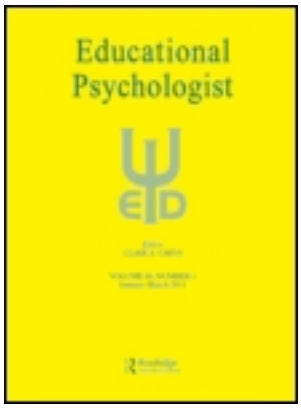


Educational Psychologist
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hedp20

Learning Through Case Comparisons: A Meta-Analytic Review

Louis Alfieri [a], Timothy J. Nokes-Malach [a] & Christian D. Schunn [a]
[a] Learning Research and Development Center, University of Pittsburgh
Version of record first published: 20 Apr 2013.

EDUCATIONAL PSYCHOLOGIST, *48*(2), 87–113, 2013

ISSN: 0046-1520 print / 1532-6985 online
DOI: 10.1080/00461520.2013.775712

Routledge
Taylor & Francis Group

Learning Through Case Comparisons: A Meta-Analytic Review

Louis Alfieri, Timothy J. Nokes-Malach, and Christian D. Schunn
Learning Research and Development Center
University of Pittsburgh

Over the past 20 years, there has been much research on how people learn from case comparisons. This work has implemented comparison activities in a variety of different ways across a wide range of laboratory and classroom contexts. In an effort to assess the overall effectiveness of case comparisons across this diversity of implementation and contexts and to explore what variables might moderate learning outcomes, we conducted a meta-analysis of 57 experiments with 336 tests. Random effects analyses of the 336 tests revealed that case comparison activities commonly led to greater learning outcomes than other forms of case study including sequential, single case, and nonanalogous, as well as traditional instruction and control (d = .50), 95% CI [.44, .56]. Of 15 potential moderators, four variables were found to reliably moderate the effectiveness of case comparisons: the objective of the comparison, the presentation of a principle, the content, and the lag between the comparison and testing. Asking learners to find similarities between cases, providing principles after the comparisons, using perceptual content, and testing learners immediately are all associated with greater learning. We conclude with a discussion of the theoretical and practical implications of these results for cognitive theory and classroom practice.

Analogies have long been recommended as an effective instructional strategy to be incorporated into classroom pedagogy to improve student learning (Gee, 1978; Lewis, 1933; Schustack & Anderson, 1979; Webb, 1985; Weller, 1970). They continue to be recommended as pedagogical tools that can help teachers convey new, complex information by drawing parallels to more familiar concrete examples and cases (Buehl, 2008; Druit, 1991; Iding, 1997; Loewenstein & Thompson, 2000; Richland, Zur, & Holyoak, 2007; Siegler et al., 2010; Treagust, Druit, Joslin, & Lindauer, 1992). These recommendations are typically based on the large research literature that has shown that analogies can support problem solving, learning, creativity, and explanation, among other educationally relevant practices and outcomes (National Research Council, 2000). For example, prior work has shown that analogies can help students focus on the key features of new information (e.g., Christie & Gentner, 2010; Cummins, 1992; Gentner, Loewenstein, & Thompson, 2003; Gick & Holyoak, 1983; Mason, 2004; Mundy, Honey, & Dwyer, 2009) and generate appropriate inferences for that content (e.g., Clement & Gentner, 1991; Kurtz, 2005; Kurtz, Miao, & Gentner, 2001). Analogies can also help students acquire a more abstract understanding of the to-be-learned content. Through analogical comparison, students can extract the features that are in common, and the resulting representation can then facilitate transfer to new examples and situations that share the same underlying structure but differ in the specific features (Chen & Daehler, 1989; Gentner et al., 2003; Gerjets, Scheiter, & Schuh, 2008; Gick & Holyoak, 1980, 1983; Loewenstein, Thompson, & Gentner, 2003; Richland & McDonough, 2010; Schuh, Gerjets, & Scheiter, 2005).

The two ways in which analogies are typically incorporated into instruction are through direct instruction and students' activities such as problem solving and case comparison. However, as many instructors know, analogies used in instruction are not always effective in promoting the learning of the target content. For example, Richland and colleagues (2007) illustrated through cross-cultural comparisons that analogies used in direct instruction need to be accompanied by appropriate supportive cues in order for them to be effective. Specifically, they emphasized that if an instructor intends to use an analogous comparison (e.g., of equations to balance scales), it is important for the instructor to help students see the corresponding features that make them

Correspondence should be addressed to Louis Alfieri, Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260. E-mail: alfieri@pitt.edu

analogous (the equal symbol like a fulcrum, numerical expressions like weights, etc.). Cueing students to why and how the two are analogous by highlighting those corresponding features, instead of assuming students will fully recognize the analogy independently, makes the analogy more likely to be effective. If such cues are not presented, analogical reasoning might result in not only less learning but also learning something other than what was intended. It follows that instructional support and scaffolding is also likely to be helpful during students' case comparison activities. Although research on case comparisons has explored specific ways in which case comparisons can facilitate learning, no prior work has broadly examined the diversity of ways that case comparisons have been implemented and tested in different contexts. What do supportive cues look like when students are making sense of analogous comparisons in classroom activities? More importantly, which of these factors are critical for students' success in using such analogies? The current meta-analysis examines whether case comparison tasks consistently lead to learning and which factors facilitate or impede learning from such exercises.

ANALOGY AND THE COMPARISON OF CASES

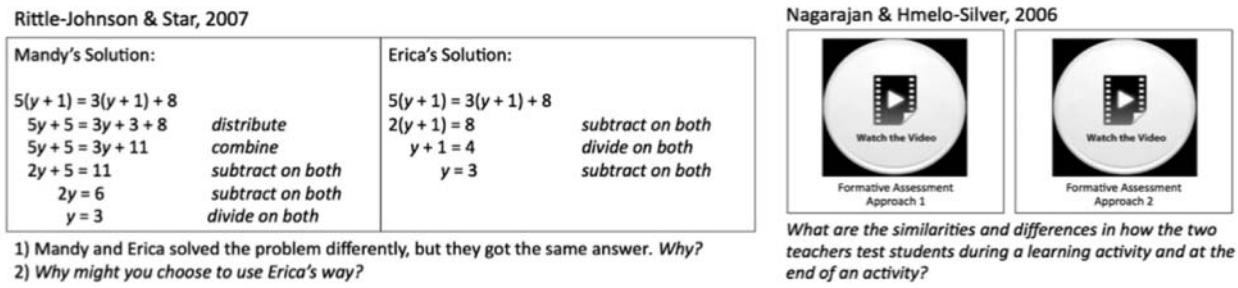
Analogical thinking is the ability to identify similar features and relationships between those features across cases or examples (Gentner, 2003, 2010). We define cases as situations, events, or things (concepts, procedures, etc.), which have been experienced by, depicted for, or described to the person who is drawing the analogy (or comparing the cases). This ability is hypothesized to be a defining and distinguishing characteristic of human cognition (e.g., Gentner, 2010; Goldstone, Day, & Son, 2010; Hofstadter, 2001; Holyoak, 2012; Rittle-Johnson & Star, 2011). Indeed, analogical processes occur in a variety of cognitive activities including perception, categorization, explanation, and problem solving (Goldstone & Barsalou, 1998; Hofstadter, 2001; James, 1890). For example, when human information processing is said to be similar to the processing of a personal computer, or atomic structure is said to be similar to that of the solar system, we are utilizing our ability to recognize the similarities in the relationships between the features (e.g., objects, events, agents, and consequences) of these phenomena.

An example of such comparisons can be found in the categorization of objects during word learning (Namy, Gentner, & Clepper, 2007). Namy and colleagues found that 4-year-olds show better learning when provided with two exemplars (two hats) for a given label (*blicket*) than with one. Through comparison of the cases, the learner can determine how the cases might be related and the features that they share (Boroditsky, 2007; Gentner & Markman, 1997; Rittle-Johnson & Star, 2011). After the cases are aligned, the learner has the opportunity to construct a more abstract understanding by focusing on the underlying relations across the cases (Gentner, 2010; Holyoak, 2012).

In the situation designed by Namy and colleagues, two exemplars of a *blicket* (baseball cap and fedora) allowed those children to compare items and to notice the features that they share (oval in shape, similar in color, the fronts of the *blickets* protrude downward, both are worn on the head). Through aligning these common features, children were afforded the opportunity to recognize the interrelations of a *blicket*'s features and to infer/build category abstractions (Gentner & Namy, 1999; Namy et al., 2007). These category abstractions are schemas. Schemas are cognitive representations of the structures or relational systems shared by cases/exemplars as have been highlighted through structural alignments (Gentner, 1983, 2010; Gick & Holyoak, 1983; Markman & Gentner, 1993; Namy & Gentner, 2002). After the schema (abstract representation) has been acquired, learners can then use this knowledge to infer missing information in a new situation or for a new case (Gentner, 2010). During the subsequent posttest, the children in the two-exemplar condition more readily extended the label of *blicket* to a sombrero (a type of hat) than to an igloo (similar in shape and color but not function), whereas children in the one-exemplar condition still demonstrated uncertainty as to the meaning of *blicket*. Thus, the knowledge acquired from the comparison went beyond a surface (perceptual/object) representation and led to a deeper relational understanding of the category (conceptual), which in turn afforded inferences to new members of the category that had different surface features but similar relations.

IMPLEMENTATIONS WITHIN CLASSROOMS

Many researchers have begun to examine the potential benefits of explicit case comparisons for academic learning across a variety of contexts (e.g., Gadgil & Nokes, 2009; Gentner, Loewenstein, Thompson, & Forbus, 2009; Mason, 2004; Michael, Klee, Bransford, & Warren, 1993; Nagarajan & Hmelo-Silver, 2006; Rittle-Johnson & Star, 2011; Schwartz & Bransford, 1998). The following two examples, presented in Figure 1, were selected to illustrate the range of implementations of comparison activities used in the literature. Rittle-Johnson and Star (2007) asked seventh-grade students to either compare or study sequentially two different solution procedures to the same equation/problem. The target principle of the content (i.e., the method of using composite variables to solve the problems) was not provided to students. However, students were familiarized with the conventional method, which the teacher presented before and during study. Students in the comparison condition compared the two solution procedures and then explained why both approaches obtained the same answer and why one approach was preferable. Not only were the important procedural steps (features) of each example labeled, but supporting questions also directed the learner's attention to how the solutions



Variable	Rittle-Johnson & Star, 2007	Nagarajan & Hmelo-Silver, 2006
Content of subject matter	•conceptual understandings of problems •procedural know-how to solve	•conceptual understanding of formative assessments •procedural understanding of best administrations
The presentation of a unifying principle	•taught conventional solution method (e.g., Mandy's) before and during study but not provided with the principle of how to solve using composite variables (e.g., Erica's)	•no provided principle of when/how to administer assessments
An outline of cases' key features	•labeled solution steps for each case	•key features of assessments determined by learners
The objective of the task	•to find similarities and differences between the solutions	•to find similarities and differences between approaches to formative assessment
Lag between study and test phases	•tested on a subsequent day	•tested on the same day and a subsequent day

FIGURE 1 Two examples of case comparison tasks within the sample.

differed. These comparisons facilitated learning both how to consider/reason about equations containing composite variables (conceptual) and methods for solving such multistep problems (procedural).

Nagarajan and Hmelo-Silver (2006) asked undergraduates enrolled in an introductory educational psychology course to compare two videos of teachers administering formative assessments to their classes and then to explain the similarities and differences in how the two teachers assessed student learning. Students in the other conditions were asked to watch the videos and then answer either neutral questions or those probing their affective or metacognitive states. In this situation, learners were not provided with a general principle as to how formative assessments should be administered, nor were they provided with the key features explicitly labeled within the cases.

As shown in Figure 1, both of these studies intended to convey both conceptual and procedural content and both studies asked learners to consider both similarities and differences between the cases. Overall however, Rittle-Johnson and Star's design provided learners with a more guided case comparison by having outlined the key features within the cases and used directive questions to guide learners' attention to the differences between solutions. In comparison, Nagarajan and Hmelo-Silver required learners to recognize best practices for formative assessments without providing a principle before the case comparison or explicitly labeling key features for learners to look for while watching the videos. These two studies illustrate some of the ways that classroom-based research has varied in the implementation of case comparison. We consider these implementation features as "process" variables because we hypothesize their manipulation influences the cognitive processes involved in comparing the cases.

Other process variables include the types of cases (e.g., the minimal worked examples of the first study vs. the richly detailed video recordings of the second), the types of instructions that introduce learners to the task, and the scaffolds in place to assist learners (e.g., the guided questions for students to answer in the first vs. the general question posed in the second). Other variables can be considered context variables: the domain of the subject matter (math vs. science), the age of participants (seventh graders vs. undergraduates), and the setting (both examples are from classroom-based investigations as opposed to laboratory-based studies).

Finally, some variables can be considered measurement variables: the lag between study and test phases (testing conducted on a subsequent day vs. testing on the same day and then again on a subsequent day), the type of dependent measure of learning (i.e., near vs. far transfer), and the type of learning condition to which case comparisons are being

compared (sequentially studying worked examples or cases vs. more traditional reading with questions that do not prompt comparisons).

A meta-analysis is particularly useful to determine how robust effects are for case comparisons across process, context, and measurement variations. In addition, a meta-analysis would enable an exploration of which variables moderate learning outcomes and which types of learning scenarios are associated with the largest effects. That is, what differentiates the better case comparisons from the rest?

A MODEL OF THE LEARNING PROCESS

To conceptualize more formally the examined case comparison variables, we describe a process model of case comparisons. This model outlines a general set of cognitive and behavioral processes hypothesized to occur as learners engage in case comparison. The model is informed by prior theory and research on analogy making and case comparison and was used as a top-down theoretical tool to focus our selection of the most promising features of case comparison activities that are likely to affect learning outcomes. Such a model informs both the selection of relevant potential moderators and the interpretation of those variables should they in fact be found to moderate learning outcomes.

Our process model, shown in the center of Figure 2, assumes that the presentation of cases is simultaneous. The model proposes only five steps of the process, but each contains several variables that could lead to different learning outcomes. Although many potentially important details of the task stand out, we limited which variables we coded for by considering the steps within the process—the initiation of the comparison, the effortful search for commonalities, the alignment of the target features, and the retrieval of the analogical case for alignment with a new case for poststudy measure. We chose those variables that seemed most likely to affect if/how learners would reach these steps and their learning outcomes. Most of these variables are outlined for the examples within Figure 2.

