pedocs
Open Access Erziehungswissenschaften

DIPF

Schwichow, Martin Geert; Croker, Steve; Zimmerman, Corinne; Höffler, Tim Niclas; Härtig, Hendrik

Teaching the control-of-variables strategy: A meta-analysis

Developmental Review 39 (2016) 1, S. 37-63

Empfohlene Zitierung/ Suggested Citation:
Schwichow, Martin Geert; Croker, Steve; Zimmerman, Corinne; Höffler, Tim Niclas; Härtig, Hendrik: Teaching the control-of-variables strategy: A meta-analysis - In: Developmental Review 39 (2016) 1, S. 37-63 - URN: urn:nbn:de:0111-pedocs-126966

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: http://creativecommons.org/licenses/by/4.0/deed.de - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License:
http://creativecommons.org/licenses/by/4.0/deed.en - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor.
By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
Deutsches Institut für Internationale Pädagogische Forschung (DIPF)
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

Developmental Review 39 (2016) 37–63

Contents lists available at ScienceDirect

Developmental Review

journal homepage: www.elsevier.com/locate/dr

Teaching the control-of-variables strategy: A meta-analysis

Martin Schwichow [a], Steve Croker [b], Corinne Zimmerman [b,*], Tim Höffler [a], Hendrik Härtig [a]

[a] IPN – Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24098, Kiel, Germany
[b] Department of Psychology, Illinois State University, Campus Box 4620, Normal, IL 61790, USA

ARTICLE INFO

Article history:
Received 3 February 2014
Revised 4 December 2015
Available online 24 December 2015

Keywords:
Control-of-variables strategy
Meta-analysis
Experimentation skills
Inquiry skills
Scientific reasoning
Science instruction

ABSTRACT

A core component of scientific inquiry is the ability to evaluate evidence generated from controlled experiments and then to relate that evidence to a hypothesis or theory. The control-of-variables strategy (CVS) is foundational for school science and scientific literacy, but it does not routinely develop without practice or instruction. This meta-analysis summarizes the findings from 72 intervention studies at least partly designed to increase students' CVS skills. By using the method of robust meta-regression for dealing with multiple effect sizes from single studies, and by excluding outliers, we estimated a mean effect size of $g = 0.61$ (95% CI = 0.53–0.69). Our moderator analyses focused on design features, student characteristics, instruction characteristics, and assessment features. Only two instruction characteristics – the use of cognitive conflict and the use of demonstrations – were significantly related to student achievement. Furthermore, the format of the assessment instrument was identified as a major source of variability between study outcomes. Implications for teaching and learning science process skills and future research are discussed.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

* Corresponding author. Department of Psychology, Illinois State University, Campus Box 4620, Normal, IL 61790, USA. Fax: 309-438-5789.
E-mail address: czimmer@ilstu.edu (C. Zimmerman).
http://dx.doi.org/10.1016/j.dr.2015.12.001
0273-2297/© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

In science, controlled experiments are crucial for drawing valid inferences about causal hypotheses. Valid inferences are only possible if an experiment is designed in a way that alternative causal effects or interactions can be excluded. Therefore, all variables except the one being investigated should ideally be held constant (or "controlled") across experimental conditions (Dewey, 2002; Popper, 1966). The cognitive and procedural skills associated with being able to select or conduct controlled experiments have been of interest to both science educators and psychologists who are interested in the development of scientific thinking. Descriptions of the specific skill of controlling experiments include "isolation of variables" (Inhelder & Piaget, 1958), "vary one thing at a time" (VOTAT; Tschirgi, 1980), and the "control of variables strategy" (Chen & Klahr, 1999). For the remainder of this paper, we will refer to this critical science process skill as the control-of-variables strategy (CVS).

Resulting from its fundamental importance in science, CVS is also addressed in standards and curriculum materials for science education. In particular, the *Framework for K-12 Science Education* (National Research Council, 2012) makes a distinction between the concepts and processes of science, outlining various scientific and engineering practices related to CVS such as asking questions, conducting investigations, and interpreting and using evidence. The *Next Generation Science Standards* (NGSS; NGSS Lead States, 2013) are defined in the context of science and engineering practice. Furthermore, scientific process skills such as CVS are required for learning through inquiry as they enable students to conduct their own informative investigations. Reasoning on the basis of unconfounded evidence is crucial not only in science but in all argumentation about causality. Again, current science standards focus on skills such as the ability to construct arguments and to argue on the basis of evidence (NGSS, 2013; NRC, 2012), which require students to produce interpretable evidence. Hence, an understanding of the importance and principles of unconfounded evidence is required for critical thinking in general and is linked to broader educational goals, such as inquiry skills and argumentation (Kuhn, 2005a). The control of variables strategy, therefore, plays a supporting role in many of the science and engineering practices that are the focus of current science education reform.

The prominent role of CVS in scientific reasoning and science education has made it the focus of much research. The domain-general adaptability of CVS has also made it an ideal task for developmental psychologists to study cognitive development in children. For example, Inhelder and Piaget's (1958) theory that children's thinking develops from concrete to abstract was based, in part, on observations of children's performance on tasks that involve manipulating and isolating variables (e.g., pendulum task, ramps task). Consequently, investigations of people's ability to design and interpret controlled experiments can be classified as either *investigative studies*, in which the development of skill on CVS tasks is correlated with other measured skills or individual differences (e.g., Cloutier & Goldschmid, 1976; Linn, Clement, & Pulos, 1983), or *intervention studies*, which explore the impact of instruction on students' achievement on CVS tasks (e.g., Chen & Klahr, 1999; Lawson & Wollman, 1976).

Investigative studies show that even elementary students are able to *select* controlled experiments and to interpret unconfounded evidence when the experimental data are consistent with students' beliefs and preconceptions (e.g., Croker & Buchanan, 2011; Schulz & Gopnik, 2004; Sodian, Zaitchik, & Carey, 1991). However, it is also evident that students (Bullock & Ziegler, 1999; Croker & Buchanan, 2011; Kuhn, Garcia-Mila, Zohar, & Anderson, 1995; Schauble, 1996; Tschirgi, 1980) and even adults (Kuhn, 2007) perform poorly on tasks when the task domain includes information that conflicts with their current beliefs and preconceptions. Across many studies, it is evident that most students and even some adults do not have a generalized understanding of CVS because their ability to identify, select, or design controlled experiments depends on the task content or situational factors (Koslowski, 1996; Linn et al., 1983; for a review see Zimmerman & Croker, 2013). Additionally, Siler and Klahr (2012) outline the procedural misconceptions about controlling variables that have been identified. For example, students often over-extend a "fairness schema" to produce experiments that are completely equivalent (i.e., identical), they often have trouble making the distinction between a variable and the variable levels, and they often misunderstand the goal of the task as to be one that is consistent with engineering an outcome rather than finding out about the causal status of a single variable.

Decades of research on the development of scientific thinking in general, and on experimentation skills in particular, show a long trajectory that requires educational scaffolding (Klahr, Zimmerman, & Jirout, 2011; Kuhn, Iordanou, Pease, & Wirkala, 2008; see also Sodian & Bullock, 2008 for a collection

of papers; see Zimmerman, 2000, 2007 for reviews). Investigative studies have done much to add to our basic understanding of the developmental and educational factors that influence how individuals select or design experiments and interpret evidence from controlled or uncontrolled experiments. Such findings can be used to inform the design of intervention studies (e.g., Klahr & Li, 2005).

Intervention studies, in contrast, investigate whether and how students' ability to design controlled experiments can be improved by instruction. The first intervention studies were conducted by developmental psychologists to test Inhelder and Piaget's (1958) claim that the acquisition of formal reasoning strategies such as CVS cannot be accelerated by instruction (e.g., Case & Fry, 1973; Siegler, Liebert, & Liebert, 1973). Evidence from those studies demonstrated that accelerating students' understanding of CVS is indeed possible. Numerous intervention studies were conducted between 1973 and 1988. These studies were quite variable with respect to instructional methods, student populations, type of achievement test used, and findings. For example, Case and Fry (1973) report a significant advantage of six-year-old students receiving CVS training over students in a control group, whereas Padilla, Okey, and Garrard (1984) found no influence of CVS training on the achievement of 14-year-old students. To make sense of the variability in research methods and findings, Ross (1988a) conducted a meta-analysis on this set of training studies.

Ross's (1988) meta-analysis

The meta-analysis conducted by Ross (1988a) summarized the results of 65 intervention studies conducted between 1973 and 1988. The studies were carried out to answer theoretical questions and to evaluate new science curricula and programs. Accordingly, the meta-analysis included experimental laboratory studies and quasi-experimental classroom studies. The methods used to instruct treatment groups range from providing explicit lectures about CVS (e.g., Linn, 1978) to asking students to discover the principles of CVS on their own (e.g., Purser & Renner, 1983). The tests used to measure treatment effects differ between and within studies in format, content, and range. Studies that included a control group comparison and focused at least partly on CVS during instruction and testing were included in the meta-analysis. A mean effect size of $d = 0.73$ (95% CI = 0.54–0.92) estimated by Ross (1988a) shows that interventions aimed at teaching CVS can be effective.

Ross identified several differences between studies that moderated their outcomes. He found that published studies had larger effect sizes than non-published reports or dissertations and that studies focusing only on teaching CVS showed larger effect sizes than studies teaching additional skills. Studies that provided practice opportunities using both school and out-of-school contexts were more effective than studies in which students practiced CVS skills in either context alone. When students were given feedback about their performance on training tasks there were larger effect sizes compared to when students received no feedback. In addition, studies using an assessment designed for that particular study showed larger effect sizes than studies using assessments that had been developed by other researchers. Larger effect sizes were evident when students were assessed on the same tasks that were used during instruction, relative to studies that used novel tasks to assess instructional effectiveness. Furthermore, when an assessment identified the relevant independent variables for the participants, effect sizes were smaller when compared to more challenging assessments in which the participants had to encode the variables for themselves.

The current meta-analysis

During the past 25 years, a second wave of intervention and investigative studies on CVS has been conducted. These studies differ from those included in Ross's (1988a) meta-analysis in a number of ways, including the use of computerized instructional materials, computerized performance tests, and the inclusion of younger students as participants. The second wave of research was less concerned with testing the details of Piagetian theory (e.g., whether children not yet in the formal operations stage could be taught the control-of-variables strategy), and focused more on determining which types of interventions work best. Research questions include, for example, whether particular types of instruction are more effective (Chen & Klahr, 1999; Dean & Kuhn, 2007; Klahr, 2005; Klahr & Nigam,

2004; Kuhn, 2005b; Kuhn & Dean, 2005), and whether hands-on activities and virtual training tasks are equally effective in teaching CVS (e.g., Klahr, Triona, & Williams, 2007).

Because a large body of research has been conducted since Ross's (1988a) meta-analysis – 42 studies published after 1988 are included in the current meta-analysis – and because these studies pose different questions and use different methods and populations, we conducted a new meta-analysis focusing on intervention studies. The goal of this meta-analysis was to identify features of effective instruction, features of assessment instruments, and characteristics of students that moderate the study outcome. In addition, we investigated whether Ross's (1988a) findings would be replicated with newer meta-analytical approaches. For example, new methods allowed us to investigate whether Ross's findings depended on the inclusion of outliers or the methodological approaches he used. Analyzing the effect of outliers is important for two key reasons. First, excluding outliers provides a more precise estimate of treatment effect sizes. Second, the identification of accurate (or more conservative) effect sizes will prevent the frustration of teachers and researchers who may implement reported interventions and/or assessments. Current approaches to meta-analysis include procedures for handling outliers (Huffcutt & Arthur, 1995) and dependency of effect sizes due to multiple effect sizes from single studies (Hedges, Tipton, & Johnson, 2010). In the following sections we present a review of our moderator variables before describing the methods and results.

Moderator variables

We examined the potential reasons for variance between study outcomes by coding studies with respect to design features, student characteristics, instruction characteristics, and assessment characteristics. At the most global level, we coded the *publication type.* It is well known that studies with large, significant effects are more likely to be published than studies with non-significant or small effects (Lipsey & Wilson, 2001). Therefore, we coded publication type and made efforts to find non-published reports (see Methods). Studies were coded into one of two categories: (a) peer-reviewed journal articles and book chapters, or (b) unpublished reports, theses, dissertations, or published conference proceedings.

