Review of Educational Research
June 2015, Vol. 85, No. 2, pp. 275–314
DOI: 10.3102/0034654314551063

Strategies for Teaching Students to Think Critically: A Meta-Analysis

Philip C. Abrami, Robert M. Bernard, Eugene Borokhovski, David I. Waddington, C. Anne Wade, and Tonje Persson
Concordia University, Canada

Critical thinking (CT) is purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference, as well as explanations of the considerations on which that judgment is based. This article summarizes the available empirical evidence on the impact of instruction on the development and enhancement of critical thinking skills and dispositions and student achievement. The review includes 341 effects sizes drawn from quasi- or true-experimental studies that used standardized measures of CT as outcome variables. The weighted random effects mean effect size (g+) was 0.30 ($p < .001$*). The collection was heterogeneous (*$p < .001$*). Results demonstrate that there are effective strategies for teaching CT skills, both generic and content specific, and CT dispositions, at all educational levels and across all disciplinary areas. Notably, the opportunity for dialogue, the exposure of students to authentic or situated problems and examples, and mentoring had positive effects on CT skills.*

Keywords: critical thinking, instructional practices, learning processes/strategies

Toward the mid- to late 1920s, John Dewey became significantly more pessimistic in his outlook. Discouraged by the intellectual vacuity and corruption of the Harding and Coolidge administrations and by a faith-based free market approach to social and economic problems, Dewey (1925) underlined, again and again, the importance of critique. The final chapter of *Experience and Nature*, which is recognized as one of Dewey's most important philosophical works, is dedicated to an expansive and passionate defense of the power of critique in all aspects of our lives. Intelligence, said Dewey, was "critical method applied to goods of belief, appreciation and conduct, so as to construct free and more secure goods" and was "the stay and support of all reasonable hopes" (p. 437). Critical thinking (henceforth, in this article, abbreviated CT), for Dewey, was something all citizens needed to engage in on a regular basis, and the role of the philosopher

was not to postulate and reify eternal truths but rather to provide systematic critique of the beliefs generated through everyday critical (and not-so-critical) thinking.

Most philosophers within the Western tradition have emphasized critique, but by philosophical standards, at least, the emphasis on the development of the critical faculties of individual citizens is relatively recent and linked to the rise of liberal democracy. The importance of CT is now recognized in many North American schools, given the centrality of CT to conceptions of liberal democracy, especially in its more Jeffersonian iterations. It is not surprising, then, that it is highly valued by numerous educators, parents, and policymakers (e.g., Bloom & Watt, 2003; Jefferson, 1829; Tsui, 2002).

Contemporary research interest in CT dates to the work of Edward Glaser (1941a) whose dissertation was titled *An Experiment in the Development of Critical Thinking*. At the outset of the introduction of the monograph based on his dissertation, Glaser (1941b) wrote,

> One hundred and fifty years of public education in the United States have resulted in a largely literate electorate . . . Our public education has not resulted, however, in the development of a sufficient proportion of citizens who can evaluate critically what they read. (pp. 4–5)

Glaser (1941b) continued by saying that competent citizenship in a democracy calls for a good deal more than the ability to read and write. Among other things, it requires the ability to think critically, and only those who possess the degree of social understanding and critical-mindedness can make intelligent judgments about public issues.

Yet, some 70 years after Glaser's pioneering efforts, we still need to better understand how CT can be supported by instruction and curriculum reforms. This is an important task: In spite of the popularity of CT, there is widespread concern that educational institutions have failed to successfully instill CT skills and dispositions in students (Case & Wright, 1999; Hyslop-Margison, 2003). In a 1997 study of faculty at 38 public and 28 private colleges and universities in California, 89% of the teachers and instructors viewed the teaching of CT as important, yet only 9% felt that they were teaching CT on a regular basis (Paul, Elder, & Bartell, 1997).

This gap between teachers' values and their practices with respect to CT indicates that a synthesis of effective strategies for teaching CT could be useful. This is the task of the comprehensive review of the empirical evidence on CT that is undertaken in this article. However, before the empirical evidence is addressed, it is necessary to briefly touch on a number of important conceptual questions: What is CT? Is CT a generic skill set or does it vary depending on the subject matter context? What are some promising approaches to teaching CT? What have other reviews of CT instruction found? It is the first of these questions that will now be addressed.

Researchers have offered many definitions of CT (Ennis, 1962, 1987; Facione, 1990a, 1990b; Kurfiss, 1988; Lipman, 1991; Paul & Binker, 1990; Scriven & Paul, 1996; Siegel, 1988).

One leading definition of CT, developed by a Delphi consensus panel of 46 experts, including leading scholars such as Ennis, Facione, and Paul, and organized by the American Philosophical Association (APA), recognizes this complexity by offering a broad definition:

> We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. . . . The ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry permit. (Facione, 1990b, p. 2)

For the purposes of this review, we decided that the APA's report (issued by Facione, 1990b) is clear and inclusive enough to serve as a framework for deciding which empirical studies have adequately captured elements of CT. The authors describe six CT skills (including 16 "subskills") and 19 CT dispositions as listed in Appendix A.

Within Appendix A, the skills/dispositions divide is worth a small additional discussion. The Delphi panel maintained that it was possible to possess the cognitive skills necessary to carry out CT but lack the affective dispositions—the general habits and attitudes—to exercise these skills. Consider the example of a person who possessed the cognitive skills associated with CT but who lacked the disposition to learn about or discuss social issues—it would be difficult to call this individual an effective critical thinker. Hence, the panel held that the development of both skills and dispositions was critical in the education of well-rounded critical thinkers (Facione, 1990b).

There are a number of potential objections to the APA's consensus definition, one of which is that it is insufficiently broad (Alston, 2001). For example, Barbara Thayer-Bacon (2000) argues that good CT (she uses the term *constructive thinking*) broadens the definition to include more than just traditional rational skills and dispositions. Pushing beyond the traditional tools of reason, Thayer-Bacon argues convincingly that intuition, emotion, and imagination also play important roles in effective constructive conversations. She further argues that constructive thinking is more of a social practice than a skill or disposition inculcated in individuals. For Thayer-Bacon, good constructive thinking is analogous to a quilting bee in which one has many different tools at one's disposal along with multiple participants who are diverse in their backgrounds and orientations. An even more profound objection to the APA definition is that it is not only insufficiently broad but also

fundamentally inadequate. Two versions of this argument are available in a special issue of *Studies in Philosophy of Education*—one by Gert Biesta and Geert Stams (2001) and another by James Marshall (2001).

Biesta and Stams (2001) note that a specific conception of "criticality" always underlies any CT that is undertaken. Two of the approaches to criticality that they discuss are "critical dogmatism" and Derridean deconstruction. Critical dogmatism conceives of the activity of critique as "the application of a criterion to evaluate a particular state of affairs" (Biesta & Stams, 2001, p. 60). Given that the APA definition suggests critical thinkers should apply a variety of criteria in order to assess propositions, one can locate the definition of CT used in this article firmly within this camp. Biesta and Stams point out that a significant problem for the position of critical dogmatism is that it rests on an unjustifiable belief in the validity of its own method; when one tries to look back behind the criteria that are postulated, there is no further justification of them available that is not circular or that does not imply an infinite regress. Therefore, thinkers who hold this position reflectively actually stop at this point and make a deliberate choice to hold to their criteria "dogmatically" since no further justification is possible.

Biesta and Stams's (2001) central objection to critical dogmatism is its "totalizing" tendency—its pretence to be a foundational point from which all critique can proceed and with which all critique must align. By contrast, their preferred framework of Derridean deconstruction abandons all pretentions of trying to establish a foundation of this type and, in fact, attacks critical dogmatism on this basis, thus serving as a critique of the standard approach to critique. Instead of the traditional approach, deconstruction places a heavy emphasis on what is made invisible and excluded:

> Deconstruction thus tries to open up the system in the name of that which cannot be thought of in terms of the system. . . . Deconstruction is an affirmation of what is wholly other, of what is unforeseeable from the present . . . It is from this concern for what is totally other, a concern to which Derrida sometimes refers as *justice*, that deconstruction derives its right to be critical, its right to deconstruct. (Biesta & Stams, 2001, p. 68)

It is difficult to say exactly how an approach like this would translate into CT teaching practice, but at any rate, a deconstructionist approach would imply a style of CT that would depart dramatically from most of the studies synthesized in this meta-analysis.

Marshall's (2001) account ventures less far afield in its critique of standard approaches to CT but is nonetheless more caustic. Like Biesta and Stams (2001), Marshall (2001) is dubious of the APA definition of CT cited above but for a different set of reasons. Marshall thinks that traditional conceptions of CT view it as a neutral set of conceptual tools that "lack the force of 'critical' to be found in, e.g., the Frankfurt school" (p. 75). Like the critical theorists of this group, Marshall is interested in a more robust style of critique that will enable students to question foundational aspects of the social systems they inhabit.

In the course of providing an articulation of this style of critique, Marshall (2001) traces the history of the concept of the self in Western philosophy from Descartes to Foucault. It is the Foucaultian conception that Marshall finds to be particularly promising. For Foucault, Marshall (2001) maintains, the fundamental question for the self is "how one practices one's freedom" (p. 83). In this regard, Foucault (1984/1998) emphasizes the importance of a kind of self-care, which involves "an exercise of the self on the self by which one attempts to develop and transform oneself" (p. 282). This kind of reflective, critical practice, which also includes care-for-others by extension, allows for the development of positive "practices of freedom" that go beyond simply the negative but important acts of liberating people from oppressive social systems (e.g., oppressive sexuality; Foucault, 1984/1998, p. 283).

These counternarratives to the APA definition of CT are important, and there is no intellectually honest way to wholly reconcile our analysis with them. It is, for example, virtually impossible that the kinds of studies captured by this meta-analysis (or, indeed, any other meta-analysis of CT) are going to correspond to the kind of criticality that Biesta and Stams (2001) and Marshall (2001) would hope to see. They would likely object to the studies' generally short duration, the standardized assessments of CT, and the conventional and narrow definition of CT that is endorsed in most of them. Furthermore, all of the authors, Thayer-Bacon (2000) included, would like to see an approach that puts a greater emphasis on *working with* the student as opposed to *intervening on* him or her, and all of them have at least some sympathy for a position that would teach students to question social systems.

Some of the questions raised by these counternarratives will be further addressed in the "Outstanding Questions" section toward the end of the article, but for now, let us give three responses to these challenges to the APA definition of CT that we have endorsed. First, we are, in fact, sympathetic to some of these objections. We appreciate the inclusiveness of Thayer-Bacon's (2000) conception of constructive thinking, and some of us feel Marshall (2001) may well be right in saying that a deeper and less interventionist approach to the promotion of criticality is desirable. However, as meta-analysts, this question is outside the competence of our method; our task is not to forge new pathways in conceptualizing CT but rather to describe and pool existing findings. The specificity of the APA definition, with its enumeration of specific skills and dispositions, serves well to synthesize the kinds of interventions that are typically used in CT research.

Second, although this is a question that would ultimately need to be resolved empirically, we believe that even if we adopted a broader definition of CT that attempted to account for some of these counternarratives, it would have made little to no difference in terms of the results of the meta-analysis. This is due to the fact that the kinds of new approaches to CT in which Thayer-Bacon (2000), Biesta and Stams (2001), and Marshall (2001) are interested are unlikely to be investigated using the types of quantitative approach that can be synthesized by a meta-analysis. We recognize that this raises legitimate questions about the limitations of the method we have chosen, which we address briefly in the final section.

Third, although there is a sense in which these authors, particularly Biesta and Stams (2001) and Marshall (2001), are working in an alternate paradigm of criticality

vis-à-vis the APA definition, the paradigms are not totally incommensurable. There is some potential common ground between even the narrowest effort to promote the development of CT skills and the kinds of efforts that a scholar who is sympathetic to critical theory (e.g., Marshall) might hope to promote. Let us take Foucault's (1978/1997) article—"What Is Critique?"—as an illustration in this case. Foucault (1978/1997) locates the beginnings of the "critical attitude," which he defines as "the art of not being governed quite so much," in the early efforts to establish the consistency and truth of scripture (p. 45). He also identifies the efforts to question and move beyond justifications rooted merely in authority as an important precursor to the modern critical movement.