On each side of our model in Figure 2, we have outlined a case comparison task (Gentner & Namy, 1999, Experiment 2; Star & Rittle-Johnson, 2009) to draw attention to the different implementation variables that could be influential in successful completion of the task. On the left side of the figure is a word/category learning case comparison (Gentner & Namy, 1999, Experiment 2).[1] The intention was for these children to recognize both cases as means of transportation and therefore facilitate the selection of the skateboard as a member of the same category. It stands in contrast to the task of Star and Rittle-Johnson in which they asked fifth- and sixth-grade students to learn about approximation strategies by comparing different methods of approximating in the form of worked examples.

As a formal learning task, case comparisons typically begin with a *prompt to compare* (Step 1). After initial instructions, the learner's task of comparing begins with the focused, *effortful search for commonalities* (Step 2; Seifert, McKoon, Abelson, & Ratcliff, 1986). When beginning such a search, the type of instructions, the objective, the type of cases, and the learners' prior experience could influence how successful they are in finding common features and relations across the cases. The *instructions* to compare cases might be only a general prompt (e.g., *See how these are the same kind of thing?*), or might be more guided (e.g., *How is Allie's way similar to Claire's way? Would Claire's way give a different estimate for 21*43 than Allie's way?*). Notice how the latter example draws the learners' attention to the important features of the cases (i.e., how both accomplish approximation and the relationship between the procedure and the outcome). Also, notice that the *objective* of the comparison in this example is to identify both the similarities and the differences. In contrast, the objective in the category learning example was to find only similarities. These two variables are hypothesized to affect both which and how many features are identified across the cases.

The *type* of cases and the *experience* level of learners can influence the amount of information to be considered. Cases that consist of many details may include spurious commonalities that may interfere with finding the target features and relations (structural or relational similarities; Gentner, 2010) or lead to learning irrelevant information (e.g., superficial matching of similar objects; Markman & Gentner, 1993). This rich type of case is hypothesized to entail a different search process than would be expected when cases only present the target details. For example, in the category learning task, rich cases could have led learners to focus on aligning surface, perceptual similarities (e.g., circles and metal frames; leading them to eyeglasses as a categorical match) instead of functional/conceptual similarities (e.g., modes of transport; leading them to the skateboard as a categorical match). In such a scenario, the learner would have missed the deeper, conceptual connection between the cases (that bicycles and tricycles are both vehicles) because the richness of the cases introduced ambiguity as to which were important similarities (perceptual *or* conceptual).[2] In contrast, the cases in the approximation comparison were minimal because the learners needed to consider only the solution methods' steps and their effects/products, both of which were explicitly provided. A second factor that could affect how much and which information in the cases is considered is the learner's prior

[1]Notably, only the comparison condition under which children were not provided with a word/label for the category was included, because it included an explicit prompt for children to compare (i.e., *See how these are the same kind of thing?*). Thus, this condition involves category learning and not word learning.

[2]Of course, the perceptual similarities/distances between cases also would factor into this possibility (Namy et al., 2007).

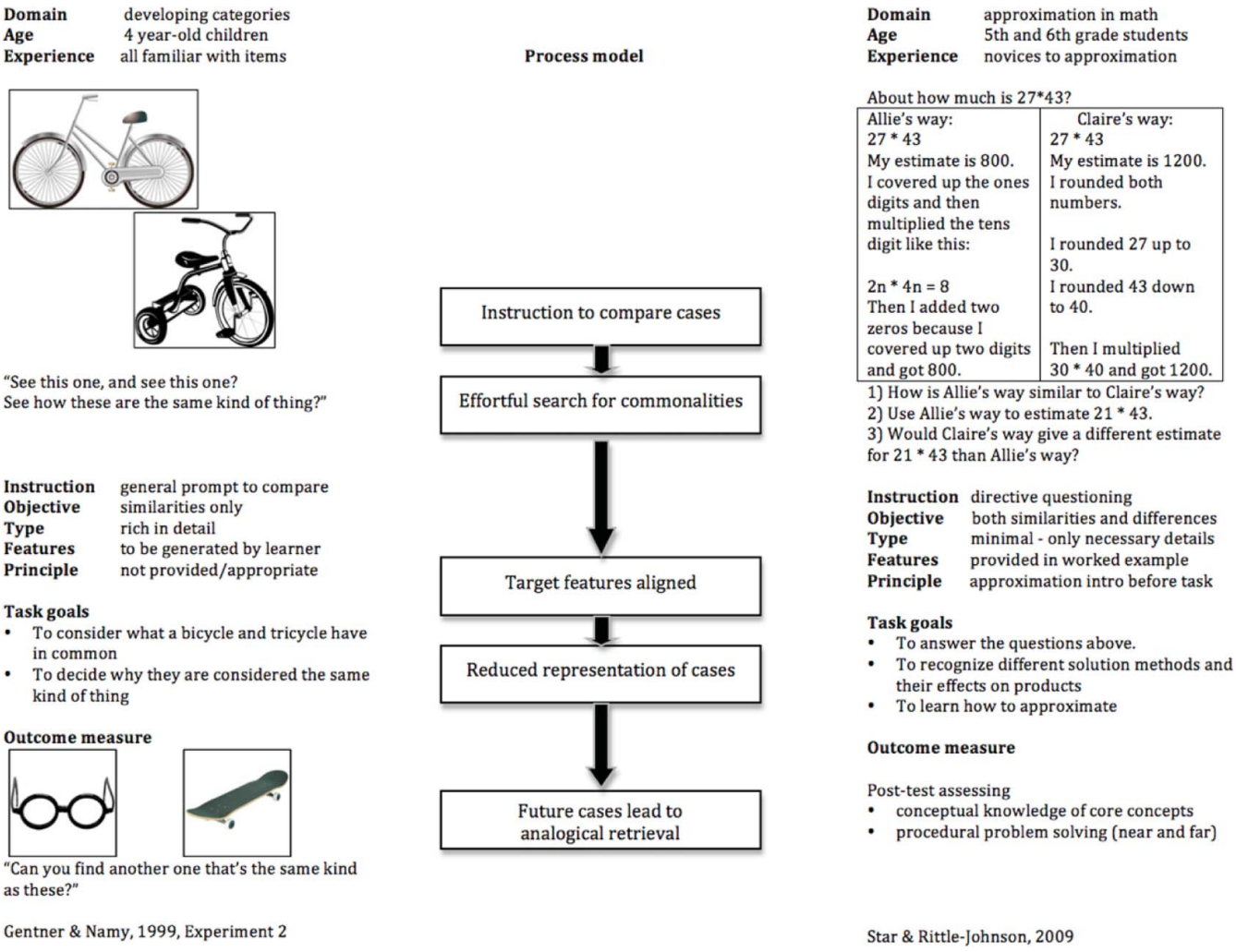


FIGURE 2 Our process model flanked by two example tasks from the literature (color figure available online).

experience with the to-be-learned content. The learners in the category learning study were generally familiar with the items (tested after participation to ensure they could identify items by name or function), and this familiarity could have affected which features participants noticed and attended to. Similarly, learners comparing approximation methods, although not familiar with all approximation methods studied, were able to do some basic rounding. That small base of familiarity could have helped learners recognize the newer methods and answer questions comparing them. Thus, more experience may help to focus a search for those features that have been previously represented in the task or domain as relevant (e.g., Rittle-Johnson, Star, & Durkin, 2009).

Toward the end of the effortful search for commonalities, the next pivotal step in the process is *the alignment of the target features* (Step 3)—the relevant alignable details of the cases. Whether target features have been explicitly provided or principles have been presented both could affect this step. When *features* of at least one case are labeled explicitly, the learner can use that list to identify the corresponding features in the other case and align the two cases in that way. When features are not provided, other components of the learning task (more directive instructions, only asking learners to find similarities, providing minimal cases) might replace or supplement that support.

The examples within Figure 2 illustrate the differences between whether or not features are provided. The features within the approximation example (multiplying only the tens digit, rounding, each strategy's effects, its product) were both displayed numerically and explained at each step. This explicit provision of features could have influenced both learners' search for commonalities and their success in aligning the target features. In contrast, the features within the categorization example (wheels, self-propelled, etc.) were not provided, which had the potential of making the search for commonalities difficult and consequently required the alignment of target features to be all the more selective.

If feature alignment is an important intermediary step for successful learning from comparisons, then providing features to at least one case should help to facilitate the

identification and alignment of those features in the second case. However, if the constructive aspect of having to identify and infer which features of the cases should be aligned is a useful learning activity, then providing such features might oversimplify the learning task. It is important to note that it might depend on whether learners construct explanations for why those features align (Chi, 2009). Again, the experience level of the learner might also factor into whether it is better to present at least some of the features. Familiarity might reduce the need for features to be highlighted for learners because their experience has prepared them to attend to the relevant features.

Being presented with a *principle* that connects the two cases can affect both search and alignment processes and the resulting representations depending on if and when it was presented. A principle gives learners a verbal description that connects cases to some larger concept or procedure. In the approximation example, a principle was provided before the task during a 10-min introduction to approximation during which learners were introduced to the idea of estimation and the strategy/procedure of *trunc* (truncation of multiplicands). Knowing that estimation is the process of getting an approximate answer (possible through a number of methods) could help students identify the commonalities between the cases (both are strategies/methods of approximating). In contrast, the categorization example did not provide learners with a principle—perhaps because none seemed appropriate without providing the explanation of the category, thereby making the task too easy. However, providing a principle before could also potentially increase the learner's cognitive load (Paas, Renkl, & Sweller, 2010) because they would need to maintain it within working memory while interpreting and comparing the cases. It is also possible that presentation before could change what is being compared if doing so leads learners to compare each case sequentially to the principle instead of the cases to one another.

Alternatively, the principle could be presented after the case comparison task. This order of activities could affect how students process the principle in relation to the case features. By comparing cases before reading the principles, students could learn about the key features of the target principle, which might better prepare them to understand more deeply the principle when they are presented with it. Holyoak (2012) suggested that learners' schemas following comparisons might only be tentatively considered. Under this theory, providing a principle after the task could modify and/or reinforce more appropriate schematic representations. In such instances, the comparison could be considered a preparation for future learning (Bransford & Schwartz, 1999). Again as exemplified by the categorization example (Gentner & Namy, 1999, Experiment 2), a principle might not have been provided because learners were to construct that principle for themselves through comparison. In such cases, the principle could not easily be incorporated into instruction (e.g., Cummins, 1992; Gick & Holyoak, 1983, Experiment 4; Graham, Namy, Gentner, & Meagher, 2010, Experiment 1a). However, there might be other cases for which there are no high-level principles (e.g., Kurtz, 2005, Experiments 1 & 2; Markman & Gentner, 1993, Experiments 1a, 1b, 2, & 3; Mundy et al., 2009, Experiments 1 & 2).

Following the alignment of target features, we propose that there is a *reduction in the representation of cases* (Step 4). Prior work has shown that analogy making across multiple cases affords the opportunity of acquiring a schematic representation that includes the common features of the cases but does not include features unique to each given case (Gick & Holyoak, 1983; Hummel & Holyoak, 1997, 2003). The result is a representation that is more easily transferable to future cases because the unimportant, case-specific details can be removed. An alternative hypothesis is that comparisons lead learners to develop deeper, concrete understandings of the individual cases. Comparing cases could highlight the underlying relations that could help the learner make sense of the purpose of the specific features in the cases and result in an improvement in the comprehension and understanding of those individual cases (especially for initially unfamiliar cases; e.g., the cases of convergence analogous to the radiation problem; Gick & Holyoak, 1983). Analysis of effect sizes for near and far transfer might help to begin sorting out these two alternatives. If comparisons lead to abstracted schemas, then we should see larger effects for far transfer than near transfer (when such tasks are compared to other learning conditions) because more abstract representations of problems should make them more readily transferable (National Research Council, 2000). If, however, comparisons lead to deeper concrete understandings of cases, we should see benefits only for near transfer to similar cases. Thus, whether the dependent measure is of near or far transfer is a concern connected both to the step of retaining a reduced representation of cases and to the step of *retrieving analogs when confronted with future cases* (Step 5).

CATEGORIES OF VARIABLES THAT COULD MODERATE LEARNING OUTCOMES

Our general model illustrates many important variables that may influence processes within case comparisons. Many of these variables can be considered manipulations of the amount of instructional support provided to learners: the instructions (i.e., the type of instructions, the objective of the task) and the information provided to learners (the alignable target features of cases, the unifying principle). We consider these to be process variables. Others are contextual variables: the setting (class or laboratory), the content of the to-be-learned material (conceptual, procedural, or perceptual), the age of participants (children, adolescents, or adults), and the domain of the subject matter (science, math, or other).

The contextual variables of setting and content deserve further explanation for how they could moderate learning outcomes. Whether a study is conducted in the classroom

or laboratory may affect the success of the manipulation or intervention. Laboratory studies may reveal larger effects because there may be more control over the critical manipulation and less variation in how the comparison task is implemented compared to classroom settings in naturalistic contexts. Laboratory studies may also use more precise measurement instruments (i.e., artificial tasks that focus on testing specific features) and control for other variables such as prior knowledge. Setting may also impact learning outcomes by capturing student motivation to engage in the comparison. For example, the learning outcomes from comparing cases in a classroom-based study might affect the students' grade for the course (or at least is information required for the course), whereas task performance in a laboratory setting typically does not have consequences for students' grades. Instead, students' participation in the lab usually results in participation credit or payment. These different reward structures may affect students' intrinsic or extrinsic motivation for completing the case comparisons. Furthermore, these different settings may trigger different motivational goals. For example, classrooms may be more likely to involve mastery goals (learning the content) rather than performance goals (get the credit/money), which may influence the effectiveness of learning-oriented tasks (Belenky & Nokes-Malach, 2012).

In regard to content, case comparisons have been used as learning tasks for perceptual, conceptual, and procedural subject matter. Although some details within the process are likely to differ across these different types of comparison, our model attempts to capture the similarities in process across content type. However, because comparisons highlight features (Gentner, 2010), case comparisons with perceptual content might yield the greatest effects on learning. Our last two context variables (domain and age) not only address issues of generalizability but also are associated with the process. Domain might moderate findings because of the complex interrelationships between the subject matter and other variables (content, dependent measures, etc.). For example, it seems likely that studies within math will intend to convey formal procedures more than science, which will more often focus on conceptual knowledge. Age could be related to the amount of support required to make comparisons effective, with younger learners requiring more support than older learners.

It may also be important to consider measurement variables: the lag between study and test phases, the dependent measure used, and the learning conditions to which case comparisons were compared. These variables do not influence learning per se but may affect the sizes of the case comparison effects found because they often influence the magnitude of learning effects. For example, one might expect larger effect sizes for studies comparing case comparisons to control situations (learners were not provided with any study phase) than comparing case comparisons to sequential conditions (learners studied the same cases sequentially).