In the current meta-analysis we included research using two main types of *study design*: experimental designs, typically done in the laboratory, and quasi-experimental designs, typically done in the classroom. In experimental designs, students are randomly assigned to either a control or a treatment group. In quasi-experimental designs, it is common for whole classes to be allocated to the intervention or control condition, and thus systematic differences other than the treatment could influence the outcome. For example, in a study by Ross (1986), teachers could decide whether they wanted to teach the treatment condition or the control condition. It is possible that more enthusiastic teachers chose to teach the treatment condition. Hence, differences related to teachers may have been responsible for some of the achievement differences. However, classroom studies are relevant because, in addition to being more ecologically valid, they are more likely to influence the praxis of teaching than laboratory studies (Hofstein & Lunetta, 2004), and are therefore included in our analysis.

As we are interested in examining the effects of instructional interventions relative to a control, it is important to consider the nature of the *control group activity*. We coded the activities that the control or comparison group engaged in while the treatment group(s) received CVS instruction. For example, in some laboratory studies, the control group received no instruction of any kind (e.g., Lawson & Wollman, 1976). In contrast, some laboratory studies and most classroom studies used a comparison group that received some kind of non-CVS instruction while the treatment group(s) received CVS instruction. For example, a comparison group may receive instruction on the same content domain of the tasks used by those receiving CVS instruction (e.g., Zohar & David, 2008). In other cases, the comparison group may use the same equipment the CVS group uses, but without any CVS-related instruction (e.g., Keselman, 2003).

The remainder of our review of potential moderator variables is organized in three subsections: (a) student characteristics, (b) instruction characteristics, and (c) assessment characteristics. Each section includes a brief rationale for the inclusion of the moderator variables and a preview of how they were coded.

Student characteristics

Among the student characteristics that might moderate the study results, *age* is most commonly investigated. Piaget's early research and theorizing led to the prediction that children would not be able to use CVS until reaching adolescence (Inhelder & Piaget, 1958). However, many studies since then have shown that teaching CVS to elementary school children is possible (e.g., Chen & Klahr, 1999; Grygier, 2008; Sodian, Jonen, Thoermer, & Kircher, 2006). To investigate whether learning is age dependent, some cross-sectional studies have compared different age groups who are instructed and tested on the same materials. Cross-sectional studies with elementary school children (Chen & Klahr, 1999; Dejonckheere, van de Keere, & Tallir, 2011) as well as with secondary school children (Danner & Day, 1977; Goossens, Marcoen, & Vandenbroecke, 1987) found a significantly larger learning effect in the older groups. However, in all of these studies, the treatment groups do significantly better than the control groups even in the younger cohort. Therefore, it is necessary to examine age as a potential moderator.

Another potential source of variance between study outcomes is the general *achievement level* of the students. Although teachers often believe that only high-achieving students are capable of higher-order thinking skills such as CVS (Raudenbush, Rowan, & Cheong, 1993), Klahr and Li (2005) showed that low- and high-achieving students are equally able to learn CVS. Zohar and colleagues found that the pretest-posttest gains of low-achieving students were higher than the gains of high-achieving students in a laboratory study (Zohar & Peled, 2008) and a classroom study (Zohar & David, 2008). However, this effect was not replicated by Lorch et al. (2010). In some studies, information about socioeconomic status (SES) was used as a proxy for achievement level (e.g., Case & Fry, 1973) because of the correlation between SES and school outcomes (e.g., Sirin, 2004). To preview, despite the importance of achievement level as a potential moderator, this variable proved difficult to code because of the lack of information provided. We return to this issue in the Discussion section.

Instruction characteristics

The settings, materials, and methods used to instruct students vary widely between studies. Obviously, studies differ in *instruction or treatment duration*. Single intervention studies often last from a few minutes to a few hours (e.g., Chen & Klahr, 1999; Siegler et al., 1973). Microgenetic studies involve repeated instruction sessions over the course of several weeks (Kuhn et al., 1995; Schauble, 1996) whereas curriculum studies can take several years (e.g., Adey & Shayer, 1990; Bowyer & Linn, 1978). However, investigating the moderator effect of treatment duration is problematic, as longer and shorter interventions differ also with respect to design (experimental versus quasi-experimental), the number of teachers involved, and the quantity of additional instructional objectives. Despite these potential problems, treatment duration was considered and was recorded as a continuous variable, measured in minutes (see Methods).

A related moderator variable is the *focus of the instruction*. That is, an intervention might focus only on CVS, or – in the case of longer interventions – it may include additional instructional objectives such as content knowledge or other science process skills such as observation, measurement, or the evaluation of evidence (Adey & Shayer, 1990; Amos & Jonathan, 2003). Therefore, we coded whether a study had a CVS-focus or was more focused on general science skills and knowledge.

Instruction type is clearly an important characteristic and one that has received a lot of attention. However, there are issues with potential misunderstandings based on the everyday connotations of the labels that are used to describe intervention types (for an extended discussion, see Klahr, 2009; see Klahr & Li, 2005 for a discussion of media reactions to intervention studies that used particular labels such as "discovery learning" or "direct instruction"). Ross (1988a) referred to this characteristic as either "the amount of support provided to the problem solver" (p. 406) or "level of intensity" (p. 421). For the minimal amount of support, Ross included treatments that "consist of practice in designing experiments, sometimes in large amounts, without providing specific direction to students in how to benefit from the practice" (p. 421). Such interventions may or may not involve teacher feedback. In contrast, Ross (1988a) described instruction that includes much student support as "the *rules provided* type" (p. 423, emphasis in original). Students are typically given explicit rules about how to design controlled experiments, and teachers may illustrate and explain the use of those rules with

example experiments. Although Ross (1988a) found different effect sizes for these instruction types, the differences failed to reach statistical significance.

In the newer wave of intervention studies published since 1988, some studies provide evidence that explicit explanations of CVS are more efficacious than learning with lower levels of support (Chen & Klahr, 1999; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008), but other studies do not replicate this finding (Dean & Kuhn, 2007; Kuhn & Dean, 2005). Additionally, evidence from microgenetic studies shows that students do improve their experimentation strategies when working on multivariable tasks for a longer period of time and are, therefore, able to learn appropriate knowledge of CVS with extended practice opportunities (Kuhn & Phelps, 1982; Kuhn, Schauble, & Garcia-Mila, 1992). Given the importance of this issue in the literature, we coded whether an instructional intervention included the explicit mention of a rule for how to design a controlled experiment or not.

Another difference between studies concerns the *use of training tasks* during instruction. The type of equipment used during instruction is a potential moderator variable. In some studies, students are trained on experimentation skills using real equipment (e.g., Ford, 2005; Lawson & Wollman, 1976) or virtual experimental setups (e.g., Kuhn & Dean, 2005; Lin & Lehman, 1999). Other types of instruction, however, do not include training on performance tasks at all. A study by Padilla et al. (1984) reports the advantage of a group that received a demonstration plus practical training over a group that received the demonstration alone. Recent research shows that training tasks have a positive impact on students' CVS achievement but that it does not matter if the tasks are virtual or physical (Klahr et al., 2007; Smetana & Bell, 2012; Triona & Klahr, 2003). For the purposes of our analysis, we considered whether the instruction did or did not include any type of training task.

Studies differ in whether or not students receive *feedback* on their performance on training tasks (Huppert, Lomask, & Lazarowitz, 2002; Lawson & Wollman, 1976). Because of the evident power of feedback in supporting students' learning in general (Hattie, 2008; Hattie & Timperley, 2007), this moderator variable might be correlated with higher student achievement when teaching CVS.

Demonstrations of controlled and uncontrolled experiments are common (e.g., Matlen & Klahr, 2013; Padilla et al., 1984). A demonstration is a didactic presentation of a controlled experiment by the teacher. Demonstrations were sometimes used to support verbal explanations of CVS (Strand-Cary & Klahr, 2008). Demonstrations are not used in all CVS instruction (e.g., Day & Stone, 1982; Zion, Michalsky, & Mevarech, 2005), such as interventions using minimal support (e.g., Bowyer & Linn, 1978).

Additionally, we coded for the presence of an instructional technique known as *cognitive conflict*. This concept has roots in Piagetian theory (Limón, 2001; McCormack, 2009), with many researchers explicitly working within a Piagetian theoretical framework (e.g., Bredderman, 1973; Lawson & Wollman, 1976; McCormack, 2009). For example, in the Ross (1988a) meta-analysis, cognitive conflict was described as such: "In this strategy student conceptions and expectations were overtly challenged to create disequilibrium" (p. 419). The key idea is that the teacher presents discrepant or anomalous information, typically in the form of an uncontrolled comparison, with the goal that the student will notice "the inherent indeterminacy of confounded experiments" (Chen & Klahr, 1999, p. 1098). In more recent work, within a broadly defined constructivist framework, cognitive conflict is defined with reference to the activity of the teacher and its intended goal on student learning. Limón (2001) operationally defines the cognitive conflict paradigm: The teacher must first identify the student's current knowledge and then explicitly confronts the student with contradictory information. To assess the effectiveness of the technique, the student's ideas before and after the intervention are compared. This technique is used in science education to promote conceptual change about specific phenomena, in particular, those subject to misconception.

In the context of CVS instruction, however, what the teacher is drawing attention to is whether or not a particular (confounded) comparison allows one to draw conclusions about the effect of a particular variable. The teacher tries to induce cognitive conflict in students by drawing attention to a current experimental procedure or interpretation of empirical data (set up by either the experimenter or the student) in an attempt to get the student to notice that the comparison or conclusion is invalid (e.g., Adey & Shayer, 1990). For example, Lawson and Wollman (1976) asked students to predict which of two different balls would bounce higher. To test the students' predictions, the teacher conducted an unfair experiment in which the ball type and the height from which it was released were confounded. This procedure continued until the students recognized that everything other than the variable

under investigation needed to be consistent across comparisons. Strand-Cary and Klahr (2008) induced cognitive conflict by asking students whether they could tell *for sure* whether the variable under consideration had an effect, after (a) the students had conducted an experiment, and (b) the experimenter had provided examples of both confounded and unconfounded experiments. This procedure required students to reflect on their experimental design and whether the results would or would not be informative. Studies were coded for the presence or absence of instructional techniques designed to challenge students' existing misconceptions about controlling variables via cognitive conflict. Although the idea behind this instructional technique originated within the Piagetian theoretical framework, our coding focused on the actions taken by the teacher, rather than the putative cognitive mechanism (e.g., disequilibrium, accommodation). Interestingly, in many cases, cognitive conflict was induced via the use of demonstrations. Although we coded the presence or absence of both cognitive conflict and demonstrations separately and independently, these two instructional features often co-occur. We return to this point in the Results and Discussion sections.

The *contexts* of training tasks, demonstrations, and lectures also vary among studies. The current meta-analysis is limited to intervention studies using at least some content related to the natural sciences, as we want to be able to draw conclusions for implementing effective CVS instruction in science classes. The majority of studies used content related to physics, biology, chemistry, or geo-sciences, but some studies used content related to the everyday life of students. For example, Lawson and Wollman (1976) demonstrated the difference between good and bad experiments on bouncing balls, and Beishuizen, Wilhelm, and Schimmel (2004) used simulation tasks about the impact of food on the health of an imaginary person. It is possible that such everyday life contexts are more meaningful for students and increase instructional efficacy. Therefore, we coded for school science versus out-of-school contexts.

Assessment characteristics

Another potential source of variance comes from the variety of assessment instruments used to measure the treatment effect. The impact of test characteristics on the scores of single students (Staver, 1984) and across study outcomes (Ross, 1988a) is evident. For instance, Staver (1984) found significant differences between students tested using individual clinical interviews and students tested with group-administered tests. Additionally, Staver (1986) found that students' scores on multiple-choice tests were higher than their scores on open-response tests when four or more independent variables had to be considered. Thus, one potential moderator variable is the *test format*. The intervention studies summarized in this meta-analysis used paper-and-pencil tests in either a multiple-choice or open-response format, or they used performance tasks. Additionally, when performance tasks were used, we coded whether they were virtual or hands-on performance tests.

