Cognitive Task Analysis–Based Training: A Meta-Analysis of Studies

Colby Tofel-Grehl and **David F. Feldon**, University of Virginia

Cognitive task analysis (CTA) is enjoying growing popularity in both research and practice as a foundational element of instructional design. However, there exists relatively little research exploring its value as a foundation for training through controlled studies. Furthermore, highly individualized approaches to conducting CTA do not permit broadly generalizable conclusions to be drawn from the findings of individual studies. Thus, examining the magnitude of observed effects across studies from various domains and CTA practitioners is essential for assessing replicable effects. This study reports the findings from a meta-analysis that examines the overall effectiveness of CTA across practitioners and settings in relation to other means for identifying and representing instructional content. Overall, the effect of CTA-based instruction is large (Hedges's $g = 0.871$). However, effect sizes vary substantially by both CTA method used and training context. Though limited by a relatively small number of studies, the notable effect size indicates that the information elicited through CTA provides a strong basis for highly effective instruction.

Keywords: cognitive task analysis, methods, training, cognitive engineering

INTRODUCTION

Cognitive task analysis (CTA) is a core cluster of cognitive engineering tools frequently applied to understand work processes, inform the design of decision support systems, and develop ergonomically sound tools to effectively support human performance (Woods & Roth, 1988). CTA techniques elicit from experts the knowledge and processes they use to perform complex tasks and analyze them to derive representations that can be used for various purposes (Crandall, Klein, & Hoffman, 2006). Increasingly, these tools are also used to contribute to the design of training and instruction by providing detailed information to learners about how to perform target tasks at a high level of proficiency. These approaches often provide measurably greater quantities of useful information about the effective execution of tasks than other methods of identifying information, such as the observation of task performance alone and self-generated explanations provided by subject matter experts. Empirical assessments suggest that CTA contributes between 12% (Chao & Salvendy, 1994) and 43% (Clark & Estes, 1996; Crandall & Getchell-Reiter, 1993) more information for documenting performance-relevant processes than approaches that are not CTA based.

Task analysis in general has been an important part of the instructional systems design process since the 1980s (Reigeluth, 1983). However, Jonassen, Tessmer, and Hannum (1999, p. 5) note,

> The value accorded to task analysis is often low. Even when designers are skilled in performing task analysis, time constraints prevent them from undertaking any kind of analysis. Project managers do not perceive the need or importance of adequately articulating tasks, preferring to begin development in order to make the process more efficient.

Many approaches to task analysis differ from CTA, which emphasizes the cognitive components of effective task performance. Typically, analyses are conducted with domain experts to identify those strategies, decisions, and procedures that are highly effective for performing target tasks in authentic contexts and can provide an appropriate foundation for the design of

Address correspondence to David F. Feldon, University of Virginia, Curry School of Education, P.O. Box 400273, 206B Bavaro Hall, Charlottesville, VA 22904-4261, USA, dff2j@virginia.edu.

Journal of Cognitive Engineering and Decision Making
Vol. 7, No. 3, September 2013, pp. 293–304
DOI:10.1177/1555343412474821

instruction (Guimond, Sole, & Salas, 2012). However, CTA is more expensive than other methods of obtaining instructional content, because the identification of cognitive processes can be time-consuming, be labor-intensive, and require the participation of experts, which precludes the performance of their usual responsibilities (Clark & Estes, 1996; Clark, Feldon, van Merriënboer, Yates, & Early, 2008).

Consequently, other approaches to task analysis are more commonly used for instructional development, including job analysis, subject matter analysis, and anthropological analyses (Jonassen et al., 1999). Job analysis utilizes behavioral task analyses that emphasize directly observable activities rather than their underlying cognitive precursors. For example, determined using behavioral task analyses, the job specifications of a computer programmer might include the preparation of flowcharts to illustrate sequences of operations within a piece of software, but they would not specify any of the complex cognitive tasks that are necessary to develop the programming structures that the flowcharts represent (Clark & Estes, 1996; Cooke, 1992). Subject matter analysis emphasizes the examination of the structural nature of knowledge and the ways in which relevant concepts are related to one another (e.g., the hierarchical relationships among tasks or categories). Other types of task analysis include anthropological methods that emphasize the situated nature of task performance as part of cultural and social human activities (Jonassen et al., 1999). For instructional purposes, these forms of task analysis commonly utilize direct observations, content analysis of existing documents (e.g., manuals, policy handbooks), and interviews or focus groups with experts (Loughner & Moller, 1998).

STUDY PURPOSE

Despite the growing popularity of CTA, there exists relatively little research that quantifies its value for instructional design compared to other approaches used to create training in terms of stronger posttraining performance. Furthermore, practitioners of CTA may employ idiosyncratic combinations of CTA methods, which limit the extent to which the findings from individual studies might inform expectations for effects in other projects (Yates & Feldon, 2011). Thus, examining the magnitude of observed training benefits across studies from various domains and CTA practitioners is essential for determining a more generalizable estimate of its value as part of the instructional design process.