Aside from being considered variables of measurements, these variables can also be used in conjunction with process variables to begin to inform the interpretation of results and implications for theory. When a measurement variable is confounded with a process or context variable across studies (e.g., if most studies that provided the principle before also compared case comparison tasks to control conditions) and both influence effect sizes, this can skew estimates of the impact of both variables. To investigate such potential confounds within meta-analysis, we had to create a new method of analysis.

A METHODOLOGICAL INNOVATION IN META-ANALYSIS

As Lipsey (2003) pointed out, a great risk within meta-analysis is not to recognize when variables are confounded. For example, if we were to find no difference between the effect sizes on learning with rich or minimal cases, we might conclude that both are equally effective for learning. However, that would be an incomplete explanation if the type of cases was confounded with the type of instructions presented to learners (i.e., general prompts vs. guided questioning). It could be that minimal cases with general prompts are equivalent to rich cases with guided questioning and that the interaction of the type of cases (rich vs. minimal) and instruction type (prompted vs. guided) is driving the null overall effect. If the sample included more instances of all the possible combinations (rich with prompts, minimal with prompts, rich with guided questioning, minimal with guided questioning), analyses might have revealed differences.

To tease apart the effects of partially confounded moderators, we propose using a simple procedure of examining to what extent each moderator's effect is maintained across levels of other confounded/contingent variables. This new method of analysis also highlights which variables need to be investigated further because they either are always confounded with other variables across studies or suffer from low power in critical distinguishing cells.

In our new methodological approach, we first examine whether the sample has confounded variables through contingency analyses. For each pair of moderator variables identified as significantly confounded, we examine how the moderator effects hold up across levels of the confounded variable. For example, if this method found a confound between the type of case and the instruction type, we would then investigate whether all combinations of types of cases and instruction types yielded functionally equivalent effect sizes (i.e., the overall trend). If they did not, we then would be able to report which cells (conditional combinations of the two moderator variables in question; rich/prompted, rich/guided, minimal/prompted, minimal/guided) do not show this effect and/or which might suffer from low power (i.e., too few studies with those conditional combinations).

The new method also provides a general sense of where most of the variation in moderators occurs across studies and which variables future work should examine. For example, two of the previous example studies (Nagarajan & Hmelo-Silver, 2006; Rittle-Johnson & Star, 2007) asked learners to find similarities and differences between cases and did not measure learning immediately. If this pattern holds across most classroom-based studies, then future class-based investigations might consider including an immediate posttest and perhaps asking learners to find only the similarities.

CURRENT STUDY

Our model tests 15 variables that may affect learning outcomes. Those identified as process variables (the type of instructions, the objective, the type of cases, the presentations of principles, the provision of features, etc.) are of particular interest in this investigation. Generally, we propose that conditions that act as supportive cues for facilitating analogical comparison will result in greater learning outcomes. We included a large number of possible moderators to ensure that potentially confounded variables were not influencing effect sizes. Our new meta-analytic method enables us to examine whether variables are confounded with one another and provides a more comprehensive picture of the effects of the various variables and their interrelationships.

METHOD

Literature Search

Articles examining analogical case comparisons were identified through a variety of sources. The majority of the articles were identified using the Web of Science, PsycINFO, ERIC, and Google Scholar computerized literature searches by searching for particular authors' names and/or by searching using terms like *analogical comparison*, *analog*, *structural alignment*, *schematic learning*, and so on, either alone or in combination. Studies were also identified using forward and backward searches and through e-mail correspondences with authors. Dissertations, unpublished theses, and conference proceedings were also considered for inclusion (Rothstein & Bushman, 2012). The selection criterion was that studies had to test directly for differences between a condition employing case comparisons and a condition that involved sequential case study, single case study, nonanalogous case study, a control/baseline group (no study phase), or more traditional instruction (lecture and/or problem solving).

Exclusion criteria prevented the addition of some potentially relevant studies. Articles with incomplete statistical information or those that analyzed only qualitative data alone were not included. Before excluding any study that did not provide useable statistical information, we contacted authors to request information that could be included in the analysis. We also excluded studies that did not consistently maintain the instructional manipulation because they equivocated groups prior to measures of learning.

Units of Analysis and Data Sets

Analyses were conducted both at the level of experiments and at the level of tests. Analysis at the level of *experiments* refers to the inclusion of individual experiments with different participants each as its own entry with effect sizes averaged across its measures. Thus, if a single article reported on multiple experiments, then those experiments each have their own average effect size and are included individually. Analysis at the level of *tests* refers to the inclusion of each individual statistical comparison as an independent contribution. Although multiple comparisons reported for a single sample violate assumptions of independence, analysis at this level was required to test for the effects of potentially moderating variables. For example, each comparison required individual inclusion when examining the dependent measure because, within the same study, some measures were of near transfer and others were of far. Consequently, articles that include many tests have more weight in the overall computation of the effect than those that run fewer. Whenever possible, potential moderators that did not vary between tests were considered at the level of experiments. Because many potential moderators did differ between tests, only five could be analyzed at the level of experiments: publication rank, domain, age, setting, and duration.

Because several potential moderators were significantly correlated with one another, interactions needed to be considered. Whenever the potential moderator was correlated with another that varied between tests, further analyses had to be conducted at the level of tests. The current analysis investigates the results of 57 experiments with 336 comparisons. See Table 1 for a complete listing of the experiments included.

Variables Coded as Potential Moderators

Fifteen moderators were used for blocking purposes, including publication rank to investigate potential publication biases. Seven of the remaining 14 were considered process variables, four were considered context variables, and three were considered measurement variables. See Table 2 for a complete listing. All potential confounds were examined.

For codes of publication rank, journals were categorized as *top-tier* if they earned an impact factor greater than 2.0 based on the listings of impact factors (2009 Journal Citation Reports® Social Sciences Edition; Thomas Reuters). Journals ranked below 2.0 were coded as *second-tier*. Studies published in *conference proceedings* were coded separately.

Codes for the process variables were entered as follows: The type of instructions was coded as either *prompted* or

TABLE 1
Sample Included in the Meta-Analysis of Case Comparisons Learning Tasks

Experiment	*Year*	*CC n*	*OLS n*	*Cohen's d*	*95% CI*	*Objective*	*Principle*	*Features*	*Content*
Catrambone & Holyoak (Experiment 1)	1989	19	58	0.35	−.12/.81	Similarities	Not	Generated	Procedural
Catrambone & Holyoak (Experiment 2)	1989	25	23	0.25	−.34/.84	Similarities	Not	Generated	Procedural
Catrambone & Holyoak (Experiment 3)	1989	16.5	16	0.15	−.58/.87	Similarities	Not	Generated	Procedural
Catrambone & Holyoak (Experiment 4)	1989	74	16	0.74	.29/1.19	Both	Not	Provided	Procedural
Chen & Daehler	1989	46	70	1.11	.69/1.53	Similarities	After-Not	Generated	Procedural
Christie & Gentner (Experiment 1)	2010	28	28	1.28	.64/1.92	Similarities	Not	Generated	Perceptual
Christie & Gentner (Experiment 2b)	2010	15	30	1.51	.75/2.27	Similarities	Not	Generated	Perceptual
Clement & Gentner (Experiment 1)	1991	24	24	0.94	.30/1.59	Both	Before	Provided	Conceptual
Clement & Gentner (Experiment 2)	1991	16	16	1.20	.35/2.05	Both	Before	Provided	Conceptual
Clement & Gentner (Experiment 3)	1991	24	24	0.87	.23/1.50	Both	Before	Provided	Conceptual
Cummins (Experiment 1)	1992	24	24	0.81	.18/1.44	Similarities	After-Not	Generated	Conceptual
Cummins (Experiment 2)	1992	36	36	0.53	.04/1.02	Similarities	After-Not	Generated	Conceptual
Cummins (Experiment 3)	1992	24	24	0.36	−.24/.95	Similarities	After-Not	Generated	Conceptual
Gadgil & Nokes	2009	20.4	40.8	–0.01	−.52/.51	Both	Before	Generated	Conceptual-Procedural
Gentner, Loewenstein, & Thompson (Experiment 2)	2003	64	64	0.58	.22/.95	Similarities	Not	Generated	Procedural
Gentner, Loewenstein, & Thompson (Experiment 3)	2003	40	80	0.07	−.30/.43	Similarities	Not	Generated	Conceptual-Procedural
Gentner, Loewenstein, & Thompson (Experiment 1)	2004	51.7	51.7	0.51	.11/.92	Similarities	Not	Generated	Conceptual-Procedural
Gentner, Loewenstein, Thompson, & Forbus (Experiment 1)	2009	53.3	50	0.51	.11/.92	Similarities	Not	Generated	Conceptual-Procedural
Gentner, Loewenstein, Thompson, & Forbus (Experiment 2)	2009	19	17	1.05	.28/1.82	Similarities	Not	Generated	Conceptual
Gentner, Loewenstein, Thompson, & Forbus (Experiment 3)	2009	18	32	0.65	.05/1.26	Similarities	Not	Generated	Conceptual

(Continued on next page)

TABLE 1
Sample Included in the Meta-Analysis of Case Comparisons Learning Tasks *(Continued)*

Experiment	*Year*	*CC n*	*OLS n*	*Cohen's d*	*95% CI*	*Objective*	*Principle*	*Features*	*Content*
Gentner & Namy (Experiment 2)	1999	20	60	0.01	−.44/.45	Similarities	Not	Generated	Conceptual
Gerjets, Scheiter, & Schuh (Experiment 2)	2008	15	16	0.65	−.13/1.43	Both	Before	Provided	Procedural
Gick & Holyoak (Experiment 4)	1983	51	94	0.65	.31/1.00	Similarities	Not	Generated	Procedural
Graham, Namy, Gentner, & Meagher (Experiment 1a)	2010	64	64	1.94	1.46/2.43	Similarities	Not	Generated	Conceptual
Kotovsky & Gentner (Experiment 4)	1996	11	11	1.63	−.16/3.41	Both	Not	Provided	Perceptual
Kurtz (Experiment 1)	2005	51	105	0.28	−.04/.60	Both	Not	Generated	Conceptual
Kurtz (Experiment 2)	2005	67	120	0.31	.02/.60	Both	Not	Generated	Conceptual
Kurtz & Loewenstein (Experiment 1)	2007	76	79	0.60	.27/.93	Similarities	Not	Provided	Procedural
Kurtz, Miao, & Gentner (Experiment 1)	2001	40	40	0.53	.07/1.00	Similarities	Not	Generated-Provided	Conceptual
Kurtz, Miao, & Gentner (Experiment 2)	2001	10	10	1.20	.09/2.31	Similarities	Not	Provided	Conceptual
Loewenstein & Gentner (Experiment 2)	2001	24	24	0.63	.02/1.25	Similarities	Not	Provided	Perceptual
Loewenstein & Gentner (Experiment 3)	2001	24	24	0.71	.09/1.33	Similarities	Not	Provided	Perceptual
Loewenstein, Thompson, & Gentner (Experiment 2)	1999	27	31	0.74	.18/1.31	Similarities	Before-Not	Generated	Procedural
Loewenstein, Thompson, & Gentner	2003	141	375	0.55	.37/.72	Similarities	Not	Generated	Procedural
Markman & Gentner (Experiment 1a)	1993	12	36	0.42	−.17/1.02	Similarities	Not	Provided	Perceptual
Markman & Gentner (Expriment 1b)	1993	24	48	0.54	.06/1.03	Similarities	Not	Provided	Perceptual
Markman & Gentner (Experiment 2)	1993	24	24	0.48	−.12/1.08	Similarities	Not	Provided	Perceptual
Markman & Gentner (Experiment 3)	1993	16	32	0.22	−.36/.81	Similarities	Not	Provided	Perceptual
Mason	2004	41	58	0.69	.27/1.12	Similarities	Not	Generated-Provided	Conceptual
Michael, Klee, Bransford, & Warren	1993	11	11	0.65	−.30/1.59	Both	Before	Provided	Conceptual
Mundy, Honey, & Dwyer (Experiment 1)	2009	8	8	0.60	−1.23/2.43	Both	Not	Generated	Perceptual

TABLE 1
Sample Included in the Meta-Analysis of Case Comparisons Learning Tasks *(Continued)*

Experiment	*Year*	*CC n*	*OLS n*	*Cohen's d*	*95% CI*	*Objective*	*Principle*	*Features*	*Content*
Mundy, Honey, & Dwyer (Experiment 2)	2009	12	12	1.05	−.42/2.53	Both	Not	Generated	Perceptual
Nagarajan & Hmelo-Silver	2006	42	39	0.25	−.20/.69	Both	Not	Generated	Conceptual-Procedural
Namy, Gentner, & Clepper	2007	24	12	0.51	−.19/1.22	Similarities	Not	Generated	Conceptual
Nokes, VanLehn, & Belenky,	2008	11.25	24.25	0.34	−.36/1.04	Both	Before	Provided	Conceptual-Procedural
Richland & McDonough (Experiment 2)	2010	26	50	0.54	.06/1.01	Both	Not	Provided	Conceptual-Procedural
Rittle-Johnson & Star	2007	36	34	0.25	−.23/.74	Both	Not	Provided	Conceptual-Procedural
Rittle-Johnson, Star, & Durkin	2009	158	78	−0.10	−.36/.15	Differences-Both	Not	Generated	Conceptual-Procedural
Scheiter, Gerjets, & Schuh	2004	84	84	−0.10	−.40/.21	Both	Not	Generated	Conceptual
Schuh, Gerjets, & Scheiter	2005	30	29	0.54	.00/1.09	Both	Before	Provided	Procedural
Schwartz & Bransford (Experiment 1)	1998	21	21	1.71	.89/2.54	Similarities	After	Provided	Conceptual
Schwartz & Bransford (Experiment 2)	1998	18	18	1.77	.86/2.68	Similarities	After	Provided	Conceptual
Schwartz & Bransford (Experiment 3)	1998	12	24	1.46	.62/2.31	Similarities	After	Provided	Conceptual
Seufert	2003	34	52	−0.21	−.64/.23	Both	Not	Generated	Conceptual
Spencer & Weisberg (Experiment 1)	1986	77	163	0.33	.08/.59	Similarities	Not	Generated	Procedural
Star & Rittle-Johnson	2009	82	75	0.23	−.09/.55	Both	Before	Provided	Conceptual-Procedural
Thompson, Gentner, & Loewenstein	2000	44	44	0.91	.24/1.58	Similarities	Not	Generated	Procedural

Note. When experiments included conditions with varying moderator levels (e.g., content in Star & Rittle-Johnson, 2009), both codes have been included within the table (conceptual-procedural) but each test was coded at a single level of that moderator. CC = case comparisons; OLS = other learning situation; CI = confidence interval.

guided to reflect the degree to which learners received instructional guidance toward the common features and/or unifying concept/procedure. Instructions that *prompted* case comparisons were like two of the examples presented in Figures 1 and 2 in the introduction (Nagarajan & Hmelo-Silver, 2006, and Gentner & Namy, 1999, Experiment 2, respectively). Those instructions were considered prompts because the important details of the cases were left to be identified by learners. Some other examples of prompted instructions are (a) asking learners to compare analogous scenarios to find commonalities, and then describe what is happening and why, or (b) asking learners to compare two procedural analogs and take notes on their similarities (Kurtz et al., 2001, Experiment 1, and Catrambone & Holyoak, 1989, Experiments 1–3, respectively). Again, these examples are coded as prompts because the instructions provided did not provide specific cues as to where learners should begin their searches for commonalities or information as to what the important details of the cases are. Other conditions considered to be *prompted* included having learners compare illustrations to judge their similarity (e.g., Markman & Gentner, 1993, Experiment 1a).