Furthermore, *the number of independent variables* that students were required to consider in the assessment task has the potential to moderate how challenging tasks are, because the cognitive load of tasks increases with an increasing number of variables. This again could influence the measured group differences because treatment groups have been trained to focus on all variables.

During some tests, *variable identification* is done for the participants (e.g., Rosenthal, 1979) whereas in other tests the participants have to identify the variables on their own (e.g., Day & Stone, 1982). To identify and encode variables (and variable levels) is challenging for students because it requires encoding strategies as well as content knowledge about the independent variables (Morris, Croker, Masnick, & Zimmerman, 2012) and hence might influence task difficulty and moderate the treatment effect. Interestingly, Ross (1988a) found that assessment instruments in which students had to identify the relevant variables for themselves had larger effect sizes compared to instruments where the potential variables were identified for the student. Therefore, we coded the assessment's variable identification with respect to whether variables were identified for the participants or whether participants had to identify the relevant variables for themselves.

The *consistency between instruction and assessment content* was another factor we considered. Instruction effects are often smaller when the assessment content differs from the instruction content (e.g., Greenbowe et al., 1981; Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008). This moderator could also explain Ross's (1988a) finding that studies using self-developed tests show significantly larger

effect sizes than studies using tests from external sources. Indeed, the question of how well students can transfer their CVS skills to new tasks is highly relevant because of the general educational benefits expected from knowing CVS.

As previously mentioned, the *origin of the test instrument* might explain variance between study outcomes. Ross (1988a) found significantly larger effect sizes when the assessment was created for the particular study, relative to those using a standard or previously used instrument. A possible reason for this finding is better consistency between instructional and assessment tasks. Tests from external sources may be those used by other researchers or standardized test instruments that have been psychometrically validated, such as the Test of Integrated Process Skills (TIPS; Dillashaw & Okey, 1980).

Equally important for educational praxis is how long-lasting treatment effects are. Depending on the *time delay* between the instruction and assessment, treatment effects can decline to zero, as follow-up assessments that occur a year or more after the instruction show (Shayer & Adey, 1992; Strand-Cary & Klahr, 2008). Therefore, it is important to investigate whether instruction in CVS can produce long-lasting effects so that students may benefit from their skills in future school or out-of-school inquiry projects.

Methods

In the following section we present our inclusion and coding criteria, and describe the methods used to calculate effect sizes and analyze the data. We also describe the procedure for detecting and excluding outliers and handling of dependency between effect sizes.

Literature search and inclusion criteria

All studies analyzed by Ross (1988a) were included in our sample of potential relevant studies. We started the literature search by adding all 65 studies analyzed by Ross (1988a) to a database. Next, we used various search tools and databases, including SSCI, ERIC, PsychInfo, Google Scholar, FIS-Bildung (a German educational research database), and Dissertation Abstracts International to search for potentially relevant studies. We searched for published journal articles and book chapters, research reports, theses, and dissertations. The keywords for this search were *control of variables strategy*, *experimentation*, *science process skills*, *cognitive development*, *inquiry learning*, and variations of these. We did not restrict the search to studies published after 1988 because the quality of databases has increased since Ross carried out his work and hence we were able to detect additional studies from the earlier research period. Further sources of studies were the reference lists in reviews (e.g., Zimmerman, 2000, 2007; Lawson, 1992) and in relevant studies, as well as the forward citation history of relevant articles in Google Scholar. After checking titles and abstracts, we found 414 studies that fit our keyword criteria and added these to the database.

Next, all of these studies were assessed for whether they met the following inclusion criteria:

1. They were intervention studies at least partly designed to increase students' ability to control variables. Studies that measured students' CVS skills but did not include an intervention were excluded. Studies where CVS skill was measured, but the intervention itself did not focus on CVS at all were also excluded.
2. The content of the instruction was at least partly related to school science. Studies using only abstract and content-free reasoning tasks (e.g., Scardamalia, 1976) or games such as *Mastermind* (e.g., Thomas, 1980) were excluded because our goal is to find implications for the praxis of traditional science teaching and learning.
3. The achievement of the treatment group was contrasted to a control or comparison group. Control and comparison groups included those that received regular classes, no specific instruction, practice tasks, or a treatment concerning only content knowledge of the intervention tasks used in the CVS treatment group.
4. In the assessment test, students had to demonstrate their understanding of CVS, either by choosing an adequate design from a set of confounded and unconfounded experiments, correcting a confounded experiment, or designing an unconfounded experiment. The results of assessment tests

asking students only to state a general rule and not to demonstrate their understanding of the rule were excluded.
5. The reported test values are not confounded with measures of other science process skills. For example, we excluded studies reporting students' scores based on multiple-choice tests that include additional skills not related to CVS (e.g., tasks requiring an understanding of other process skills such as measuring, interpreting data, or drawing graphs).
6. The quantitative data necessary for calculating the effect size were reported. If the data were not given, we requested them from the authors. This procedure worked well for studies published within the last 12 years but not for older studies.
7. The treatment and control group were comparable with respect to pretest measures or general school achievement. Studies were only excluded when group differences were explicitly reported. For example, when significant pretest differences or differences in overall school achievement were reported (e.g., study 5 in Klahr & Li, 2005) we were able to make this determination. Many studies do not report whether there were pretest differences between groups and thus were not excluded.
8. The participants were students without learning disabilities.
9. The study was available in English or German.

The inclusion criteria for our analysis differ from those of Ross's (1988a) analysis in a number of ways. We excluded studies that (a) conflated CVS skills with other science process or reasoning skills, (b) had treatment content not related to the natural or physical sciences, (c) had contrasts with non-comparable pretest groups, and/or (d) included students with learning disabilities. Additionally, we included studies only available in German that met the previously outlined criteria. Of the 414 studies found during the literature research, 76 fulfilled all of the inclusion criteria and thus were further coded and analyzed (26 of these studies were also included in Ross's analyses). Appendix A includes the list of all studies fulfilling the inclusion criteria. A summary of the study selection procedure is presented in Fig. 1.

Study coding and interrater reliability

All studies were coded by the first author. A second rater coded a random subsample of 41 studies (10% of the 414 studies detected during the literature search) to determine the objectivity and reliability of the coding procedure. The inter-rater agreement was high (90%). Disagreement between coders was resolved by discussion. In addition, a random sample of 15 (20%) from the 76 studies meeting the inclusion criteria was re-coded by the second rater to estimate the inter-rater agreement on single moderator variables. The inter-rater agreement was generally high and ranged from 75% for interpretive decisions (e.g., whether an intervention included explicit rule presentation) to 100% for information explicitly reported in the papers (e.g., focus of the instruction, duration, design).

Multiple effect sizes from a single study that were due to repeated testing of the same groups (e.g., multiple-choice test vs. performance test), or to multiple treatment groups contrasted to a single control group, were coded as separate pairwise comparisons (237 pairwise comparisons from 76 studies). As a result, our dataset includes dependent effect sizes. Although this poses a potential problem due to confounded effect sizes, approaches such as merging dependent effect sizes or excluding effect sizes from studies with multiple contrasts would cause a loss of information (Scammacca, Roberts, & Stuebing, 2014). In particular, we would lose information about the test instruments as many studies use multiple tests. This information loss would be problematic because the variety of tests used in the studies is a reasonable source of the variance between study outcomes (Staver, 1984).

In addition to all 76 studies fulfilling the inclusion criteria, all available studies included in Ross's (1988a) meta-analysis that did not meet our inclusion criteria were coded. Although these additional studies were not included in our main analysis, we used the data from 19 studies included by Ross (but which did not meet our inclusion criteria) to investigate the influence of methodological differences and different inclusion criteria on the outcome of the meta-analysis.

The moderator variables were generated by the following information extracted from the studies (each is described in more detail, above):

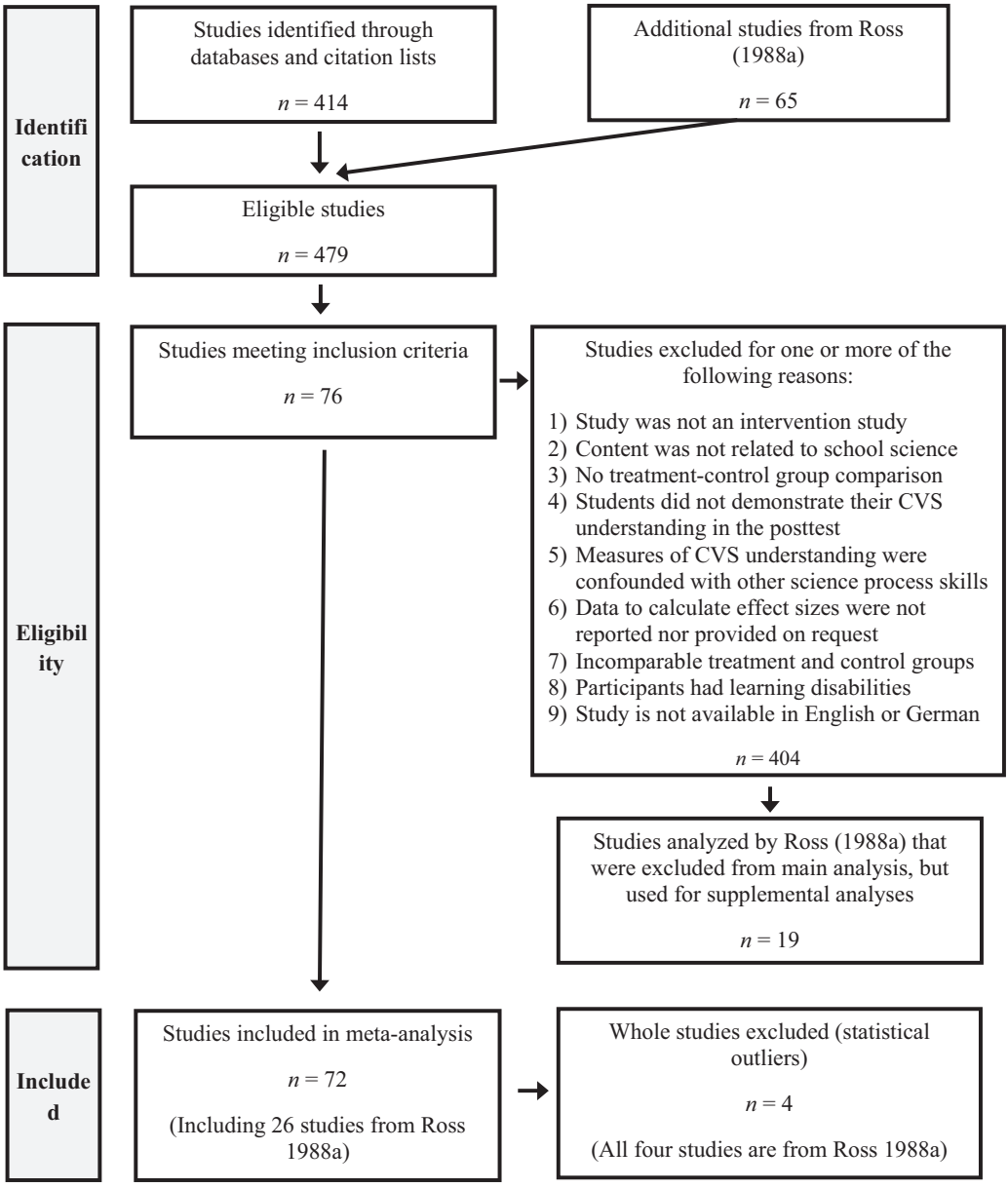

Fig. 1. Study selection flow chart. We were not able to include 16 studies from Ross (1988a) in any analyses as these studies were unavailable or did not report the required statistical data.

- Identifying information: Authors, publication year, title, journal, book or publishing institution, study identification code in literature database.
- Publication type: Journal articles and book chapters versus theses and dissertations, research reports, and conference proceedings.
- Study design: Experimental versus quasi-experimental design.

