding Prompt Type	Critical Thinking About the Concept	Prompts emphasizing key concepts and relationships critical to learning the training content.
	Self-Regulation of Learning	Prompts emphasizing that participants monitor their emerging understanding and plan/engage their learning experience.
	Strategy-Based Prompts	Prompts emphasizing underlying principles, rules, and processes required to solve the problem.
	Other	A general category for all other types of prompts.
Scaffolding Administration	Fixed	Scaffolding support removed or added on a fixed schedule independent of the participant's performance.
	Adaptive	Scaffolding support removed or added adaptively based on participant performance in training.
Scaffolding Delivery	Human Tutor	The error preventative scaffold was delivered to the trainee via a human tutor.
	Handout	The error preventative scaffold was delivered to the trainee via a handout.
	Computer/Web	The error preventative scaffold was delivered to the trainee electronically over the computer.

Appendix B

Formulas for Computing Meta-Analytic Effect Size Statistics

Computing Standardized Effects from Raw Data

The following formulas can be used to calculate the effect size statistic *d'* from raw study data, which can then be used to calculate the individual study-level Hedges' *g* effect size statistics within a *single* research study. The standardized effect size statistic *d'* can be computed using raw study-level data including: (1) descriptive statistics, (2) *t*-test output results, and *F*-test output results. The formulas for each are listed below.

Computing d' from descriptive statistics. Four different sets of descriptive statistics can be used to compute a standardized effect size. Regardless of which option is selected, the end function used to compute the standardized effect size from the raw data is:

$$d' = (\text{Mean}_{\text{treatment}}\text{-Mean}_{\text{control}})/\ \text{SD}_{\text{pooled}}$$

However, each of the following four descriptive statistics options requires four different sets of functions to get to *d'* from raw data provided. Hedges and Olkin's (1985) formulas for computing *d'* are listed below.

1. **Descriptive statistics - Option 1**. For treatment mean, treatment standard deviation, treatment sample size (n), control mean, control standard deviation, and control sample size:

   i. $$SD_{pooled} = \sqrt{\frac{\left(\left(SD^2_{treatment}(n_{treatment}-1)\right)+\left(SD^2_{control}(n_{control}-1)\right)\right)}{N_{total}-2}}$$

   ii. $$d' = \frac{(Mean_{treatment}-Mean_{control})}{SD_{pooled}}$$

2. **Descriptive statistics - Option 2**. For treatment mean, treatment standard deviation, control mean, control standard deviation, and overall sample size (N) when the sample size is equal across the treatment and control:

   i. $\text{n}_{\text{treatment}} = \text{N}/2$, $\text{n}_{\text{control}} = \text{N}/2$, and $\text{n}_{\text{treatment}} = \text{n}_{\text{control}}$

   ii. $$SD_{pooled} = \sqrt{\frac{\left(\left(SD^2_{treatment}(n_{treatment}-1)\right)+\left(SD^2_{control}(n_{control}-1)\right)\right)}{N_{total}-2}}$$

iii. $d' = \frac{(Mean_{treatment} - Mean_{control})}{SD_{pooled}}$

3. **Descriptive statistics - Option 3**. For treatment mean, treatment sample size, control mean, control sample size, and pooled standard deviation (SD_{pooled}).

   i. $d' = \frac{(Mean_{treatment} - Mean_{control})}{SD_{pooled}}$

4. **Descriptive statistics - Option 4**. For treatment mean, control mean, pooled standard deviation (SD_{pooled}), and overall sample size (N) when the sample size is equal across the treatment and control:

   i. $n_{treatment} = N/2$, $n_{control} = N/2$, and $n_{treatment} = n_{control}$
   ii. $d' = \frac{(Mean_{treatment} - Mean_{control})}{SD_{pooled}}$

Computing d' from *t*-test output. Six different sets of *t*-test output statistics can be used to compute a standardized effect size. Regardless of which of option is selected, the end function used to compute the standardized effect size from the raw data is:

$$d' = t * \sqrt{\frac{(n_{treatment} + n_{control})}{(n_{treatment} * n_{control})}}$$

However, each of the following six *t*-test output options requires six different sets of functions to get to *d'* from the raw data provided. Hedges and Olkin's (1985) formulas for computing *d'* are listed below.