In contrast, instructions that *guided* case comparisons asked questions that directed the learners' attention to the important details of the cases. The previous mathematics examples provided guided instructions (Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2009) to direct the learners' attention to the important steps in common between the solutions/strategies. Another example involved learners comparing two solved radiation analogs in order to explain the critical insight (that in all cases the problem was solved using convergence), identify the parallels between the analogs, and match the five corresponding critical features (Kurtz &

TABLE 2
Variables Coded as Potential Moderators

Variable Type	*κ*	*Potential Moderator*	*Levels*	
	1	Publication rank	Top-tier	impact factor > 2.0
			Second-tier	impact factor ≤ 2.0
			Conference	proceedings
Process	.67	Type of instructions	Prompted	asked to compare
			Guided	directed to features
	.75	Objective	Similarities	how cases are alike
			Differences	how cases differ
			Both	how alike and different
	.62	Type of cases	Rich	extraneous details
			Minimal	relevant details only
	.74	Experience	Little	unfamiliar
			Familiar	some experience
			Extensive	much experience
	1	Principle	Before	case-inclusive information before task
			After	case-inclusive information after task
			Not	not provided or not appropriate to content
	.92	Features	Provided	at least one case outlined
			Generated	neither case outlined
	1	Duration	Brief	single session < 1 hr
			Long	> 1 session or hr
Context	.75	Setting	Classroom	success impacted grade
			Laboratory	minimal/no impact
	.96	Content	Conceptual	understanding concepts, facts, schemas, etc.
			Procedural	executing steps toward a solution/goal
			Perceptual	altering how stimuli are considered or organized
	1	Age	Children	≤ 12 years old
			Adolescents	≥ 13 ≤ 17 years old
			Adults	≥ 18 years old
	.82	Domain	Science	earth or social science content
			Math	math content not imbedded in science
			Other	visual, lexical, or nonacademic content
Measurement	.95	Lag	Immediate	assessment followed study phase
			Same day	interposed delay/filler task after study phase
			Subsequent day	study and assessment on different days
	.67	Dependent measure	Near transfer	application of learned material without adaptation
			Far transfer	application required adaptation to meet demands of test task
	.68	Other learning situation	Sequential cases	same cases studied sequentially as case studies
			Single case	study of a single analogous case
			Traditional	problem solving and/or listening to lecture
			Nonanalogous	study included at least one nonanalogous case
			Control	no study phase

Loewenstein, 2007, Experiment 1). Here, instructing learners to explain the critical insight common to both directed their searches for commonalities by cuing them into the fact that there is an analogous insight to be found. Furthermore, instructing students to find five critical features directs their efforts to search for five similar features. These cues serve to make the search for commonalities much more focused. In general, *guided* instructions directed learners to search for more specific features and/or relations across the cases, whereas *prompted* instructions were more general and asked learners to simply search for similarities and/or differences.

The objective of the comparison (i.e., whether learners were asked to identify *similarities*, *differences*, or *both* during the case comparison task) was also coded. All but one of the examples from Figures 1 and 2 (Nagarajan & Hmelo-Silver, 2006; Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2009) asked learners to notice *both* similarities and differences between cases, either directly or indirectly, by asking learners what cases had in common, how they differed, and/or about the effects of those differences. In contrast, Gentner and Namy (1999, Experiment 2) asked their learners only to decide why the two exemplars were of the same kind and thereby encouraged them to focus on only the similarities. One last possibility is that the objective of the comparison could have been to find *differences* alone (e.g., Rittle-Johnson et al., 2009). Rittle-Johnson and colleagues asked learners to compare two solution methods to decide how the two are different and what needs to be true of the equation to make one method easier than the other.

The types of the cases being compared were coded as either rich or minimal in format. *Rich* cases were those that provided more information than was needed (extraneous facts, superfluous contexts, etc.) or that presented cases in potentially ambiguous forms because many other details could just as well have been considered while searching for commonalities. Nagarajan and Hmelo-Silver (2006) and Gentner and Namy (1999, Experiment 2) presented learners with cases that were rich in detail. The former's videotaped formative assessments likely presented many details of the classroom environment as well as the administrations of those assessments. The latter's line drawings shared many perceptual commonalities that were not included within the target categorization of both being, for example, types of transportation.

Minimal cases were those that highlighted the important details/key features by not including extraneous information or potentially distracting details. Rittle-Johnson and Star (2007; Star & Rittle-Johnson, 2009) provided minimal cases by presenting only the steps within the mathematical solution/strategy with labels and explanations to highlight each step. Another example of minimal cases was designed by Gentner and colleagues in which text-based cases consisted of only relevant information and that was paired with diagrammatic organizations of those facts (Gentner et al., 2003, Experiments 2 & 3). The total amount of information provided within the cases was not the focus of the context code, but instead the relation of the details provided to the target features. For example, if a case had many details but all were necessary for identifying the target features and relations, it would be considered minimal. In contrast, if the case included some salient details that were unnecessary for identifying the target features and relations, even if there were generally few details overall, these were considered rich cases because they included extraneous details to the target comparison.

The experience level of learners was coded as little, familiar, or extensive. Learners were coded as having *little* experience when they had little to no prior experience with the subject matter or target content. Their low levels of previous experience might have been because they were new students to a field (e.g., undergraduates enrolled in a communication development course learning about theory-based therapies; Michael et al., 1993) or because the task/to-be-learned material was unusual (e.g., cases were of science fiction content unique to the study or cases were complex checkerboard patterns symbolizing RNA; Clement & Gentner, 1991, Experiment 1; Mundy et al., 2009, respectively). *Familiar* learners were those who had some experience with the content/cases prior to the study (e.g., psychology students enrolled in a cognitive psychology course studying schema and encoding concepts, or undergraduates asked to recognize that a pen could be conceptualized as a container; Schwartz & Bransford, 1998, Experiment 2, and Kurtz, 2005, respectively). Learners considered to have *extensive* levels of experience were those who were professionals in the same or a related field (e.g., professional management consultants learning to negotiate contingency contracts; Gentner et al., 2009, Experiment 1).

The sixth potential moderator was whether a principle was provided explicitly as a textual/verbal description *before* or *after* the case comparisons task or *not* at all. Examples of the presentation of a principle include implementations in which learners read about a language learning theory *before* using it to critique a videotaped therapy session and then compare their critique to an expert's (Michael et al., 1993), and implementations in which learners were provided with explanations of schema and encoding concepts (Schwartz & Bransford, 1998) *after* having compared cases that embodied patterns explained by such concepts. Furthermore, if learners were first questioned as to the similarities or differences between cases and then provided with a principle and asked to sort cases by that principle (Cummins, 1992), such conditions were coded as providing the principle after because there was no requirement that learners return to the comparisons in light of the principle. The code of *not* was applied when a study potentially could have provided learners with principles but did not because learners were to construct the principle for themselves through comparison (e.g., Gentner & Namy, 1999, Experiment 2; Gick & Holyoak, 1983, Experiment 4; Kurtz et al., 2001, Experiments 1 & 2; Nagarajan & Hmelo-Silver, 2006). In the category learning study described earlier, for example, the underlying principle is the conceptual category, but learners are to construct that understanding through comparison. Similarly, in Gentner and colleagues' studies on students learning negotiation, providing the principle before the cases would provide the appropriate solution prematurely (Gentner et al., 2003, Experiments 2 & 3). Learners were to reach that understanding through comparing the cases. In other cases, there was not a high-level principle to provide (e.g., Kurtz, 2005, Experiments 1 & 2; Markman & Gentner, 1993, Experiments 1a, 1b, 2, & 3; Mundy et al., 2009, Experiments 1 & 2). Markman and Gentner asked their learners to compare pairs of illustrations to determine their level of similarity and then asked them to identify cross-mapped objects (the analogous pair of components across the two illustrations), but each pair of objects shared a unique role in comparison to other pairs. Thus, there was not an appropriate principle to provide.

We also coded for features (whether provided or not) and duration of the study phase (more or less than 1 hr). We coded as to whether the key features of the cases were to be identified by learners (*learner-generated*) or were *provided* explicitly within the study materials (e.g., in the form of a list for a matching task). Even if learners were only provided with a list of features specific to one of the two cases and asked to generate a corresponding list for the other (e.g., Clement & Gentner, 1991; Kurtz et al., 2001, Experiment 1), such studies were coded as having provided features. If studies did not provide a list of features from at least one case, then cases were coded as having learner-generated features—even in the absence of a requirement for learners to state

explicitly what those features were. Features of cases might have been procedural steps or conceptual details and both are distinguishable from principles because no individual feature or corresponding pair of features across cases captures the entire procedure, concept, or perception. To draw on previous examples, each step of approximating is a feature of the procedure but not the principle of approximating, and each evidential detail of the items for categorization was a feature but not the principle of transportation. For the duration of the study phase, studies were coded as *brief* if study sessions lasted for only a single session of less than 1 hr. Studies were coded as *long* if study sessions were extended over multiple days and/or they were longer than 1 hr.

Codes for the context variables were entered as follows: for setting, experiments were coded as *classroom* studies when learners' performances would impact their grades (as reported by authors) and/or content was expected to be learned for the purposes of the class. Studies were coded as *laboratory* studies when participation was not part of a class, or when participation was part of a class but content was not critical to course completion or performances did not impact learners' grades.

The content of the to-be-learned material was coded as *conceptual*, *procedural*, or *perceptual*. Conceptual content included phenomena (e.g., heat transfer; Kurtz et al., 2001, Experiments 1 & 2; Mason, 2004), definitions (e.g., a jiggy is a type of spatial relationship between drawings of animals and not a type of animal itself; Christie & Gentner, 2010, Experiments 1 & 2b), and/or categories (e.g., things are categorized by function and not form; Gentner & Namy, 1999, Experiment 2). Procedural content included the execution of a series of steps to a solution (e.g., dividing and dispersing forces to converge, truncating each multiplicand then multiplying and supplementing zeros, negotiating contingency contracts; Gick & Holyoak, 1983, Experiment 4; Star & Rittle-Johnson, 2009; and Gentner, Loewenstein, & Thompson, 2004, Experiment 1, respectively). Perceptual content included relationship(s) between stimuli that were not captured by declarative representations (e.g., identifying cross-mapped objects; Markman & Gentner, 1993, Experiments 1a, 1b, 2, & 3).

In many studies, coding of the content required a consideration of the subject matter as well as the target task. For example, Gentner et al. (2004, Experiment 1) asked learners to study cases of contingent contracts so that they reached a conceptual understanding of them. Their measures of learning included the quality of participants' schemas, the quality of recalled examples of contingency contracts, and how many dyads resolved the test case by implementing a contingent contract in face-to-face negotiations. The measures of schema quality and the quality of recalled personal experiences were coded as containing *conceptual* content, whereas the frequency measure of dyadic resolutions that implemented contingent contracts was coded as containing *procedural* content. Similarly in the example provided of conceptual content, although the term *jiggy* referred to a visual, spatial configuration, the target was to build a category captured by the term *jiggy* (a conceptualization that captures that configuration). Thus, the semantic content was conceptual and not perceptual.

For age, learners were considered *children* if they were 12 years old or younger, *adolescents* if they were between 13 and 17 years old, and *adults* if they were 18 years old or older. If the same comparison included a range of ages, the mean age of the sample was used for coding purposes. If the exact age of the sample of learners was not provided but its grade level was, the sample was coded as *children* through sixth grade, as *adolescents* from seventh through 12th grades, and as *adults* thereafter.

The domain of *science* included studies that ranged from problem solving within physics (Gadgil & Nokes, 2009) and recognizing the process of heat transfer (Kurtz et al., 2001; Experiment 1; Mason, 2004) to understanding formative assessments (Nagarajan & Hmelo-Silver, 2006) and understanding schema and encoding concepts (Schwartz & Bransford, 1998, Experiment 1). The domain of *math* included studies that ranged from solving multistep mathematics problems and solving algebraic equations that were set within the fields of biology, chemistry, and political science (Rittle-Johnson & Star, 2007, and Gerjets et al., 2008, Experiment 2, respectively) to sorting cases by their statistical/algebraic content and understanding the proper use of the linearity assumption (Cummins, 1992, Experiments 1–3; Scheiter, Gerjets, & Schuh, 2004, and Richland & McDonough, 2010, Experiment 2, respectively). The example studies from Gadgil, Gerjets, and their colleagues highlight that when coding for domain, what was to be learned was the focus. Thus, although math is heavily involved in physics, the focus remained on learning about physics concepts and was consequently coded as science. In contrast, Gerjets and colleagues contextualized algebraic equations within science domains to have learners appreciate how to solve such equations in different contexts. Because the focus was in solving algebraic equations and not on the subject matter within the different contexts, it was coded as math. Content considered to be within the domain of *other* ranged from solutions to physical problems (e.g., getting a bead out of a cylinder or solving a problem using the convergence solution; Chen & Daehler, 1989, and Catrambone & Holyoak, 1989, respectively), to the negotiation of contracts between disputing parties (Gentner et al., 2004, Experiment 1), to the perception of visual stimuli (Markman & Gentner, 1993).

We also coded measurement variables. The lag between the case comparisons task and the assessment of learning was coded as *immediate* if the contents of the study phase were potentially still in working memory during the test phase (i.e., the assessment was made during or immediately following study). It was coded as *same day* if a filler task or time delay interposed the study and test phases but the test was administered before the end of the same day. Or it was coded

as *subsequent day* if the test phase was administered on a day after that of the study phase.

