Förderung der Variablen-Kontroll-Strategie im Physikunterricht

DISSERTATION

ZUR ERLANGUNG DES DOKTORGRADES

DER MATHEMATISCH-NATURWISSENSCHAFTLICHEN FAKULTÄT

DER CHRISTIAN-ALBRECHTS-UNIVERSITÄT ZU KIEL

VORGELEGT VON

MARTIN GEERT SCHWICHOW

KIEL, 2015

1. Gutachter: Prof. Dr. Hendrik Härtig

2. Gutachter: Prof. Dr. Knut Neumann

3. Gutachterin: Prof. Dr. Beate Sodian

Tag der mündlichen Prüfung: 10.11.2015

Zum Druck genehmigt: 10.11.2015

gez. Prof. Dr. Wolfgang J. Duschl, Dekan

Zusammenfassung

Die Förderung experimenteller Kompetenz wird sowohl in den deutschen Bildungsstandards als auch in zahlreichen anderen nationalen Standards als ein wesentliches Ziel naturwissenschaftlicher Bildung genannt. Als zentrale Teilfähigkeit experimenteller Kompetenz sollten Schülerinnen und Schüler die Planung, Durchführung und Interpretation kontrollierter Experimente beherrschen. Diese als Variablen-Kontroll-Strategie (VKS) bezeichneten Fähigkeiten und Fertigkeiten sind von elementarer Bedeutung beim Experimentieren. Nur kontrollierte Experimente, in denen zwei Versuchsbedingungen verglichen werden, die sich ausschließlich in einer Variablen unterscheiden, liefern eindeutige Ergebnisse über kausale Zusammenhänge. Im Rahmen der vorliegenden Arbeit werden daher in vier zur Publikation eingereichten Studien Methoden zur Förderung der VKS im Physikunterricht entwickelt und ihre Wirksamkeit untersucht.

Da Methoden zur Förderung der VKS Gegenstand zahlreicher Interventionsstudien mit heterogenen Befunden sind, wurden zunächst die vorhandenen empirischen Forschungsbefunde in einer Meta-Analyse zusammengetragen. Die unter Einbeziehung von 72 Studien berechnete mittlere Effektstärke von g = 0,61 verdeutlicht, dass eine unterrichtliche Förderung der VKS möglich und effektiv ist. Eine detaillierte Analyse der Studien zeigt, dass zur Förderung der VKS eine Induktion kognitiver Konflikte sowie der Einsatz von Demonstrationsexperimenten besonders effektiv sind. Hingegen ist der Einsatz von Schülerexperimenten mit tendenziell kleineren Interventionseffekten verbunden. Darüber hinaus hängen die Studienbefunde vor allem vom verwendeten Testinstrument ab.

In einer Anschlussstudie wird daher ein neues Testinstrument entwickelt und erprobt. Die Ergebnisse der Pilotierungsstudie zeigen, dass die Schwierigkeit von VKS-Aufgaben nicht von Fachinhalt, aber von der operationalisierten Teilfähigkeit der VKS abhängt. In zukünftigen Studien sollten daher Testinstrumente eingesetzt werden, die sämtliche VKS-Teilfähigkeiten erfassen, um eine Über- bzw. Unterschätzung der Probandenfähigkeit zu vermeiden.

Eine weitere Anschlussstudie untersucht den zwar nicht signifikanten, aber für die Unterrichtspraxis relevanten Trend der geringen Effektivität von Schülerexperimenten zur Förderung der VKS. Die quasi-experimentelle Studie vergleicht die Wirkung von Schülerexperimenten und Papier-und-Bleistift Übungsaufgaben beim Erwerb der VKS. Die Studienergebnisse zeigen, dass Schülerexperimente im Vergleich zu Papier-und-Bleistift Aufgaben weder effektiver noch weniger effektiv sind. Daher wurden in einer vierten Studie VKS-Übungsexperimente zum Thema Leitfähigkeit entwickelt und im Unterricht erprobt. VKS-Übungsexperimente unterscheiden sich deutlich von klassischen Schülerexperimenten zum Fachwissenserwerb, da sie die Schülerinnen und Schüler mit mehr Experimentiermaterialien konfrontieren. Trotz der erhöhten Komplexität der Experimente haben die Schülerinnen und Schüler erfolgreich experimentiert. Die Relevanz der Studienergebnisse für zukünftige Forschung und die Unterrichtspraxis wird am Ende der Arbeit diskutiert.