1. ***T*-test output – Option 1**. For *t* value, Treatment sample size, and control sample size:

   i. $d' = \pm\, t\ value \sqrt{\frac{(n_{treatment} + n_{control})}{(n_{treatment} \times n_{control})}}$

2. ***T*-test output – Option 2a**. For *t* value and degrees of freedom (df) when $n_{treatment} = n_{control,}$ and samples are *independent*:

   i. $n_{treatment} = (df+2)/2$, $n_{control} = (df+2)/2$, and $n_{treatment} = n_{control}$

ii. $d' = \pm\ t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

***T*-test output – Option 2b**. For *t* value and degrees of freedom (df) when $n_{treatment}$ = $n_{control,}$ and samples are *dependent*:

i. $n_{treatment}$ = df+1, and $n_{treatment}$ = $n_{control}$

ii. $d' = \pm\ t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

3. ***T*-test output – Option 3**. For *t* value and overall sample size (N) when $n_{treatment}$ = $n_{control}$:

i. $n_{treatment}$ = N/2, $n_{control}$ = N/2, and $n_{treatment}$ = $n_{control}$

ii. $d' = \pm\ t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

4. ***T*-test output – Option 4a**. For *t*-test p value, treatment sample size, control sample size, and when samples are *independent*:

i. df = ($n_{treatment}$ + $n_{control}$)-2

ii. Reference t value from a table of Student’s t values using df and p value.

iii. $d' = \pm\ t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

***T*-test output – Option 4b**. For *t*-test p value, treatment sample size, control sample size, and when samples are *dependent*:

i. df = ([$n_{treatment}$ + $n_{control}$]/2)-1

ii. Reference t value from a table of Student’s t values using df and p value.

iii. $d' = \pm\ t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

5. ***T*-test output – Option 5a**. For *t*-test p value and degrees of freedom (df) when $n_{treatment}$ = $n_{control,}$ and samples are *independent*:

i. Reference t value from a table of Student’s t values using df and p value.

ii. $n_{treatment}$ = (df+2)/2, $n_{control}$ = (df+2)/2, and $n_{treatment}$ = $n_{control}$

iii. $d' = \pm\, t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

***T*-test output – Option 5b**. For *t*-test p value and degrees of freedom (df) when $n_{treatment}$ = $n_{control,}$ and samples are *dependent*:

i. Reference t value from a table of Student's t values using df and p value.
ii. $n_{treatment}$ = df+1, $n_{control}$ = df+1, and $n_{treatment}$ = $n_{control}$
iii. $d' = \pm\, t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

6. ***T*-test output – Option 6a**. For *t*-test p value and overall sample size (N) when $n_{treatment}$ = $n_{control}$, and samples are *independent*:

i. df = ($n_{treatment}$ + $n_{control}$)-2
ii. Reference t value from a table of Student's t values using df and p value.
iii. $n_{treatment}$ = N/2, $n_{control}$ = N/2, and $n_{treatment}$ = $n_{control}$
iv. $d' = \pm\, t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

***T*-test output – Option 6b**. For *t*-test p value and overall sample size (N) when $n_{treatment}$ = $n_{control}$, and samples are *dependent*:

i. df = ([$n_{treatment}$ + $n_{control}$]/2)-1
ii. Reference t value from a table of Student's t values using df and p value.
iii. $n_{treatment}$ = N/2, $n_{control}$ = N/2, and $n_{treatment}$ = $n_{control}$
iv. $d' = \pm\, t\ value \sqrt{\frac{(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

Computing d' from F test output. Six different sets of *F*-test output statistics can be used to compute a standardized effect size. Regardless of which of option is selected, the end function used to compute the standardized effect size from the raw data is:

$$d' = \sqrt{\frac{F*(n_{treatment}+n_{control})}{(n_{treatment}*n_{control})}}$$

However, each of the six *F*-test output options requires six different sets of functions to get to *d'* from the raw data provided. Hedges and Olkin's (1985) formulas for computing *d'* are listed below.