Meta-analysis provides the ability to combine the findings of multiple, independent studies to assess aggregate effects of an independent variable (CTA-based elicitation of instructional content, in this case). Results from individual studies are converted into standardized units (i.e., effect sizes) that can be pooled to compute both descriptive and inferential statistics (Lipsey & Wilson, 2000). In this way, the range, average magnitude, and variance in outcomes associated with CTA-based instruction can be determined. Furthermore, the effects of differences in study design or implementation can be tested to better understand which factors influence the variable's effectiveness (Cooper, Hedges, & Valentine, 2008). For these reasons, a meta-analysis of CTA-based training studies is useful in drawing more generalizable conclusions about the value of CTA as a component of training design to enhance human performance. In addition to identifying an aggregate magnitude of effect size, this study also disaggregates effects to reflect differences in CTA technique, training setting, and types of training outcomes assessed.

RESEARCH QUESTIONS

The research literature on CTA-based instruction demonstrates much promise for its effectiveness as an approach to capturing knowledge for use during instructional development (Clark et al., 2008). However, the lack of standardization across CTA methods and individual studies leaves several broad questions unanswered. First, the aggregate effect of CTA on learning outcomes is not known, so it is difficult to determine if outcomes from a specific study are typical of the results that might be expected. Second, the variation in CTA methods may lead to differing levels of effectiveness for the resulting instruction. Third, the use of different outcome measures across domains leaves open the possibility that certain types of learning outcomes may be affected differently by CTA-based

instruction. Therefore, this study addresses the following research questions (RQs):

RQ1: What is the overall level of effectiveness of using CTA as the basis for instructional content compared to other approaches?
RQ2: Does the use of different CTA methods lead to differing magnitudes of effect on learning outcomes?
RQ3: Does training delivered in different contexts lead to different magnitudes of effect for CTA-based training?
RQ4: Does the magnitude of effect for CTA-based instruction vary as a function of the type of learning outcome measured?

REVIEW OF THE LITERATURE

Although there are many different models used to guide the design and development of training, the generic form typically entails the following phases in sequence: analysis, design, development, implementation, and evaluation (ADDIE). The sequence is so ubiquitous that the ADDIE acronym is "virtually synonymous with instructional systems development" (Malenda, 2003, p. 34). The first phase, analysis, identifies and characterizes (a) the instructional goals (i.e., what learners will be able to do after receiving the developed training), (b) the knowledge and skills necessary to be imparted for learners to meet the instructional goals, (c) the parameters of the contexts in which the new skills must be utilized, and (d) the capabilities and existing knowledge of the people to be trained (Dick, Carey, & Carey, 2005).

CTA can be used to conduct the first three types of analysis. Typically its implementation follows the following five steps (Clark et al., 2008, p. 580):

1. Collect preliminary knowledge
2. Identify knowledge representations
3. Apply focused knowledge elicitation methods
4. Analyze and verify data acquired
5. Format results for intended application

However, within this sequence, a variety of approaches can be utilized, as described in the following section.

Types of Cognitive Task Analysis

There are three main categories of CTA as defined by Cooke (1994) and one additional category established by Wei and Salvendy (2004). These categories are (a) "observation and interview," (b) "process tracing," (c) "conceptual techniques," and (d) "formal models."

Observation and interview techniques tend to be informal in nature, thus providing analysts high adaptability in gathering and analyzing data (Cooke, 1994). However, there is some variation in the application of these techniques, such that some may use highly structured protocols and others might be more open-ended. Process tracing methods capture expertise during task performance behavior within an actual problem-solving context. They use real tasks as a means to explicate the path taken by experts when completing a procedure. In addition to fine-grained, frequently instrumented observations of task performance, these methods may also include various types of think-aloud protocols. Conceptual techniques are those CTA methods that attempt to identify hierarchical relationships among knowledge relevant to task performance within a domain (e.g., card sorting or concept mapping tasks). Formal models of CTA are computational models that generate simulated instances of targeted tasks. The simulated performance is then compared to human (expert) performance to assess the completeness of the model (Wei & Salvendy, 2004).

Although many CTA efforts incorporate multiple tools from one or more of these categories, certain named approaches are common in the CTA literature (Yates & Feldon, 2011). Two of the more frequently cited are the critical decision method (CDM; Klein, Calderwood, & MacGregor, 1989) and PARI (precursor, action, result, interpretation; Hall, Gott, & Pokorny, 1995). These approaches employ semistructured interview techniques to focus experts' recall on specific facets of their relevant knowledge. CDM elicits information about the relevant cues and strategies used by an expert in a specific problem-solving instance that was atypical or highly challenging. In contrast, PARI's protocol focuses on the identification of the elements that compose its acronym for tasks as typically performed.

Expert Cognition and Self-Report Accuracy

Experts are important sources of information about how to perform tasks effectively. Thus, they can be invaluable resources for developing instruction to train others to perform similar tasks. Experts have extensive, well-organized knowledge in their domains and excellent recall of the concepts that govern their respective domains (Glaser & Chi, 1988). However, their ability to recall the procedures that they use to perform tasks is less robust (Feldon, 2007).

Experts typically have at least a decade of effortful (deliberate) practice in their fields (Ericsson & Charness, 1994). However, as cognitive skills are practiced, they require decreasing levels of mental effort to execute and regulate (Blessing & Anderson, 1996). As a result of this skill automaticity, experts conserve most of their cognitive resources to accommodate complexities that nonexperts would be unable to navigate successfully. However, as decision making in these situations becomes automatic, it also becomes more difficult to notice and articulate the decision points and strategies used.