Summary

The ability to design and interpret controlled experiments is an important scientific process skill and a common objective of science standards worldwide. To run a valid experiment that isolates the causal effects of variables, students should contrast conditions, which differ only with respect to the investigated variable. The skills required for designing, running and interpreting such controlled experiments are summarized under the term control-of-variables strategy (CVS). This thesis presents four independent studies that investigate how CVS can be introduced in physics lessons.

The prominent role of CVS in science and science education causes numerous intervention studies that investigate how CVS can be introduced to students. However, as the results of these studies are heterogeneous the first study summarized the outcome of 72 intervention studies in a meta-analysis. An estimated mean effect size of $g = 0.61$ shows that teaching CVS is possible and effective. Detailed analyses of potential moderator variables yield that studies, which introduce cognitive conflicts and studies that use demonstrations of CVS have significant larger effect sizes than studies not sharing these characteristics. We observed a tendency for studies utilizing hands-on activities to produce smaller effect sizes compared to those that not use these. Furthermore, the study outcome depends on the instrument utilized to evaluate the intervention effect.

To further analyze the dependency of student measures from test instruments study two presents a new instrument that utilizes a border spectrum of relevant CVS sub-skills. The results of a pilot study show that the item-difficulty depends on the utilized sub-skill and not on the item content. Accordingly, future studies should apply instruments which utilize all relevant CVS sub-skills to avoid an over- or underestimation of student skills.

Study three investigates the non-significant, but for science teaching relevant negative effect of hands-on activities on students CVS skills. A quasi-experimental intervention study compares the effect of hands-on and paper-and-pencil CVS training tasks on student's CVS achievement. The results show that hands-on activities are neither necessary nor obstructive for learning CVS. Therefore, study four presents CVS hands-on training task on the conductivity of wires. The outcome of a pilot study shows that students were able to plan and run

experiments successfully, although the CVS hands-on tasks are more complex compared to traditional hands-on task, which focus on content knowledge. The relevance of all results for further research and the praxis of science teaching is discussed.

Inhaltsverzeichnis

1. Einleitung

Die Förderung experimenteller Kompetenz wird sowohl in deutschen Standards für den naturwissenschaftlichen Unterricht (Biologie: KMK, 2005a; Chemie: KMK, 2005b; Physik: KMK, 2005c), als auch in zahlreichen anderen nationalen Standards (z. B. USA: NGSS Lead States, 2013; Singapur: Curriculum Planning & Development Division, 2007) als ein wesentliches Ziel naturwissenschaftlichen Unterrichts benannt. Zurückzuführen ist dies u. a. auf die zentrale Funktion von Experimenten innerhalb der Naturwissenschaften. Ziel moderner Naturwissenschaften ist es Theorien zu finden, welche die Natur beschreiben und einer empirischen Prüfung standhalten. Dadurch grenzen sich Naturwissenschaften von rein theoretischen (philosophischen) Wissenschaften ab (Hacking, 1996; Popper, 1966; Weizsäcker, 1951). Schülerinnen und Schüler, die eine weiterführende naturwissenschaftliche Bildung anstreben, sollten daher experimentelle Kompetenz als Schlüsselqualifikation besitzen. Aber auch Schülerinnen und Schüler, die keine weiterführende naturwissenschaftliche Bildung anstreben, sollten eine basale experimentelle Kompetenz aufweisen, da sie ihnen ermöglicht wissenschaftliche Erkenntnisse, die Relevanz für persönliche und gesellschaftliche Fragestellungen haben, kritisch zu hinterfragen (Abd-El-Khalick et al., 2004; Rutherford & Ahlgren, 1990). Darüber hinaus besteht eine enge Verbindung zwischen experimenteller Kompetenz und Allgemeinbildung, da sie „die Selbsttätigkeit, (die) kritischen Erkenntnis-, Urteils- und Handlungsfähigkeit (sowie die) Fähigkeit aus eigener Initiative weiter zu lernen“ fördert (Klafki, 1996, S.145). So können Experimente z. B. auch außerhalb der Naturwissenschaften (z. B. in den Sozialwissenschaften) genutzt werden, um kausale Zusammenhänge zu identifizieren. Die empirische Sichtweise naturwissenschaftlicher Experimente lässt sich ferner auf die kritische Prüfung von Argumenten z. B. in gesellschaftlichen Fragestellungen anwenden.