1. ***F*-test output – Option 1**. For *F* value, treatment sample size, control sample size:

   i. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

2. ***F*-test output – Option 2a**. For *F* value and degrees of freedom (df) when $n_{treatment} = n_{control}$, and samples are *independent*:

   i. $n_{treatment} = (df+2)/2$, $n_{control} = (df+2)/2$, and $n_{treatment} = n_{control}$
   ii. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

   ***F*-test output – Option 2b**. For *F* value and degrees of freedom (df) when $n_{treatment} = n_{control}$, and samples are *dependent*:

   i. $n_{treatment} = df+1$, $n_{control} = df+1$, and $n_{treatment} = n_{control}$
   ii. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

3. ***F*-test output – Option 3**. For *F* value and overall sample size (N) when $n_{treatment} = n_{control}$:

   i. $n_{treatment} = N/2$, $n_{control} = N/2$, and $n_{treatment} = n_{control}$
   ii. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

4. ***F*-test output – Option 4a**. For *F*-test p value, treatment sample size, control sample size, and when samples are *independent*:

   i. $df_{numerator} = 1$
   ii. $df_{denominator} = (n_{treatment} + n_{control})-2$

iii. Reference F value from an F value distribution table using df and p value.

iv. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

***F*-test output – Option 4b**. For *F*-test p value, treatment sample size, control sample size, and when samples are *dependent*:

i. $df_{numerator} = 1$

ii. $df_{denominator} = ([n_{treatment} + n_{control}]/2)-1$

iii. Reference F value from an F value distribution table using df and p value.

iv. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

5. ***F*-test output – Option 5a**. For *F*-test p value and degrees of freedom (df) when $n_{treatment} = n_{control}$, and samples are *independent*:

i. Reference F value from an F value distribution table using df and p value.

ii. $n_{treatment} = (df+2)/2$, $n_{control} = (df+2)/2$, and $n_{treatment} = n_{control}$

iii. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

***F*-test output – Option 5b**. For *F*-test p value and degrees of freedom (df) when $n_{treatment} = n_{control}$, and samples are *dependent*:

i. Reference F value from an F value distribution table using df and p value.

ii. $n_{treatment} = df+1$, $n_{control} = df+1$, and $n_{treatment} = n_{control}$

iii. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

6. ***F*-test output – Option 6a**. For *F*-test p value and overall sample size (N) when $n_{treatment} = n_{control}$, and samples are *independent*:

i. $n_{treatment} = N/2$, $n_{control} = N/2$, and $n_{treatment} = n_{control}$

ii. $df_{numerator} = 1$

iii. $df_{denominator} = (n_{treatment} + n_{control})-2$

iv. Reference F value from an F value distribution table using df and p value.

v. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

***F*-test output – Option 6b**. For *F*-test p value and overall sample size (N) when $n_{treatment}$ = $n_{control}$, and samples are *dependent*:

i. $n_{treatment}$ = N/2, $n_{control}$ = N/2, and $n_{treatment}$ = $n_{control}$
ii. $df_{numerator}$ = 1
iii. $df_{denominator}$ = ([$n_{treatment}$ + $n_{control}$]/2)-1
iv. Reference F value from an F value distribution table using df and p value.
v. $d' = \pm\sqrt{\frac{F\ value(n_{treatment}+n_{control})}{(n_{treatment}\times n_{control})}}$

Computing Individual Study-Level Hedges' *g* Effect Sizes

To address the upward bias of *d'*, the following formulas can be used to calculate the individual study-level Hedges' *g* effect sizes within a *single* study from *d'*.

Computing model independent quantities. Model independent quantities represent the set of Hedges' *g* effect size statistics computed for an individual study-level contrast. Computation of the model independent quantities only utilizes the information from the specific contrast to compute the statistics listed below. These statistics are considered "model independent" because they don't consider information from other contrasts for their computations. Conversely, when model independent quantities are combined, or modeled, to produce summary-level effects across a range of contrasts, then additional statistics need to be computed that are dependent upon the type of model used for the summary-level effects (i.e., a fixed effects model and a random effects model). The sets of analysis dependent quantities for an individual study that are used in computation of fixed effects or random effects model will be presented after this sub-section.

There are five model independent quantities. Hedges and Olkin's (1985) and Bornstein et al.'s (2009) formulas for computing these model independent quantities are listed below.