Arguably, a similar commonality can be found between the kinds of relatively modest efforts to teach CT skills and dispositions that are synthesized in this meta-analysis and the more ambitious research programs detailed by scholars like Thayer-Bacon, Biesta, and Marshall. A short intervention that endeavours to enhance students' CT skills is not what these philosophers would prefer to see, but it may nonetheless enhance students' openness to "not be governed quite so much," which could be useful in getting students ready to think about the more expansive conceptions of criticality that these scholars wish to investigate further. In other words, the mere fact that we have opted for the APA definition of CT skills and dispositions does not mean that the analysis we offer has nothing to offer for those who would endorse another definition. The kinds of very general techniques—dialogue, posing authentic problems, mentorship—that our meta-analysis indicates are valuable in the development of CT skills and dispositions will likely also have some value for those who have alternate conceptions of CT.

Are CT Skills Generic or Context-Specific?

Settling on a working definition of CT is a good first step, but there are other important questions that remain outstanding even if one is able to convene on the APA definition. For example, in his pioneering research on the development of CT, Glaser (1941a, 1941b) highlighted a difficulty that has remained outstanding ever since: If CT has aspects that are appropriately regarded as skills-based, do these skills represent generic traits and skills or are they context-bound? Glaser (1941b) also found the changes in CT to be somewhat general in character:

> The aspect of critical thinking which appears most susceptible to general improvement is the attitude of being disposed to consider in a thoughtful way the problems and subjects that come within the range of one's experience. An attitude of wanting evidence for beliefs is most subject to general transfer. Development of skills in applying the methods of logical inquiry and reasoning, however, appear to be specifically related to, and in fact limited by, the acquisition of pertinent knowledge and facts concerning the problem of subject matter towards which the thinking is to be directed. (p. 175).

Some 70 years after Glaser, there is little consensus about whether CT is a set of generic skills that apply across subject domains (engineering, arts, science) or depends on the subject domain and context in which it is taught (Ennis, 1989).

Prevailing psychological views tend to favor the generic traits approach; learning to think critically is understood as gaining mastery over a significant series of *discrete skills* or mental operations and dispositions that can be generalized across a variety of contexts. These skills include concepts such as interpreting, predicting, analyzing, and evaluating. The primary appeal of the meta-cognitive skills discourse on CT involves its presumptive transfer between contexts. According to Woolfolk (1998), higher order thinking such as CT requires consciously applying abstract knowledge, heuristics, or procedures learned in one context to some novel circumstance or situation.

Many philosophers like Ennis (1989), Siegel (1988), Govier (1985), and Paul (1985) also view CT skills as rather general (applicable across subject domains) and conjecture (a) that general CT skills might apply to more than one subject area, despite the fact that CT always involves thinking about some specific context; and (b) that the existence of general skills does not imply the nonexistence of context-specific knowledge. The *generalist* view supported by Siegel (1988) contends that the ability to identify informal errors of reasoning (e.g., post hoc and hasty generalization fallacies) is easily transferable between different contexts.

Alternatively, the *specifist* view (perhaps best represented in the position of McPeck, 1981) argues against general CT skills or capacities due to the fact that thinking is always tied to a subject domain. McPeck also argues against the existence of general skill on the grounds that (a) all thinking is thinking about something; (b) general CT ability is not possible because knowledge of a subject is necessary for CT; and (c) CT varies greatly from field to field. McPeck (1981) remarks,

> [I]t makes no sense to talk about critical thinking as a distinct subject and therefore cannot profitably be taught as such. To the extent that critical thinking is not about a specific subject X, it is both conceptually and practically empty." (p. 5)

Strategies for Teaching CT Skills and Dispositions: What Has Previous Research Found?

The generic versus content-specific debate has significant practical implications for education—if CT is generic, then it could be fruitfully taught in specialized courses that focus on CT skills (Royalty, 1995; Sá, Stanovich, & West, 1999), and if it is dependent on subject matter, then it might best be learned by tackling concrete problems in specific disciplines (Halliday, 2000; Smith, 2002). But the generic versus content-specific debate is not the only outstanding debate about *how* CT skills can be taught. Generally speaking, there are a number of ways in which CT skills are usually taught in university and K-12 contexts. In previous analyses of current approaches (e.g., Abrami et al., 2008), Ennis's (1989) CT typology of four courses, *generic, infusion, immersion, and mixed*, was used for classifying and describing various instructional interventions. The typology breaks down as follows: in *generic* courses, CT skills and dispositions are the course objective, with no specific subject matter content. In contrast, content is important in both the *infusion* and *immersion* approaches. CT is an explicit objective in the *infusion* course but not in the *immersion* course. In the *mixed* approach, CT is taught as an independent track within a specific subject content course.

According to Ennis (1989), the general approach attempts to teach CT skills and dispositions separately from the presentation of the content of existing subject matter offerings, with the purpose of teaching CT. Examples of the general approach usually do involve some content but do not require that there be content. The infusion of CT requires deep, thoughtful, and well-understood subject matter instruction in which students are encouraged to think critically in the subject. Importantly, in addition, general principles of CT skills and dispositions are made explicit. In the immersion approach, subject matter instruction is thought-provoking, and students do get immersed in the subject. However, in contrast to the infusion approach, general CT principles are not made explicit. The mixed approach consists of a combination of the general approach with either the infusion or immersion approach. Under it, students are involved in subject-specific CT instruction, but there is also a separate thread or course aimed at teaching general principles of CT.

This mix of approaches to developing CT skills and dispositions is mirrored in the empirical literature on CT; Norris (1985) testified to this when he remarked,

> [T]here is little, if any, evidence on the long-term impact of instruction in critical thinking, despite the fact that the vision of such impact is central to the justification of critical thinking instruction . . . If diagnosis and remediation of specific flaws in reasoning are goals of critical thinking instruction, then more fine-grained information on the effects of particular teaching strategies will have to be sought. (pp. 44–45)

Norris's (1985) concerns remain valid today, and they can be coupled with the fact that most reviews have either been inconclusive in determining whether particular teaching strategies seem to enhance CT skills and dispositions or restricted their focus of attention to very specific issues. Some examples: Adams (1999) summarized studies of the impact of nursing education programs on the CT skills of professional nursing students; Allen, Berkowitz, Hunt, and Louden (1997, 1999) studied the impact of various methods of improving public communication skills on CT; Assessment and Learning Research Synthesis Group (2003) at University of London reviewed in great detail the impact on students and teachers of the use of ICT (information and communication technologies) for assessment of creative and CT skills; Bangert-Drowns and Bankert (1990) reported some effects of explicit instruction for CT on measures of CT; Follert and Colbert (1983) analyzed research on debate training and CT improvements; Follman (1987, 2003) documented the difficulty of teaching CT, especially to adults; McMillan (1987) considered the effects of instructional methods, courses, programs, and general college experiences on changes in college students' CT; and Pithers and Soden (2000) reviewed the research literature on teaching CT skills (methods and conceptions of teaching likely to inhibit or enhance CT). Otherwise, although not a comprehensive review, Abrami et al. (2008) found 117 studies based on 20,698 participants, which yielded 161 effect sizes with an average effect size of $g+ = +0.34$ and a standard deviation of 0.61, suggesting a modest but robust effect favoring the instructional viability hypothesis that is at the heart of the current review. Finally,

no review has examined comprehensively the impact that CT instructional interventions might have on subject matter achievement.

In sum, despite a number of significant efforts to collate and review the results of previous empirical research on CT, the question of effective teaching strategies for CT remains outstanding. Thus, in undertaking the current review of research on instructional interventions affecting CT, several basic questions will be addressed: (a) Can CT skills and dispositions be taught? (b) What are some promising strategies for teaching students to think critically? (c) Which students benefit from CT instruction? (d) Are there curricular areas for which CT instruction works best?

Method

Meta-Analysis

Meta-analysis is a quantitative research synthesis methodology that is applied in this study to investigate instructional interventions and their effect on CT outcomes. The overall purpose is to estimate the average effects in the population of learners who are exposed or not exposed to various instructional interventions, and to explore the variability that exceeds sampling error through moderator analysis of categories of study features. The generally accepted steps in conducting a meta-analysis are as follows: (a) define the problem; (b) establish inclusion/exclusion criteria; (c) search for, retrieve, and select studies for inclusion; (d) extract effect sizes; (e) code study features; (f) synthesize effect sizes; (g) explore variability in effect sizes through moderator analysis; and (h) interpret outcomes.

Problem Definition

This article is a meta-analysis that summarizes the empirical evidence on CT skills and dispositions in educational contexts. First, the evidence on the impact of various instructional strategies on CT skills and dispositions will be summarized. Second, this review will examine whether certain methodological aspects of individual studies (e.g., research design, type of CT measure) moderate the magnitude of this impact. Third, the role of several substantive study features will be analyzed. In particular, this review will examine how different types of instructional interventions affect CT skills and dispositions, what impact pedagogical background (e.g., instructor training) has, and how calculated effect sizes vary with age (educational level), subject matter, and treatment duration. Finally, the impact of CT instruction on achievement outcomes will be summarized.

Review Procedures

Abstracts and full-text research reports were each coded for inclusion or exclusion by two raters. An individual coder's ratings, at each stage, were specified to range from 1 (*the study is definitely unsuitable for the purposes of the project*) to 5 (*the study is definitely suitable for the purposes of the project*); the midpoint of 3 (*doubtful but possibly suitable*) was designated as a vote in favor of including the study. In other words, ratings from 3 to 5 suggested either the retrieval of the full-text document (at the abstract review stage) or inclusion of the study in further analyses (at the full-text review stage), whereas ratings of 1 or 2 meant the

elimination of the study from further consideration. Interrater agreement rate at these two stages of the review was calculated and reported in two different ways: (a) as a correlation coefficient between ratings given by independent coders across all reviewed papers and (b) as a percentage of studies, with respect to which both coders agreed whether to reject the study or to continue analyzing it (expressed as Cohen's kappa).

The extent of uniformity between coders was also documented with regard to effect size extraction and to the coding of study features. Each effect size was coded by two raters, and two agreement rates were produced: (a) a number between 50 and 100 was assigned to each study to reflect the degree of agreement between the raters with regard to how many effect sizes should be extracted from each study, and this number was averaged across studies; and (b) a similar procedure was applied with regard to agreement as to which calculation procedures should be used to determine each effect size. As for study features coding, each study was assigned a rating according to the percentage of the features on which the raters initially agreed; all disagreements were discussed until a final accord was negotiated. All agreement rates were averaged across studies, and the average rates are presented in the Results section.

Inclusion/Exclusion Criteria

Decisions regarding whether to retrieve an article were based on a review of study abstracts. Decisions about whether to include studies in the review were based on reading the full text of the article. Both of these decisions were made by two reviewers, working independently, who then met to discuss their judgments and to document their agreement rate. The following inclusion criteria were used: (a) accessibility—the study must be publicly available or archived; (b) relevance—the study addresses the issue of CT development, improvement, and/or active use; (c) presence of intervention—the study employs some kind of defined instructional intervention; (d) comparison—the study compares outcomes that resulted from different types or levels of treatment (e.g., control group and experimental group, pretest and posttest, etc.); (e) quantitative data sufficiency—measures of relevant dependent variables are reported in a way that enables effect size extraction or estimation; (f) duration—the treatment in total lasted at least 3 hours; and (g) age—participants were no younger than 6 years old. If any of these criteria were not met, the study was rejected.

Along with studies employing experimental and quasi-experimental designs, preexperimental designs (e.g., one-group pretest–posttest designs) were included at the outset. The intent was not to eliminate methodologically less sound studies a priori but instead to code for research methodology, explore the differences empirically, and decide at the analysis stage whether and how to include preexperiments. In fact, the bulk of research on CT does not consist of true experiments. This is largely because CT research has been conducted in classrooms, where randomization is difficult to achieve.