We coded the dependent measure as to whether the measure was of near or far transfer. Separate definitions of near or far needed to be specified for conceptual and procedural content. *Near* transfer of procedural content asked learners to transfer an unmodified solution learned from study to a superficially different problem. Such measures ranged from the appropriate use of mathematical algorithms (Schuh et al., 2005) to the appropriate use of convergence solutions (Catrambone & Holyoak, 1989). *Far* transfer of procedural content required that learners modify the solution to accommodate structural changes in variables or features within the new case. Far transfer of procedural content ranged from the selection and implementation of an equation from among possible alternatives when faced with physics problems with extraneous variables (Gadgil & Nokes, 2009) to having to solve for a different variable than at study (Nokes, VanLehn, & Belenky, 2008). Near transfer of conceptual content required recognition or application of studied concepts without adaptation. Such measures ranged from scores that reflected the causal relevance of learners' descriptions of cases (Mason, 2004) to scores that reflected the recognition of concepts and the reconstruction of studied processes (to be understood, not executed; Seufert, 2003). Far transfer of conceptual content ranged from sorting novel mathematics word problems into new problem categories (Cummins, 1992, E1, E2) to making inferences and analogies on a subset of posttest questions designed to assess how much learners would extend what they know about iron's and vitamin C's contributions within human metabolism to other scenarios (Seufert, 2003).

The type of learning condition to which case comparisons were compared (the other learning situation [OLS]) was coded as having no study phase (baseline group or *control*), a *nonanalogous* study task, a *sequential* case study task, a *single-case* study task, or more *traditional* instruction. It should be noted that the *sequential* OLS presented learners with the same cases as those presented to the learners in the condition of case comparisons, but in these sequential conditions, learners studied them one after another instead of simultaneously. For example, Rittle-Johnson and colleagues (Rittle-Johnson et al., 2009) asked learners in the case comparison condition to compare Abby's solution method to Patrick's to determine one way they were the same, one way they were different, and why the first step of Patrick's solution was different than Abby's. In contrast, they asked learners in the sequential condition first to study Abby's solution and explain why she combined like terms in her first step and then to study Patrick's solution (on the next page) and explain why Patrick's solution might be (or might not be) applicable to many different kinds of problems.

Nonanalogous OLS might have asked learners to compare two cases (one analog, one non) during a simultaneous presentation or to consider each sequentially, but because both cases were not the same as those used in the case comparison condition, they were coded as *nonanalogous*. For example, Catrambone and Holyoak (1989, Experiment 1) asked one group of learners to compare a case involving the convergence solution (The General or The Fire Chief) to a case that did not involve the convergence solution (The Wine Merchants); this condition controlled for the amount of time and processing during the study phase. Although it did present learners with one of the same analogs (The General or The Fire Chief) as was presented to learners conducting case comparisons, the other case was not analogous. Whether these cases were compared simultaneously as they were in this example or studied sequentially with one on each page did not change its coding. However, if learners were only to study either The General or The Fire Chief, that condition would have been coded as a single case.

A *single-case* OLS presented learners with only one analogous case to study. For example in Christie and Gentner's (2001) study (Experiment 1), they provided one group of young learners in their categorization task with only one of the two exemplars presented to those in the comparison condition. *Traditional* OLS asked learners to solve problems and/or attend a lecture(s). For example, Michael et al. (1993) provided traditional instruction in the form of lectures and readings to one group while another group was asked to compare videotaped implementations of language therapies that exemplified theories of language development. Richland and McDonough (2010, Experiment 1), in another example, taught the correct use of the linearity strategy immediately to learners and then gave them a worksheet of four questions probing what they learned about the strategy along with a final item that asked them to solve a set-up proportion. Then they were again shown how to solve the problem. The group implementing case comparisons compared two different solutions (one with and one without the proper use of the linearity function). Thus, the code of traditional instruction covered a variety of instructional approaches that did not employ case studies/comparisons. Learning conditions coded as *control* included no type of instruction or study phase.

Reliability of Moderator Coding

The coding scheme was developed by using the process model as a guide to identify relevant variables that capture aspects of a broad range of case-based learning situations, completely and yet concisely. The first author coded all studies, and the second author coded 25% of the studies. Coding reliability of all the moderators was at least adequate, and overall coding reliability was high, with an average Cohen's kappa of .80. See Table 2 for reliability measures for each potential moderator. Any disagreements were resolved through a discussion of how best to classify the variable in question both within the context of the experiment and for the purposes of analysis.

Computation and Analysis of Effect Sizes

We used the *Comprehensive Meta-analysis, Version 2* (CMA) software (Borenstein, Hedges, Higgins, & Rothstein, 2005) to calculate effect sizes for each test and to run all analyses. Overall effects are reported under both fixed and random effects models, but subsequent analyses are reported only from the random effects model. A random effects model was appropriate because our sample included such diverse methodologies, samples, and potential effect sizes.

Effect sizes. Comparisons were entered into the CMA program and effect sizes for each comparison were first calculated using the program's computation formulae. When the only statistics available in the papers were F values and group means, DSTAT (Johnson, 1993) allowed us to convert those statistics to a common metric, r, which is the correlation coefficient respective to that F value with such a sample size. For studies that reported differences that were unreliable but did not provide exact statistics, those comparisons were entered into DSTAT with an F value of 1. Those r scores and sample sizes were then entered into the CMA program.

Because g values may "overestimate the population effect size" when samples are small (Johnson, 1993, p. 19), standardized differences in means (*Cohen's d*) values are reported here as calculated by the CMA program. Effect sizes between .20 and .50 indicate a small effect size, between .50 and .80 indicate a medium effect, and greater than .80 indicate a large effect (Cohen, 1988). Of course, the effect size alone does not determine significance, and we determined that effect sizes were reliably different than zero (no effect) based on the p values of the resultant Z scores.

Statistical Approach

Comparisons of learning situations involving case comparisons to other learning situations were subjected to two separate meta-analyses, one at the level of experiments and one at the level of tests. Table 3 displays the results overall at both levels and for both fixed and random effects models. Positive effect sizes indicate superior performances for case comparisons, whereas negative effect sizes would have indicated superior performances in the other learning situations.

TABLE 3
Summary of Effect Sizes

	Cohen's d	95% CI	*Z*	*Q*	*k*	I^2
Fixed						
Experiments	.45	.38 / .53	12.10*	136.73*	56	59.05
Tests	.37	.34 / .41	23.61*	1, 048.47*	335	68.05
Random						
Experiments	.60	.47 / .72	9.29*			
Tests	.50	.44 / .56	16.74*			

*$p < .001$.

Tables of effect sizes are displayed with the following columns from left to right: Cohen's d statistics, 95% confidence interval, Z value, Q value, and k (the degrees of freedom within the Q statistic or the number of comparisons within that row minus 1). Cohen's d measures are average effect sizes and, in the current analyses, almost all will be positive, indicating superior performances following case comparisons. Next, the width of the confidence interval will provide a sense of the spread of effect sizes. The Z value can be used to determine whether the average effect size is significantly different from zero (i.e., there is no effect). Thus, any average effect with a confidence interval that includes zero will not have a significant Z value ($p > .05$).

The Q statistic for overall analyses determines whether there is significant heterogeneity within the sample and, consequently, whether violating statistical assumptions of independence is justified to consider moderators at the level of tests to determine under which specific conditions case comparisons are best. The I^2 statistic for the overall analyses further assesses heterogeneity and can be read as the percentage of the variability that is due not to mere sampling error but to true heterogeneity between experiments (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). The Q statistic is also used at the level of tests to determine whether levels of a moderator are statistically different from one another. Again, k provides the number of comparisons within that row minus 1. For moderator tables, summing k values for that moderator will also indicate whether analyses were performed at the level of experiments (k values add to almost 57 but are short by the number of levels of the moderator) or at the level of tests (k values add to almost 336 but are short by the number of levels of the moderator). Whenever possible, analyses were run at the level of experiments to avoid violating statistical assumptions of independence.

Investigations of Moderators and Post Hoc Comparisons

Analysis of the overall sample found significant heterogeneity in the effect sizes (as indicated by Q and I^2; Huedo-Medina et al., 2006; Johnson, 1989). To further investigate this heterogeneity, effect sizes were grouped by one potential moderator at a time. When significant heterogeneity among different levels of that moderator warranted still further analysis, each level was then compared to all others within the CMA program to determine if the effect sizes of the different levels were reliably different from one another. Before doing so, post hoc p values were adjusted for the number of comparisons conducted. For example, to compare all of the levels of the potential moderator, publication rank, three post hoc comparisons were required and therefore alpha was adjusted

TABLE 4
Moderators Found to be Consistent Across Confounded Variables

Moderator	*Cohen's d*	*95% CI*	*Z*	*Q*	*k*
Objective					
Similarities	.68	.60/.75	17.83**		197
Differences	−.19	−.30/ − .07	−3.06*		6
Both	.28	.20/.36	6.66**		130
				154.33**	2
$p < .016^{\dagger}$	Differences	Both			
Similarities	146.21**	48.60**			
Differences		40.08**			
Principle					
Before	.37	.24/.50	5.41**		56
After	1.18	.93/1.44	9.09**		28
Not	.47	.41/.54	14.37**		250
				31.90**	2
$p < .016^{\dagger}$	After	Not			
Before	30.78**	1.84			
After		28.18**			
Content					
Conceptual	.54	.46/.61	13.48**		210
Procedural	.40	.31/.49	8.55**		99
Perceptual	.72	.49/.96	6.00**		24
				8.89*	2
$p < .016^{\dagger}$	Procedural	Perceptual			
Conceptual	5.03	2.15			
Procedural		$6.27^{\dagger}$			
Lag					
Immediate	.57	.50/.64	16.50**		254
Same day	.44	.25/.64	4.39**		21
Subsequent	.22	.11/.34	3.94**		58
				27.32**	2
$p < .016^{\dagger}$	Same Day	Subsequent			
Immediate	1.55	27.21**			
Same day		3.57			

Note. CI = confidence interval.
*p <.05. **p < .001.

to .016. Again, the Q statistic was used to determine whether levels of each moderator were different from one another. The Q statistic is the number that appears in the cells of post hoc tables; whether levels are considered to be reliably different from one another is determined by the adjusted alpha found in the upper leftmost cell. As can be seen in Table 4, beginning with the analysis of the moderator, objective, the Q value when comparing the objective of differences to the objective of similarities is 146.21. The two asterisks indicate that this difference between levels is reliable, $Q(1) = 146.21$, $p < .001$. For objective, the adjusted alpha was set to .016. Thus, differences between the levels of objective would have been considered significant only if the p value were less than .016.

Interdependence Among Moderators

To ensure that variables were truly moderating results and were not confounded (or interacting) with other variables, cross-tabulations of all of the potential moderators determined to what extent variables were dependent on one another. We then focused on the statistically significant moderators that shared contingency coefficients greater than .4 (i.e., large enough to cause significant indirect relationships). We begin with a presentation of the variables that consistently moderated findings because their patterns held up across levels of all correlated variables. We then present inconsistent moderators, possible explanations, and moderators that require subsequent research to provide ample power.

RESULTS

Overall Effects

A total of 336 tests from 57 experiments compared learning situations involving case comparisons to other learning situations. Table 3 lists the effects across the entire sample. Under the random effects model, the 57 experiments had a medium effect size with a tight confidence interval favoring case comparison ($d = .60$), 95% confidence interval (CI) [.47, .72], indicating that there is high certainty that case comparison is, in general, moderately effective. To rule out potential publication bias for only studies that find positive effects, we calculated fail-safe total sample sizes at the level of experiments and at the level of tests with alphas set to .05, two-tailed. At the level of experiments, 2,795 unpublished experiments would be needed to alter the results so that the average benefit of case comparisons would no longer be statistically significant. At the level of tests, 5,244 unpublished results would be needed to reduce the effect to nonsignificance.

However, the overall effects of case comparison were found to be heterogeneous across samples, $Q(56) = 136.74$, $p < .001$, with approximately 60% of the variability among effect sizes caused by true heterogeneity between studies ($I^2 = 59.05$). Therefore, it is important to examine moderators of the effect.

Moderators

Table 5 lists all of the variables and classifies them in terms of whether they consistently moderated, inconsistently moderated, or did not moderate learning outcomes or suffered from low power. Of the 15 potential moderators investigated, five did not significantly moderate: the process variables of type of instructions, the experience levels of learners, whether or not features were provided, and the context variables of learners' ages and setting. That is, contrasting cases produced on average consistent medium effect sizes across those variations in implementation and context (see Figure 3).

The variables that did significantly moderate or that require further explanation are presented in groups according to variable type (process, context, or measurement). It should

TABLE 5
General Findings of Investigated Moderators, Including Confounds

	Process Variables	*Context Variables*	*Measurement Variables*
Moderators	Objective: Similarities > Both > Differences Principle: After > Not = Before	Content: Perceptual > Procedural	Lag: Immediate > Subsequent Day
Inconsistent moderators	Type of cases: Rich > Minimal *But* in science and other domains, Rich = Minimal	Domain: Other > Math *But* in 2nd-tier publications, Other = Science = Math When study is brief, Science = Math = Other When learners are familiar, Other = Math = Science	DM: Near > Far *But* in math, Far > Near
Issues of low power	Duration: Only 2 studies with long durations within laboratories		OLS: Majority are sequential
Nonmoderators	Type of instructions Experience Features	Age Setting	

Note. DM = dependent measure; OLS = other learning situation.

also be mentioned before proceeding that, whereas publications in top- and second-tier journals reported similar effect sizes (Table 6), conference proceedings reported lower effect sizes than did second-tier journals.