Im Unterricht werden Experimente jedoch nur selten zum Testen von Theorien, sondern überwiegend zur Demonstration von Phänomenen bzw. zur Veranschaulichung von Gesetzen, sprich zur Vermittlung von Fachwissen, eingesetzt. (Hofstein & Lunetta, 2004; Tesch, 2005, pp. , S. 103). Ohne eine explizite unterrichtliche Förderung experimenteller Kompetenz ist ein Großteil der Schülerinnen und Schüler jedoch nicht in der Lage selbständig zu experimentieren (Zimmerman, 2007; Zimmerman & Croker, 2013). Ziel dieser Arbeit ist es daher, Möglichkeiten zur Förderung experimenteller Kompetenz im Physikunterricht zu identifizieren und ihre Wirksamkeit empirisch zu überprüfen.

Im Folgenden wird zunächst die Funktion von Experimenten in den Naturwissenschaften diskutiert, bevor Erkenntnisse zu vorunterrichtlichen Schülerfähigkeiten und zu Fördermöglichkeiten vorgestellt werden. Anschließend erfolgt eine Vorstellung der einzelnen Publikationen sowie ihres Zusammenhangs zum übergeordneten Ziel der Arbeit.

1.1 Experimente in den Naturwissenschaften

Ein Wesensmerkmal moderner Naturwissenschaften ist die enge wechselseitige Beziehung zwischen Theorie und Experiment. Im Allgemeinen werden in den Naturwissenschaften Theorien bevorzugt, die einer umfangreichen experimentellen Überprüfung standhalten (Popper, 1966, S. 73). Umgekehrt sind für die Naturwissenschaften jene Experimente relevant, deren Ergebnisse sich auf Theorien beziehen oder die zumindest eine theoretische Betrachtung eines Phänomens initiieren (Hacking, 1996, S. 258ff). Die modernen Naturwissenschaften sind folglich weder rein theoretisch (Philosophie) noch rein empirisch, sondern durch eine wechselseitige Beziehung zwischen Theorie und Experiment gekennzeichnet. Der Theoriebezug von Experimenten ist jenes Merkmal, das Experimente von rein handwerklichen Tätigkeiten abgrenzt. (Simonyi, 2001, S. 37; Weizsäcker, 1951, S. 171). Im weitesten Sinne schließen Experimente sämtliche Handlungen, die mit der Prüfung von Theorien über die Natur verbunden sind, wie z. B. das Beobachten oder das Entwickeln von Messgeräten ein (Franklin, 1981, Hacking, 1996, S. 260; Höttecke & Rieß, 2015).

Weite Definitionen des Experimentierens beschreiben jedoch weder den Prozess des Experimentierens noch gehen sie darauf ein, wie möglichst eindeutige Aussagen über Theorien zu gewinnen sind (Woodward, 2003). Eine solche differenziertere Betrachtung des Experimentierens erfolgt in engeren Definitionen, die eine Unterscheidung zwischen Experimenten und Beobachtungen vornehmen. Beobachtungen beschränken sich auf das Messen und Wahrnehmen der Natur ohne aktiv den Gegenstand der Beobachtung zu beeinflussen bzw. zu manipulieren. Beim Experimentieren hingegen greift der Wissenschaftler aktiv in die Natur ein, um möglichst eindeutige Antworten auf seine Fragen zu erhalten (Rousmaniere, 1906, Weizsäcker, 1951, S.169ff). Das systematische Manipulieren der Natur (Operationalisiert als Variablen) ermöglicht es, Ursache-Wirkungszusammenhänge eindeutig zu identifizieren. Dazu sollte beim Experimentieren eine vermutete Ursache (unabhängige Variable) manipuliert und Veränderungen der vermuteten Wirkung (abhängige Variable) beobachtet werden. Dies reicht jedoch nicht aus, da der Einfluss alternativer Variablen auf die abhängige Variable nicht ausgeschlossen werden kann. Es ist daher notwendig, dass bei einem Experiment alle

weiteren Variablen, die einen potentiellen Einfluss auf die abhängige Variable haben können (potenziell konfundierende Variablen), konstant gehalten werden (Lehrer & Schauble, 2006; Schulz & Wirtz, 2012; Shadish, Cook, & Campbell, 2001; Woodward, 2003).