- Control-group activity: We distinguished between control and comparison groups that do activities not related to CVS (e.g., no instruction or regular science classes) and groups doing activities with the same experimental equipment that the treatment group used, but without any instruction related to CVS.
- Mean age of students and grade: If only grade levels were reported we predicted students' age by a linear regression based on studies reporting both types of information. The regression equation had the expected form of: age = 6 years + grade number.
- Total instruction or treatment duration in minutes: For classroom studies we estimated the treatment duration from the combination of information provided about the number of science classes per week, the duration of science lessons, and the total duration of the intervening instruction in weeks.
- Focus of the instruction: Treatments focusing only on CVS versus treatments having additional instructional objectives such as other science process skills or content knowledge.
- Instruction type: Instruction that includes the explicit presentation of a rule that can be used to solve typical CVS tasks at any time during the instruction versus no explicit rule presentation.
- Experimental training tasks: Use of either virtual or real experimental training tasks versus no use of training tasks.
- Feedback: Providing feedback to performance on training tasks (either written or verbal) versus no feedback.
- Use of demonstrations: Demonstrations by a researcher or teacher of correct experimental procedures with either real or virtual experiments, versus no demonstrations.
- Use of cognitive conflict: Instruction was coded as using cognitive conflict when the teacher scaffolded student recognition that some of their experimental strategies were inadequate, without making explicit reference to CVS (for examples, see the section on instruction characteristics above).
- Context: We coded whether the instruction content was presented in a school or an out-of-school context. For instance, topics such as bouncing balls, rocket design, and running contests were coded as out-of-school contexts. Topics such as extension of springs and reproduction of bacteria are examples that were coded as school contexts.
- Test format: Multiple-choice, open-response, performance task using real equipment, or performance task using virtual tasks.
- Number of independent variables: For real or virtual performance tasks, the number of variables to be controlled was classified as either three or fewer or four or more.
- Variable identification: Explicit identification of variables to be controlled during the post-test (either verbally or by text or pictures) versus tests for which students received no hints about relevant variables.
- Consistency of test and training content: Identical content used for instruction and assessment versus different content.
- Origin of the test: Pre-existing tests versus tests developed for the purposes of the study. If no external source was mentioned the test was categorized as developed for a specific study.
- Time delay between instruction and assessment: Same day versus more than one-day delay.In addition to coding the moderator variables, we also gathered statistical data for calculating the effect size on post-test measures (means, standard deviations, and sample sizes of treatment and control groups, or *t*- or *F*-values, reported effect sizes, or percentage of successful students in both groups). Coding information can be found in Appendix B.

Calculation of effect sizes and study variance

We estimated effect sizes as the standardized mean difference between treatment and control groups (Cohen's d) using the formula: $d = (M_T - M_C)/sd_p$ where M_T is the mean of the treatment group, M_C is the mean of the control group and sd_p is the pooled standard deviation (Borenstein, Hedges, Higgins, & Rothstein, 2010). A positive effect size indicates an advantage of the treatment over the control group. We used the pooled standard deviation instead of the pure standard deviation of the control group to consider changes in the variance in consequence of the treatment. We estimated effect sizes by alternative methods from t, F and χ^2 statistics in cases where means and standard deviations were not

reported. If only odds ratios were reported we computed the effect size using the arcsine transformation. If the only outcome measure given was a non-dichotomous allocation of students to different levels of CVS expertise we estimated means and standard deviations from this distribution (see Lipsey & Wilson, 2001 for details of alternative effect size estimations). To correct a slight upward bias in small sample sizes we transformed *d* values to *Hedges g* by multiplying them with the factor: $J = 1 - 3/(4(N-2)-1)$ (N = sum of the sample sizes of the treatment and control group). In addition to effect sizes, we estimated the study variance by: $\text{var} = (n_1 + n_2)/n_1 n_2 + d^2/2(n_1 + n_2)$ where n_1 and n_2 are the size of the treatment and control groups, respectively. Again we applied a correction for the small sample bias by multiplying the study variance with the factor J^2. The study variance served as source for the calculation of weights of the effect sizes. This procedure ensures that effect sizes based on larger samples – and thus more precise estimators of the underlying treatment effect – are weighted more heavily than effect sizes based on smaller samples (Borenstein et al., 2010). In contrast to meta-analyses of independent effect sizes we did not weight effect sizes by the inverse study variance. Instead, we weighted effect sizes by a factor considering dependency between effect sizes (see data analysis).

Ross (1988a) corrected effect sizes for pre-test differences. We did not apply this correction because pre-test results were reported in only 56% of the studies, and a restricted correction of only studies reporting pre-test results would cause confounded estimations of effect sizes. Alternatively, we could have only included studies in which pre-test results were reported, but this would have reduced the study sample drastically. However, group differences prior to instruction can have a huge impact on the post-test measure. Therefore, we excluded studies from the final sample that reported significant group differences prior to instruction or different learning abilities of students (see inclusion criteria). To avoid the analysis being dominated by unreported pre-instructional group differences, we also excluded studies with outlying effect sizes, as described in more detail in the next section.

Detecting and handling of outliers

We detected outlying studies that had a disproportionate impact on the mean effect sizes by computing the *sample-adjusted meta-analytical deviancy* (SAMD) statistic (Huffcutt & Arthur, 1995). We computed SAMD values for all pairwise comparisons by dividing the deviation between the effect size of the pairwise comparison *i*, and the mean effect size without *i*, by the sampling standard error. Thus, high SAMD values indicate studies that have a large impact on the mean effect size by large *g* values and small sampling standard errors. To determine cut-offs, we rank-ordered the values from the highest to the lowest and plotted them over the rank-position (see Fig. 2). The first SAMD-value divergent from the flat, gradual slope is the cut-off value (Huffcutt & Arthur, 1995).

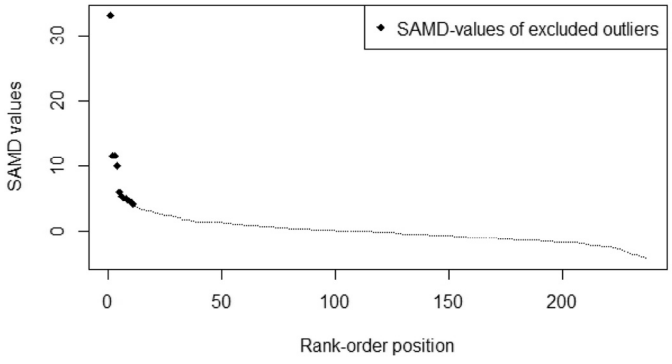

Fig. 2. Sample-adjusted meta-analytic deviancy (SAMD) values over rank-order position. The SAMD cut-off value is estimated by identifying the first value divergent from the flat gradual line of SAMD values (Huffcutt & Arthur, 1995). Outlying SAMD-values represent pair-wise comparisons with an unreasonable large impact on the mean overall effect size.

Table 1
Statistical characteristics of excluded outliers.

	N	Sampling standard error	*g*	SAMD
Ross (1988b, comparison 1)*	168	0.16	5.97	33.1
Ross (1988b, comparison 2)*	186	0.15	2.45	11.58
Zohar and David (2008)	59	0.27	3.83	11.56
Ross (1986)*	153	0.17	2.35	9.93
Case and Fry (1973)*	30	0.39	2.98	5.93
Strawitz (1984)	56	0.28	2.16	5.29
Lawson and Wollman (1976)	32	0.38	2.55	4.98
Peterson (1977)*	50	0.3	2.16	4.98
Zion, Michalsky, & Mevarech (2005)	199	0.15	1.36	4.62
Rosenthal (1979)	27	0.41	2.47	4.35
Tomlinson-Keasey (1972)	30	0.39	2.23	3.99

Note. SAMD is the sample-adjusted meta-analytical deviancy (Huffcutt & Arthur, 1995). *N* is equal to the sum of the participants in the specific pairwise comparisons within each paper excluded from further analysis. Ross (1988b) appears twice because two different treatment groups are contrasted to one comparison group in his second study. The sample size of 30 for Tomlinson-Keasey (1972) is an estimate based on interpolation of the data, as insufficient information is provided in the original paper. Papers for which the entire data set was excluded (and not just a specific pairwise comparison) are denoted with an asterisk.

We excluded 11 (4.6%) of the 237 pairwise comparisons and 4 whole studies (see Table 1). Thus, the final sample consisted of 226 pairwise comparisons from 72 independent studies (on average 3.2 effect sizes per independent study). A post-hoc assessment of the outlying studies revealed possible causes for the large effect sizes. A check of the statistical data transcribed from the studies showed we made no transcription errors. Possible reasons for the large effect sizes include coding student responses to two open-ended questions using criteria favoring the treatment group (Ross, 1988b), small sample sizes (e.g., $N = 30$; Case & Fry, 1973), a sample with extreme demographic characteristics (e.g., low SES), and non-random assignment of the teachers to instructional conditions (Ross, 1986). Although we did not find plausible explanations for all outliers, we excluded them all on the grounds that unknown or unreported measurement errors, pre-test differences, range restriction, or test restrictions could have caused the unusually large effect sizes. Of course, outliers should be included when they are caused by a large sampling error that can occur by chance when students are randomly drawn from a population (Hunter & Schmidt, 2004). However, large sampling errors are unlikely compared to possible study weaknesses and thus an exclusion of outliers results in a more accurate estimation of the treatment effect (Huffcutt & Arthur, 1995). To determine the impact of the inclusion of outliers we calculated the mean effect sizes with and without outliers for our sample of studies and all available studies from Ross's (1988a) sample.

Data analysis

The final sample includes dependent effect sizes due to multiple testing of the same groups of participants and contrasting multiple treatment groups with one control group. We included all pairwise comparisons meeting the inclusion criteria to avoid any loss of information either by merging dependent effect sizes or by considering only one effect size from studies with multiple group contrasts (Scammacca et al., 2014). Instead, we dealt with dependency among effect sizes by applying a robust meta-regression. This procedure handles dependency among effect sizes by adjusting the weights *W* (inverse variance of effect sizes) of dependent effect sizes by calculating $W_{ij} = 1/[(V_i + \tau^2)(1 + (k_j\text{-}1)\,\rho)]$ for each effect size i within each study j where V_i is the mean variance for each study i, τ^2 the component of the between-study variance, k_j the number of dependent effect sizes in study j and ρ an estimate of the common correlation between dependent effect sizes (Tanner-Smith & Tipton, 2014).

The advantage of this procedure is that it requires only the common correlation between all dependent effect sizes and not the correlations between single dependent effect sizes. Although we do

not know the common correlation coefficient, simulation studies show that its impact to the meta-regression is only marginal (Hedges et al., 2010; Tanner-Smith & Tipton, 2014). To control for the impact of the common correlation between dependent effect sizes on the results of the meta-analysis, we computed all analyses with multiple correlations ($\rho = 0.2, 0.5, 0.8, 1$) and found only marginal differences. Hence, we present only results computed with a correlation of 1 because a correlation of 1 results in a conservative estimation of coefficients (Hedges et al., 2010; Tanner-Smith & Tipton, 2014).

To investigate possible relations between the moderator variables and the study effect sizes we applied regression analyses with the weighted effect size estimations as the dependent variable and the moderator variables as independent variables. Further, we calculated *t*-values for the estimated regression coefficients from their standard errors to test whether they differ significantly from zero (Cohen, 2010, p. 42). The corresponding *p*-values were calculated from a *t*-distribution with *m*-2 degrees of freedom, where *m* is the number of studies (not pairwise comparisons) used to estimate the coefficients (Hedges et al., 2010). Other than student age and treatment duration, all moderator variables were categorical variables and were dummy coded as either 1 when the feature was present or 0 when the feature was not present in each comparison within a study. We conducted separate analyses of all moderator variables instead of conducting a single meta-regression model because the exclusion of studies with missing values in a single moderator variable would cause a huge reduction of the sample when combining multiple moderator variables. However, this approach could result in a misleading interpretation of the data when moderator variables are correlated. For example, several studies that used cognitive conflict to motivate students also used demonstrations of valid experiments. Therefore, it is impossible to distinguish which instruction features caused the moderator effects. The effect could be due to a single moderator, a combination of the two, or a third unknown moderator that is correlated with both features. In order to examine the possible combined effects of moderator variables, we also computed the total number of studies sharing both features.