Process Variables

As can be seen in Table 5, the process variables of the objective and the presentation of a principle consistently moderated learning outcomes. Table 4 displays the results and post

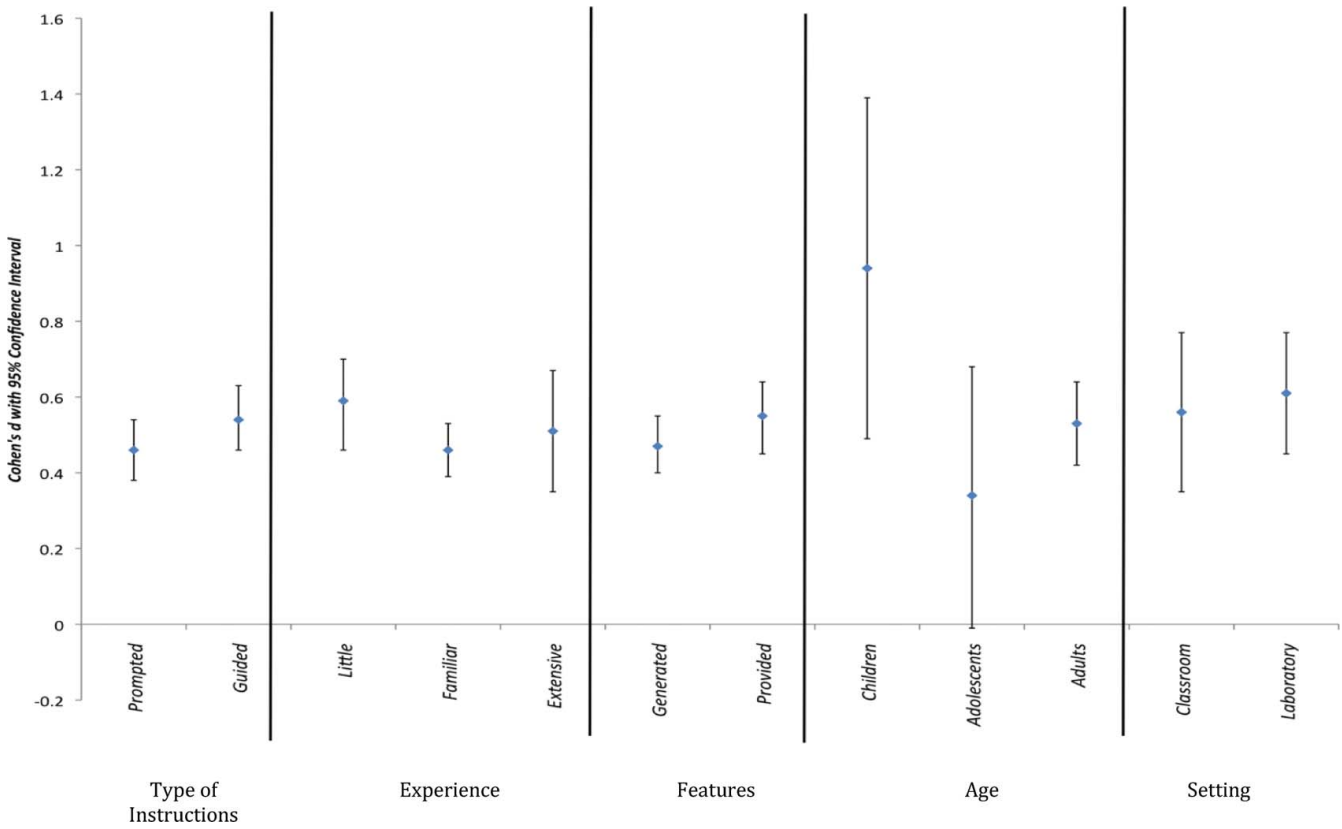


FIGURE 3 Consistent benefits to case comparisons tasks found across levels of variables that did not moderate (color figure available online).

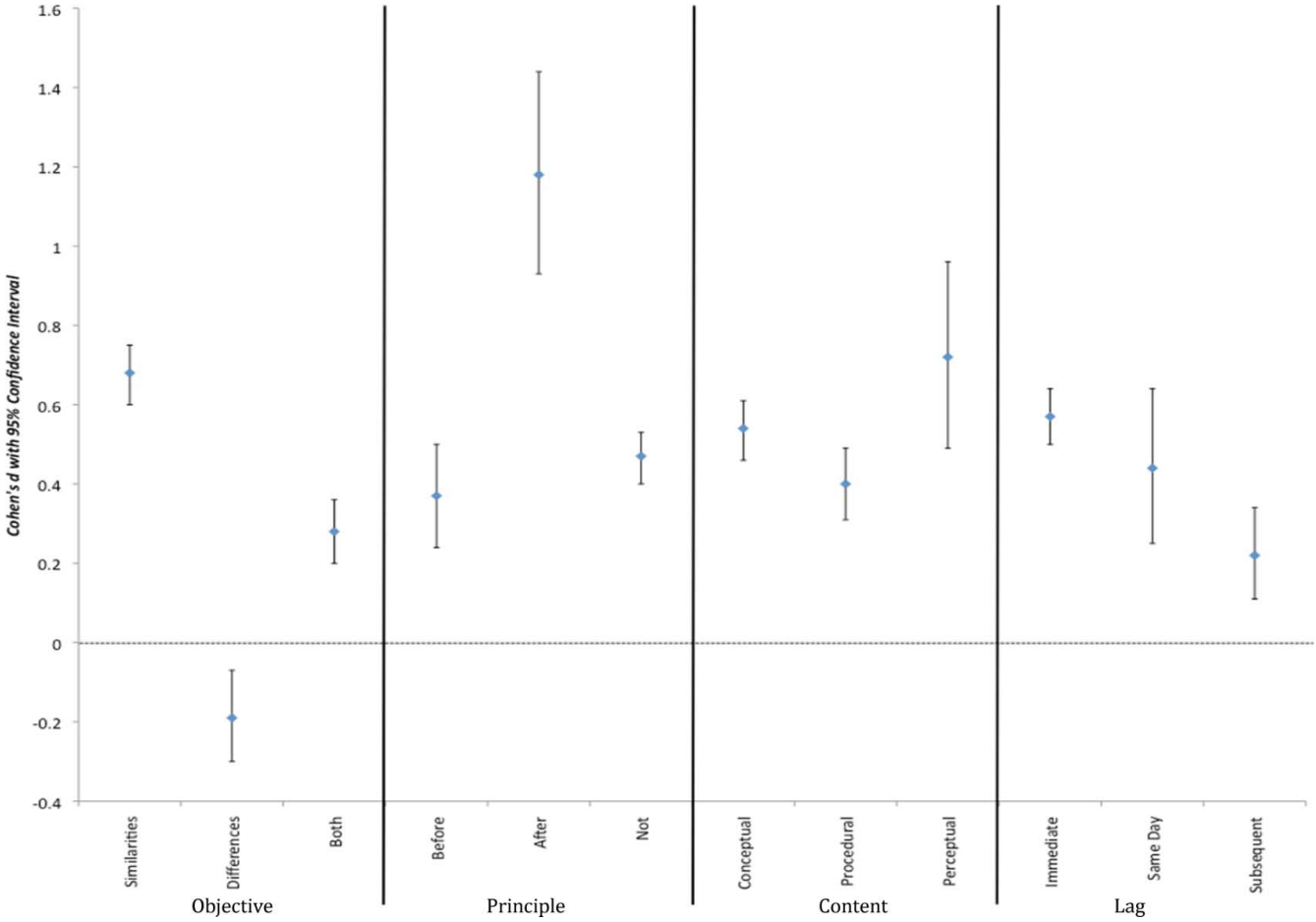


FIGURE 4 Moderators found to do so consistently (color figure available online).

hoc analyses of these consistent moderators and Figure 4 displays their effect sizes with confidence intervals. The type of cases was an inconsistent moderator (Table 7) with different effects observed across the levels of variables with which it was confounded (domain). Last, analysis of the duration of study revealed issues of low power (Table 6).

As can be seen at the top of Table 4, the objective to find similarities across the cases led to greater learning than the objective to find similarities and differences, which in turn led to greater learning than the objective to find only differences. Although the learning deficit for only finding differences should be considered tentative because that level includes only seven tests from a single experiment, the more general pattern that finding only similarities leads to greater learning outcomes than finding similarities and differences is consistent across correlated variables and has ample power. This effect is consistent across all levels of the correlated variables of principle (contingency coefficient $C = .50$), publication rank ($C = .51$), and lag ($C = .45$).

For the moderator of principle, studies that provided learners with the principle after the case comparison yielded greater learning outcomes than studies that either did not provide a principle or provided the principle before the comparison (Table 4). In addition, there was no difference between studies that provided the principle before case comparisons and those that did not provide a principle at all. Further analyses indicated that the moderator of principle was potentially confounded with the objective of the learning task ($C = .50$), the type of instructions ($C = .49$), and the duration of the task ($C = .41$) but the overall effect was consistent across levels of the correlated variables with ample power.

There is a particular finding to note between the two previous patterns of effects within the objective and principle variables. When learners are asked to find only similarities and are provided with the principle after case comparisons ($d = 1.18$), 95% CI [.93, 1.44], the average effect size for those 29 tests indicate even greater benefits than the average effect size of similarities indicated. Thus, when the objective is for learners to find only similarities when comparing cases and they are provided with the principle after case comparisons, the effects on learning are large.

The type of cases was an inconsistent moderator with rich cases leading to greater effects than minimal cases (Table 7). Whereas this trend did hold up across levels of the type of instructions ($C = .42$), this trend did not hold up across levels of domain ($C = .33$). Rich cases within the domain of science

TABLE 6
Moderators That Suffer From Low Power

Moderator	*Cohen's d*	*95% CI*	*Z*	*Q*	*k*
Publication					
Top	.59	.41/.76	6.50**		30
Second	.73	.50/.95	6.31**		19
Conference	.28	.05/.51	2.38*		5
				7.85*	2
$p < .016^{\dagger}$	Second	Conference			
Top	.92	4.33			
Second		7.43†			
Duration					
Brief	.58	.43/.72	7.76**		39
Long	.66	.40/.91	5.00**		16
				.27	1
Age					
Children	.94	.48/1.39	4.04**		9
Adolescents	.34	−.01/.68	1.92		4
Adults	.53	.42/.64	9.55**		41
				4.27	2
OLS					
Control	.69	.58/.80	12.57**		63
Nonanalogous	.82	.55/1.08	6.06**		23
Sequential	.37	.31/.44	11.11**		201
Single case	.92	.57/1.28	5.11**		16
Traditional	.49	.24/.74	3.87**		29
				37.59**	4
$p < .005^{\dagger\dagger}$	Non	Sequ	Single	Trad	
Control	.79	24.60††	1.56	1.98	
Non		10.35††	.23	3.06	
Sequential			9.08††	.87	
Single				3.81	

Note. CI = confidence interval; OLS = other learning situation.
*$p < .05$. **$p < .001$.

TABLE 7
Moderators that Varied Between Confounded Variables

Moderator	*Cohen's d*	*95% CI*	*Z*	*Q*	*k*
Type					
Minimal	.34	.26 / .43	7.74**		128
Rich	.60	.52 / .67	15.52**		206
				18.84**	1
Domain					
Science	.72	.36 / 1.08	3.94**		10
Math	.23	.05 / .42	2.44*		9
Other	.66	.50 / .81	8.24**		35
				12.93*	2
$p < .016^{\dagger}$	Math	Other			
Science	5.51	.09			
Math		11.46†			
DM					
Near	.52	.46 / .58	16.11**		287
Far	.34	.20 / .48	4.64**		47
				5.08*	1

Note. CI = confidence interval; DM = dependent measure.
*$p < .05$. **$p < .001$.

($d = .52$), 95% CI [.41, .64], were found to be equivalent to minimal cases within science ($d = .42$), 95% CI [.29, .55), $Q(1) = 1.39, p = .24$. Similarly, rich cases within the domain of other ($d = .69$), 95% CI [.57, .80], were found to be equivalent to minimal cases within other ($d = .64$), 95% CI [.34, .94], $Q(1) = .08, p = .78$.

Although duration did not moderate learning outcomes (Table 6), our new method for investigating possible confounds revealed that part of the reason might be that it is confounded with setting ($C = .64$). Moreover, the cell containing the average effect of long durations within laboratory settings ($d = .91$), 95% CI [.66, 1.17], contains only 11 tests from four experiments of only two studies. In contrast, 153 of the 164 tests of long durations were within classroom settings, and 156 of the 172 of the brief durations were within laboratory settings. Thus, the average effect size for the 153 tests of long durations within classrooms ($d = .41$), 95% CI [.33, .49], is being inflated when effect sizes from tests of long durations within laboratory settings are added to them. In contrast, the effect sizes for brief durations are about the same in both laboratory and classroom settings ($d = .56$), 95% CI [.46, .65], and ($d = .55$), 95% CI [.44, .66], respectively, despite many more tests coming from the laboratory than from the classroom. Therefore, greater numbers of laboratory studies of long durations and classroom studies of short durations are needed to further examine this potential interaction.

Context Variables

Of the context variables, only content was found to consistently moderate findings (see Table 5). Findings and post hoc analyses of content can be found in Table 4 and a presentation of the effects with their confidence intervals in Figure 4. The domain of study inconsistently moderated findings (see Table 7) as effects changed across levels of correlated variables of publication rank, duration, and experience.

Content consistently moderated findings with larger learning outcomes for perceptual content than for procedural content and no differences between conceptual and perceptual or procedural content (Table 4). Analyses indicated that the moderator of content was potentially confounded with domain ($C = .47$), the other learning situation (OLS; $C = .46$), and experience ($C = .44$), but the pattern of effects was generally consistent across all levels of those other moderators. The only exception to the pattern was found within the cell of procedural content within the science domain ($d = .10$), 95% CI [–.03, .23], showing a nonsignificant effect size, but the 30 tests included within that cell were all from only three conference proceedings.

The academic domain inconsistently moderated learning outcomes. Comparisons implemented in other domains led to significantly greater learning outcomes than comparisons implemented in mathematics. Similarly, comparisons

implemented in science led to marginally greater effects than comparisons in math. However, the smaller effect sizes in math disappear in a number of cells across levels of moderators potentially confounded with domain: OLS ($C = .56$), publication rank ($C = .55$), duration ($C = .53$), experience ($C = .53$), and setting ($C = .51$). In second tier publications, math ($d = .55$), 95% CI [.36, .73], is statistically equivalent to both science ($d = .70$), 95% CI [.61, .79], and other ($d = .72$), 95% CI [.52, .92], $Q(2) = 2.30$, $p = .32$. In studies that were brief in duration, learners' performances in math ($d = .47$), 95% CI [.35, .60], were equivalent to learners' performances in science ($d = .31$), 95% CI [.12, .50], and in other ($d = .65$), 95% CI [.54, .77]. Furthermore, with samples of familiar learners, math ($d = .45$), 95% CI [.33, .58]; science ($d = .46$), 95% CI [.37, .54]; and other ($d = .45$), 95% CI [.29, .61], were again equivalent, $Q(2) = .004$, $p = .99$. Thus, the moderator of domain does not show consistent trends because in many cells, science, math, and other domains were statistically equivalent.

It should also be noted that the moderator of age also suffers from low power for studies with adolescents, which evidenced a null effect size. As can be seen in Table 6, there were only five experiments that sampled adolescents. Despite the null effect size within the adolescent subsample, age was not found to significantly moderate findings. Adult learners and young learners (children) both reliably benefit and roughly to the same extent from case comparisons.