Das Kontrollieren potenziell konfundierender Variablen (alternative Ursache-Wirkungsbeziehungen) gewährleistet eine eindeutige Identifizierung kausaler Zusammenhänge und begründet die höhere Aussagekraft (Validität) von Experimenten im Vergleich zu reinen Beobachtungen. Die engere Definition naturwissenschaftlicher Experimente ist daher vergleichbar mit der Definition sozialwissenschaftlicher Experimente. In den Sozialwissenschaften wird ebenfalls nur dann von einem Experiment gesprochen, wenn eine Kontrolle alternativer Ursache-Wirkungsbeziehungen u. a. durch eine zufällige Aufteilung der Probanden auf unterschiedliche Bedingungen erfolgt (Bortz & Döring, 2002, S. 58). Das Ausschließen alternativer Ursache-Wirkungsbeziehungen durch konstant halten potenziell konfundierender Variablen wird auch als Variablenkontrolle bezeichnet (Chen & Klahr, 1999; Schulz& Wirtz, 2012).

Die Variablenkontrolle hat in der engen Definition des Experimentierens eine zentrale Funktion, da sie jenes Merkmal ist, das Experimente von anderen empirischen Methoden abgrenzt. Sie ist jedoch nicht nur für Experimente im engeren Sinne von Bedeutung. Vielmehr werden Strategien zur Kontrolle alternativer Ursache-Wirkungsbeziehungen ebenfalls in zahlreichen Experimenten im weiteren Sinne angewendet. Wurden z. B. auf Basis experimenteller Befunde und theoretischer Überlegungen neue physikalische Objekte identifiziert (z. B. Elektronen), so wird in der Regel versucht die Objekte mit unterschiedlichen Messmethoden nachzuweisen. Dadurch wird geprüft, ob die Existenz physikalischer Objekte von der verwendeten Messmethode unabhängig ist. Die Logik dieses Vorgehens ist mit der Variablenkontrolle vergleichbar, da ebenfalls alternative Erklärungen (ein Effekt des Messgerätes) für die, dem physikalischen Objekt zugeschriebenen Effekte ausgeschlossen werden. Selbst theoretisch erklärte und bereits experimentell geprüfte Effekte sind noch Gegenstand weiterer Experimente. Das Ziel solcher Experimente ist häufig systematische Fehler bei der Messung zu minimieren, sprich den Einfluss alternativer Variablen zu verringern (Höttecke & Rieß, 2015). Zwar vergleichen die beschriebenen Experimente im Sinne der weiteren Definition keine kontrollierten Bedingungen, doch sind die Motive für ihre Durchführung und die Interpretation der Befunde mit der Variablenkontrolle vergleichbar. Die Variablenkontrolle ist somit eine entscheidende

Voraussetzung für erfolgreiches Experimentieren sowohl im Sinne der engeren als auch im Sinne der weiteren Definition von Experimenten.

Eine umfassende Operationalisierung der Variablenkontrolle beschränkt sich daher nicht auf die Planung und Durchführung kontrollierter Experimente. Vielmehr werden auch die Interpretation kontrollierter Experimente, die Unterscheidung zwischen kontrollierten und unkontrollierten Experimenten, sowie ein Verständnis der fehlenden Aussagekraft unkontrollierter Experimente der Variablenkontrolle zugeordnet und unter dem Begriff Variablen-Kontroll-Strategie (VKS) zusammengefasst (Chen & Klahr, 1999). Die VKS ist ein Werkzeug, das auf zahlreiche experimentelle Fragestellungen angewendet werden kann und ein Konzept, das dazu befähigt, Experimente im Sinne des weiteren Experimentierbegriffs zu verstehen. Im folgenden Kapitel werden empirische Befunde zu den vorunterrichtlichen Schülerfähigkeiten bezüglich VKS vorgestellt. Diese Befunde können als Ausgangspunkt für die Planung und Gestaltung von Unterricht zur Förderung der VKS dienen.