We used a random-effects model instead of a fixed-effect model to compute the mean effect size because a common treatment effect of all included studies seems unreasonable when studies are diverse with respect to participants, treatment procedures, and test instruments. Furthermore, we want to be able to generalize our findings beyond the sample of studies included in the analysis, in order to inform future research and practice on teaching CVS. For investigating moderator effects, we applied a mixed-effects model that recognizes heterogeneity between study outcomes due to moderator variables and sampling error (Borenstein et al., 2010). All statistical analyses were conducted using the open source package "robumeta" in R (Fischer & Tipton, 2015; R Core Team, 2013).

Results

The 72 studies included in this meta-analysis were published between 1972 and 2012 (see Fig. 3). The majority (41) were conducted in the USA. Of the remainder, eight were conducted in Germany, seven in Israel, two each in Australia, Canada, and Belgium, and one each in Great Britain, China, Ireland, Finland, Italy, Austria, Pakistan, South-Africa, and the Netherlands. For one study (Wollman & Chen, 1982), no country is reported; it is not possible to guess the country as the authors were located in the USA and Israel. Except for eight studies only available in German, all studies are in English. The final sample of studies includes 19 (26%) studies included in Ross's (1988a) meta-analysis. It also includes 17 (24%) studies that were either published in conference proceedings, or were dissertations or theses. In 55% of the studies, individual students were randomly assigned to either a treatment or a control group, whereas in all other studies whole student groups were assigned to treatment or control conditions. The sample size varied, with studies ranging from 14 to 318 students; half of the studies used 40 or fewer students. Overall, 5355 students participated in the intervention studies included in the analysis. The age of students ranged from 6 to 24, but 50% of the studies used students aged 12 or younger.

Overall mean effect size

The overall mean weighted effect size of all 226 pairwise comparisons extracted from 72 independent studies is $g = 0.61$ ($SE = 0.04$; 95% CI = 0.53–0.69). The distribution of the effect sizes (see Fig. 4)

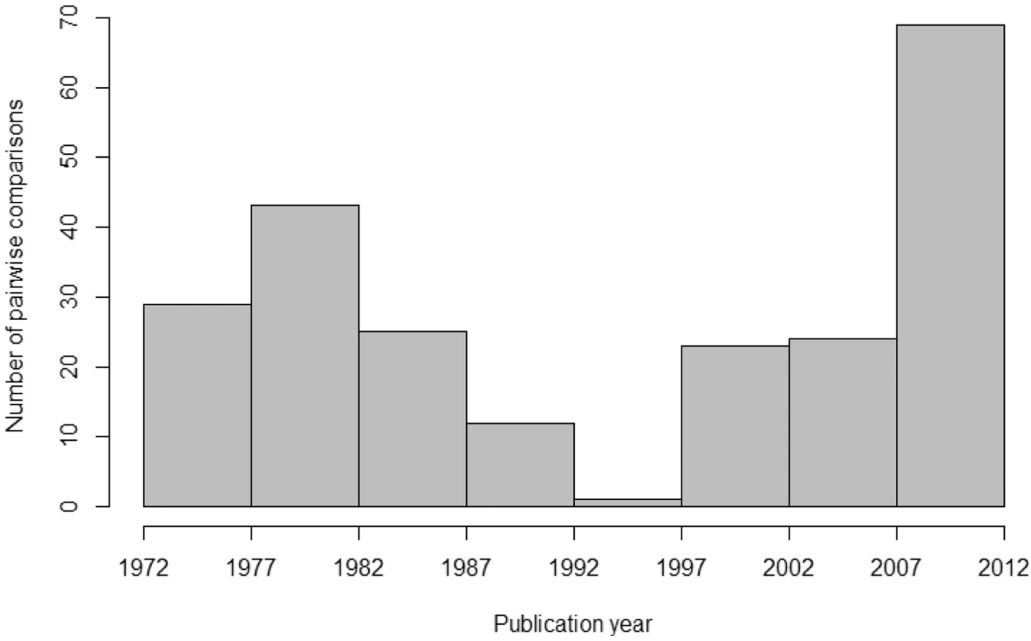

Fig. 3. Number of pairwise comparisons over publication year.

shows a general positive influence of interventions on student achievement. Furthermore, the heterogeneity of the study results is apparent: the outcome of single studies range from negative to large treatment effects. To investigate whether a common effect size underlies all included studies, we calculated the weighted sum of the deviations between single study effect sizes and the mean effect size (Q-value, see Borenstein et al., 2010). A significant Q-value of 186.51 ($p < 0.001$) indicates that it is unreasonable to expect a common underlying intervention effect for all studies summarized.

A comparison of meta-analyses of samples with and without outliers (see Table 2) shows a considerable impact of excluded outliers. Recall that Ross (1988a) found an overall effect size of $d = 0.73$ (95% CI = 0.54–0.92). When our new sample of studies was analyzed with identified outliers included, the effect size ($g = 0.77$, 95% CI = 0.61–0.93) is similar to that found by Ross. An exclusion of the 11 pairwise comparisons with outlying effect sizes reduced our mean effect size by 20%. When the sample of studies used by Ross was reanalyzed with outliers excluded, the resulting effect size was the same found in our meta-analysis ($g = 0.61$, 95% CI = 0.50–0.72).

Publication bias

A publication bias may occur because studies with statistically significant findings are preferred for publication. Thus, meta-analyses that include only published studies may cause an overestimation of the mean effect size. To avoid a publication bias, we searched Google Scholar and Dissertation Abstracts International databases for relevant unpublished studies. As a result, we included 16 unpublished studies (22%) in the meta-analysis. However, even an in-depth literature search does not necessarily avoid a publication bias because unpublished studies are hard to detect (Lipsey & Wilson, 2001). A comparison of the mean effect sizes of published and unpublished studies in our sample shows the expected larger effect sizes of published studies, but the group difference was non-significant (see Fig. 4). This finding contrasts with that of Ross (1988a), who did find a significant publication bias. In addition, to detect a potential publication bias in our meta-analysis we

Variable	g	95%-CI	m	k
Overall mean effect size	0.61	[0.53-0.69]	72	226
Publication type:				
Published studies	0.63	[0.51-0.75]	55	171
Unpublished studies	0.55	[0.31-0.79]	16	53
Design features:				
Design				
Experimental studies	0.61	[0.49-0.73]	40	154
Quasi-experimental studies	0.61	[0.47-0.75]	32	72
Control group activity				
Control-group uses same equipment	0.59	[0.45-0.73]	23	97
Other control group activity	0.62	[0.50-0.74]	50	129
Instruction features:				
Focus of the instruction				
Focus on CVS	0.63	[0.51-0.75]	43	166
Additional (non-CVS) content	0.58	[0.44-0.72]	30	60
Instruction type				
Teaching explicit CVS rule	0.58	[0.46-0.70]	44	137
No explicit rule	0.65	[0.51-0.79]	32	84
Use of training tasks				
Training on real or virtual tasks	0.59	[0.49-0.69]	66	183
No training	0.74	[0.58-0.90]	14	43
Feedback				
Use of feedback	0.66	[0.48-0.84]	33	91
No use of feedback	0.58	[0.48-0.68]	46	135
Demonstrations				
Use of demonstrations	0.69	[0.57-0.81]	44	133*a
No use of demonstrations	0.48	[0.32-0.64]	24	69*b
Cognitive conflict				
Use of cognitive conflict	0.80	[0.62-0.98]	22	51*a
No use of cognitive conflict	0.53	[0.43-0.63]	54	175*b
Context				
School science context	0.56	[0.46-0.66]	52	150
Out-of-school context	0.72	[0.56-0.88]	23	76
Assessment features:				
Test format				
Multiple-choice	0.52	[0.42-0.62]	20	42*a
Open-response	0.65	[0.53-0.77]	27	46
Performance tests (real)	0.74	[0.64-0.84]	32	98*b
Performance tests (virtual)	0.42	[0.32-0.52]	11	25*a
Number of independent variables				
Four or more variables	0.68	[0.50-0.86]	36	103
Three or fewer variables	0.75	[0.44-1.06]	11	19
Variable identification				
Identification of relevant variables	0.59	[0.47-0.71]	65	180
No identification of relevant variables	0.71	[0.40-1.02]	9	19
Consistency between instruction and assessment				
Identical content to training	0.61	[0.43-0.79]	23	59
Different content to training	0.57	[0.47-0.67]	51	139
Origin of the test				
Test developed for study	0.49	[0.35-0.63]	27	72
Test from external sources	0.66	[0.54-0.78]	52	145
Time delay				
Test on the last day of the treatment	0.64	[0.50-0.78]	48	156
Test later than one day	0.51	[0.31-0.71]	12	23

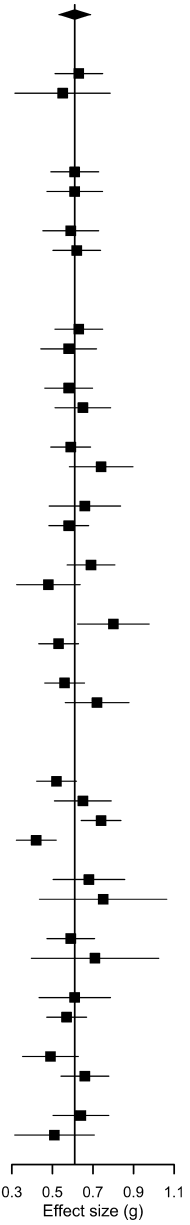

Fig. 4. Forest plot of the moderator effects (g) and 95% confidence intervals (CIs). *Note.* Items with subscripts *a* and *b* mark groups differing from each other at $p < 0.05$. The column *m* refers to the number of studies, and *k* represents the total number of pairwise comparisons.

created a funnel plot (Borenstein et al., 2010) that shows the relationship between effect size and corresponding standard error for every included pair-wise comparison (see Fig. 5). There is no evidence that studies with small effect sizes (typical unpublished studies) are missing, as the plot shows a symmetrical distribution of effect sizes.

Table 2
Comparison of mean effect sizes calculated using different meta-analytical approaches.

	Ross (1988a) with outliers	New analysis with outliers	Ross (1988a) without outliers	New analysis without outliers
Number of studies[a]	65	76	44	72
Percentage of studies included in Ross's analysis	100%	38%	100%	35%
Mean effect size *g*	0.73	0.77	0.61	0.61
95% CI	0.54–0.92	0.61–0.93	0.50–0.72	0.53–0.69

Note. The term *new* labels study samples based on our literature search and inclusion criteria whereas the term *Ross* labels studies based on Ross's (1988a) literature research and inclusion criteria.
[a] Datasets are not identical due to differing inclusion criteria.

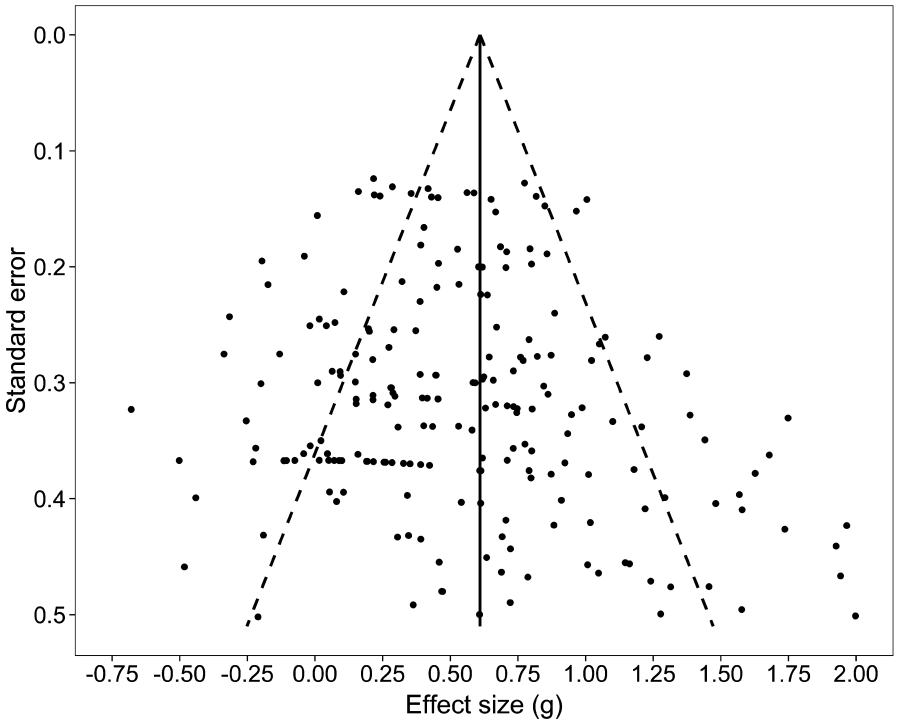

Fig. 5. Funnel plot showing the mean weighted effect size in relation to the corresponding standard error for every pair-wise comparison. The dotted line indicates the 95% confidence interval.