Consequently, experts are often unable to share fully what it is they do and how they do it (Blessing & Anderson, 1996; Feldon, 2010). Comparing the explanations provided by experts with direct observations of their performance identifies substantial disconnects between their actions and their descriptions of them. Across multiple studies, the rate of omission is approximately 70% (Clark, 2009). For example, Cooke and Breedin (1994) asked a number of experts in physical mechanics to predict the trajectories of various objects and explain how those estimates were generated. The researchers used the explanations provided to attempt a replication of the predictions made. However, the trajectories computed from the explanations did not correlate to the original trajectory estimates. Similarly, a study of scientific reasoning during laboratory meetings in leading research laboratories found that even when scientific breakthroughs were made during discussions, the scientists participating in those discussions were unable to accurately recall the reasoning processes that led to their insights (Dunbar, 2000).

This phenomenon also surfaces during instruction. For example, Sullivan and colleagues (2007) found that of the 26 identified steps in a surgical procedure taught to medical residents, individual expert physicians articulated only 46% to 61% when teaching the procedure. Similarly, content analysis of instruction to train undergraduate biology students in scientific problem solving based on unguided report by an expert was less specific than the instruction based on CTA in nearly 40% of the content covered by both versions (Feldon & Stowe, 2009).

Accuracy and Instruction

The negative effects of such omissions on instructional effectiveness of results are measurable. Gaps in instructional content require learners to allocate more cognitive resources than otherwise necessary to learn a skill with complete information available to them. This extraneous effort leads to lower recall and poorer task performance (Kirschner, Sweller, & Clark, 2006; van Merriënboer & Sweller, 2005). However, when steps are taken to increase the completeness of instructional materials, student performance increases.

Studies of training in a variety of domains, including radar troubleshooting (Schaafstal & Schraagen, 2000), spreadsheet use (Merrill, 2002), and medicine (Sullivan et al., 2007), reflect significantly better posttraining performance for learners receiving instruction where efforts are made to fully articulate experts' strategies compared to "business as usual" instructional conditions. Other studies have reported higher levels of self-efficacy (Campbell et al., 2011), less time necessary for task performance (Velmahos et al., 2004), and deeper conceptual knowledge related to the task (Schaafstal & Schraagen, 2000).

METHOD

Inclusion and Exclusion Criteria

A comprehensive search of the literature was conducted to ascertain studies appropriate for this meta-analysis. Searches were conducted

using the Boolean search phrase ("cognitive task analysis" or "knowledge elicitation") and (training or instruction). Based on a review of the prior literature in the area, these terms were determined to be the ones that would cast the broadest net for successfully finding studies in which CTA was used as a tool for further training or learning agenda. This search yielded 467 articles from the following databases: PsycINFO, ERIC, Education Research Complete, ProQuest Dissertations and Theses, Medline, PubMed, and ISI Web of Science.

In an attempt to augment the literature attained through database search, several individuals recognized as major contributors to the development and study of CTA were contacted to request any additional studies or technical reports that had not been published. These individuals were Richard Clark, Maura Sullivan, Robert Hoffman, Jan Schraagen, Beth Crandall, Roberta Calderwood, Robert Pokorny, and Gary Klein. These snowball techniques yielded two additional unpublished dissertations, four technical reports, one conference paper, and one additional peer-reviewed article that were appropriate but not obtained through the database search.

The selection of articles was winnowed by selecting criteria that would allow for the most direct analysis of studies related to the current study's RQs. In selecting studies for this meta-analysis, three excluding criteria were used. First, studies were excluded if they were theoretical articles or literature reviews that lacked empirical data. Because of the questions being asked in this study, theoretical pieces, although beneficial for understanding the depth and breadth of thinking on CTA, were not going to possess the data necessary to make the methodological and instructional comparisons sought in this study. These articles (n = 301) were removed from consideration of the overall literature search.

The remaining literature (n = 166) was reviewed and evaluated for appropriate inclusion based on their robustness to other exclusion criteria. The second factor for exclusion considered for the remaining studies was the type of research study. Because of the nature of qualitative research and the descriptive questions asked therein, case studies and pieces focusing on CTA process or technique explication were also excluded. Although explications of the CTA methods would provide insight into the differences between those methods, their data would not provide the quantifiable information necessary to determine effect size differences. Specifically, this left studies reporting quantitative measures of training outcomes.

We also excluded studies that did not use a comparison in the form of a control group or historical baseline established using non-CTA instruction. Studies containing a control or comparison group allow for the most robust examination of the effects of CTA compared to other methods of instruction development. This factor left a small corpus of literature (n = 20) yielding 56 comparisons on specific variables that could directly examine use of CTA as an instructionally effective practice.

Coding

All articles included in this meta-analysis were reviewed by both authors. All areas of disagreement were discussed until consensus was reached. Intercoder agreement prior to discussion was established at 92%. Studies were coded for each specifically measurable outcome: tests of conceptual knowledge (declarative), task performance (procedural), self-efficacy, and time required for task completion. The specific measures employed varied between studies, as the content on which participants trained differed. However, these categories of assessment are typical classes of training outcomes (Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997). Tests of conceptual knowledge were traditional tests asking multiple-choice or free-response questions about the principles or concepts relevant to performance of the target task. Task performance measures included checklist-based scores of live performance, counts of errors made, or attempts required to complete the target task. Self-efficacy measures utilized surveys of learners in which they were asked to estimate their confidence in their ability to perform the task in the future using a numeric scale (e.g., 1–5, with 1 representing *very low confidence* and 5 representing *very high confidence*). Time necessary

for task completion following training was measured in hours, minutes, or seconds, depending on the nature of the target task.