One of the greatest challenges was to determine what constituted a control condition. The best candidate for such a condition would be a course with identical subject matter and basic materials but without any instructional activities that are explicitly intended to promote CT development. But sometimes the

description of the control condition in the studies either was unclear or contained some degree of CT-oriented instruction.

In the majority of reviewed studies, the difference between treatment conditions was self-evident and reflected the research hypothesis—the experimental condition offered some treatment that was expected to influence CT development more than the corresponding control condition, which was usually described as traditional instruction without the activities, methods, approaches, technologies, and so on featured in the experimental condition.

Studies with a one-group pretest–posttest design were retained for the initial analyses. For these studies, the pretest scores were treated as the control group scores, and these data were compared with the posttest results, presumably reflecting the effect of a CT-oriented treatment although it is also obvious that threats to internal validity may have occurred.

There were a few cases in which several treatment conditions offered different CT instructional interventions. When faced with such cases, reviewers decided which treatment was more relevant for the purposes of promoting CT development. Several criteria were used in making this decision: (a) the degree to which CT skills and/or dispositions were addressed explicitly (i.e., reflecting Ennis's classification; Ennis, 1989), (b) the relative strength of the treatment (more activities, exercises, etc.) that seemed more likely to influence CT, and (c) the presence of activities and exercises that included more of the basic CT skills of interpretation, analysis, evaluation, inference, explanation, and/or self-regulation. In every study with several interventions, the condition that satisfied one (or any combination) of these criteria was designated as experimental and the other condition was designated as control. When a judgment could not be made or reviewers strongly disagreed but pretest data were available, each condition was coded as a one-group pretest–posttest design and labeled as a preexperiment. Eventually all these studies were removed from the final data set.

Whenever none of the above solutions was possible, the study was rejected and excluded from the review, as well as studies with incompatible units of analysis (where interventions, group compositions or outcome measures could not be matched). Last, studies were rejected when the description of the treatment was very unclear, lacking any specifics that could link the treatment(s) to an expected change in CT skills or dispositions. The reasons for exclusion were documented.

Outcome and Test Types

The issue of measuring CT is a complex one. First, the APA Delphi panel clearly distinguished between CT-linked cognitive skills and affective dispositions relevant to CT, and as a result, each must be measured differently. In addition to the skills/dispositions divide, researchers sometimes focus on "content-specific" thinking skills that are linked to particular fields of study (e.g., Test of Critical Thinking in Biology; McMurray, Beisenherz, & Thompson, 1991). As a result, this analysis considers three particular outcome types: (a) generic CT cognitive skills, (b) content-specific CT skills, and (c) affective dispositions relevant to CT.

In determining whether and how to synthesize evidence on specific aspects of CT, we built on the Bernard et al. (2008) study that factor-analyzed CT subscale weighted means from 60 data sets for the Watson–Glaser Critical Thinking

Appraisal (WGCTA) (G. Watson & Glaser, 1980). In this study, a strong general factor emerged for two versions of the WGCTA and so this review focuses on global indices of CT only, instead of individual subscale means that are often reported in the literature where the WGCTA is used. Although a similar analysis on other popular standardized CT measures was not performed, the findings and decision as to how to proceed with the WGCTA suggested that this strategy should be followed uniformly.

To further differentiate among CT assessment tools, five categories were used:

Standardized tests: These are well-established measures of CT or particular thinking skills and dispositions: WGCTA (G. Watson & Glaser, 1980), Cornell Critical Thinking Test (Ennis & Millman, 1985), California Critical Thinking Skills Test (CCTST; Facione, 1990a), and California Critical Thinking Disposition Inventory (Facione, Facione, & Giancarlo, 1996).

Tests/evaluations developed by a teacher: This category includes, for example, the content analysis of students' responses to interview questions, and open-ended and essay-type tasks teachers used to address CT skills development in their students.

Tests developed by researchers (i.e., one or more of a study's authors): These are nonstandardized measures developed by a researcher for use in a particular study. For example, Bonk, Angeli, Malikowski, and Supplee (2001) and VanTassel-Baska, Zuo, Avery, and Little (2002) developed a measure of CT that they used in their research.

Tests developed by researchers who also taught the courses in question: These are developed by a researcher who was also the teacher or instructor. For example, Zohar and Tamir (1993) developed the Critical Thinking Application Test to evaluate performance in reasoning skills. One of the researchers served as the instructor in the CT condition.

Secondary-source measures: These instruments are usually adopted from other sources with or without modifications. Researchers may use previously developed (standardized or unstandardized) instruments or modify them to meet the requirements of their research setting. For example, Feuerstein (1999) adapted the Language and Media Test developed by an Australian research group (Quin & McMahon, 1993) to suit her Israeli study according to the local media content.

In addition to the three categories of CT outcomes, measures of student achievement (e.g., final exam scores, course grades), when reported in the reviewed studies, were extracted and analyzed.

Literature Searches

To identify and retrieve primary empirical studies relevant to the major research questions, extensive literature searches without date restrictions were performed. Search strategies were customized according to the field being searched (affects terminology) and the supplier of the database being searched (affects nature of the search strategy). Generally speaking, keywords used in the search strategies were divided into two primary concepts: domain and treatment type. In some database

searches, a third concept, context, was added to increase the precision of the search. Searches were not limited to a particular age group or special population of learners. A combination of controlled vocabulary and keywords was used as appropriate, as summarized in Appendix B.1.

The teaching of CT skills is addressed in many different disciplines, and so a multidisciplinary approach was taken with regard to database selection. The subject and multidisciplinary databases used are listed in Appendix B.2.

A strategy was also developed to locate "gray literature" (e.g., unpublished dissertations, private and governmental reports, unpublished manuscripts, and prepublications from institutional repositories). The primary tools used for retrieval of gray literature were Web search engines Yahoo and Google, both of which do not have any controlled vocabulary; it was therefore decided to run a series of searches using different combinations of keywords and to browse and select from the first 200 records found by each. As a further step, the "open-access library OAIster was searched, as was the Ed/ITLib digital library produced by the Association for the Advancement of Computing in Education. These were useful in locating conference papers and dissertations/theses not available in ProQuest Dissertations & Theses Full Text.

Attempts were made to specifically target material from across the English-speaking world, with searches of American, Canadian, British, and Australian databases.

Furthermore, approximately 60 review articles and previous meta-analyses were used for "branching" (i.e., their bibliographies were scanned for other relevant studies). A citation search was also conducted on many of these same review articles using the Web of Science (ISI) database to locate publications that had cited them; citations searches were also conducted on the main CT tests (Cornell, California, Watson–Glaser).

Searches were originally conducted in 2003, and updated in 2005, 2008, and 2009, before being finalized in 2010 (the cutoff date for records is 2009). The original search strategies and a bibliography of the studies included in the review are available on request.

Of the 7,524 records identified through searches of the literature, 2,332 were retrieved for full text review. Of this number, 684 met the inclusion criteria outlined above.

Data and Effect Size Extraction

The final collections (i.e., generic CT skills, content-specific CT skills and CT dispositions) contained a combination of standardized and nonstandardized measures. The majority of the studies in each collection reported only one outcome measure. However, in some cases two measures of the same type were found. When several measures representing the same outcome type were reported in a given study, reviewers always chose a standardized measure over a locally produced measure (e.g., teacher-made). When two standardized (e.g., WGCTA and CCTST) or two locally produced measures were encountered, effect sizes were first calculated for each test separately and then the effect sizes were averaged (Scammacca, Roberts, & Stuebing, 2013). For example, in the collection of generic outcomes, 20 composites out of 341 effect sizes were created in this way.

Also, whenever the same group of participants was used repeatedly (i.e., the same control group compared to two different treatment groups), the group sample size was reduced proportionally in order to avoid its overrepresentation in the final data set.

Effect size is a standardized metric expressing the difference in two group means (usually a control and a treatment). Cohen's (1988) *d* (Equation 1) is the biased estimator of effect size.

$$d = \frac{\bar{X}_e - \bar{X}_c}{SD_{\text{Pooled}}}. \tag{1}$$

There are also two modifications of this basic equation: one for studies reporting pretest data for both experimental and control groups and another for a single-group pre–posttest design. In other cases (e.g., *t* tests, *F* tests, *p* levels), effect size is estimated using conversion formulas provided by Glass, McGaw, and Smith (1981) and Hedges, Shymansky, and Woodworth (1989).

To correct for bias in small samples, *d* was converted to the unbiased estimator *g* (Hedges & Olkin, 1985), as follows:

$$g \cong \left(1 - \frac{3}{4N - 9}\right) d. \tag{2}$$

The standard error of *g* was calculated using Equation (3).

$$SE_g = \sqrt{\frac{1}{n_e} + \frac{1}{n_c} + \frac{g^2}{2N}} \left(1 - \frac{3}{4N - 9}\right). \tag{3}$$

Comprehensive Meta-Analysis (Version 2.2.064; Borenstein, Hedges, Higgins, & Rothstein, 2005), a dedicated statistical software package, was used for all primary, moderator, publication bias, and sensitivity analyses.

Study Features

To explain variability in effect sizes, coded study features were assessed individually and, when appropriate, in combinations. The following methodological features were tested: (a) type of research design (preexperimental, quasi-experimental, or true experimental), (b) type of CT measure (standardized, teacher-made, researcher-made, teacher-/researcher–made, or secondary source measures), and (c) *ES* extraction method (calculated from descriptive statistics, estimated with no assumptions made, or estimated with assumptions).

The following substantive features were coded: (a) educational level and age of participants (elementary or 6–10 years, early secondary or 11–15 years, high school or 16–18 years, undergraduate education, graduate education, or adult learners outside of formal school settings), (b) intervention type according to Ennis's (1989) classifications (General, Infusion, Immersion, or Mixed), (c) subject matter (science, technology, engineering, and math or STEM classification; health education or other, non-STEM disciplines), and (d) treatment duration

(short-duration interventions lasting from several hours to 2 days, medium-duration interventions lasting several days or longer, long-duration interventions of about a semester or term, and extended-duration treatments lasting more than a semester). In addition, this review included further coding of the nature of CT instruction.

Turning now to more complex questions of taxonomy, the Abrami et al. (2008) analysis of CT instructional approaches used Ennis's (1989) CT typology of four courses (General, Infusion, Immersion, and Mixed) for classifying and describing various instructional interventions. In the general course, CT skills and dispositions are learning objectives, without specific subject matter content. In contrast, content is important in both the infusion and immersion approaches. CT is an explicit objective in the infusion course but not in the immersion course. In the mixed approach, CT is taught as an independent track within a specific subject matter. These four approaches, general, infusion, and immersion and mixed, will be assessed in this review for their instructional efficacy.

Although Ennis's (1989) taxonomy has proven its usefulness, the aim in this study was to expand the analysis beyond a single instructional classification scheme. A more fine-grained approach, which might explain more of the variability in CT outcomes and highlight especially effective instructional approaches, was sought. A new set of four major categories was developed: individual study, dialogue, authentic or anchored instruction, and coaching. Subcategories were also developed for the major categories and a substantial number of the effect sizes ($k = 95$) were coded for their presence.

A rating scale (0–3) describes the extent to which each category was represented in a given study:

0 = *not present*
1 = *slightly present*
2 = *moderately present*
3 = *strongly present*

Category 1: Individual Study

The initial coding included a category for individual study. Individual study includes instructional techniques and learning activities that are based on students' individual work. It takes place whenever students study *alone* by engaging in reading, watching, listening to a teacher's explanations, reflecting on new information, and solving abstract problems on their own. This category was not used in the main set of moderator variable analyses because in the vast majority of cases it did not distinguish between experimental and control conditions: In either group, student individual work was represented to a similar extent.

Category 2: Dialogue

This category is characterized by learning through discussion. The idea that dialogue facilitates CT has a long history, dating back to the Socratic method, in which concepts were clarified through one-on-one interactions.

When engaged in critical dialogue, individuals are *discussing* a particular problem together. The dialogue may be adversarial or it may be quite

cooperative, but in either case, some sort of question is under consideration. Critical dialogue can take multiple forms, including whole-class debates, within-group debates, within-group discussions, whole-class discussions, and online discussion forums.