Furthermore, we computed the fail-safe *N* to estimate how many undetected studies with an effect size of zero would need to be added to the sample to reduce the mean effect size to 0.15 (definition of low effect size by Hattie, 2008). According to Orwin (1983), fail-safe *N* is computed using $N_{fs} = N_0(d_0\text{-}d_c)/d_c$, where N_0 is the number of studies included, d_0 the estimated mean effect and d_c the criterion effect size. We estimated that 693 additional pairwise comparisons with an effect size of zero (217 studies with on average with 3.2 pairwise comparisons per study) would be necessary to decrease the mean effect size to 0.15. Further, the potential for an inflated effect size resulting from a bias toward published studies is reduced, because excluded outliers that had a large impact on mean effect size were all from published studies. In sum, the funnel plot and the fail-safe *N* calculation show that a

potential publication bias does not mask an overall null effect of CVS instruction, even though we cannot avoid the potential effect of publication bias on the results of the meta-analysis.

Analysis of categorical moderator variables

The estimated mean effect size of $g = 0.61$ (95% CI = 0.53–0.69) indicates that teaching CVS is possible and can be very effective. However, to understand why study outcomes vary and why some studies report larger effect sizes than others, we conducted an analysis of moderator effects to determine which features affected study outcomes. We begin with a discussion of the categorical moderator variables (see Fig. 4). Neither of the design features (*study design*, *control group activity*) had an impact on effect sizes.

We found two instruction characteristics that significantly moderated the effect size. Instructional interventions that employed *demonstrations* of good experiments showed significantly larger effect sizes ($g = 0.69$, 95% CI = 0.57–0.81) than studies that did not use demonstrations ($g = 0.48$, 95% CI = 0.32–0.64). Interventions using procedures to induce *cognitive conflict* in students had larger effect sizes ($g = 0.80$, 95% CI = 0.62–0.98) than interventions not using such procedures ($g = 0.53$, 95% CI = 0.43–0.63). A closer examination of both variables indicates that all studies except one (i.e., Tomlinson-Keasey, 1972) that used procedures to induce cognitive conflict also used demonstrations. In terms of pairwise comparisons, 22% included both instruction characteristics, whereas in 33% neither of them was present. Taken together, 55% of the pairwise comparisons have identical values for both moderator variables, and in 45% demonstrations were used but no procedure to induce cognitive conflict was used. The co-occurrence of these two instructional features is discussed in more detail below.

The difference between studies focusing on CVS ($g = 0.63$, 95% CI = 0.51–0.75) and studies teaching additional content ($g = 0.58$, 95% CI = 0.44–0.72) was found to be non-significant. Most instruction including additional content either taught further science process skills such as drawing graphs (e.g., Lazarowitz & Huppert, 1993) or other content knowledge (e.g., Zimmerman, Raghavan, & Sartoris, 2003), but only three studies taught three or more additional content areas.

In contrast with findings from single intervention studies (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004) we found studies that explicitly taught a CVS rule to have effect sizes ($g = 0.58$, 95% CI = 0.46–0.70) no different from studies in which CVS rules were not explicitly taught ($g = 0.65$, 95% CI = 0.51–0.79). The effect sizes for studies in which students were trained on virtual or real performance tasks ($g = 0.59$, 95% CI = 0.49–0.69) were not significantly different from studies that did not train students on performance tasks ($g = 0.74$, 95% CI = 0.58–0.90). Even though this difference was not significant, this finding stands in contrast to the commonly held belief that hands-on activities support student learning (Haury & Rillero, 1994), as they were used in 81% of the pairwise comparisons. Most studies that did not use real or virtual hands-on activities during instruction trained students in CVS with paper-and-pencil tasks (e.g., Goossens et al., 1987), which proved to be as effective as performance tasks.

Interestingly, we found no significant difference between studies in which students received verbal or written feedback on their performance on training tasks ($g = 0.66$, 95% CI = 0.48–0.84) and studies in which students received no feedback ($g = 0.58$, 95% CI = 0.48–0.68). The use of feedback procedures was a significant moderator in Ross's (1988a) meta-analysis.

The difference between studies using at least one out-of-school context ($g = 0.72$, 95% CI = 0.56–0.88) and studies using only school contexts ($g = 0.56$, 95% CI = 0.46–0.66) was non-significant. Ross (1988a) found significantly larger effect sizes when students were given opportunities to practice CVS in a mix of both in-school and out-of-school contexts, compared to either type alone. In our meta-analysis, however, we coded the context of the main instruction rather than the context of any post-instruction practice sessions.

Of the assessment characteristics investigated in our moderator analysis only the *test format* was found to moderate study outcomes. Studies assessing student achievement with real performance tests show larger effect sizes ($g = 0.74$, 95% CI = 0.64–0.84) than studies using multiple-choice items ($g = 0.52$, 95% CI = 0.42–0.62), or virtual performance tasks ($g = 0.42$, 95% CI = 0.32–0.52), but were not different from open-response assessments ($g = 0.65$, 95% CI = 0.53–0.77). Tests with different formats also tend to differ with respect to task demands. In multiple-choice tests, students have

Table 3
Summary of moderator effects of continuous moderator variables.

		SE	*N*	*K*
Student age: intercept	0.59	0.19	71	225
b_1	0.001	0.015		
Treatment duration [min]: intercept	0.63	0.063	65	210
b_1	-1.86×10^{-5}	0.0001		

to select an unconfounded experimental design from a range of experimental designs (recognition), whereas in open-response or performance tasks students have to design an experiment (free recall).

We found no moderation of study effects by the identification of relevant variables or the number of variables in tests. We recorded the number of variables used in virtual or real performance tasks and found that in most tests students have to control four or five variables. Only 17 pairwise comparisons are based on tests using fewer than 4 variables and no tests used more than five variables. Hence, the low variability in the number of variables makes it hard to detect an impact of the number of variables on study outcomes.

We also found no differences between studies using the same content during training and test and studies using different content or an impact of any study feature. This finding contrasts with Ross (1988a), who found significantly larger effect sizes when students were assessed on the same type of task that they were trained on.

Ross (1988a) found that studies that used self-developed tests had larger effect sizes than studies that used previously existing tests. However, we found no significant differences between self-developed and pre-existing tests. It is possible that Ross' finding is based on the inclusion of studies with large effect sizes. The largest outlier in our analysis (Ross, 1988b) was a study that used a self-developed test and was included in Ross' meta-analysis. A moderator analysis with the dataset including outliers supports this possibility, as studies using self-developed tests had descriptively larger effect sizes, although the difference remained non-significant. Finally, no differences in effect sizes were detected when there was or was not a time delay between instruction and assessment.

Analysis of continuous moderator variables

Our investigation of the two continuous moderator variables (see Table 3) shows that neither the mean age of the students nor the treatment duration significantly moderates the study outcome. The mean age of students in the studies ranges from 6 years to 24 years. In 65% of the pairwise comparisons the students were 10–15 years old, in 23% of the comparisons the students were younger than 10 years, and in 12% the students were older than 15 years. Only eight studies directly compare intervention effects on students of different ages.

The treatment duration varied between 25 minutes and 35 hours but in 66% of the studies students were instructed for a maximum of 4 hours. As noted previously, long and short interventions differ with respect to many other features. For example, 68% of the studies lasting longer than 4 hours, but only 13% of the studies lasting less than 4 hours, are quasi-experimental studies. Moreover, the mean number of additional content items taught during instruction is 0.1 in studies lasting 4 or fewer hours whereas in studies lasting longer, an average of 1.1 additional content items were taught.

Discussion

First, we will discuss the comparison of different meta-analytical procedures and their impact on the mean effect sizes. After this, the results of the moderator analysis and implications for further research and teaching are discussed.

Impact of methodological approaches

The mean effect size of $g = 0.61$ (95% CI = 0.53–0.69) estimated in the current meta-analysis is smaller than the mean effect size of $d = 0.73$ (95% CI = 0.54–0.92) estimated by Ross (1988a). When comparing both estimations we have to consider the differences in methodological approaches between the two analyses. We used (a) different inclusion criteria, (b) different methods of estimating effect sizes, and (c) statistical techniques for excluded outliers. Importantly, we analyzed the data using a robust meta-regression instead of a traditional meta-analytical analysis of variance. Given these differences, however, the effects sizes are similar when we compare Ross's (1988a) findings to our sample of studies with outliers included (g values of 0.73 [95% CI = 0.54–0.92] and 0.77 [95% CI = 0.61–0.93], respectively). Furthermore, we found the same mean effect size ($g = 0.61$, 95% CI = 0.53–0.69) in our meta-analysis and in a re-analysis of the sample of all available studies from Ross's (1988a) analysis when we excluded outliers ($g = 0.61$, 95% CI = 0.50–0.72). Although our analysis differs from Ross's in several ways, by far the most influential difference is the exclusion of outliers. An exclusion of only 5% of the pairwise comparisons resulted in a 20% reduction in effect size. We discussed previously why the exclusion of outliers results in a more precise estimation of the mean treatment effect (see Methods). As noted previously, an additional argument for excluding outliers is that the more conservative estimation of effect sizes will prevent frustration of teachers and researchers who implement previously used interventions or assessments to try to replicate findings.

Moderator variables

We considered the role of 18 variables that could moderate the effect size of a CVS intervention. We classified these variables with respect to design features, student characteristics, instruction characteristics, and assessment characteristics.

Design features

Experimental and quasi-experimental studies did not differ systematically from each other. Accordingly, classroom studies are appropriate to study treatment effects even though they have a lower internal validity. The lack of a difference is relevant because of the higher ecological validity of classroom studies. Moreover, classroom studies have a larger impact relative to laboratory studies, in part because they are more likely to influence the praxis of teaching (Hofstein & Lunetta, 2004). The nature of the control or comparison group activity did not influence the effect size. Again, this lack of a significant difference has pragmatic implications for classroom practice and future research, in that the effect of an intervention does not depend upon the comparison to an impoverished control group activity. That is, a control group can be engaged in relevant activities and/or content domain knowledge without conferring the benefits of specific CVS instruction.

Student characteristics

At the outset, we intended to examine age and achievement level as two potential student characteristic moderators. As mentioned previously, existing literature suggests that general school achievement level could moderate effect sizes (e.g., Zohar & David, 2008; Zohar & Peled, 2008) but the information required to allow this variable to be coded was rarely reported. In the few cases when information was reported, it was based on different criteria across different studies. Therefore, we could not systematically investigate this potential moderator variable. Future research should investigate the interplay between achievement level and the effect of instruction on student achievement because low-achieving children may need to be taught differently than high-achieving children. Thus, a potential future research question is whether low-achievers require similar instruction to average ability students. Such investigations are also important in order to be able to answer questions about aptitude-treatment interactions.

The mean age of students was the only characteristic of participants left in our moderator analysis. We found no systematic impact of student age on study outcomes. As a result, there is no evidence that teaching CVS is more effective or appropriate for students of a specific age. In fact, elementary school students through to college students benefit from CVS instruction. However, this finding is

primarily based on between-study comparisons because only six studies investigated the effect of an identical treatment on students of different ages. Accordingly, we cannot generalize this finding to conclude that the same treatments work equally in students of different ages. Instead, the treatments may be adapted to the age of the participants. However, out-of-school content, for example, is not more prevalent in studies with younger participants than in studies with older participants. We found no evidence that treatments were adapted to the age of students. Hence, a direction for future research is to investigate how treatments can be adapted to the learning requirements of younger and older students. To have meaningful comparisons, studies should compare the achievement of different age groups after receiving an identical treatment. The age groups should cover K-12 students because inquiry skills are now part of the curriculum during all school years (National Research Council, 2012). For example, an interesting research question is whether the quantity of scaffolding can be decreased without negative consequences on the achievement of older students.