Each comparison made between CTA training outcomes and non-CTA training outcomes was treated as an individual case. Some studies reported multiple outcomes, resulting in more than one case from a single report represented in the meta-analysis.

To maintain consistency in the treatment of effect size data across cases and studies, two data transformations were applied when appropriate. The first involved data reflecting counts of errors, problem-solving attempts, and performance duration (time). These were coded negatively (i.e., $-1 \times$ attained effect sizes) to maintain consistency with the rest of the data entered into analysis for which larger effect sizes reflect better performance. This transformation was applied because more time necessary to perform a task and higher numbers of errors/attempts are negative indicators of training outcomes. In the second transformation, a study reported only gains from pretest to posttest without reporting the scores themselves; since equivalence was established for participants across study conditions prior to training, pretests were recorded as scores of zero and posttest scores were recorded as equal to the reported gains to maintain consistency with the rest of the data points reported.

When identifying effect sizes from studies, several standards were applied to the selection of appropriate comparisons and computations. The overarching principle used in these instances was to adopt the most conservative approach. For example, some studies reported multiple delayed posttests. Reporting every comparison between the pretest and each posttest would have severely overrepresented certain studies in the sample, so only the longest delay comparison was used. It was anticipated that these gains would be smaller than immediate posttest results but more durable. Furthermore, if exact p values were not reported, the value used for p in computations of effect size was the identified critical value reported. As such, significance reported as $p < .05$ was recorded as $p = .05$. Thus, effect sizes reported here may slightly underestimate effect sizes actually obtained.

Analyses

In examining the CTA literature, several challenges presented themselves. Most prominent among them was the highly divergent and relatively sparse nature of the literature. CTA began being used for training only in the mid-1980s (Glaser et al., 1985). Because of this, there are few empirical articles appropriate for analysis. Many of the studies found during the literature search were noted later to not provide all of the information necessary for computing effect sizes. This created relatively small cell sizes for more detailed examinations. Effect sizes were initially computed using Cohen's d since the RQs being asked were trying to determine the effect of CTA generally (RQ1) and types of CTA specifically (RQ2). However, further examination led to the conclusion that the relatively small sample sizes within the studies analyzed might be skewing the d values (Hedges & Olkin, 1985). Therefore, Cohen's d values were transformed into Hedges's g, which is an effect size measure that accounts for the inflation of effect size inversely related to sample size. According to Cohen's (1988) guidelines for interpreting measures of effect size, a Hedges's g value of 0.2 is considered to be small, 0.5 is a medium effect, and 0.8 or greater is large.

For each of the RQs discussed in this paper, separate analyses were computed. Because the data were not uniform in reporting standards, several of the ANOVAs were run on less than the full data set. For some analyses, the sample did not pass tests for homogeneity of variance and normality. However, one-way fixed effects ANOVAs are robust to Type I errors under conditions of nonnormality (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992), so they did not present limitations for the analyses.

RESULTS

RQ1: What is the overall level of effectiveness of using CTA as the basis for instructional content compared to other approaches? Analyses compare the effects of CTA-based instruction to the effects of instruction developed using content *derived through means other than CTA* (e.g., behavioral task analysis, unguided expert

self-report). All studies using a non-CTA control group and a CTA-based instruction treatment group were used to compute a mean effect size for the overall treatment effects of CTA-based instruction. Across the 56 comparisons identified, the overall value of Hedges's g is 0.871 (SD = 0.846), as presented in Table 1.

RQ2: Does the use of different CTA methods lead to differing magnitudes of effect on learning outcomes? Unfortunately, the majority of studies within the literature reviewed failed to report the type of CTA used in the development of training materials, making a meaningful comparison of CTA methods impossible. In fact, of the more than 100 types of CTA discussed by Cooke (1999), only 2 varieties of CTA were reported within the empirical findings of these studies. Of the 56 comparisons identified within this meta-analysis, only 15 were associated with a reported CTA method.

With that caveat in mind, a one-way fixed effects ANOVA yields significant differences between CTA method types, $F(2, 53) = 6.566$, $p = .003$ (see Table 2). The mean effect sizes of the reported methods are $g = 0.329$ ($SD = 0.308$) for CDM and $g = 1.598$ ($SD = 0.993$) for the PARI method. The mean effect size for studies with unidentified CTA methods is $g = 0.729$ ($SD = 0.731$). Tukey's honestly significant difference (HSD) multiple comparison follow-up test produced a statistically significant difference between the CDM method and the PARI method of CTA-based training ($p = .018$). CDM and other CTA methods did not differ significantly ($p = .586$). PARI method outcomes differed significantly from other CTA (non-PARI) techniques ($p = .005$).