1. **Hedges' g**:

Hedges' *g* = $g_i = d\left(1 - \frac{3}{4N-9}\right)$, where N = n_t + $n_{c,}$

2. **Within study variance of Hedges’ *g***:

$$\text{Variance of } g_i = \hat{\sigma}_i^{\,2} = \left(\frac{1}{n_t} + \frac{1}{n_c} + \frac{g_i^{\,2}}{2(n_t+n_c)}\right) \times \left(\mathbf{1} - \frac{3}{4(n_t+n_c)-9}\right)^2$$

3. **Standard error of Hedges’ g**:

$$SE \text{ of } g = \hat{\sigma}_i = \sqrt{\frac{1}{n_t} + \frac{1}{n_c} + \frac{g_i^{\,2}}{2(n_t+n_c)}}\left(\mathbf{1} - \frac{3}{4(n_t+n_c)-9}\right)$$

or

$$SE \text{ of } g = \hat{\sigma}_i = \sqrt{\hat{\sigma}_i^{\,2}}$$

4. **Lower limit of the 95% confidence interval around Hedges’ *g***:

$$95\% \text{ C.I. of } g = g_i - (1.96 * \hat{\sigma}_i)$$

5. **Upper limit of the 95% confidence interval around Hedges’ *g***:

$$95\% \text{ C.I. of } g = g_i + (1.96 * \hat{\sigma}_i)$$

Computing fixed effect quantities. Fixed effect quantities represent the additional set of Hedges’ *g* effect size statistics (used in conjunction with the model independent quantities) computed for an individual study-level contrast that are used in computation of fixed effects model summary statistics. Fixed effects models assume a single source of variation between effect sizes due *only* to sampling error (i.e., within-study variance). The assumption of a fixed effects model is that each contrast’s effect size is a sample from a distribution of effect sizes around a true population mean. Consequently, the only source of variation around the true population mean is random variation of each contrast’s estimate from the true mean. A fixed effects model is typically chosen when the modeler assumes that each research study is an equivalent context of training assessment; that is, each study is assumed to be estimating the same true population effect rather than different population effects associated with between-study differences in training environment, sample characteristics, training tasks, etc. The single variation parameter associated with the assumption a single true population mean and random variation of each contrast’s estimate around that mean is used to compute the set of fixed effects quantities that are later used to compute the fixed effect model summary-level statistics. There are five fixed effects quantities. Bornstein et al.'s (2009, p. 88-89) formulas for computing these fixed effects quantities are listed below.

1. **Fixed effect weight of *g***:

$$\text{Fixed Effect Weight (inverse variance)} = w_i = \frac{1}{\hat{\sigma}_i{}^2}$$

2. **Fixed effect relative weight of *g***:

$$\text{Fixed Effect Relative Weight} = \frac{w_i}{\sum_{i=1}^{k} w_i}$$

3. **Fixed effect weighted g**:

$$\text{Fixed Effect Weighted } g = (g_i)(w_i)$$

4. **Fixed effect weighted g^2**:

$$\text{Fixed Effect Weighted Squared } g = (g_i)^2(w_i)$$

5. **Fixed effect squared weight of g**:

$$\text{Fixed Effect Squared Weight} = (w_i)^2$$

Computing random effects quantities. The random effects model is considered a more advanced modeling alternative because the random effects model makes more assumptions about how effect size estimates vary. To be consistent with the general logic of TARGET, the simplest high-level information is always presented first and then more advanced more restricted options are made available depending upon the individual user's preference and comfort level. As such, the fixed effects model is the default model in TARGET.

Random effects quantities represent the set of Hedges' *g* effect size statistics computed for an individual study-level contrast that are used in computation of random effects model summary-level statistics. In contrast to the fixed effects model, the assumption of a random effects model is that each contrast's effect size is a sample from a distribution of effect sizes around a unique population mean. Consequently, *two* sources of variation exist for the random effects model; random variation of the sample from the distribution around the unique population mean, and random variation of the sample from the mean of the distribution of unique population means.

A random effects model is typically chosen when the modeler assumes that each research study represents a different context of training assessment; that is, each study is assumed to be estimating a different population mean due to between-study differences in training environment, sample characteristics, and/or training tasks. The random effects model summary statistic is a calculation of the mean of the distribution of different unique population parameters. Therefore, random effects models assume two sources of variation: within-study variance (as with the fixed effects model) and variation due to true differences between effect size estimates (i.e., between-study variance). The two variation parameters are used to compute the set of random effects quantities that are later used to compute the random effects summary-level statistics. There are five random effects quantities. Bornstein et al.'s (2009, p. 72-73, 88-89) formulas for computing these random effects quantities are listed below.