Instruction characteristics

Teaching CVS is possible and can be effective, as the mean effect size of $g = 0.61$ indicates. In our moderator analysis of what makes some instruction more effective than others, we considered seven features. Although 81% of the pairwise comparisons involved instruction that incorporated the use of hands-on or virtual training tasks, this feature was not significantly related to student achievement. We found a trend (albeit non-significant) of lower effect sizes for studies using hands-on or virtual training activities compared to those without such training tasks.

The lack of a difference between instruction with and without training tasks may reflect that CVS is a *cognitive* strategy; therefore, the manual or virtual manipulation of variables may not bear directly on students' understanding of CVS. Instead, "hands-on" activities (whether they are manipulations of physical apparatus or computer simulations) may actually have a negative impact on student understanding. When running experiments, students have to attend to additional challenges such as measuring and recording data. Thus, it may be the case that students think less about CVS while running experiments than they do in instruction that does not require a hands-on training task. However, we do not mean to imply that students cannot learn adequate experimental strategies when working on training tasks; evidence from many microgenetic studies (e.g., Kuhn & Phelps, 1982; Kuhn et al., 1992) shows that learning just may be more time consuming and challenging. The pattern in our meta-analysis is supported by Renken and Nunez's (2010) finding that students who learned about a physical concept conflicting with their beliefs by running their own experiments performed worse on a content knowledge test than students who learned by reading about the experiment. Taken together, it seems that students may not learn from the manipulation of a physical or virtual apparatus *per se*, but rather by thinking about data or evidence and reflecting on experimental strategies. Subsequently, there is no specific additional advantage to student learning using hands-on or virtual training tasks. It may be the case that carefully constructed hands-on training tasks could be developed with the sole purpose of CVS instruction. Such tasks would require that measurement and data recording are made as simple as possible. Moreover, such training tasks would not be concerned with developing content knowledge or other process skills, which may lead to better student achievement on CVS assessments.

Although the issue of instruction type, particularly with respect to the degree to which students are scaffolded or supported, has been a major topic of discussion within the literature, neither our meta-analysis nor the one conducted by Ross (1988a) showed significantly different effect sizes for the amount of support or self-directedness with which CVS instruction is implemented. It is important to note that our operationalization of explicit rule teaching is not the same as other definitions of "direct instruction." Whereas some definitions of direct instruction include additional elements such as telling students what they will learn and why they will learn it, or training tasks that give students feedback on their achievement (Hattie, 2008), we only coded whether students were or were not explicitly told how to solve typical CVS tasks. The lack of a difference is notable, again, largely because of the amount of attention paid to this issue in the literature.

One feature of instruction that did moderate effect size was our finding that studies using demonstrations of good experiments had larger effect sizes than studies not using demonstrations of CVS. By following a demonstration of a controlled experiment, students receive similar information to that received by conducting their own experiments, but without needing to attend to the additional challenges

described above (e.g., measuring outcomes, recording data). In addition, the teacher can draw students' attention to the design of the experiment by, for example, contrasting good and weak experimental designs.

Further, we found that studies using procedures to induce cognitive conflict in students had significantly larger effect sizes than studies not using such strategies. Ross (1988a) found that for a small number of studies ($n = 9$), there was a large effect ($ES = 1.00$) of cognitive conflict. Although statistically non-significant, Ross concluded, "the effectiveness of treatments was enhanced by using … cognitive conflict" (p. 427). Cognitive conflict involves the teacher directing students' attention to their experimental strategies in order to prompt them to think about the validity of their strategies rather than on the task content or measurement problems. This finding may lend support to the argument made above that the additional attentional demands required of students conducting their own experiments may be detrimental to learning. In using cognitive conflict, the teacher scaffolds the student by focusing attention on the problematic aspects of an experimental design or to a conclusion drawn from a confounded comparison. This approach may be especially effective for teaching CVS because even elementary school children already have some intuitive understanding of "fair" or good experiments without instruction (Schulz & Gopnik, 2004; Sodian et al., 1991). Hence, it may be ideal to teach CVS using cognitive conflicts because the conflicts address a reasoning strategy familiar and meaningful to the students (Limón, 2001).

Additionally, this effect of cognitive conflict could explain why we found no advantage for studies in which students were given an explicit rule to use to solve typical CVS tasks over studies in which students were not explicitly given such a rule. Students need not be taught what unambiguous evidence looks like; rather, they need to be reminded to apply a reasoning strategy they may already know when carrying out experimental tasks. However, if students are exposed to hands-on training tasks (without explicit instruction) they have to make the connection between their understanding of unambiguous evidence and the design of valid experiments on their own. This pattern of findings may explain why discovery learning requires more time than instruction offering some scaffolds. It may be the case that a scaffold, such as reminding students to focus on only one variable at a time, as Kuhn and Dean (2005) did, works to accelerate learning in the absence of more explicit forms of instruction (e.g., demonstrations).

Interestingly, we found that studies often include instructional interventions that used both demonstrations and procedures to induce cognitive conflict. In particular, nearly all studies in which cognitive conflict was induced also used demonstrations, either for inducing this conflict or for resolving it. One potential reason for the co-occurrence of both instruction features is that demonstrations are often used as the method to induce a cognitive conflict in students (e.g., Lawson & Wollman, 1976). In other studies, probe questions about the experiments designed by the students are used to induce a cognitive conflict (Strand-Cary & Klahr, 2008), but these studies still use demonstrations subsequently to assist the student to resolve the conflict. As cognitive conflict and demonstration are currently so conflated, further research is required to investigate the impact of demonstrations and cognitive conflicts both separately and in combination.

Assessment characteristics

Our moderator analysis included six features of assessments used to measure student achievement. With respect to test format, studies using real (hands-on) performance tests as assessments had significantly larger effect sizes than studies using either virtual performance tests or multiple-choice paper-and-pencil tests. At first glance, this finding of larger effects with hands-on assessment tasks seems to conflict with the previous finding that use of hands-on training tasks during instruction resulted in nonsignificantly smaller effect sizes relative to when such training tasks were not used. Ross (1988a) also found a seemingly counter-intuitive finding with respect to assessments. Larger effect sizes were evident when the assessment was more *demanding*. Our results are consistent with Ross's (1988a) findings and with the idea that challenging assessments are more sensitive to treatment effects. That is, even though all of the types of assessment tasks require an understanding of CVS, in some cases (e.g., multiple-choice) the assessment tasks provide students with a constrained search space. Therefore, some assessment tasks are less challenging and, consequently, less sensitive with respect to differentiating between trained and untrained students.

Physical or hands-on performance assessments may do a better job at differentiating between instructed and uninstructed students because they are more cognitively demanding and require the physical manipulation of an apparatus. Even compared to virtual performance tasks, physical tasks have more degrees of freedom (e.g., a computer simulation may have a constrained problem space with respect to variables to manipulate and the levels of those variables, it may be restricted in the number of choices to click on, and may provide additional scaffolds or cues). Additionally, students in a control condition using a physical assessment, who may be unfamiliar with the task apparatus, may understand the request to manipulate the equipment as a prompt to produce an effect instead of investigating causality (e.g., Schauble, Klopfer, & Raghavan, 1991). In the other types of assessments, various constraints (e.g., multiple choice, limited choices to click on in a virtual environment) may facilitate students in the control condition selecting the correct answers, thus resulting in smaller effect sizes between instruction and control conditions.

Nevertheless, the significant impact of assessment characteristics challenges our knowledge about student learning and understanding of CVS. It could mean that assessments using different formats are not measuring the same underlying construct. Although there are studies that investigated the effect of test format on student scores in CVS tasks, they did not include performance tasks (Staver, 1984, 1986). Further research is also needed to explore the interplay between student content knowledge and inquiry strategies, as we know that beliefs and preconceptions influence how students choose strategies and interpret evidence (Koslowski, 1996).

We did not replicate Ross's (1988a) finding that self-developed tests are related to larger effect sizes compared to more widely used tests. It seems possible that Ross's finding may have been based on the inclusion of outliers that used self-developed tests. In addition, we found no differences between studies in which the relevant variables of the test were identified for the students and studies not doing so. The trend seems consistent with Ross (1988a), such that the trend is toward larger (but nonsignificant) effect sizes when the students have to do the challenging variable identification work for themselves. This suggests that students search independently for variables to be controlled when they know CVS.

We also found no evidence for limitations of student performance due to a higher cognitive load in tasks with four or more variables. However, based on this meta-analysis we cannot say whether this is because performance on CVS tasks depends solely on the ability to apply CVS and not on the ability to remember all relevant variables, as we have little variability in the number of variables in the achievement tests. In order to investigate the impact of the number of variables, future research should use a larger range of variables and consider possible differential effects on performance on tasks of different formats. Notably, future research is also needed in order to draw any strong conclusions about the timing of the assessment, as an indicator of whether treatment effects are long lasting. The majority of the comparisons (87%) only assessed student learning on either the last day of the treatment or the day after the treatment. Longitudinal studies are rare, but more are necessary, in order to investigate different "transfer distances" (Strand-Cary & Klahr, 2008).

Conclusions

This meta-analysis summarizes relevant intervention studies on teaching CVS conducted within the last four decades. We found unexpected moderator effects that have yet to be investigated systematically. Moreover, we found that particular moderators that have received attention in the research literature were not as effective as expected. Accordingly, this work is an example of the benefits of using meta-analytical methods to summarize research as it gives us a more precise picture of patterns across a wide range of studies, and therefore provides suggestions for what further research should focus on. Furthermore, we show that meta-analyses need to be conducted carefully to avoid being dominated by a few studies with outlying effect sizes. However, this analysis does have limitations and does not include all research on CVS instruction, as we only summarize studies that met the inclusion criteria. Nevertheless, studies using no control groups, studies that do not report adequate statistical data, and – most importantly – studies not published in English or German may also be relevant.

It is important to note, when discussing meta-analytical results, that the analyses depend on the research available. One consequence is that moderator variables are often confounded. For example,

many studies using demonstrations to instruct students also used cognitive conflict. Thus, a meta-analysis cannot determine whether one variable, the other variable, or a combination of the two caused the significantly larger treatment effects in studies sharing both characteristics. Hence, further studies are required to investigate the effects of both instruction characteristics independently of one another. This example illustrates how the results of a meta-analysis can provide concrete suggestions for future research.

Unfortunately, we cannot investigate all moderator variables of interest. For example, evidence from single studies suggests that the general achievement level of students moderates the treatment effects. Many studies do not investigate (or at least do not report) the achievement level of their participants. However, as this idea is relatively new to the field, future researchers may decide to measure and report relevant information about achievement levels of their samples. Such studies would allow conclusions to be drawn about aptitude–treatment interactions, which is clearly important when trying to meet the needs of diverse student populations. For example, in the current meta-analysis, we focused our attention on traditional school topics. However, Kuhn and Dean (2005; Dean & Kuhn, 2007) have reported the results of a number of promising interventions with at-risk student populations that teach CVS and inquiry skills in non-traditional science domains, such as the factors that influence the sale of CDs. The goal is to teach low-achieving students that there are things that can be "found out" or investigated, using inquiry skills and experimentation.

An additional challenge in conducting a meta-analysis is that even when information regarding a variable is reported regularly, the validity of findings is challenged when the variable varies only between studies and not within. For instance, most studies only investigate the effect of a treatment in one age group. Thus, we cannot say whether the same treatments work equally well within all age groups or if treatments should be adapted to the age of the participants. Taken together, the dependence of meta-analysis on reported studies limits the validity of the findings. However, the reporting of a meta-analysis brings to light some of the limitations in a research area that may not have been detected otherwise. This, in turn, will allow the next wave of researchers to further focus their efforts on findings that can be used to improve the science of intervention research and the classroom practice of teaching and learning science.

Appendix: Supplementary material

Supplementary data to this article can be found online at doi:10.1016/j.dr.2015.12.001.