Dialogue does not have to be oral; it can be written as well. If written, the activity must be characterized by multiple back-and-forth interactions to count as dialogue (e.g., online discussion forums).

The dialogue category also has numerous subcategories, which fall into three broad groups: question asking (2.1 and 2.2), discussion (2.3–2.8 and 2.10), and debate/Socratic dialogue (2.9 and 2.11). The full list of subcategories is as follows:

2.1. Teacher poses questions to students
2.2. Students question their teacher
2.3. Student dyads (no/minimal teacher participation)
2.4. Whole-class discussion (no/minimal teacher participation)
2.5. Group discussions (no/minimal teacher participation)
2.6. Student dyads (teacher-led)
2.7. Whole-class discussion (teacher-led)
2.8. Group discussions (teacher-led)
2.9. Formal debate
2.10. Student presentation with a follow-up discussion
2.11. Socratic dialogue

See Appendix C.1 for illustrative examples of the category codes.

Category 3: Authentic or Anchored Instruction

This category is characterized by an effort to present students with genuine problems or problems that make sense to them, engage them, and stimulate them to inquire. This approach to CT is somewhat newer than dialogue but still has a lengthy genealogy. At the very least, it can be traced back to Jean-Jacques Rousseau (1762) who, in *Emile*, criticized teaching methods that relied on fact memorization. Instead, Rousseau argued that children should learn to solve problems that appealed to them—in one of the lengthy examples he sketched in *Emile*, he suggested that if a child wished to learn to use a compass, it would be best for him or her to go outside and actually use the compass in order to navigate unknown terrain.

John Dewey (1902) later echoed this sentiment in *The Child and the Curriculum*, an essay within which he distinguished between what he called the "logical" and "psychological" aspects of knowledge. The logical aspect of knowledge was the finished product, the type of knowledge that is written down in a textbook. Dewey suggested that this type of knowledge is comparable to a completed map of a country. The psychological aspect of knowledge, however, is analogous to the process through which the finished map was constructed—it is analogous to the experience of the explorer walking through the undiscovered country. The challenge of teaching, Dewey argued, was to "psychologize" knowledge for students,

to make students into explorers who would work through genuine questions in order to rediscover, for themselves, the finished structure of knowledge.

Given this view, it is not surprising that Dewey, like Rousseau, criticized rote learning in which children are able to recite concepts and facts, but do not really understand the meaning of these facts or their conceptual significance. Dewey (1907) felt that in order to be able to use an idea as a tool for thinking (and, by extension, for CT), it was essential that one learned the significance of the idea in the context of a problem that appeared genuine to the student. As he noted, acerbically, "There is all the difference in the world between having something to say and having to say something" (p. 67).

A wide variety of approaches can be categorized under the banner of authentic/anchored instruction. Simulations are, perhaps, some of the most powerful examples, since they bring the problem to life in the strongest possible way. Role-playing is also a very strong fit for this category, as are various kinds of dilemmas (e.g., ethical, medical) that are presented to students. In sum, whenever there is a well-defined real-world problem that the students are analyzing, one has an example of authentic/anchored instruction.

The authentic/applied category has numerous subcategories:

3.1. Applied problem solving (including some hypothetical problems with high applied value for students, e.g., ethical dilemmas)
3.2. Case studies
3.3. Simulations (may overlap with case studies; include manipulations of the content that is more formalized and, often, computer-based)
3.4. Playing games (similar to simulations, but the content is more abstract or problems are more hypothetical)
3.5. Role-play (e.g., nurse/patient in nursing education)

See Appendix C.2 for illustrative examples of the category codes.

Category 4: Mentoring

One-on-one mentoring, tutoring, coaching, apprenticeship or modeling are, arguably, the oldest forms of teaching. The key component of mentoring is one-on-one interaction between an expert, or more generally someone with more expertise, and a novice, or more generally someone with less expertise. Mentoring emphasizes one-on-one modeling and error correction based on critical analysis. An advisor talking to a student, a physician modeling a procedure for a medical student, an employee correcting an intern are all paradigmatic examples of mentoring.

Three subcategories of mentoring were developed:

4.1. One-on-one teacher-student interaction
4.2. Peer-led dyads
4.3. Internship (an experienced professional coaching a younger colleague, e.g., medical internship).

See Appendix C3 for illustrative examples of the category codes.

Results

The results are divided into four sections of unequal size. The first and largest section is devoted to an examination of generic CT outcomes. A second section briefly presents the evidence on the effects of CT instruction on achievement outcomes. The third section presents the results for content-specific outcome measures. The final section is devoted to the effects of CT instruction on CT dispositions.

Agreement rates at various stages of the review were as follows:

- Screening abstracts: 92.449% (Cohen's $\kappa = 0.85$) or $r = 0.78$, $p < .001$
- Full-text manuscript inclusion/exclusion: 89.45% ($\kappa = 0.79$) or $r = 0.87$, $p < .001$
- Decisions on the data source for and the number of effects: 91.93% ($\kappa = 0.84$)
- Effect size extraction: 97.82% ($\kappa = 0.96$)
- Study features coding: 89.55% ($\kappa = 0.79$)
- Instructional dimensions coding: 79.84% ($\kappa = 0.60$)

Generic CT Outcomes

Publication Bias

Several measures were taken to examine potential bias due to publication source. First, a funnel plot of the effect sizes by standard error (i.e., sample size) was examined to determine if the collection was relatively balanced on either side of the average random effect size. This examination yielded a reasonably symmetrical funnel with no obvious biases. Second, a classic fail-safe analysis and Orwin's adaptation yielded the following results: (a) 434,819 studies are needed to bring the observed *p* value to an alpha above .05 and (b) 1,833 null-effect studies would be required to reduce the average effect in this meta-analysis to a "trivial" average effect size value of 0.10. Both of these analyses were conducted within the analysis module of Comprehensive Meta-Analysis, and they indicate the robustness of the collection.

Date of Publication

The formal development of standardized instruments to measure CT skills began in the 1940s along with the first experimental studies of CT instruction. The examination of the instructional viability and nature of CT teaching methods continues to this day. The frequency of empirical included studies increases in the decades from the late 1930s to 2009, with the greatest frequency in the past two decades (i.e., 1990–1999, 26.6%; 2000–2009, 44.6%) for a combination of more than 70%. Research on the topic has increased steadily over time, which is likely indicative of the importance that educators and researchers have come to attach to the teaching of generic CT skills.

Characteristics of the Complete Collection

Table 1A shows the summary statistics for the collection of 867 effect sizes included in the original collection. This average effect size is on the lower boundary of a moderate average effect size according to Cohen's (1988)

TABLE 1

Analysis of generic CT skills: (A) overall mean effect size for the fixed and random models; (B) mixed-effects moderator analysis of categories of research design

		Effect size		95% Interval		
Overall outcome	*k*	*g*+	*SE*	Lower	Upper	Test of null, *z*
A						
Fixed effects	867	0.33	0.01	0.31	0.34	47.42*
Heterogeneity		*Q*-Total = 5,275.62, *df*(*Q*) = 866, $p < .001$, $I^2 = 83.58$				
Random effects	867	0.39	0.02	0.35	0.42	21.39*
B						
Preexperiments	363	0.39	0.03	0.34	0.45	13.00*
Quasi experiments	361	0.33	0.03	0.28	0.38	11.00*
True experiments	128	0.30	0.04	0.21	0.38	7.50*
Q-Between = 5.25, *df* = 2, $p = .07$						

Note. CT = critical thinking; *df* = degrees of freedom.
*$p < .001$.

categories of effect size magnitude. This positive average effect size is suggestive of the potential to teach generic CT skills.

Analysis of Outliers, Research Design, and Quality of Outcome Type

Because of the magnitude of this collection, it was prudent to examine it first in terms of the potentially biasing effects of extreme outliers, research design, and the quality and nature of the outcome measures. Fifteen outliers, whose magnitude was $g = +2.5$ or higher were identified and removed. The average random effect size ($k = 852$) dropped to $g+ = 0.35$, $p < .001$.

Table 1B shows the breakdown of studies categorized by research design. Although the average mixed effect for research design approached significance ($p = .07$), Bonferroni-adjusted post hoc analysis revealed that preexperiments were significantly higher than both true experiments and quasi experiments, and that these categories were not significantly different from each other ($z = 0.82$, $p = .21$). The decision was made to use only true and quasi experiments in further analyses. Also, studies that were long in treatment duration (e.g., over a semester) were excluded if the specifics of the treatment could not be confidently identified and coded.

This left 425 effect sizes with a random effects average of $g+ = 0.35$, $z = 14.17$, $p < .00$. Under the fixed-effects model, heterogeneity was significant ($Q_T = 2{,}059.43$, degrees of freedom [*df*] = 424, $p < .001$) and I^2, a descriptive measure of between-study heterogeneity exceeding sampling error was 79.41%.

To further determine if these initial results stood up to scrutiny when the larger collection of evidence was reduced to the set of better evidence, nonstandardized outcome measures (i.e., those that are of indeterminate reliability and validity)

TABLE 2

Analysis of generic CT skills with standardized measures only: (A) overall mean effect size for the fixed and random models; (B) mixed-effects moderator analysis of categories of research design

		Effect size		95% Interval		
Overall outcome	*k*	*g*+	*SE*	Lower	Upper	Test of null, *z*
A						
Fixed effects	341	0.25	0.01	0.23	0.28	21.09*
Heterogeneity		*Q*-Total = 1,225.89, *df*(*Q*) = 340, *p* = .001, I^2 = 72.27				
Random effects	341	0.30	0.02	0.25	0.34	12.06*
B						
Quasi experiments	243	0.29	0.03	0.23	0.34	9.67*
True experiments	98	0.32	0.05	0.23	0.41	6.40*
Q-Between = 0.31, *df* = 1, *p* = .58						

Note. CT = critical thinking; *df* = degrees of freedom.
*$p < .001$.

were removed. After removing 84 effect sizes derived from nonstandardized measures, the collection was further reduced to 341 effect sizes. Removing nonstandardized measures also reduced the random-effects average effect size, but only minimally (Table 2A). The collection, however, remained heterogeneous (Q_T = 1,225.89, *df* = 340, *p* = .001) and I^2 remained high at 72.27%.

After removing nonstandardized outcomes, high-quality quasi-experimental designs and true experiments were compared and were found to be not significantly different (Table 2B), suggesting that these two categories of effect sizes could be combined. The final reduced sample, on which all other comparisons are based, was *k* = 341. This number represents 39.33% of the effect sizes originally included in the study.

A frequency histogram of this reduced collection is shown in Figure 1. This distribution is slightly leptokurtic, but essentially normal, with an un-weighted mean of 0.33 and a standard deviation of 0.55. Of the 341 effect sizes, 261 (76.5%) were zero or above, and 80 (23.5 %) were below zero.

Even with the dramatic reduction in the number of effect sizes retained for analysis, the average effect size and heterogeneity statistics did not change dramatically. In the remaining analyses only the subset of effect sizes considered to represent the best and most methodologically sound evidence on instructional interventions was included.