RQ3: Does training delivered in different contexts lead to different magnitudes of effect for CTA-based training? Five distinct settings are identified in the included studies as environments where CTA-based training has been implemented and reported. Specifically, these are military, government (nonmilitary), academic, medical, and private industry. Of the 56 comparisons of CTA-based training, almost half are conducted within medical settings ($n = 25$). A Shapiro–Wilk test indicates two of the five groups do not meet the threshold value for normality ($p > .05/5$). A one-way fixed effects ANOVA indicates that study setting significantly affects training outcomes, $F(4, 51) = 4.257$, $p = .005$. Post hoc Tukey's HSD tests reveal significant differences between the military setting and the government (nonmilitary) setting ($p = .004$), and approach significance between the military and the medical setting ($p = .056$). The mean effects of all settings are reported in Table 3.

RQ4: Does the magnitude of effect for CTA-based instruction vary as a function of the type of learning outcome measured? Four specific types of outcome measures are used in assessing CTA-based training outcomes: procedural knowledge gains, declarative knowledge gains, self-efficacy, and performance speed. Mean Hedges's g effect size scores for each measure type are $g = 0.872$ for procedural knowledge, $g = 0.926$ for declarative knowledge, $g = 0.893$ for self-efficacy, and $g = 0.611$ for performance speed (time). A one-way fixed effects analysis of variance does not indicate a significant difference between the outcome measures, $F(3, 52) = 0.072$, $p = .975$. Follow-up pairwise comparisons among the methods using Tukey's test confirm a lack of significant differences between groups (see Table 4).

DISCUSSION

Much of the CTA literature advocates its use for effectively eliciting a more complete set of procedural directions from which better instruction can be derived. In fact, a prior, unpublished meta-analysis reported effect size gains triple those of non-CTA-based training (Lee, 2005). Although the effect sizes seen in this analysis are not as large, the gains measured here do robustly support the claim that CTA-based training is more effective than training not based on CTA. Furthermore, by correcting for the inflating effects of small sample size on effect size estimates and including a larger sample of studies, the large effect size obtained supports the assertion that CTA-based training yields highly effective results even under conservative analyses.

Analyses determined that the PARI method yields the largest effects. However, many

TABLE 1: Effect Sizes of All Identified Comparisons in Included Studies of Cognitive Task Analysis–Based Instruction

Article	Knowledge Type	Training Setting	Cohen's *d*	Hedges's *g*	*N*
Bathalon et al. (2004)	Procedural	Medical	1.197	1.164	29
Biederman and Shiffrar (1987)	Procedural	Academic	2.870	2.806	36
Campbell (2010)	Procedural	Medical	0.835	0.815	33
Campbell (2010)	Procedural	Medical	0.916	0.894	33
Crandall and Calderwood (1989)	Procedural	Medical	3.330	2.664	6
DaRosa et al. (2008)	Declarative	Medical	0.697	0.686	48
DaRosa et al. (2008)	Procedural	Medical	0.032	0.031	48
DaRosa et al. (2008)	Procedural	Medical	0.491	0.483	48
DaRosa et al. (2008)	Procedural	Medical	0.118	0.116	48
Feldon et al. (2009)	Procedural	Government	–0.292	–0.286	41
Feldon et al. (2009)	Procedural	Government	–0.142	–0.139	41
Feldon et al. (2009)	Procedural	Government	–0.038	–0.037	41
Feldon et al. (2009)	Procedural	Government	–0.225	–0.221	41
Feldon et al. (2010)	Procedural	Academic	0.270	0.269	298
Feldon et al. (2010)	Procedural	Academic	0.310	0.309	298
Feldon et al. (2010)	Procedural	Academic	0.270	0.269	298
Feldon et al. (2010)	Procedural	Academic	0.230	0.229	298
Gott (1998)	Procedural	Military	1.286	1.262	41
Gott (1998)	Procedural	Military	1.170	1.147	41
Gott (1998)	Declarative	Military	0.870	0.835	41
Gott (1998)	Procedural	Military	0.960	0.941	41
Gott (1998)	Declarative	Military	0.760	0.745	41
Green (2008)	Declarative	Academic	0.091	0.052	4
Hall et al. (1995)	Procedural	Military	1.084	1.056	32
Lajoie (2003)	Procedural	Academic	0.757	0.725	20
Merrill (2002)	Procedural	Industry	0.685	0.680	98
Merrill (2002)	Time	Industry	0.685	0.680	98
Park et al. (2010)	Procedural	Medical	0.909	0.872	21
Park et al. (2010)	Procedural	Medical	0.909	0.872	21
Roth et al. (2001)	Procedural	Industry	1.372	1.267	12
Roth et al. (2001)	Procedural	Industry	1.472	1.359	12
Roth et al. (2001)	Procedural	Industry	1.330	1.227	12
Roth et al. (2001)	Procedural	Industry	1.258	1.161	12
Schaafstal and Schraagen (2000)	Procedural	Military	2.231	2.142	21
Schaafstal and Schraagen (2000)	Procedural	Military	2.461	2.362	21
Schaafstal and Schraagen (2000)	Procedural	Military	3.848	3.694	21
Schaafstal and Schraagen (2000)	Declarative	Military	2.865	2.750	21
Schaafstal and Schraagen (2000)	Declarative	Military	0.671	0.644	21
Staszewski and Davison (2000)	Procedural	Military	0.888	0.854	22
Staszewski and Davison (2000)	Procedural	Military	0.888	0.854	22
Staszewski and Davison (2000)	Procedural	Military	0.888	0.854	22

(continued)

TABLE 1: (continued)