Measurement Variables

Of the measurement variables, only the lag between study and testing consistently moderated findings (Tables 4 and 5). The dependent measure inconsistently moderated findings (Table 7). The analysis of the OLS (other learning situation) revealed a disproportionately large sample of sequential case studies.

Analyses of the lag between study and testing indicated that immediate testing led to greater outcomes than did testing on a subsequent day (Table 4). However, testing on the same day after a filler task or another activity was not found to lead to statistically different outcomes in comparison to either immediate or subsequent day lags but that is at least partly due to the large confidence interval (standard error) for same day which contains only 22 tests. However, the benefit of having a briefer lag holds true across levels of the OLS ($C = .49$) and objective ($C = .45$).

For the moderator of dependent measure, near transfer leads to greater effect sizes than far transfer (Table 7). However, this trend reverses in the domain of math where far transfer ($d = .41$), 95% CI [.22, .60], leads to greater effects sizes than near transfer ($d = .18$), 95% CI [.07, .28], $Q(1) = 4.67$, $p = .03$. This reversal is not due to low power (i.e., 32 tests of far transfer from seven experiments of five studies and 39 tests of near transfer from 10 experiments of eight studies).

The three significant differences found between types of other learning situations were all in comparison to sequential case studies, which composes the majority of our OLS sample (Table 6). That is, comparisons of case comparisons to control conditions ($d = .69$), 95% CI [.58, .80], and to nonanalogous conditions ($d = .82$), 95% CI [.55, 1.08], led to greater effect sizes than did comparisons of case comparisons to sequential case studies ($d = .37$), 95% CI [.31, .44]), $Q(1) = 24.60$, $p < .005$, and $Q(1) = 10.35$, $p < .005$, respectively. Arguably with greater numbers of tests for the other types of OLS, analyses would have been able to distinguish OLS further.

DISCUSSION

Overall our meta-analysis found that case comparisons are generally effective for learning under many different types of implementations, contexts, and measurements. However, the average effect is only medium in size, and there was significant heterogeneity of effects across tasks, contexts, and test measurements, which highlights the importance of investigating which variables moderate the learning outcomes to inform both theory and educational practice.

Reviewing Our Process Model

Our analysis revealed four consistent moderators: objective, principle, content, and lag. In contrast, the type of instruction, experience levels, the provision of features, age, and setting did not moderate findings. The type of cases, domain, and dependent measure initially appeared to moderate findings but then were found to do so inconsistently when confounded variables were examined. Furthermore, conclusions to be drawn from publication rank, duration, and OLS are tentative due to low sample power. We use our process model (Figure 2) to help interpret these findings.

We hypothesized that the first major step in carrying out case comparisons is the *effortful search for commonalities* between the cases. With this in mind, we identified four variables that might affect the success of this search process—the type of instructions provided to learners, the objective of the task, the types of cases, and the experience levels of learners. The type of instructions and the experience levels of learners did not moderate effects. These results suggest for that both prompted and guided instructions are equally beneficial and for learners with varying degrees of experience.

The result that experience level did not moderate findings suggests that case comparisons can be productive with different levels of background knowledge ranging from novice students first being introduced to the topic to more expert students who have already spent much time and practice learning about the domain. Perhaps a more fine-grained measure of expertise or a different operational measure would have led to moderated effects, but the current analysis suggests that case

comparisons can be effective across a range of experience levels.

The type of cases, whether rich or minimal, inconsistently moderated outcomes and only seemed to matter within the domain of math where rich cases yielded greater effects than did minimal cases. It is unclear why the type of cases would matter specifically for mathematics, but the finding raises an important question as to whether the added detail can support the search for commonalities by providing additional context to help identify key features.

Analyses revealed that effect sizes were largest when learners were asked to find the similarities between cases (objective) compared to finding similarities and differences. This result is consistent with the hypothesis that searching for similarities helps students focus on finding the critical features of the cases. In contrast, focusing on differences may highlight the superficial similarities that are not relevant for finding the target features. Focusing on superficial differences may also lead to a higher cognitive load than focusing on only similarities potentially making it more difficult for students to encode the common features of the target content.

This interpretation depends on how the cases are constructed, because an experimenter or instructional designer could create cases where a focus on differences in fact highlights at least some of the key to-be-learned target relations between the two cases (e.g., Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2009). In both studies, Rittle-Johnson and Star wanted learners to recognize and understand not only that the two cases of mathematics procedures were appropriate for the same type of problem but also how they differed and which was more appropriate/efficient. Most studies in the current review utilized case comparisons to illustrate only a single procedure or concept but not to highlight the differences between two types procedures or concepts (for exceptions, see also Gentner et al., 2003; Kurtz et al., 2001; Mason, 2004; Rittle-Johnson et al., 2009; Thompson, Gentner, & Loewenstein, 2000). Furthermore, most posttest measures did not explicitly require learners to distinguish between multiple concepts or procedures. Thus, finding only similarities seems best when learning about single concepts and/or procedures, but further investigations should examine the effects of objective on learning multiple concepts or procedures.

The equivalent effect sizes for the different type of instructions prompted further analysis because it seemed a fundamentally important variable in how comparisons are carried out. We first suspected that the equivalent effect sizes for prompted and guided instructions might have been the result of having the other process variables tailored to the type of instruction. For example, it seemed possible that when instructions were prompted and not guided, the principle and features might have been provided along with the objective only to find similarities in an attempt to simplify the comparison. Therefore, we investigated the frequencies of guided and then prompted studies to determine if rates differed as to the presentation of a principle, the objective, and the provision of features (all variables that moderated), but analyses did not reveal any explanatory patterns. To further investigate, we then cross-tabulated the type of instructions with all other variables (even those that did not moderate) to determine with which other variables it was potentially confounded. The type of instructions was found to be confounded with domain ($C = .47$) and the type of cases ($C = .42$).

Analyses revealed that prompted tasks were most frequently in the domain of other (87 of 165 comparisons) and with rich cases (139 of 165), which were both conditions found to yield the greatest effect sizes. In contrast, guided comparisons were not frequently implemented within the domain of other (22 of 171) or with rich cases (68 of 171). Thus, our functionally equivalent effects for prompted and guided case comparisons might be the result of imbalanced numbers of rich and minimal cases across the different domains—a pattern that was not revealed until we investigated for the purposes of further explanation. Specifically, it may be that the type of prompt interacts with amount of details presented in the case. We hypothesize that guided instruction may be particularly helpful with rich cases where students could potentially focus on irrelevant details, whereas simple prompts may be sufficient for minimal cases. Future work should further help to fill out these cells to test this hypothesis.

As a potentially influential part of both/either the *effortful search for commonalities* and/or the *alignment of target features*, we investigated whether having a principle provided before or after case comparisons led to greater learning. Analyses revealed that effect sizes were greatest when the principle was provided after case comparisons as compared to providing them before or not at all. Providing principles after the comparison may serve as an instructional resource either to reify the emerging knowledge representation acquired or to modify it (Holyoak, 2012). This result is consistent with an emerging literature on the benefits of providing learning resources such as worked examples and direct instruction after an instructional intervention to "prepare students for future learning." Case comparisons may serve as particularly effective instructional interventions to promote such learning. Although our sample had ample power of 29 comparisons from three experiments in which the principle was provided after, all three experiments were from a single paper (Schwartz & Bransford, 1998), and we therefore hope that future work will take advantage of the promising effects of providing the principle after and investigate it further.

The hypothesis that providing the principle before the cases helps students identify and align the key features of the task was not supported by our finding of functionally equivalent effect sizes for both before and not provided conditions. This finding should be further examined, as this is the typical type of instructional scenario in math and science classes where an abstract concept is introduced, which is then followed up with cases illustrating that concept. If this result holds up to further experimental investigation and replication, it may have large implications for STEM instruction.

We also hypothesized that providing the learners with the features of at least one of the two cases could also influence learners' progress through the second major step of *aligning the target features*. Analyses did not reveal differences in effect sizes between studies that provided learners with the features and those that required learners to generate them. This finding suggests that aligning the target features can be accomplished without having to provide the learner with explicit labels of one case to scaffold the process. It also shows that there is not any additional benefit for generating all of the relevant features across the cases, which is in contrast to what might be expected from research on generative versus passive instruction (Chi, 2009). However, it is possible that providing the features could promote more efficient alignment of the features and thereby reduce time on task. In future studies, time measures should be more consistently reported in order to further assess this hypothesis.

Alternative hypotheses regarding the provision of principles and features were concerned with extraneous cognitive load (Paas et al., 2010). It was a concern that providing the principle before comparisons could increase cognitive load, thereby hindering comparisons and subsequent learning, but because analyses did not reveal differences between providing the principle before or not providing a principle, that concern seems unwarranted. The results also do not support the prediction that providing some features reduces extraneous cognitive load, thereby promoting better learning.

After the alignment of target features, the next step within our model is the acquisition of *reduced representations of the cases* such as schemas without extraneous details, because those details are not shared across the cases. However, we also presented an alternative hypothesis that case comparisons could foster deeper concrete understandings of the individual cases and therefore better memory for extraneous details. We expected that these different types of knowledge representations would facilitate different types of performance on the transfer tasks. That is, prior work on transfer has shown that more abstract knowledge representations would be more likely to support far transfer than more specific knowledge, which might not be recalled because it is not seen as similar to the new situation (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi & VanLehn, 2012; Fong, Krantz, & Nisbett, 1986; Fong & Nisbett, 1991). Unfortunately, our results do not provide support for either hypothesis as it revealed inconsistent effects for near and far transfer across confounded variables. For domains other than math, we see support for the hypothesis that comparisons foster concrete, case-specific representations. However, we urge caution in such an interpretation because, although we see greater outcomes for near transfer than for far transfer, we still see benefits of case comparisons over other instruction for far transfer as well, suggesting some degree of abstraction. Furthermore, within the math domain, we found support for the hypothesis that comparisons lead to reduced representations of cases because far transfer led to greater effects than near transfer—indicating representations of the cases were minimal and abstract enough to be readily transferable to different contexts that required modification.

The finding that case comparisons benefit subsequent performances also speaks to the step of *future cases leading to analogical retrieval*. Although we did not find greater effects for far transfer than for near transfer in all domains, effect sizes for both measures clearly indicated a benefit to case comparisons. Thus, what was learned through comparisons was activated and retrieved when learners encountered subsequent cases or related subject matter. We also found greater effect sizes for immediate testing (lag) than for testing on a subsequent day, which implies that temporal proximity also influences the likelihood that analogs/schemas are retrieved for transfer.

The moderators of lag and principle presentation might also be considered together to investigate the possibility that case comparisons lead learners to only tentative conclusions (Holyoak, 2012) in preparation for future learning (Bransford & Schwartz, 1999). The finding that immediate testing is best is consistent with Holyoak's claim that the conclusions drawn from analogy are tentative and require further support. If further support is not found, these tentative conclusions from comparison are forgotten. Of course, it could also be merely an effect of the forgetting curve (Ebbinghaus, 1913). One way to sort out these two possibilities is to run analyses only with conditions that provided the principle after case comparisons to see if lag continues to moderate findings. When analyzed with the current sample, the difference between the effect sizes of immediate testing and subsequent testing is no longer significant.[3] This finding lends support for Holyoak's claim that these inductive conclusions from comparison are tentative but may become less so when supported by further evidence (e.g., the presentation of a principle). Furthermore, when analyses only included conditions that provided the principle after and effect sizes were grouped by whether they were assessments of near or far transfer, the benefits of near transfer also disappear and are found to be functionally equivalent to far transfer.[4] The presentation of the principle after case comparisons allows learners to modify and/or retain their reduced representations of cases, therefore equally facilitating both near and far transfer. Our observed advantages of having the principle provided after the comparisons task lend further support for preparation-for-future-learning perspectives on case comparisons (Bransford & Schwartz, 1999; Schwartz & Bransford, 1998), albeit with a relatively

[3]Analysis of only conditions with the principle provided after ($N = 29$) with effect sizes grouped by the lag in testing reveals no significant difference between immediate testing ($d = 1.07$), 95% CI [.79, 1.36], and testing on subsequent days ($d = 1.60$), 95% CI [1.14, 2.06], $Q(1) = 3.67$, $p = .056$.

[4]Analysis of only conditions with the principle provided after ($N = 29$) with effect sizes grouped by the type of dependent measure reveals no significant difference between near transfer ($d = 1.30$), 95% CI [.97, 1.63], and far transfer ($d = .98$), 95% CI [.62, 1.35], $Q(1) = 1.65$, $p = .20$.

"near" future focus. Again, more designs that present the principle after would better inform these conclusions.

Generalizing Findings Across Contexts

In regard to the context variables, we found that case comparisons are effective in both classroom and laboratory settings, for all content within all domains, and for both children and adults (i.e., the low sample of adolescents does not show reliable effects). However, we do have some reservations about drawing strong conclusions about those variables because setting is confounded with duration, and further analyses revealed the need for more laboratory studies with long durations. Furthermore, the initial moderator analysis revealed smaller effects within the math domain, but further investigation revealed that in conditions in which the duration was brief or the learners were familiar with the subject matter, the domains all showed similar benefits of case comparisons.

Last, we turn to our measurement variables. As was previously discussed, shorter lags between study and test yielded greater effect sizes. Although we did not find overall greater effects for far transfer than for near, we did find such a pattern within the domain of math. Analyses also revealed that when case comparisons were compared to sequential case studies, effect sizes were smaller than when case comparisons were compared to nonanalogous cases and single-case conditions. Although it is possible that learners in sequential conditions could spontaneously compare cases, prior research has shown that this is unlikely to occur (e.g., Rittle-Johnson & Star, 2007). However, as Rittle-Johnson and Star emphasized, sequential conditions are favorable to control groups because they differ from the comparisons condition only in that cases are studied in succession. In that light, the greater effects found in comparison to single case, nonanalogous, and baseline-control conditions may be affected by other factors (e.g., receiving less information, opportunities for practice, etc.).

Recommendations for Classroom Implementation

Many instructors already provide cases following the discussion of a principle or the presentation of a concept in hopes that such cases are illustrative and enhance students' understandings. Instructors aim to provide clarifying cases under the assumption that students will consider the case(s) in light of the principle. The current results suggest that principles presented before may not have a strong effect on what is learned and alternatively that presenting principles after case comparisons may better promote learning the principle. Case comparisons might be a way to encourage students to construct their own explanations and engage with the subject matter in meaningful ways. Furthermore, case comparisons may trigger students' motivational responses to want to better understand why the cases are similar and thereby promote a deeper processing of the principle when presented (Belenky & Nokes-Malach, 2012).