Course-Level Moderator Variables

This section begins the exploration of demographic factors, instructional variables, and other characteristics that may relate to the effect sizes. As we

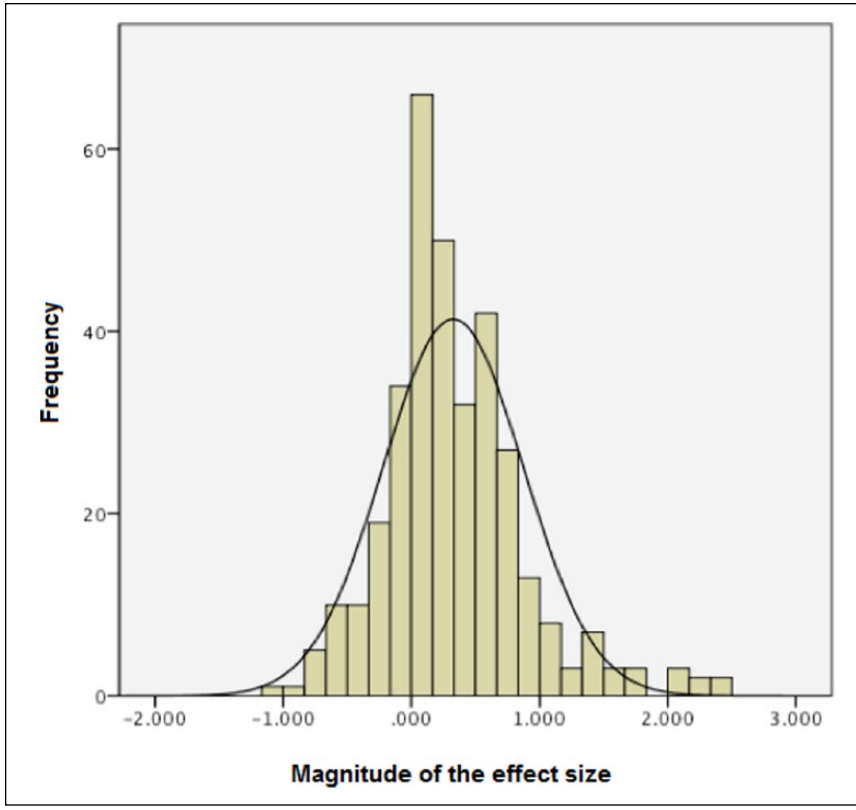

FIGURE 1. *Distribution of unweighted effect sizes for generic critical thinking skills ($k = 341$, $\bar{X} = 0.33$, $SD = 0.55$).*

will outline below, variations in educational level, subject matter, and treatment duration did not generate significant differences in outcome. As a result, there are likely no confounds with these variables that could compromise subsequent substantive analyses.

Educational level. Table 3A shows the 341 effects broken down by educational level. There are no significant differences among the levels, suggesting that a skills-based approach to CT improvement can achieve some level of success at all grade levels. Furthermore, with the exception of the last category (i.e., graduate and adult students) the average effect size of CT instruction is significantly greater than zero.

Subject matter. No significant differences among different broad types of subject matter were observed as shown in Table 3B. In fact, the confidence intervals for STEM and non-STEM overlap almost perfectly, with a $g+$ of 0.31 for the former

TABLE 3

Mixed moderator variable analysis of course demographic variables: (A) educational level; (B) subject matter; (C) duration of treatment/control

		Effect size		95% Interval		
Educational levels (age, years)	*k*	*g*+	*SE*	Lower	Upper	Test of null, *z*
A						
Elementary school (6–10)	49	0.37	0.08	0.22	0.52	4.63*
Middle school (11–15)	78	0.37	0.06	0.26	0.48	6.17*
High school (16–18)	71	0.25	0.05	0.15	0.35	6.25*
Undergraduate	126	0.26	0.04	0.19	0.33	6.50*
Graduate and adult students	17	0.21	0.13	–0.05	0.46	1.62
Q-Between = 4.83, *df* = 4, *p* = .30						
B (*k* = 269; "Missing" and "Other" categories removed)						
Health/medical education	29	0.20	0.09	0.02	0.37	2.22*
STEM subjects[a]	73	0.31	0.05	0.21	0.41	6.20*
Non-STEM subjects[b]	167	0.29	0.04	0.22	0.36	7.25*
Q-Between =1.25, *df* = 2, *p* = .53						
C (*k* = 338)[a]						
Short (1 hour to 2 days)	13	0.66	0.26	0.15	1.17	2.54*
Medium (2 days to < one semester)	99	0.33	0.05	0.24	0.43	6.60*
One semester	130	0.27	0.03	0.21	0.34	9.00*
>One semester	96	0.23	0.04	0.15	0.32	5.75*
Q-Between = 4.54, *df* = 3, *p* = .21						

Note. df = degrees of freedom.

[a]STEM subjects are science, technology, engineering, and mathematics.

[b]Non-STEM subjects are all other subject areas, except for health/medical education.

*$p < .05$

and 0.29 for the latter. Furthermore, the average effect size of CT instruction was significantly greater than zero for all subject matters.

Duration of instruction in treatment/control. Duration of the intervention did not significantly affect CT effect sizes, suggesting that the effects of CT instruction do not vary whether instruction is brief or not. More importantly, the average effect size for CT instruction was significantly greater than zero for each category of treatment duration (Table 3C).

The analyses of educational level, subject matter, and treatment duration point toward the generalizability of CT instructional effects, notwithstanding the fact that there remains significant variability among the effect sizes.

TABLE 4

*Mixed-effects moderator analysis of Ennis's (1989) classification of instruction (*k = *341)*

		Effect size		95% Interval		
Overall outcome	*k*	*g*+	*SE*	Lower	Upper	Test of null, *z*
Direct instruction	44	0.26	0.06	0.14	0.38	4.33*
Infusion	152	0.29	0.04	0.22	0.36	5.80*
Immersion	61	0.23	0.05	0.13	0.32	4.60*
Mixed	84	0.38	0.06	0.26	0.51	6.33*
Q-Between = 4.10, degrees of freedom = 3, *p* = .25						

Instructional Moderator Variables

Having dealt with these basic study features, we now turn to an analysis of instructional variables related to generic CT outcomes based on the above-described two taxonomies (Ennis, our three-category scheme) and resultant coding schemes.

Instructional variables based on Ennis's (1989) classification. All four of the methods in Table 4, coded according to Ennis's (1989) classification, produced significantly positive average effect sizes, but the categories did not differ from one another.

Instructional variables based on three-category scheme. The Three-Category Scheme includes coding for Dialogue, Authentic Instruction, and Mentoring. The analyses of each instructional category are based on a subset of the data because not all types of instruction were present in each study. To further refine the analyses, findings where there is no difference in type of instruction between the experimental and control groups (no difference in instruction) are contrasted with findings where the experimental group had higher amounts of the type of instruction compared to the control condition (instruction favoring the experimental group).

Two of the three categories, Dialogue and Authentic or Anchored Instruction, produced significantly larger positive effect sizes for instruction favoring the experimental group. This did not hold true for the Mentoring category, perhaps due to the smaller number of effect sizes for the latter. However, as Table 5 details, all three categories of instruction were significantly different from zero in terms of instruction favoring the experimental group.

Single and combined effects of instructional interventions. The following analysis (Table 6) describes certain combinations of instructional strategies (e.g., authentic instruction plus dialogue) when these strategies exceeded those in the control condition. Other combinations were excluded because of low cell frequency. Therefore, this result constitutes a refinement of the result reported in

TABLE 5

Mixed-effects moderator analysis of categories of the three-category classification: (a) dialogue; (b) authentic or anchored instruction; (c) mentoring, coaching, or tutoring

		Effect size		95% Interval		
Categories	*k*	*g*+	*SE*	Lower	Upper	Test of null, *z*
A (*k* = 201)						
No difference	87	0.19	0.04	0.11	0.27	4.75*
experimental > control	114	0.32	0.05	0.23	0.42	6.40*
Q-Between = 4.53, *df* = 1, *p* = .03						
B (*k* = 214)						
No difference	118	0.22	0.04	0.14	0.29	5.50*
Experimental > Cont.	96	0.34	0.05	0.24	0.44	6.80*
Q-Between = 4.07, *df* = 1, *p* = .04						
C (*k* = 266)						
No difference	238	0.26	0.03	0.20	0.31	8.67*
Experimental > control	28	0.39	0.08	0.23	0.55	4.88*
Q-Between = 2.58, *df* = 1, *p* = .11						

Note. df = degrees of freedom.
*$p < .05$.

TABLE 6

Mixed-effects moderator analysis of category combinations based on the three-category classification

		Effect size		95% Interval		
Categories	*k*	*g*+	*SE*	Lower	Upper	Test of null, *z*
Authentic Instruction (A)	22	0.25	0.10	0.05	0.46	2.50*
Dialogue (D)	43	0.23	0.08	0.07	0.39	2.87*
A + D	45	0.32	0.08	0.17	0.47	4.00*
A + D + Mentoring	19	0.57	0.10	0.38	0.77	5.70*
Q-Between = 8.19, degrees of freedom = 3, *p* = .04						

*$p < .05$.

Table 5. This analysis represents the unique contributions of authentic instruction and dialogue as well as both the outcome of the dual combination of authentic instruction and dialogue and the triple combination of authentic instruction, dialogue, and mentoring (A + D + M). As an example, the 19 effect sizes represented in the A + D + M row of Table 6 are the only studies in which all three dimensions

were higher in the experimental groups than in the control groups. A comparison between "authentic instruction + dialogue + mentoring" and "authentic instruction + dialogue" revealed a significant difference ($z = 1.98$, $p = .024$). A combination of all three strategies produced the highest effect size.

Instructional subcategories. For each subcategory, studies were identified where the experimental group exceeded the control group on the quality or quantity of each specific feature. When the experimental group equaled or exceeded the control group for at least five comparisons (i.e., $k \geq 5$), an analysis was conducted to determine whether a subcategory explained effect sizes significantly ($p < .05$). The following Dialogue subcategories were significant and favored the experimental group: teacher poses questions ($g+ = 0.38$, $k = 19$, $p < .001$), whole-class discussions led by the teacher ($g+ = 0.42$, $k = 16$, $p < .001$), and small group discussions led by the teacher ($g+ = 0.41$, $k = 14$, $p < .001$). Two subcategories of Authentic and Anchored Instruction significantly favored the experimental group: applied problem solving ($g+ = 0.35$, $k = 31$, $p < .001$) and role-playing ($g+ = 0.61$, $k = 5$, $p < .001$). For mentoring, no subcategory significantly favored the experimental group.

Achievement Outcomes

Many of the studies that were reviewed contained achievement measures of course content (e.g., chemistry) in addition to measures of generic CT skills. By and large, these measures were teacher-made and so their psychometric properties remain unknown.

The results of this analysis yielded a significantly positive average effect size of $g+ = 0.33$, $k = 140$, $p < .05$. The collection was significantly heterogeneous ($Q = 826.10$, $df = 139$, $p < .001$) and I^2 was 83.17.

This result is very similar to the findings reported for 341 generic CT outcomes (i.e., $g+ = 0.30$). In addition, there were no important differences ($p > .05$) found among categories of either instructional or demographic study features.

Content-Specific CT Outcomes

The same analysis approach that was used in the previous section on generic CT outcomes was also applied to studies that attempted to promote content-specific CT. Content-specific CT outcomes are those skills that are assessed using measures specifically designed to relate thinking skills to the content that is being taught in a course. The analysis began with 198 effect sizes. Table 7A shows the results after the removal of preexperiments. It is immediately evident that content-specific outcomes produced a higher average effect size compared to generic CT outcomes. The finding of $g+ = 0.57$ ranks at the low end of Cohen's category of effects that are deemed to be of a moderate magnitude. However, as with generic effects, the distribution of 97 effect sizes was significantly heterogeneous (Table 7A), suggesting a large degree of between-study variation based on an I^2 value of 82.36%.

Source of Measure With Preexperiments Removed

The results shown in Table 7B examine the difference in average effect sizes between standardized and non-standardized outcome measures. The average

TABLE 7

*Overall mean effect size (*g+*) and statistics for content-specific CT outcomes: (A) content-specific ct outcomes with pre-experiments removed; (B) mixed moderator variable analysis for measure source (*k = 97*)*

		Effect size		95% Interval		
Overall outcome	*k*	*g+*	*SE*	Lower	Upper	Test of null, *z*
A						
Fixed effects	97	0.53	0.02	0.49	0.57	25.24*
Heterogeneity		*Q*-Total = 544.32, *df*(*Q*) = 96, $p < .001$, $I^2 = 82.36$				
Random effects	97	0.57	0.05	0.47	0.68	10.75*
B						
Standardized	31	0.40	0.07	0.26	0.53	5.71*
Nonstandardized	66	0.65	0.07	0.52	0.78	9.29*
Q-Between = 6.67, *df* = 1, *p* = .01						

Note. CT = critical thinking; *df* = degrees of freedom.
*$p < .05$.

effect fell to $g+ = 0.40$ for the 31 effect sizes derived from standardized measures. Regardless of the measure source, the average effect sizes for CT instruction were significantly greater than zero for this category of outcome.