References

Adey, P., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school students. *Journal of Research in Science Teaching*, *27*(3), 267–285.
Amos, A. M. S., & Jonathan, S. M. (2003). The effects of process-skill instruction on secondary school students' formal reasoning ability in Nigeria. *Science Education International*, *14*(4), 51–54.
Beishuizen, J., Wilhelm, P., & Schimmel, M. (2004). Computer-supported inquiry learning: effects of training and practice. *Computers & Education*, *42*(4), 389–402.
Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). *Introduction to meta-analysis (Reprinted)*. Chichester: Wiley.
Bowyer, J. B., & Linn, M. C. (1978). Effectiveness of the science curriculum improvement study in teaching scientific literacy. *Journal of Research in Science Teaching*, *15*(3), 209–219.
Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12. Findings from the Munich longitudinal study* (pp. 38–54). Cambridge: Cambridge University Press.
Case, R., & Fry, C. (1973). Evaluation of an attempt to teach scientific inquiry and criticism in a working class high school. *Journal of Research in Science Teaching*, *10*(2), 135–142.
Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.
Cloutier, R., & Goldschmid, M. L. (1976). Individual differences in the development of formal reasoning. *Child Development*, *47*(4), 1097.
Cohen, J. (2010). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
Croker, S., & Buchanan, H. (2011). Scientific reasoning in a real-world context: the effect of prior belief and outcome on children's hypothesis-testing strategies. *British Journal of Developmental Psychology*, *29*, 409–424.
Danner, F. W., & Day, M. C. (1977). Eliciting formal operations. *Child Development*, *48*(4), 1600–1606.
Day, M. C., & Stone, C. A. (1982). Developmental and individual differences in the use of the control-of-variables strategy. *Journal of Educational Psychology*, *74*(5), 749–760.

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: the long view. *Science Education*, *91*(3), 384–397.
Dejonckheere, P., van de Keere, K., & Tallir, I. (2011). Are fourth and fifth grade children better scientists through metacognitive learning? *Electronic Journal of Research in Educational Psychology*, *9*(1), 133–156.
Dewey, J. (2002). *Logik: Die Theorie der Forschung [Logic: The theory of inquiry]*. Frankfurt/Main: Suhrkamp.
Dillashaw, G., & Okey, J. (1980). Test of the integrated science process skills for secondary science students. *Science Education*, *64*(5), 601–608.
Fischer, Z., & Tipton, E. (2015). *Package robumeta*. Retrieved from: https://cran.r-project.org/web/packages/robumeta/robumeta.pdf.
Ford, M. J. (2005). The game, the pieces, and the players: generative resources from two instructional portrayals of experimentation. *Journal of the Learning Sciences*, *14*(4), 449–487.
Goossens, L., Marcoen, A., & Vandenbroecke, G. (1987). Availability of the control-of-variables strategy in early adolescence: elicitation techniques revisited. *The Journal of Early Adolescence*, *7*(4), 453–462.
Greenbowe, T., Herron, J. D., Lucas, C., Nurrenbern, S., Staver, J. R., & Ward, C. R. (1981). Teaching preadolescents to act as scientists: replication and extension of an earlier study. *Journal of Educational Psychology*, *73*(5), 705–711.
Grygier, P. (2008). *Wissenschaftsverständnis von Grundschülern im Sachunterricht [Epistemological understanding of elementary students participating in science classes]*. Bad Heilbrunn: Klinkhardt.
Hattie, J. (2008). *Visible learning. A synthesis of meta-analyses relating to achievement*. London: Routledge.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.
Haury, D. L., & Rillero, P. (1994). *Perspectives of hands-on science teaching*. Columbus, Ohio: ERIC Clearinghouse for Science, Mathematics and Environmental Education,.
Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods 1* (1), 39–65.
Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: foundations for the twenty-first century. *Science Education*, *88*(1), 28–54.
Huffcutt, A. I., & Arthur, W. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology*, *80*(2), 327–334.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis. Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
Huppert, J., Lomask, S. M., & Lazarowitz, R. (2002). Computer simulations in the high school: students' cognitive stages, science process skills and academic achievement in microbiology. *International Journal of Science Education*, *24*(8), 803–821.
Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence. An essay on the construction of formal operational structures*. London: Routledge and Kegan Paul.
Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, *40*(9), 898–921.
Klahr, D. (2005). Early science instruction: addressing fundamental issues. *Psychological Science*, *16*(11), 871–873.
Klahr, D. (2009). To every thing there is a season, and a time to every purpose under the heavens": What about direct instruction? In S. Tobias & T. M. Duffy (Eds.), *Constructivist theory applied to instruction: Success or failure?* (pp. 291–310). New York, London: Routledge.
Klahr, D., & Li, J. (2005). Cognitive research and elementary science instruction: from the laboratory, to the classroom, and back. *Journal of Science Education and Technology*, *14*(2), 217–238.
Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.
Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science Teaching*, *44*(1), 183–203.
Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, *333*(6045), 971–975.
Koslowski, B. (1996). *Theory and evidence: the development of scientific reasoning*. Cambridge, MA: MIT Press.
Kuhn, D. (2005a). *Education for thinking*. Cambridge, MA: Harvard University Press.
Kuhn, D. (2005b). What needs to be mastered in mastery of scientific method? *Psychological Science*, *16*(11), 873–874.
Kuhn, D. (2007). Jumping to conclusions: can people be counted on to make sound judgments? *Scientific American*, *18*(1), 44–51.
Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*(11), 866–870.
Kuhn, D., Garcia-Mila, M., Zohar, A., & Anderson, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, *60*(4).
Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: what needs to develop to achieve skilled scientific thinking? *Cognitive Development*, *23*(4), 435–451.
Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. In H. W. Reese (Ed.), *Advances in child development and behavior* (pp. 1–44). New York: Academic Press.
Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, *9*(4), 285–327.
Lawson, A. E. (1992). The development of reasoning among college biology students. *Journal of College Science Teaching*, *21*, 338–344.
Lawson, A. E., & Wollman, W. T. (1976). Encouraging the transition from concrete to formal cognitive functioning-an experiment. *Journal of Research in Science Teaching*, *13*(5), 413–430.
Lazarowitz, R., & Huppert, J. (1993). Science process skills of 10th-grade biology students in a computer-assisted learning setting. *Journal of Research on Computing in Education*, *25*(3), 366–382.
Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: a critical appraisal. *Learning and Instruction*, *11*(4–5), 357–380.
Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching*, *36*(7), 837–858.
Linn, M. C. (1978). Influence of cognitive style and training on tasks requiring the separation of variables schema. *Child Development*, *49*(3), 874–877.

Linn, M. C., Clement, C., & Pulos, S. (1983). Is it formal if it's not physics? (The influence of content on formal reasoning). *Journal of Research in Science Teaching*, *20*(8), 755–770.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.
Lorch, R. F., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: contributions of explicit instruction and experimentation. *Journal of Educational Psychology*, *102*(1), 90–101.
Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: is it all in the timing? *Instructional Science*, *41*(3), 621–634.
Morris, B. J., Croker, S., Masnick, A., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B.J. Morris, & J.L. Amaral (Eds.), *Current topics in children's learning and cognition* (pp. 61–82). Rijeka, Croatia: InTech.
National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies.
NGSS Lead States (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational and Behavioral Statistics*, *8*(2), 157–159.
Padilla, M. J., Okey, J. R., & Garrard, K. (1984). The effects of instruction on integrated science process skill achievement. *Journal of Research in Science Teaching*, *21*(3), 277–287.
Popper, K. R. (1966). *Logik der Forschung [The logic of scientific discovery]*. Tübingen: J.C.B. Mohr.
Purser, R. K., & Renner, J. W. (1983). Results of two tenth-grade biology teaching procedures. *Science Education*, *67*(1), 85–98.
R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org/>.
Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: class, teacher, and school influences. *American Educational Research Journal*, *30*(3), 523–553.
Renken, M. D., & Nunez, N. (2010). Evidence for improved conclusion accuracy after reading about rather than conducting a belief-inconsistent simple physics experiment. *Applied Cognitive Psychology*, *24*(6), 792–811.
Rosenthal, D. (1979). The acquisition of formal operations: the effect of two training procedures. *Journal of Genetic Psychology*, *134*, 125–140.
Ross, J. A. (1986). Cows moo softly: acquiring and retrieving a formal operations schema. *European Journal of Science Education*, *8*(4), 389–397.
Ross, J. A. (1988a). Controlling variables: a meta-analysis of training studies. *Review of Educational Research*, *58*(4), 405–437.
Ross, J. A. (1988b). Improving social-environmental studies problem solving through cooperative learning. *American Educational Research Journal*, *25*(4), 573–591.
Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, *84*(3), 328–364.
Scardamalia, M. (1976). *The interaction of perceptual and quantitative load factors in the control of variables*. Dissertation, York University, York.
Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102–119.
Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*(9), 859–882.
Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176.
Shayer, M., & Adey, P. S. (1992). Accelerating the development of formal thinking in middle and high school students II: postproject effects on science achievement. *Journal of Research in Science Teaching*, *29*(1), 81–92.
Siegler, R. S., Liebert, D. E., & Liebert, R. M. (1973). Inhelder and Piaget's pendulum problem: teaching preadolescents to act as scientists. *Developmental Psychology*, *9*(1), 97–101.
Siler, S. A., & Klahr, D. (2012). Children's explicit and implicit misconceptions about experimental design. In R. W. Proctor & E. J. Capaldi (Eds.), *Psychology of science. Implicit and explicit processes* (pp. 137–180). New York, NY: Oxford University Press.
Sirin, S. R. (2004). *The relationship between socioeconomic status and school outcomes: Meta analytic review of research 1990–2000*. Dissertation, Boston College, Boston.
Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: a critical review of the literature. *International Journal of Science Education*, *34*(9), 1337–1370.
Sodian, B., & Bullock, M. (2008). Scientific reasoning – where are we now? *Cognitive Development*, *23*(4), 431–434.
Sodian, B., Jonen, A., Thoermer, C., & Kircher, E. (2006). Die Natur der Naturwissenschaften verstehen. Implementierung wissenschaftstheoretischen Unterrichts in der Grundschule. [Understanding the nature of science. Implementing epistemological instruction in elementary schools]. In M. Prenzel (Ed.), *Untersuchungen zur Bildungsqualität von Schule* (pp. 147–160). Münster, Germany: Waxmann.
Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*(4), 753–766.
Staver, J. R. (1984). Effects of method and format on subjects' responses to a control of variables reasoning problem. *Journal of Research in Science Teaching*, *21*(5), 517–526.
Staver, J. R. (1986). The effects of problem format, number of independent variables, and their interaction on student performance on a control of variables reasoning problem. *Journal of Research in Science Teaching*, *23*(6), 533–542.
Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: instructional effectiveness and path independence. *Cognitive Development*, *23*(4), 488–511.
Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30.
Thomas, W. E. (1980). *The effects of playing the game of master mind on the cognitive development of concrete-operational college students*. Dissertation. University of Missouri.
Tomlinson-Keasey, C. (1972). Formal operations in females from eleven to fifty-four years of age. *Developmental Psychology*, *6*(2), 364.
Triona, L. M., & Klahr, D. (2003). Point and click or grab and heft: comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction*, *21*(2), 149–173.

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Development*, *51*(11), 1–10.
Wollman, W. T., & Chen, B. (1982). Effects of structured social interaction on learning to control variables: a classroom training study. *Science Education*, *66*(5), 717–730.
Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99–149.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223.
Zimmerman, C., & Croker, S. (2013). Learning science through inquiry. In G. Feist & M. Gorman (Eds.), *Handbook of the psychology of science* (pp. 49–70). New York, NY: Springer.
Zimmerman, C., Raghavan, K., & Sartoris, M. (2003). The impact of the MARS curriculum on students' ability to coordinate theory and evidence. *International Journal of Science Education*, *25*(10), 1247–1271.
Zohar, A., & David, A. B. (2008). Explicit teaching of meta-strategic knowledge in authentic classroom situations. *Metacognition Learning*, *3*(1), 59–82.
Zohar, A., & Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learning and Instruction*, *18*(4), 337–353.