If instructors are interested in incorporating case comparisons into lesson plans, we would advise presenting cases simultaneously—as was the focus of this investigation. Cases do not have to be simplistic, as this analysis revealed either no difference between rich and minimal cases or effect sizes favoring rich cases. Although this sample included few studies with adolescent learners, case comparison tasks work well for both children and adults with any level of experience. One caveat to this generalization is for those working within the domain of math: It seems best to ensure that students are at least familiar with the subject matter and that the session is brief in duration. If the instructor feels that students would benefit from guidance, then providing either directive instructions initially that will guide students to the relevant information or the features of at least one case would not detract from the benefits of case comparisons; however, merely prompted case comparison tasks and those requiring students to generate features seem equally as effective.

Asking students to find only the similarities appears to be most effective when the cases are both illustrating the same concept or procedure. Few studies in our sample attempted to contrast concepts or procedures. However, searching for similarities in order to align cases should also highlight the differences between the cases (Gentner, 2010; Gentner & Sagi, 2006; Ming, 2009) even when the objective does not include differences. Therefore, it might be prudent for instructors to begin with the objective of finding the similarities between cases before shifting attention to the differences and how the cases exemplify other, more nuanced categories. For example, if a biology instructor intends to highlight the characteristics that distinguish the plantae (e.g., a cactus) and fungi (e.g., yeast) kingdoms within the classification system, (s)he might start by having students consider why both kingdoms belong to the eukarya domain (their superordinate classification). While looking for what is true about both a cactus and yeast, students will also notice differences between the cases, which can subsequently be attended to in order to further explore their subordinate classifications (plantae vs. fungi).

Case comparisons are not only used as learning interventions. Some textbooks and teachers use case comparisons after lessons as assessments of learning, and our analyses suggest that learners would learn from those comparisons as well (i.e., as indicated by effect sizes for when principles were provided prior to comparisons) but to a lesser extent than when they precede a lesson. However, if case comparisons are being used as categorization tasks or formative assessments following lessons, then some of our suggestions might not be appropriate given the shift to assessment rather than instructional goals. For example, providing features might be too leading or revealing. In such cases, teachers might decide not to provide features so that students' alignment and/or categorization of cases are not dictated by the task's provided

details and are instead left up to the student to identify. Not providing such cues should not reduce the potential for learning from those comparisons.

Future Work

There are still several pieces of our process model that require further investigation because they are either unresolved after the current meta-analysis or, more importantly, not directly addressable with the information currently provided within the literature. First, what is the average, subjective cognitive load during case comparisons? Measuring learners' cognitive load during a variety of case comparison tasks (prompted or guided, with features provided or generated, with the objective to find similarities or both similarities and differences, etc.) would provide insight into what the major challenges are for learners across a variety of implementations. How do the demands of the task directly impact reported cognitive load?

Second, what types of knowledge representations are derived from such tasks? Are they mostly schema based or case based, or might memory include both and the characteristics of the test task determine which is retrieved? Measures of recall and/or recognition of cases that can sort out these different types of memories would also help in investigating these underlying mechanisms. Does encountering a new case lead learners with experience to retrieve the analogous case and to encode the current case by aligning it with the previous, or does such experience only prompt the retrieval of the relevant schema? Perhaps through think-aloud protocols we can begin to examine these more long-term consequences of case comparison tasks.

In conclusion, case comparisons are effective learning activities that can help learners acquire more abstract concepts and procedures. The task of comparing cases utilizes the human capacity for analogy making and prompts learners to construct an understanding by figuring out how/why those cases are related. We recommend having learners attend to the similarities between cases and following comparisons with instruction of the principle.

ACKNOWLEDGMENTS

Work on this project was funded by a grant from the Institute for Educational Sciences, #R305C080009. We would like to Bethany Rittle-Johnson and the anonymous reviewers for their very helpful suggestions and comments on the article.

REFERENCES

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

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences*, *21*, 399–432. doi:10.1080/10508406.2011.651232

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

Boroditsky, L. (2007). Comparison and the development of knowledge. *Cognition*, *102*, 118–128. doi:10.1016/j.cognition.2002.08.001

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*, 61–100. doi:10.2307/1167267

Buehl, D. (2008). *Classroom strategies for interactive learning (*3rd *ed.)*. Newark, DE: International Reading Association.

*Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1147–1156. doi:10.1037//0278–7393.15.6.1147

*Chen, Z., & Daehler, M. W. (1989). Positive and negative transfer in analogical problem solving by 6-year-old children. *Cognitive Development*, *4*, 327–344. doi:10.1016/S0885-2014(89)90031-2

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*, 73–105. doi:10.1111/j.1756-8765.2008.01005.x

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanation: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145–182. doi:10.1207/s15516709cog1302_1

Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist*, *47*, 177–188. doi:10.1080/00461520.2012.695709

*Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*, 356–373. doi:10.1080/15248371003700015

*Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical comparison. *Cognitive Science*, *15*, 89–132.

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

*Cummins, D. D. (1992). Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1103–1124. doi:10.1037//0278-7393.18.5.1103

Druit, R. (1991). On the role of analogies and metaphors in learning science. *Science Education*, *75*, 649–672.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Teachers College, Columbia University. doi:10.1037/10011–000

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*, 253–292. doi:0010-0285186

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, *120*(1), 34–45. doi:10.1037/0096–3445.120.1.34

*Gadgil, S., & Nokes, T. J. (2009, July). *Analogical scaffolding in collaborative learning*. Paper presented at the annual meeting of the Cognitive Science Society, Amsterdam, The Netherlands.

Gee, B. (1978). Models as a pedagogical tool: Can we learn from Maxwell? *Physics Education*, *13*, 287–291. doi:10.1088/0031-9120/13/5/004 Retrieved from http://iopscience.iop.org/0031-9120/13/5/004

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, MA: MIT Press.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*, 752–775. doi:10.1111/j.1551-6709.2010.01114.x

*Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*, 393–405. doi:10.1037/0022-0663.95.2.393

*Gentner, D., Loewenstein, J., & Thompson, L. (2004, August). *Analogical encoding: Facilitating knowledge transfer and integration*. Paper presented at the twenty-sixth annual meeting of the Cognitive Science Society, Chicago, IL.

*Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, *33*, 1343–1382. doi:10.1111/j.1551-6709.2009.01070.x

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56. doi:10.1037//0003-066X.52.1.45

*Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*, 487–513. doi:10.1016/S0885-2014(99)00016-7

Gentner, D., & Sagi, E. (2006). Does "different" imply a difference? A comparison of two tasks. In R. Sun & N. Miyake (Eds.), *Proceedings of the twenty-eighth annual conference of the Cognitive Science Society* (pp. 261–266). Mahwah, NJ: Erlbaum.

*Gerjets, P., Scheiter, K., & Schuh, J. (2008). Information comparisons in example-based hypermedia environments: Supporting learners with processing prompts and an interactive comparison tool. *Education Technology Research and Development*, *56*, 73–92. doi:10.1007/s11423–007–9068-z

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355. doi:10.1016/0010-0285(80)90013-4

*Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38. doi:10.1016/0010-0285(83)90002-6

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*, 231–262. doi:10.1016/S0010–0277(97)00047–4

Goldstone, R. L., Day, S., & Son, J. Y. (2010). Comparison. In B. Glatzeder, V. Goel, & A. von Müller (Eds.), *On thinking: Towards a theory of thinking* (Vol. II, pp. 103–122). Heidelberg, Germany: Springer-Verlag. doi:10.1007/978-3-642-03129-8_7

*Graham, S. A., Namy, L. L., Gentner, D., & Meagher, K. (2010). The role of comparison in preschoolers' novel object categorization. *Journal of Experimental Child Psychology*, *107*, 280–290. doi:10.1016/j.jecp.2010.04.017

Hofstadter, D. R. (2001). Epilogue: Analogy as the core of cognition. In D. G. Gentner, K. J. Holyoak, & B. K. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 449–538). Cambridge, MA: MIT Press.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). New York, NY: Oxford University Press.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods*, *11*, 193–206. doi:10.1037/1082-989X.11.2.193

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466. doi:10.1037//0033-295X.104.3.427

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–263. doi:10.1037/0033-295X.110.2.220

Iding, M. K. (1997). How analogies foster learning from science texts. *Instructional Science*, *25*, 233–253.

James, W. (1890). *The principles of psychology, volume 1*. New York, NY: Holt.

Johnson, B. (1989). *DSTAT: Software for the meta-analytic review of research literature*. Hillsdale, NJ: Erlbaum.

Johnson, B. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literature: Upgrade documentation*. Hillsdale, NJ: Erlbaum.

*Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797–2822. doi:10.2307/1131753

*Kurtz, K. J. (2005). Re-representation in comparison: Building an empirical case. *Journal of Experimental & Theoretical Artificial Intelligence*, *17*, 447–459. doi:10.1080/09528130500324255

*Kurtz, K. J., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. *Memory & Cognition*, *35*, 334–341. doi:10.3758/BF03193454

*Kurtz, K. J., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*, 417–466. doi:10.1207/S15327809JLS1004new_2

Lewis, J. R. (1933). Analogies in teaching freshman chemistry. *Journal of Chemical Education*, *10*, 627–630. doi:10.1021/ed010p627

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The ANNALS of the American Academy of Political and Social Science*, *587*, 69–81. doi:10.1177/0002716202250791

*Loewenstein, J., & Gentner, D. (2001). Spatial mapping in preschoolers: Close comparisons facilitate far mappings. *Journal of Cognition and Development*, *2*, 189–219. doi:10.1207/S15327647JCD0202_4

Loewenstein, J., & Thompson, L. (2000). The challenge of learning. *Negotiation Journal*, *16*, 399–408. doi:10.1023/A:1026692922914

*Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6*, 586–597. doi:10.3758/BF03212967

*Loewenstein, J., Thompson, L., & Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning and Education*, *2*, 119–127. doi:10.5465/AMLE.2003.9901663

*Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467. doi:10.1006/cogp.1993.1011

*Mason, L. (2004). Fostering understanding by structural alignment as a route to analogical encoding. *Instructional Science*, *32*, 293–318.

*Michael, A. L., Klee, T., Bransford, J. D., & Warren, S. F. (1993). The transition from theory to therapy: Test of two instructional methods. *Applied Cognitive Psychology*, *7*, 139–153. doi:10.1002/acp.2350070206

Ming, N. (2009). Analogies vs. contrasts: A comparison of their learning benefits. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *Proceedings of the second international conference on analogy* (pp. 338–347). Sofia, Bulgaria: NBU Press.

*Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2009). Superior discrimination between similar stimuli after simultaneous exposure. *The Quarterly Journal of Experimental Psychology*, *62*, 18–25. doi:10.1080/17470210802240614

*Nagarajan, A., & Hmelo-Silver, C. (2006). Scaffolding learning from contrasting video cases. *Proceedings of the 7th International Conference on Learning Sciences*, 495–501.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, *131*, 5–15. doi:10.1037//0096–3445.131.1.5

*Namy, L. L., Gentner, D., & Clepper, L. E. (2007). How close is too close? Alignment and perceptual similarity in children's categorization. *Cognition, Brain, Behavior*, *11*, 647–659.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington, DC: National Academy Press.

*Nokes, T. J., VanLehn, K., & Belenky, D. M. (2008, July). Coordinating principles and examples through analogy and explanation. Poster presented at the Thirtieth Annual Conference of the Cognitive Science Society, Washington, DC. doi:10.1007/s10212-012-0164-z

Paas, F., Renkl, A., & Sweller, J. (2010). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*, 1–4. Retrieved from doi:10.1207/S15326985EP3801_1

*Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, *35*, 28–43. doi:10.1016/j.cedpsych.2009.09.001

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*, 1128–1129. doi:10.1126/science.1142103

*Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*, 561–574. doi:10.1037/0022-0663.99.3.561

Rittle-Johnson, B., & Star, J. R. (2011). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons.. In J. P. Mestre & B. H. Ross (Eds.), *Cognition in education* (Vol. 55, pp. 199–226). Oxford, UK: Academic.

*Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge on equation solving. *Journal of Educational Psychology*, *101*, 836–852. doi:10.1037/a0016026

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: Comment on Ferguson and Brannick (2012). *Psychological Methods*, *17*, 129–136. doi:10.1037/a0027128

*Scheiter, K., Gerjets, P., & Schuh, J. (2004, June). *The impact of example comparisons on schema acquisition: Do learners really need multiple examples?* Paper presented at the 6th International Conference of the Learning Sciences, Santa Monica, CA.

*Schuh, J., Gerjets, P., & Scheiter, K. (2005, July). *Fostering the acquisition of transferable problem-solving knowledge with an interactive comparison tool and dynamic visualizations of solution procedures*. Paper presented at the Twenty-seventh Annual Conference of the Cognitive Science Society, Stresa, Italy.

Schustack, M. W. & Anderson, J. R. (1979). Effects of analogy to prior knowledge on memory for new information. *Journal of Verbal Learning and Verbal Behavior*, *18*, 565–583. doi:10.1016/S0022-5371(79)90314-1

*Schwartz, D. L. & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*, 475–522. doi:10.1207/s1532690xci1604_4

Seifert, C. M., McKoon, G., Abelson, R. P., & Ratcliff, R. (1986). Memory connections between thematically similar episodes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *12*, 220–231. doi:10.1037//0278-7393.12.2.220

*Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction*, *13*, 227–237. doi:10.1016/S0959-4752(02)00022-1

Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., … Wray, J. (2010). *Developing effective fractions instruction for kindergarten through 8th grade: A practice guide* (NCEE #2010-4039). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://whatworks.ed.gov/ publications/practiceguides

*Spencer, R. M., & Weisberg, R. W. (1986). Context-dependent effects on analogical transfer. *Memory & Cognition*, *14*, 442–449.

*Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology*, *102*, 408–426. doi:10.1016/j.jecp.2008.11.004

*Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organizational Behavior and Human Decision Processes*, *82*, 60–75. doi:10.1006/obhd.2000.2887

Treagust, D. F., Druit, R., Joslin, P., & Lindauer, I. (1992). Science teachers' use of analogies: Observations from classroom practice. *International Journal of Science Education*, *14*, 413–422. doi:10.1080/0950069920140404

Webb, M. J. (1985). Analogies and their limitations. *School Science and Mathematics*, *85*, 645–650. doi:10.1111/j.1949-8594.1985.tb09677.x

Weller, C. M. (1970). The role of analogy in teaching science. *Journal of Research in Science Teaching*, *7*, 113–119. doi:10.1111/j.1949-8594.1985.tb09677.x