The analysis of substantive study features for content-specific CT outcomes revealed a similar pattern to the results for generic outcomes. In particular, the effects of CT instruction were significant ($p < .05$) for all of the substantive study features including educational level, subject area, duration of instruction, and type of instruction.

CT Dispositions

The original collection of included studies produced 82 effect sizes. After removing preexperiments, this number fell to 25 effect sizes with an average effect size of $g+ = 0.23$ that was significantly greater than zero (Table 8) but with moderately high heterogeneity. Due to the small number of effect sizes, methodological, assessment, and substantive features were not explored in detail.

Summary of Findings

The global findings are summarized in Table 9. The column heading "Magnitude of Average Effect" applies Cohen's (1988) qualitative criteria to the average effects of generic CT skills, content-specific CT skills, and CT dispositions. The column heading "Improvement Index" refers to the percentage of gain under the normal distribution between the experimental condition compared to the control condition. One way of interpreting this index is to consider the fact that an average student in the experimental condition (50th percentile) would

TABLE 8

Overall mean effect size for CT dispositions with preexperimental studies removed

		Effect size		95% Interval		
Overall outcome	*k*	*g*+	*SE*	Lower	Upper	Test of null, *z*
Fixed effects	25	0.19	0.04	0.11	0.28	4.40*
Heterogeneity		*Q*-Total = 82.32, *df*(*Q*) = 24, $p < .001$, $I^2 = 70.84$				
Random effects	25	0.23	0.09	0.06	0.40	2.64*

Note. CT = critical thinking; *df* = degrees of freedom.
*$p < .05$.

TABLE 9

General findings from three CT measures and content outcomes derived from the reduced collections

Outcome measure	*k*	*g*+[a]	Magnitude of average effect	Improvement index[c]	Significantly heterogeneous[d]
Generic CT Skills	341	0.30	Low[b] (0.20 to <0.50)	11.79%	Yes
Achievement	140	0.33	Low (0.20 to <0.50)	12.93%	Yes
Content-specific CT skills	97	0.57	Moderate (0.50 to <0.80	21.57%	Yes
CT dispositions	25	0.23	Low (0.20 to <0.50)	9.10%	Yes

Note. CT = critical thinking.
[a]Random-effects model.
[b]Based on Cohen's (1988) qualitative descriptions.
[c]Improvement index = ($\Phi(g+) - 0.50) \times 100$), where Φ is the cumulative frequency of *g*+ under the normal curve.
[d]Derived from the fixed-effects model.

place *X*% higher among students in the control condition (i.e., 50th percentile + *X*th percentile). Finally, the column titled "Significantly Heterogeneous" is a result of a fixed-effects analysis and indicates whether between-study variability exceeds what is expected by chance. It also indicates whether there is sufficient unexplained variation to warrant moderator variable analysis.

Discussion

This review addressed several basic questions: (a) Can CT skills and dispositions be taught? (b) What are some promising strategies for teaching students to think critically? (c) Which students benefit from CT instruction? (d) Are there curricular areas for which CT instruction works best?

At the most general level, this analysis clearly reveals that a variety of CT skills (both generic and content specific) and dispositions can develop in students through instruction at all educational levels and across all disciplinary areas using a number

of effective strategies. There are some caveats attached to this finding, but we will postpone discussion of these until our "Outstanding Questions" section.

Looking more specifically at the question of possible instructional strategies, it is clear that two general types of instructional interventions are especially helpful in the development of generic CT skills. Notably, the opportunity for dialogue (e.g., discussion) appears to improve the outcomes of CT skills acquisition, especially where the teacher poses questions, when there are both whole-class teacher-led discussions and teacher-led group discussions. Similarly, the exposure of students to authentic or situated problems and examples seems to play an important role in promoting CT, particularly when applied problem solving and role-playing methods are used.

In addition, as Table 6 indicates, it also appears as though dialogue and authentic instruction are effective in combination, particularly when mentorship is added to the mix. As our findings demonstrate, studies that featured all three types of intervention (A + D + M) generated significantly larger effect sizes than either (A + D) or A or D alone. This is particularly interesting in light of the fact that mentoring did not generate especially strong results when analyzed on its own. In light of these results, it appears as though mentoring may serve in a catalytic capacity for CT; it can augment other strategies in a powerful way but is not especially successful if pursued in isolation.

It may be illustrative to examine more closely three interventions in which all of the instructional elements of the three-category scheme are strongly present. First, Yang, Newby, and Bill (2008) studied the effectiveness of structured Web-based bulletin board discussions on CT skills of distance veterinary students. Learners were required to contribute their thoughts and ideas to a series of course exercises and were encouraged to think critically about the course content and especially about questions and feedback comments from both their peers and the instructor. In this study, all three major instructional dimensions were coded equally (+1) in favor of the experimental condition. The effect on CT skills (as measured by the CCTST) in comparison with the control students, whose Web-based discussions were unstructured and unguided, was $g = +0.53$.

Second, in a study conducted by Arrufat (1997), the high prevalence in the experimental condition of all three major instructional strategies (i.e., Discourse—coded +3, Authentic Instruction—coded +3, and Mentoring—coded +2) resulted in the effect size of $g = +0.638$. In this study, undergraduate Psychology students received an intervention intended to promote identity development. The instruction focused on fostering an increase in the use of exploration and critical problem solving with respect to making personal life choices (Authenticity) with guidance offered by the teacher, both individually (Mentoring) and in whole-class discussions (Discourse).

Third, and finally, Pellegrino (2007) conducted a study in which high school students in the experimental condition were taught American History through systematically engaging in "historical thinking" activities. Typically, students were asked to present their views on a historical period in question and offer their explanations of what developments had shaped it. Students' responses were then aligned with various sources of historical information and were asked to judge,

for instance, the relative reliability of similar events used to illustrate conflicting viewpoints. The teacher tried to relate course themes to learners' previous experience and knowledge in class discussions. Activities were designed to allow both the teacher and the students to recognize and appreciate various perceptions of historical content in its complexity, conduct an independent inquiry, find and examine documents, and build their understanding of the dynamics of historical events. There were several role-play games and series of discussions involved. The major difference between the experimental and control conditions was on the dimension of Authentic Instruction (+3) with a supplement of Discourse, and this intervention resulted in the effect size of $g = +1.13$.

In all three studies, we see large effect sizes resulting from a combination of all elements of our three-category scheme. Furthermore, it is clear from the descriptions that all three examples are compelling, longer duration studies in which a robust multifaceted intervention achieved impressive results, and the results of this meta-analysis clearly point toward a need for further pursuit, exploration and refinement of these especially successful strategies.

Outstanding Questions

Before concluding, we feel that it is worthwhile to consider three significant objections that could be levied against the analysis so far: first, that this review tacitly endorses a quasi-causal view of teaching CT in which successful instruction is simply a matter of adopting the correct instructional processes; second, that our contention that CT can be taught is subject to some significant caveats. As we will explain, the first objection is mitigated by the nature of our analysis, whereas the second objection deserves serious consideration.

The first objection holds that within some strands of educational thought, there is a prevailing misrepresentation of the teaching process that makes it seem as though good teaching is nothing more than applying the "magic recipe" of teacher behaviors that will produce the desired product. One could further argue that this mentality is present in our own approach, which has attempted to determine which types of instructional interventions yield the most promising results in terms of generating favorable results on standardized tests that purport to measure CT.

The first part of the objection has some substance and captures a line of flawed thinking that has been influential. One can see a particularly clear and virulent version of it in Skinner's (1950, 1959) educational reform efforts, and a more moderate version is present within the process–product research tradition (Gage & Needels, 1989). Furthermore, it is probable that many of the studies analyzed in this meta-analysis reflect this view to some extent. The very fact that these studies took the approach they did (in most cases, a relatively short-duration intervention followed by a standardized test) arguably indicates that they tacitly favor a process–product view of teaching.

However, as we indicated in our section on the contested definition of CT, the mere fact that this analysis synthesizes articles that occasionally take a simplistic approach to CT does not imply that we endorse such a view. We regard teaching (and, more specifically, teaching CT) as a complex and multifaceted process, in which there is no magic recipe for the "production of learner success." There are,

however, ingredients that appear to be particularly promising, and it is our task as meta-analysts to highlight these. The techniques that we have identified as being promising (i.e., discussion, authentic situations, mentorship) are not an exhaustive list of promising ingredients, but they are a useful starting point for thinking about the challenges teachers face, as well as for further research efforts in the area of CT. Even if one endorses a highly contextual "quilting bee" approach to CT, as Barbara Thayer-Bacon (2000) does, there is still some usefulness in taking stock of the fact that certain kinds of CT tools seem to be particularly helpful.

A second major objection, to the findings presented above, holds that the results presented do not show that CT can be taught, nor do they show that certain types of interventions tend to facilitate the development of CT. What the results do show, the critics maintain, is that certain pedagogical interventions are associated with better performance on a CT test that measures a narrow band of skills that have only a tenuous link to genuine CT. These critics may hold a more dispositional view of CT or they may hold a view that CT is a complex, multifaceted practice, in line with the position of Thayer-Bacon (2000) that was sketched earlier.

These critics have a point: Although a wide variety of measures are captured by our analysis (including teacher-developed and dispositional measures), there is a potential bias among the measures toward a skills-focused conception of CT, which is, arguably, to borrow from Paul's (1990) terminology, a relatively weak sense of CT. Yet, even after conceding this point, one could still argue, as we did in the section on the contested definition of CT, that these skills are important potential precursors to a more ambitious and robust form of criticality. To reject these results completely would be tantamount to discarding the good for the sake of a nonexistent and unobtainable better, at least as far as meta-analysis is concerned.

Still, one can easily imagine arguments that reject this line of reasoning as well and hold that a skills approach to CT, "effective" though it may be, is completely counterproductive. Again, regardless of the merits of this critique, we come up against the limitations of what a meta-analysis of education research can show. A meta-analysis is only capable of answering questions that have already been asked in certain very specific kinds of ways, and its claims must always be somewhat modulated. Like a crude early map, a meta-analysis of education research charts terrain that has already been visited many times and provides some modest degree of guidance for future visitors to the area. One will still see plenty of blank spots and indications to the effect that "Here there be dragons."

Beyond these two serious objections, a number of other smaller questions remain outstanding. Perhaps most significantly, although the three-category scheme revealed some important differences between types of instructional interventions, the analyses of methodological and substantive study features failed to identify factors that explained completely the heterogeneity in effect sizes. Although the findings appear generally robust across educational level, subject area, treatment duration, and type of instruction, we are stymied in our

attempts to understand why a minority of the CT instructional interventions were unsuccessful, resulting in lower levels of CT performance than in the control conditions. The frustration of the reviewer is that she/he can only explore factors that are evident in a sufficient number of studies, and code and aggregate those factors. There may, therefore, be factors other reviewers will see that were not seen by us or that the next round of new research will elucidate. In that sense, this review may act as a heuristic for future thought and research efforts.

The links between CT instruction and both content-specific CT outcomes and achievement outcomes make clear that student learning can be positively affected by the integration of CT-linked content. However, there is no guarantee of this being the case, especially if what is to be learned focuses on factual knowledge, even when such knowledge is a prerequisite for deeper, critical understanding. To the extent that there is a perceived conflict among curricular objectives, such a conflict may relate, in part, to why, despite the evidence that it may have a halo effect on other kinds of learning, CT instruction is not more widespread.

An additional outstanding question relates to age. Viewed from the perspective of developmental psychology, the fact that there is no significant difference among age levels in CT is quite surprising. According to at least some theories, young children should lack the developmental capabilities to perform well on certain types of CT tasks. One possibility here is that standardized measures (not to mention the aggregation involved in meta-analysis) simply do not capture these subtle differences particularly well, but another possibility is that young children may be more capable of CT than some have surmised. Regardless, this is a question that merits careful analysis.