1. **Between study variance of g, or $\mathbf{T^2}$**:

$$\text{T2} = \frac{\left(\sum_{i=1}^{k} w_i g_i^2 - \frac{\left(\sum_{i=1}^{k} w_i g_i\right)^2}{\sum_{i=1}^{k} w_i}\right) - (k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}$$

2. **Total variance of *g*, or $\hat{\sigma}_i^{*\,2}$**:

$$\text{Total Variance of } g_i = \hat{\sigma}_i^{*\,2} = \hat{\sigma}_i^{\,2} + \text{T2}$$

3. **Random effects weight of *g***:

$$\text{Random Effects Weight (inverse total variance)} = w_i^* = \frac{1}{\hat{\sigma}_i^{*\,2}}$$

4. **Random effects relative weight of *g***:

$$\text{Random Effects Relative Weight} = \frac{w_i^*}{\sum_{i=1}^{k} w_i^*}$$

5. **Random effects weighted *g***:

$$\text{Random Effects Weighted } g^* = (g_i)(w_i^*)$$

Computing Summary-Level Hedges' *g* Statistics

The previous section described the Hedges' *g* computations for each individual study-level contrast. To view the summary-level effects for a range of individual contrasts (i.e., the *mean* Hedge's *g*), the set of contrasts must be modeled in order to pool the contrasts into an overall effect. The purpose of this section is to provide formulas for computing summary-level Hedges' *g* statistics using either a fixed effect model or random effects model.

Computing Fixed Effects Summary Statistics

There are five fixed effects summary statistics. Bornstein et al.'s (2009, p. 66) formulas for computing these fixed effects summary statistics are listed below.

1. ***g+***: The summary effect, or mean of *g* under a fixed effects model.

$$\text{g+ = sum of weighted g / sum of weights} = \frac{\sum_{i=1}^{k} w_i g_i}{\sum_{i=1}^{k} w_i}$$

2. **Within study variance of *g+***: The within study variation in the estimate of *g+*.

$$\hat{\sigma}^2_{g+} = \left(\sum_{i=1}^{k} \frac{1}{\hat{\sigma}^2_{gi}}\right)^{-1} = \frac{1}{\sum_{i=1}^{k} \frac{1}{\hat{\sigma}^2_{gi}}} = \frac{1}{\sum_{i=1}^{k} wi}$$

3. **Standard error of *g+***: The standardized estimate of the mean deviation error of the sample around *g+*.

$$\hat{\sigma}_{g+} = \sqrt{\hat{\sigma}^2_{g+}}$$

4. **Lower limit of the 95% confidence interval around *g+***: The lower bound of the 95% confidence interval likely to contain the true *g+*.

$$g+ -\left(1.96 * \hat{\sigma}_{g+}\right)$$

5. **Upper limit of the 95% confidence interval around *g+***: The upper bound of the interval likely to contain the true *g+*.

$$g+ +\left(1.96 * \hat{\sigma}_{g+}\right)$$

Computing Random Effects Summary Statistics

There are five random effects summary statistics. Bornstein et al.’s (2009, p. 73-74) formulas for computing these random effects summary statistics are listed below.

1. ***g+***: The summary effect, or mean of *g* under a random effects model.

$$\text{g+} = \text{sum of weighted* g / sum of weights*} = \frac{\sum_{i=1}^{k} w_i^* g_i}{\sum_{i=1}^{k} w_i^*}$$

2. **Total variance of *g+***: The composite variation in the estimate of *g*+.

$$\hat{\sigma}_{g+}^2 = \frac{1}{\sum_{i=1}^{k} \frac{1}{\hat{\sigma}_i^{*2}}} = \frac{1}{\sum_{i=1}^{k} w_i^*}$$

3. **Standard error of *g+***: The standardized estimate of the mean deviation error of the sample around *g*+.

$$\hat{\sigma}_{g+} = \sqrt{\hat{\sigma}_{g+}^2}, \text{ where } \hat{\sigma}_{g+} \text{ is computed from total variance}$$

4. **Lower limit of the 95% confidence interval around *g+***: The lower bound of the 95% confidence interval likely to contain the true *g*+.

$$\text{g+} - \left(1.96 * \hat{\sigma}_{g+}\right), \text{ where } \hat{\sigma}_{g+} \text{ is computed from total variance}$$

5. **Upper limit of the 95% confidence interval around *g+***: The upper bound of the interval likely to contain the true *g*+.

$$\text{g+} + \left(1.96 * \hat{\sigma}_{g+}\right), \text{ where } \hat{\sigma}_{g+} \text{ is computed from total variance}$$