Article	Knowledge Type	Training Setting	Cohen's *d*	Hedges's *g*	*N*
Sullivan et al. (2007)	Procedural	Medical	1.476	1.413	20
Sullivan et al. (2007)	Declarative	Medical	1.617	1.548	20
Tirapelle (2010)	Self-efficacy	Medical	0.852	0.831	33
Tirapelle (2010)	Declarative	Medical	0.208	0.203	33
van Herzeele et al. (2008)	Procedural	Medical	–0.908	–0.869	29
van Herzeele et al. (2008)	Procedural	Medical	–0.103	–0.099	29
van Herzeele et al. (2008)	Procedural	Medical	1.398	1.339	29
van Herzeele et al. (2008)	Procedural	Medical	1.314	1.259	29
van Herzeele et al. (2008)	Procedural	Medical	0.076	0.073	29
van Herzeele et al. (2008)	Procedural	Medical	–0.282	–0.270	29
Velmahos et al. (2004)	Declarative	Medical	0.903	0.875	26
Velmahos et al. (2004)	Procedural	Medical	1.458	1.412	26
Velmahos et al. (2004)	Procedural	Medical	0.904	0.875	26
Velmahos et al. (2004)	Procedural	Medical	0.599	0.580	26
Velmahos et al. (2004)	Time	Medical	0.560	0.542	26

TABLE 2: Mean Effect Sizes and Standard Deviations for CTA-Based Instruction by Type of CTA Used

CTA Method	Number of Cases (*k* = 56)	Mean Effect Size	*SD*
CDM	4	0.329	0.308
PARI	11	1.598	0.993
Other (unreported)	41	0.729	0.731

Note. CDM = critical decision method; CTA = cognitive task analysis; PARI = precursor, action, result, interpretation.

TABLE 3: Mean Effect Sizes and Standard Deviations for Cognitive Task Analysis–Based Instruction by Instructional Setting

Setting	Number of Cases (*k* = 56)	Mean Effect Size	*SD*
Military	14	1.439	0.926
Government (nonmilitary)	4	–0.171	0.107
Academic/ university	7	0.666	0.966
Medical	25	0.732	0.714
Industry	6	1.062	0.303

studies did not identify which type of CTA method yielded the most significant results. In fact, some methods lack formal method names. As one researcher stated regarding his approach, "I've not named it. I should have given it a name years ago" (R. E. Clark, personal communication, February 10, 2011). Furthermore, as discussed extensively by Yates and Feldon (2011), named CTA methods may mask similar techniques used under different approaches or application of different techniques that use the same name. Future work in the area would benefit greatly from significantly more detailed reporting regarding the CTA including the type of CTA performed, its duration, and the role of subject matter experts in the training and materials development.

One of the most notable findings is the observed influence of studies' settings on the effectiveness of CTA-based training. Studies in

TABLE 4: Mean Effect Sizes and Standard Deviations for Cognitive Task Analysis–Based Instruction by Type of Outcome Measure

Knowledge Measure	Number of Cases (k = 56)	Mean Effect Size	*SD*
Declarative	9	0.926	0.805
Procedural	44	0.830	0.855
Self-efficacy	1	0.883	—
Time	2	0.611	0.097

the military setting significantly outperform those in all other categories except university settings for reasons that are not immediately evident. Further exploration of approaches to design, implementation, and characteristics of trainees is warranted.

LIMITATIONS

There are three substantial limitations to this study. First, there is only limited experimental and quasi-experimental research within the CTA literature. This causes overall effects detected to be less stable than might otherwise be seen within a larger corpus of literature. Second, most studies did not report the reliability coefficients associated with the measures used to assess training outcomes. This could be because of the use of CTA-based instruction in highly specialized domains for which previously validated assessments may not exist. However, it prevented consideration of measurement error as an indicator of study quality and when computing meta-analytic effect sizes. Third, divergent reporting practices across the many disciplines using CTA-based training prevented coding of all variables from some studies. Many of the details identified to analyze study findings were obtained through personal emails with authors and research assistants. Compilation of data in this manner, although replicable, is difficult.

CONCLUSIONS

Now, 25 years after the first training-based study of CTA effectiveness was published, strong evidence exists regarding the effectiveness of CTA-based training. Despite its high costs relative to other methods used during the instructional design process (Clark & Estes, 1996; Clark et al., 2008), the large effects it demonstrates on learning outcomes suggest that it offers great value to organizations with human performance needs. Industrial and military training outcomes included in this study reported mean effect sizes greater than 1.0 (very large), and mean effects of medical training and academic instruction were also medium to large. The success found across these diverse settings indicates that the benefits of CTA are broadly applicable and can enhance the quality of instruction in contexts critical to economic growth and human health.

Future research in this area would be strengthened by significant changes in the approach to reporting data. Specifically, more information regarding the method of CTA completed may yield significant results regarding the effectiveness of specific CTA methods. In addition, there are strong indications that various groups within various settings respond differently to CTA-based training. With additional study features reported, more nuanced understandings regarding specific effects may be identified.

REFERENCES

References marked with an asterisk indicate studies included in the meta-analysis.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology*, *50*, 341–358.

*Bathalon, S., Martin, M., & Dorion, D. (2004). Cognitive task analysis, kinesiology and mental imagery: Challenging surgical attrition. *Journal of the American College of Surgeons*, *199*(3), 73.

*Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 640–645.

Blessing, S. B., & Anderson, J. R. (1996). How people learn to skip steps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 576–598.

*Campbell, J. C. (2010). *Employing cognitive task analysis supported instruction to increase medical student and surgical resident performance and self-efficacy* (Unpublished doctoral dissertation). University of Southern California, Los Angeles.

Campbell, J. C., Tirapelle, L., Yates, K., Clark, R., Inaba, K., . . . Sullivan, M. (2011). The effectiveness of a cognitive task analysis informed curriculum to increase self-efficacy and improve performance for an open cricothyrotomy. *Journal of Surgical Education*, *68*, 403–407.

Chao, C.-J., & Salvendy, G. (1994). Percentage of procedural knowledge acquired as a function of the number of experts from whom knowledge is acquired for diagnosis, debugging, and interpretation tasks. *International Journal of Human-Computer Interaction*, *6*(3), 221–233.

Clark, R. E. (2009). How much and what type of guidance is optimal for learning from instruction? In S. Tobias & T. M. Duffy (Eds.), *Constructivist theory applied to instruction: Success or failure?* (pp. 158–183). New York, NY: Routledge, Taylor and Francis.

Clark, R. E., & Estes, F. (1996). Cognitive task analysis. *International Journal of Educational Research*, *25*(2), 403–417.

Clark, R. E., Feldon, D., van Merriënboer, J. J. G., Yates, K., & Early, S. (2008). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 577–593). New York, NY: Macmillan/Gale.

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

Cooke, N. J. (1992). Modeling human expertise in expert systems. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 29–60). Mahwah, NJ: Lawrence Erlbaum.

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies*, *41*, 801–849.

Cooke, N. J. (1999). Knowledge elicitation. In F. T. Durso (Ed.), *Handbook of applied cognition* (pp. 479–509). New York, NY: John Wiley.

Cooke, N. J., & Breedin, S. D. (1994). Constructing naive theories of motion on the-fly. *Memory and Cognition*, *22*, 474–493.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2008). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage.

*Crandall, B., & Calderwood, R. (1989). *Clinical assessment skills of experienced neonatal intensive care nurses* (Report Contract 1-R43-NR01911-01). Fairborn, OH: Klein Associates.

Crandall, B., & Getchell-Reiter, K. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the intuition of NICU nurses. *Advances in Nursing Science*, *16*(1), 42–51.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. Cambridge, MA: MIT Press.

*DaRosa, D., Rogers, D. A., Williams, R. G., Hauge, L. S., Sherman, H., Murayama, K., . . . Dunnington, G. L. (2008). Impact of a structured skills laboratory curriculum on surgery residents' intraoperative decision-making and technical skills. *Academic Medicine*, *83*(10), S68–S71.

Dick, W., Carey, L., & Carey, J. O. (2005). *The systematic design of instruction* (6th ed.). New York, NY: Pearson.

Dunbar, K. (2000). How scientists think in the real world: Implications for science education. *Journal of Applied Developmental Psychology*, *21*(1), 49–58.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725–747.

Feldon, D. F. (2007). The implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review*, *19*(2), 91–110.

Feldon, D. F. (2010). Do psychology researchers tell it like it is? A microgenetic analysis of research strategies and self-report accuracy along a continuum of expertise. *Instructional Science*, *38*(4), 395–415.

*Feldon, D. F., Hurst, M. H., Chao, J., & Jones, J. (2009). *Evaluation of the CTA-based SNAP training program*. Project report generated for the South Carolina Department of Social Services by the Center for Child and Family Studies.

Feldon, D. F., & Stowe, K. (2009). A case study of instruction from experts: Why does cognitive task analysis make a difference? *Technology, Instruction, Cognition, & Learning*, *7*, 103–120.

*Feldon, D. F., Timmerman, B. E., Stowe, K., & Showman, R. (2010). Translating expertise into effective instruction: The impacts of cognitive task analysis (CTA) on lab report quality and student retention in the biological sciences. *Journal of Research on Science Teaching*, *47*(10), 1165–1185.

Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Mahwah, NJ: Lawrence Erlbaum.

Glaser, R., Lesgold, A., Lajoie, S., Eastman, R., Greenberg, L., Logan, D., . . . Yengo, L. (1985). *Cognitive task analysis to enhance technical skills training and assessment* (Final Report, Contract No. F41689-8v3-C-0029). Brooks Air Force Base, TX: U.S. Air Force Armstrong Laboratory.

Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research*, *42*, 237–288.

*Gott, S. (1998). *Rediscovering learning: Acquiring expertise in real world problem solving tasks* (Rep. No. AL/HR-TR-1997-0009). Brooks Air Force Base, TX: U.S. Air Force Armstrong Laboratory.

*Green, R. S. (2008). *Cognitive task analyses for life science automation training program design* (Unpublished doctoral dissertation). North Carolina State University, Raleigh.

Guimond, M. E., Sole, M. L., & Salas, E. (2012). Getting ready for simulation-based training: A checklist for nurse educators. *Nursing Education*, *32*(3), 179–185.

*Hall, E. P., Gott, S. P., & Pokorny, R. A. (1995). *A procedural guide to cognitive task analysis: The PARI methodology* (Rep. No. AL/HR-TR-1995-0108). Brooks Air Force Base, TX: U.S. Air Force Armstrong Laboratory.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics*, *17*(4), 315–339.