Finally, the findings also seem to call for further investigation into the broad instructional strategies that were especially successful. An initial attempt was made to divide these categories into subcategories (see pp. 289–291) and to analyze the results according to them, but the results of this investigation, although intriguing, have been somewhat limited thus far. Some subcategories (e.g., teacher-led whole class discussions, applied problem solving, role-playing) favored the experimental group, but in the cases of many of the subcategories, the number of studies in the subcategory was insufficient to draw conclusive results. This is, in part, an inevitable difficulty in meta-analysis, but further research into more precise classification and alternative taxonomies is nonetheless advisable given the promising findings for the three general instructional strategy categories.

Despite these caveats, we are confident that the empirical evidence we were able to synthesize supports the notion that there are a number of promising teaching strategies for helping students develop CT skills and dispositions. Specifically, there are strong indications that dialogue, authentic instruction, and mentorship are effective techniques for the promotion of this goal. These techniques appear to be particularly effective when combined. Nevertheless, we must also concede there is a great deal more investigation to do in this area, and our tentative prescriptions would benefit from being embedded in robust normative frameworks for the teaching of CT.

APPENDIX A

American Philosophical Association report of CT skills and dispositions

Cognitive skills and subskills

- Interpretation: Categorization, Decoding Significance, Clarifying Meaning
- Analysis: Examining Ideas, Identifying Arguments, Analyzing Arguments
- Evaluation: Assessing Claims, Assessing Arguments
- Inference: Querying Evidence, Conjecturing Alternatives, Drawing Conclusions
- Explanation: Stating Results, Justifying Procedures, Presenting Arguments
- Self-Regulation: Self-examination, Self-Correction (p. 6.)

Approaches to specific issues, questions, or problems

- Clarity in stating the question or concern
- Orderliness in working with complexity
- Diligence in seeking relevant information
- Reasonableness in selecting and applying criteria
- Care in focusing attention on the concern at hand
- Persistence though difficulties are encountered
- Precision to the degree permitted by the subject and the circumstance (p. 13)

Approaches to life and living in general

- Inquisitiveness with regard to a wide range of issues
- Concern to become and remain generally well-informed
- Alertness to opportunities to use CT
- Trust in the processes of reasoned inquiry
- Self-confidence in one's own ability to reason
- Open-mindedness regarding alternatives and opinions
- Understanding of the opinions of other people
- Fair-mindedness in appraising reasoning
- Honesty in facing one's own divergent world views
- Flexibility in considering biases, prejudices, stereotypes, egocentric or sociocentric tendencies
- Prudence in suspending, making or altering judgments
- Willingness to reconsider and revise views where honest reflection suggests that change is warranted (p. 13)

Note. CT = critical thinking.

APPENDIX B

Vocabulary, keywords, and databases searched

B.1: *Search vocabulary and keywords*

Domain: Critical Thinking, Thinking Skills

Treatment: Experiment*, Study, Studies, Intervention*, Treatment*, Control Group, Posttest, Posttest

Context: Education, Student*, Learn*, Teach*

(continued)

APPENDIX B (continued)

B.1: *Search vocabulary and keywords*

Terms were combined within sets using the Boolean operator OR, and the sets themselves were combined using the AND operator. For example: ("critical thinking" OR "thinking skills") AND (Experiment* OR Study OR Studies OR Intervention* OR Treatment* OR "Control Group" OR Posttest OR "posttest") AND (education OR student* OR learn* OR teach*)

B.2: *Subject and multidisciplinary databases searched*

Education
ERIC (WebSpirs, CSA)
CBCA Education (ProQuest)
Education Abstracts (WilsonWeb)
British Education Index (DataStar)
Australian Education Index (DataStar)
Social Science
PsycINFO (EBSCO)
Sociological Abstracts (CSA)
EconLit (EBSCO)
Social Sciences Index (Wilson)
Social Studies Abstracts (CSA)
Multidisciplinary
Academic Search Complete (EBSCO)
Francis (CSA)
Dissertations & Theses Full Text (ProQuest)
PAIS International (OCLC)
Web of Science (Thomson Reuters)
Business
ABI/InformGlobal (ProQuest)
Medicine
Medline (PubMed)

APPENDIX C

Codes and illustrative examples of instructional coding categories

C.1: *Illustrative examples of codes for Category 2 (Dialogue)*

Code "0": dialogue not present (Rose, 1997)

The study examined the effectiveness of two explicit methods of instruction in CT skills for postsecondary students with and without learning skills. Students in the experimental group received explicit and embedded CT instruction as a part of the curriculum of a literature course. They were taught to use special icons designed to represent analogous modes of thoughts to aid cognitive processing. However, there was no evidence of any discussion or group work of any kind.

(continued)

APPENDIX C (continued)

C.1: *Illustrative examples of codes for Category 2 (Dialogue)*

Code "1": dialogue slightly present (Crawford, 1976)

The primary concern of the research was the evaluation of transferability of CT skills acquired in language art classes to analyses of social studies problems. The experimental group instruction focused on reasoning, analyses, and students' ability to defend choices. Although the latter seemed to involve some dialogue, it was not made clear by the description of the procedure how much discussion argumentation and group work was involved.

Code "2": dialogue moderately present (D. L. Watson, Hagihara, & Tenney, 1999)

In the experimental condition, students met in small groups to discuss good and poor answers to their assignments. Discussion was the major instructional strategy, but discussion did not take place in all student activities, and so the code "2" was assigned.

Code "3": dialogue strongly present (Parkinson & Ekachai, 2002)

The intervention consisted of using the "Socratic Dialogue" method in an introductory public relations course. The Socratic approach was modeled on an introductory law course where students were asked to brief the cases described in the readings and then individual students were called on in class to describe the case and answer questions about it. These questions included identification of objectives, audiences, research, legal restrictions, and public relations tactics. The questions and comments from the instructor were intended to help the students see principles that underlay the public relations problems or solutions described in the cases read.

C.2: *Illustrative examples of codes for Category 3 (Authentic or Anchored Instruction)*

Code "0": authentic instruction not present (Schulhauser, 1990)

Fourth-grade students in the treatment group were divided into literary discussion groups consisting of six students each. Groups read a text book every 3 weeks over a period of 4 months and met twice weekly with their teacher to discuss the book content. There was no evidence of anchored or authentic instruction. The focus of the intervention was on individual study, discussions, and teacher explanations.

Code "1": authentic instruction slightly present (Zohar, Weinberger, & Tamir, 1994)

CT-oriented activities in the study included meta-cognitive discussions of the particular reasoning skills and how to use them. The main premise was that the same CT skills should be transferable and may be applied in many occasions and contexts, including various applied problems in biology.

Code "2": authentic instruction moderately present (Faryniarz, 1989)

In this study, the experimental group of community college students studied the topic of ecosystems using three simulator modules. These simulations addressed real-life problems of lake pollution analysis, wastewater quality management, and population dynamics.

Code "3": authentic instruction strongly present (Hill, 2000)

There was a high degree of solving applied problems. Educational psychology students tackled difficult and contentious issues in educational psychology based on real-life scenarios and begin to understand that the aim of inquiry is to further understand and create meaning in a world of conflicting perspectives and interpretations.

(continued)

APPENDIX C (continued)

C.3: *Illustrative examples of codes for Category 4 (Mentoring)*

Code "0": mentoring not present (Sungur & Tekkaya, 2006)

10th-grade biology students in the experimental group were taught with the problem-based approach, in which they worked with ill-structured real-world problems. Thus, the main focus of the intervention was on individual study, high in elements of authentic instruction, but with no evidence of modeling or coaching on the teacher's part.

Code "1": mentoring slightly present (Solon, 2007)

In this study, community college students in introductory psychology classes received a moderate infusion of generic CT material—approximately 10 hours of class time activity that included elements of individual consultations with the instructor.

Code "2": mentoring moderately present (Kemp & Sadoski, 1991)

The focus of the intervention was on individual study, discussion but also on teacher modeling. The instruction in world history included guidelines for the formation of generalizations demonstrated through teacher modeling, guided practice in groups with ongoing individualized teacher feedback, and the use of self-monitoring checklists.

Code "3": mentoring strongly present (Housen, 2001)

This study tested the Visual Thinking Strategies curriculum for primary school students. A strong emphasis was placed on one-on-one teacher-student interactions.

Note

This research was supported by grants from the Social Sciences and Humanities Research Council of Canada and Fonds de recherche sur la Société et culture, Québec. The authors express gratitude to and acknowledge the contribution of the following (names listed alphabetically): Edward C. Bethel, Katherine Hanz, Emery J. Hyslop-Margison, David Pickup, Anna Sokolovskaya, Rana Tamim, Vivek Venkatesh, Jonathan Woods, and Dai Zhang.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M., Tamim, R., & Zhang, D. A. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage one meta-analysis. *Review of Educational Research*, *78*, 1102–1134. doi:10.3102/0034654308326084

Adams, B. L. (1999). Nursing education for critical thinking: An integrative review. *Journal of Nursing Education*, *38*, 111–119.

Allen, M., Berkowitz, S., Hunt, S., & Louden, A. (1997, November). *Measuring the impact of forensics and communication education on critical thinking: A meta-analytic summary*. Paper presented at the annual meeting of the National Communication Association, Chicago, IL. (ERIC Document Reproduction Service No. ED413625)

Allen, M., Berkowitz, S., Hunt, S., & Louden, A. (1999). A meta-analysis of the impact of forensics and communication education on critical thinking. *Communication Education*, *48*, 18–30. doi:10.1080/03634529909379149

Alston, K. (2001). Re/thinking critical thinking: The seductions of everyday life. *Studies in Philosophy and Education*, *20*, 27–40.

Arrufat, O. (1997). *The role of exploration and critical decision making and problem solving in making life choices* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9813418)

Assessment and Learning Research Synthesis Group. (2003). *A systematic review of the impact on students and teachers of the use of ICT for assessment of creative and critical thinking skills*. Retrieved from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=109

Bangert-Drowns, R. L., & Bankert, E. (1990, April). *Meta-analysis of effects of explicit instruction for critical thinking*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Bernard, R. M., Zhang, D., Abrami, P. C., Sicoly, F., Borokhovski, E., & Surkes, M. (2008). Exploring the structure of the Watson-Glaser Critical Thinking Appraisal: One scale or many subscales? *Thinking Skills and Creativity*, *3*, 15–22. doi:10.1016/j.tsc.2007.11.001

Biesta, G. J., & Stams, G. J. (2001). Critical thinking and the question of critique: Some lessons from deconstruction. *Studies in Philosophy and Education*, *20*, 57–74.

Bloom, M., & Watt, D. (2003). *Solving Canada's innovation conundrum: How public education can help*. Ottawa, Ontario, Canada: Conference Board of Canada.

Bonk, C. J., Angeli, C. M., Malikowski, S. R., & Supplee, L. (2001, August). Holy cow: Scaffolding case based conferencing on the web with preservice teachers. *Ed at a Distance, 15*. Retrieved from http://209.151.89.205/usdla.org/public_html/cms/html/journal/AUG01_Issue/article01.html

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis* (Version 2). Englewood NJ: Biostat.

Case, R., & Wright, I. (1999). Taking seriously the teaching of critical thinking. In R. Case & P. Clark (Eds.), *The Canadian anthology of social studies: Issues and strategies for teachers* (pp. 179–189). Vancouver, British Columbia, Canada: Pacific Educational Press.

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

Crawford, J. A. (1976). *An investigation of the transfer of critical thinking skills from language arts to social studies* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 7703254)

Dewey, J. (1902). *The child and the curriculum*. Chicago, IL: University of Chicago Press.

Dewey, J. (1907). *The school and the curriculum*. Chicago, IL: University of Chicago Press.

Dewey, J. (1925). *Experience and nature*. Chicago, IL: Open Court.

Ennis, R. H. (1962). A concept of critical thinking: A proposed basis for research in the teaching and evaluation of critical thinking ability. *Harvard Educational Review*, *32*, 81–111.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and skills. In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9–26). New York, NY: W. H. Freeman.

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, *18*, 4–10. doi:10.3102/0013189X018003004

Ennis, R. H., & Millman, J. (1985). *Cornell critical thinking test*. Pacific Grove, CA: Critical Thinking Books & Software.

Facione, P. A. (1990a). *The California Critical Thinking Skills Test: College level: Interpreting the CCTST, group norms and sub-scores* (Technical Report No. #4). Millbrae: California Academic Press.

Facione, P. A. (1990b). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations*. Newark, DE: American Philosophical Association. (ERIC Document Reproduction Service No. ED315423)

Facione, P. A., Facione, N. C., & Giancarlo, C. A. F. (1996). *The California Critical Thinking Disposition Inventory*. Millbrae: California Academic Press.

Faryniarz, J. V. (1989). *The effectiveness of microcomputer simulators to stimulate environmental problem-solving with community college students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9002473)

Feuerstein, M. (1999). Media literacy in support of critical thinking. *Journal of Educational Media*, *24*, 43–54. doi:10.1080/1358165990240104

Follert, V. F., & Colbert, K. R. (1983, November). *An analysis of the research concerning debate training and critical thinking improvements*. Paper presented at the 69th annual meeting of the Speech Communication Association, Washington, DC.

Follman, J. (1987). Critical thinking instruments: Instruments of plenty or plenty of instruments? *Research Bulletin*, *20*, 71–75.

Follman, J. (2003). Reliability estimates of contemporary critical thinking instruments. *Korean Journal of Thinking and Problem Solving*, *13*, 73–81.

Foucault, M. (1997). What is critique? In S. Lotringer (Ed.), *The politics of truth* (pp. 41–81). Los Angeles, CA: Semiotexte. (Original work published 1978)

Foucault, M. (1998). The ethics of the concern of the self as a practice of freedom. In P. Rabinow (Ed.), *Ethics, subjectivity and truth* (pp. 281–301). New York, NY: Free Press. (Original work published 1984)

Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal*, *89*, 253–300. Retrieved from http://www.jstor.org/stable/1001805

Glaser, E. M. (1941a). *An experiment in the development of critical thinking* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 0156200)

Glaser, E. M. (1941b). *An experiment in the development of critical thinking*. New York, NY: Teachers College, Columbia University.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Govier, T. (1985). *A practical study of argument*. Belmont, CA: Wadsworth.

Halliday, J. (2000). Critical thinking and the academic vocational divide. *Curriculum Journal*, *11*, 159–175. doi:10.1080/09585170050045182

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

Hedges, L. V., Shymansky, J. A., & Woodworth, G. (1989). *A practical guide to modern methods of meta-analysis*. (ERIC Document Reproduction Service No. ED309952)

Hill, L. (2000). What does it take to change minds? Intellectual development of preservice teachers. *Journal of Teacher Education*, *51*, 50–62. doi:10.1177/002248710005100106

Housen, A. (2001, April). *Methods for assessing transfer from an art-viewing program*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA. (ERIC Document Reproduction Service No. ED457186)

Hyslop-Margison, E. J. (2003). The failure of critical thinking: Considering a virtue epistemology pedagogy. *Philosophy of Education Society Yearbook*, 2003, 319–326.

Jefferson, T. (1829). Letter to Marquis de la Fayette. In T. J. Randolph (Ed.), *Memoirs, correspondence, and private papers of Thomas Jefferson* (pp. 311–313). London, England: Shacknell & Baylis.

Kemp, S. G., & Sadoski, M. (1991). The effects of instruction in forming generalizations on high school students' critical thinking in world history. *Literacy Research and Instruction*, *31*, 33–42. doi:10.1080/19388079109558069

Kurfiss, J. G. (1988). *Critical thinking: Theory, research, practice, and possibilities* (ASHE-ERIC Higher Education Report No. 2). Washington, DC: Association for the Study of Higher Education. (ERIC Document Reproduction Service No. ED304041)

Lipman, M. (1991). *Thinking in education*. Cambridge, England: Cambridge University Press.

Marshall, J. D. (2001). A critical theory of the self: Wittgenstein, Nietzsche, Foucault. *Studies in Philosophy and Education*, *20*, 75–91.

McMillan, J. H. (1987). Enhancing college students' critical thinking: A review of studies. *Research in Higher Education*, *26*, 3–29. doi:10.1007/BF00991931

McMurray, M. A., Beisenherz, P., & Thompson, B. (1991). Reliability and concurrent validity of a measure of critical thinking skills in biology. *Journal of Research in Science Teaching*, *28*, 183–191. doi:10.1002/tea.3660280208

McPeck, J. (1981). *Critical thinking and education*. Toronto, Ontario, Canada: Oxford University Press.

Norris, S. P. (1985). Synthesis of research on critical thinking. *Educational Leadership*, *42*, 40–45.

Parkinson, M. G., & Ekachai, D. (2002). The Socratic method in the introductory PR course: An alternative pedagogy. *Public Relations Review*, *28*, 167–174. doi:10.1016/S0363-8111(02)00123-6

Paul, R. W. (1985). The critical thinking movement: A historical perspective. *National Forum: Phi Kappa Phi Journal*, *42*, 2–3.

Paul, R. W. (1990). *Critical thinking: What every person needs to survive in a rapidly changing world*. Santa Rosa, CA: Foundation for Critical Thinking.

Paul, R. W., & Binker, A. J. A. (1990). Strategies: Thirty-five dimensions of critical thinking. In A. J. A. Binker (Ed.), *Critical thinking: What every person needs to survive in a rapidly changing world* (pp. 305–349). Rohnert Park, CA: Centre for Critical Thinking and Moral Critique, Sonoma State University.

Paul, R. W., Elder, L., & Bartell, T. (1997). *Study of 38 public universities and 28 private universities to determine faculty emphasis on critical thinking on instruction*. Retrieved from http://www.criticalthinking.org/pages/center-for-critical-thinking/401

Pellegrino, A. M. (2007). *The manifestation of critical thinking and metacognition in secondary American history students through the implementation of lesson plans and activities consistent with historical thinking skills* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3282653)

Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research*, *42*, 237–249. doi:10.1080/001318800440579

Quin, R., & McMahon, B. (1993). Monitoring standards in media studies: problems and strategies. *Australian Journal of Education*, *37*, 182–197. doi:10.1177/000494419303700206

Rose, M. M. (1997). *Critical thinking skills instruction for postsecondary students with and without learning disabilities: The effectiveness of icons as part of a literature*

curriculum (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9806188)

Rousseau, J. J. (1762). *Émile*. Amsterdam, Netherlands: Jean Néaulme.

Royalty, J. (1995). The generalizability of critical thinking: Paranormal beliefs versus statistical reasoning. *Journal of Genetic Psychology*, *156*, 477–488. doi:10.1080/00221325.1995.9914838

Sá, W. C., Stanovich, K. E., & West, R. F. (1999). The domain specificity and generality of belief bias: Searching for generalizable critical thinking skill. *Journal of Educational Psychology*, *91*, 497–510. doi:10.1037/0022-0663.91.3.497

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research, 84*, 328–364. doi:10.3102/0034654313500826

Schulhauser, C. E. (1990). *The effects of literary discussion groups on students' critical thinking ability and attitude toward reading* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9131095)

Scriven, M., & Paul, R. (1996). *Defining critical thinking: Critical thinking as defined by the National Council for Excellence in Critical Thinking, 1987*. Retrieved from http://www.criticalthinking.org/pages/defining-critical-thinking/766

Siegel, H. (1988). *Educating reason: Rationality, critical thinking, and education*. New York, NY: Routledge.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, *57*, 193–216. doi:10.1037/h0054367

Skinner, B. F. (1959). *The cumulative record*. New York, NY: Appleton-Century-Crofts.

Smith, G. (2002). Are there domain-specific thinking skills? *Journal of Philosophy of Education*, *36*, 207–227. doi:10.1111/1467-9752.00270

Solon, T. (2007). Generic critical thinking infusion and course content learning in introductory psychology. *Journal of Instructional Psychology*, *34*, 95–109.

Sungur, S., & Tekkaya, C. (2006). Effects of problem-based learning and traditional instruction on self-regulated learning. *Journal of Educational Research*, *99*, 307–317. doi:10.3200/JOER.99.5.307-320

Thayer-Bacon, B. (2000). *Transforming critical thinking: Thinking constructively*. New York, NY: Teachers College Press.

Tsui, L. (2002). Fostering critical thinking through effective pedagogy: Evidence from four institutional case studies. *Journal of Higher Education*, *73*, 740–763. Retrieved from http://muse.jhu.edu/journals/journal_of_higher_education/v073/73.6tsui.pdf

VanTassel-Baska, J., Zuo, L., Avery, L. D., & Little, C. A. (2002). A curriculum study of gifted-student learning in the language arts. *Gifted Child Quarterly*, *46*, 30–44. doi:10.1177/001698620204600104

Watson, D. L., Hagihara, D. K., & Tenney, A. L. (1999). Skill-building exercises and generalizing psychological concepts to daily life. *Teaching of Psychology*, *26*, 193–195. doi:10.1207/S15328023TOP260306

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser critical thinking appraisal*. San Antonio, TX: PsychCorp.

Woolfolk, A. E. (1998). *Educational psychology*. Boston, MA: Allyn & Bacon.

Yang, Y.-T. C., Newby, T., & Bill, R. (2008). Facilitating interactions through structured web-based bulletin boards: A quasi-experimental study on promoting learners' critical thinking skills. *Computers & Education*, *50*, 1572–1585. doi:10.1016/j.compedu.2007.04.006

Zohar, A., & Tamir, P. (1993). Incorporating critical thinking into a regular high school biology curriculum. *School Science and Mathematics*, *93*, 136-140. doi:10.1111/j.1949-8594.1993.tb12211.x

Zohar, A., Weinberger, Y., & Tamir, P. (1994). The effect of the biology critical thinking project on the development of critical thinking. *Journal of Research in Science Teaching*, *31*, 183–196. doi:10.1002/tea.3660310208

Authors

PHILIP C. ABRAMI, PhD, is a Concordia University Research Chair and the Director of the Centre for the Study of Learning and Performance, Concordia University, LB 589-2, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8; email: *abrami@education.concordia.ca*. His current work focuses on research integrations and primary investigations in support of applications of educational technology in distance and higher education, in early literacy and numeracy, and in the development of higher order thinking skills and learning strategies.

ROBERT M. BERNARD, PhD, is a professor of Education and Systematic Review Team Leader at the Centre for the Study of Learning and Performance, Concordia University, LB-578-1, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8; e-mail: *bernard@education.concordia.ca*. His research interests include distance and online learning, instructional technology, research design, and statistics and meta-analysis.

EUGENE BOROKHOVSKI, PhD, works as Systematic Reviews Project Manager at the Centre for the Study of Learning and Performance of Concordia University, LB-581, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8; e-mail: *eborokhovski@education.concordia.ca*. His professional expertise and research interests include cognitive and educational psychology, language acquisition and methodology of systematic reviews, and meta-analyses.

DAVID I. WADDINGTON is an associate professor in the Department of Education at Concordia University, LB-545-5, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8; e-mail: *waddington@education.concordia.ca*. His principal interests include Deweyan conceptions of citizenship as well as new and critical approaches toward technology in education.

C. ANNE WADE, MLIS, has been a manager and information specialist at the Centre for the Study of Learning and Performance for over 20 years, Concordia University, LB-581, 1455 de Maisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8; e-mail: *anne.wade@education.concordia.ca*. She has also served as a sessional lecturer in the Department of Education/Information Studies at Concordia for the past two decades. Her expertise is in information storage and retrieval, educational technology and research methods. She is the former president of the Quebec Library Association and the Eastern Canada Chapter of the Special Libraries Association; the Information Specialist of the Education Coordinating group, international Campbell Collaboration; and an associate of the Evidence Network, United Kingdom.

TONJE PERSSON, PhD candidate, is in the Psychology Department at Concordia University, 7141 Sherbrooke West, Montreal, QC H4B 1R6, e-mail: *tj_perss@hotmail.com*. Her research interests lie in the field of clinical psychology, psychology of creativity and social self–identity, and the methodology of systematic reviews.