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

Jonassen, D. H., Tessmer, M., & Hannum, W. H. (1999). *Task analysis methods for instructional design*. Mahwah, NJ: Lawrence Erlbaum.

Kirschner, P., Sweller, J., & Clark, R. E. (2006). Why minimally guided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. *Educational Psychologist*, *41*(2), 75–86.

*Lajoie, S. P. (2003). Individual differences in spatial ability: Developing technologies to increase strategy awareness and skills. *Educational Psychologist*, *38*(2), 115–125.

Lee, R. (2005). *The impact of cognitive task analysis on performance: A meta-analysis of comparative studies* (Unpublished doctoral dissertation). University of Southern California, Los Angeles.

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Loughner, P., & Moller, L. (1998). The use of task analysis procedures by instructional designers. *Performance Improvement Quarterly*, *11*(3), 79–101.

*Luker, K. R., Sullivan, M. E., Peyre, S. E., Sherman, R., & Grunwald, T. (2008). The use of a cognitive task analysis–based multimedia program to teach surgical decision making in flexor tendon repair. *American Journal of Surgery*, *195*(1), 11–15.

Malenda, M. (2003). In search of the elusive ADDIE model. *Performance Improvement*, *42*(5), 34–36.

*Merrill, M. D. (2002). A pebble-in-the-pond model for instructional design. *Performance Improvement*, *41*(7), 39–44.

*Park, C. S., Rochlen, L. R., Yaghmour, E., Higgins, N., Bauchat, J. R., Wojciechowski, K. G., . . . McCarthy, R. J. (2010). Acquisition of critical intraoperative event management skills in novice anesthesiology residents by using high-fidelity simulation-based training. *Anesthesiology*, *112*(1), 202–211.

Reigeluth, C. M. (1983). Current trends in task analysis: The integration of task analysis and instructional design. *Journal of Instructional Development*, *6*(4), 24–30.

*Roth, E. M., Lin, L., Kerch, S., Kenney, S. J., & Sugibayashi, N. (2001). Designing a first-of-a-kind group view display for team decision making: A case study. In E. Salas & G. Klein (Eds.), *Linking expertise and naturalistic decision making* (pp. 113–134). Mahwah, NJ: Lawrence Erlbaum.

*Schaafstal, A., & Schraagen, J. M. (2000). Training of troubleshooting: A structured, task analytical approach. In J. Schraagen, S. Chipman, & V. Shalin (Eds.), *Cognitive task analysis* (pp. 57–70). Mahwah, NJ: Lawrence Erlbaum.

*Staszewski, J., & Davison, A. (2000). Mine detection training based on expert skill. In A. C. Dubey, J. F. Harvey, J. T. Broach, & R. E. Dugan (Eds.), *Detection and remediation technologies for mines and mine-like targets V: Proceedings of Society of Photo-Optical Instrumentation Engineers 14th Annual Meeting* (Vol. 4038, pp. 90–101). Bellingham, WA: Society of Photo-Optical Instrumentation Engineers.

*Sullivan, M. E., Brown, C. V. R., Peyre, S. E., Salim, A., Martin, M., Towfigh, S., & Grunwald, T. (2007). The use of cognitive task analysis to improve the learning of percutaneous tracheostomy placement. *American Journal of Surgery*, *193*(1), 96–99.

*Tirapelle, L. A. (2010). *The effect of cognitive task analysis based instruction on surgical skills expertise and performance*. Los Angeles: University of Southern California.

*van Herzeele, I., Aggarwal, R., Neequaye, S., Darzi, A., Vermassen, F., & Cheshire, N. J. (2008). Cognitive training improves clinically relevant outcomes during simulated endovascular procedures. *Journal of Vascular Surgery*, *48*(5), 1223–1230.

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*, 147–177.

*Velmahos, G. C., Toutouzas, K. G., Sillin, L. F., Chan, L., Clark, R. E., Theodorou, D., & Maupin, F. (2004). Cognitive task analysis for teaching technical skills in an inanimate surgical skills laboratory. *American Journal of Surgery*, *187*(1), 114–119.

Wei, J., & Salvendy, G. (2004). The cognitive task analysis methods for job and task design: Review and reappraisal. *Behaviour & Information Technology*, *23*(4), 273–299.

Woods, D. D., & Roth, E. M. (1988). Cognitive engineering: Human problem solving with tools. *Human Factors*, *30*, 415–430.

Yates, K. A., & Feldon, D. F. (2011). Towards a taxonomy of cognitive task analysis methods for instructional design: Interactions with cognition. *Theoretical Issues in Ergonomics Science*, *12*(6), 472–495.

Colby Tofel-Grehl is a doctoral candidate in STEM (science, technology, engineering, mathematics) education at the University of Virginia's Curry School of Education. Her research examines factors that affect students' engagement with and success in STEM disciplines as a function of instructional practices.

David F. Feldon is associate professor of STEM education and educational psychology at the Curry School of Education and associate director for STEM initiatives at the Center for the Advanced Study of Teaching and Learning in Higher Education (CASTL-HE). His research examines the development of expertise in university and training settings, including the application of cognitive task analysis to instructional design and assessment.