1.2 Vorunterrichtliche Schülerfähigkeiten zur VKS

Schülerinnen und Schüler besitzen bereits vor der gezielten unterrichtlichen Thematisierung Vorstellungen zu Konzepten, die Gegenstand des Unterrichts sind. Diese so genannten Präkonzepte bzw. Schülervorstellungen basieren auf alltäglichen Erfahrungen und stimmen nur selten mit elaborierteren wissenschaftlichen Konzepten überein (Gilbert, Osborne, & Fensham, 1982). Ziel von Unterricht ist, dass Schülerinnen und Schüler elaboriertere Konzepte erlernen und diese anstelle ihrer Präkonzepte, zumindest in relevanten Kontexten, anwenden (Duit & Treagust, 2003). Damit dies gelingt, sollte Unterricht an den Präkonzepten der Schülerinnen und Schüler anknüpfen, systematisch elaboriertere Konzepte einführen und deren Überlegenheit gegenüber weniger elaborierten Konzepten verdeutlichen (Carey, 2000; Gilbert et al., 1982). Im Folgenden werden daher Befunde zu den vorunterrichtlichen Schülerfähigkeiten bzw. Schülervorstellungen bezüglich der VKS vorgestellt.

Aus dem, im vorangegangenen Kapitel vorgestellten, idealtypischen Vorgehen beim Experimentieren im engeren Sinne, können drei mögliche Fehler bei der Planung und Durchführung von Experimenten abgeleitet werden. Ein nicht kausal interpretierbares Experiment liegt vor, wenn 1) mehr als eine unabhängige Variable bzw. 2) keine unabhängige Variable manipuliert wurde, oder 3) keine Beobachtung der abhängigen Variable erfolgte (Schulz & Wirtz, 2012). Diese Fehler zeigen jedoch nur die Konsequenzen fehlerhafter Schülervorstellungen und sagen wenig über die ursächlichen Präkonzepte aus. Um etwas über die Präkonzepte zu erfah-

ren, reicht es daher nicht aus, die von Probanden geplanten und durchgeführten Experimente zu betrachten. Die den Fehlern zugrunde liegenden Konzepte bzw. Schülervorstellungen können in Interviewstudien durch gezielte Fragen zu den Beweggründen für das gewählte Vorgehen aufgedeckt werden (Gilbert et al., 1982).

In einer solchen Interviewstudie konnten Siler und Klahr (2012) vier typische Schülervorstellungen identifizieren, die zu fehlerhaften Experimenten führen. Schülerinnen und Schüler haben u. a. eine fehlerhafte Vorstellung bezüglich des Ziels von Experimenten. Das Ziel eines Experiments ist es ihrer Vorstellung nach nicht etwas herauszufinden, sondern einen Effekt zu erzeugen bzw. ein Gerät zum Laufen zu bringen. Sie verändern daher mehr als eine Variable und führen unkontrollierte Experimente durch um sicherzustellen, dass der erwartete Effekt eintritt. Dies erklärt auch warum sowohl Schülerinnen und Schüler (z. B. Croker & Buchanan, 2011; Schauble, 1996) als auch erwachsene Probanden (Kuhn, 2007) Probleme haben kontrollierte Experimente zu planen und unkontrollierte Experimente zu erkennen, wenn die Befunde der Experimente ihren Vermutungen widersprechen. Die Probanden versuchen einen Konflikt zwischen ihren Vermutungen und den experimentellen Befunden zu vermeiden indem sie beim Experimentieren mehrere Variablen manipulieren, so dass ein erwartetes Ergebnis eintritt (Croker & Buchanan, 2011). Widersprechen die experimentellen Befunde jedoch nicht den Erwartungen der Probanden, so sind bereits Kinder im Kindergarten- (Schulz & Gopnik, 2004) und Grundschulalter (Samarapungavan, 1992; Sodian, Zaitchik, & Carey, 1991) in der Lage die VKS anzuwenden.

Aber auch wenn Schülerinnen und Schülern bewusst ist, dass mit Experimenten etwas herausgefunden werden soll, können sie die fehlerhafte Vorstellung haben, dass in einem Experiment der Einfluss mehrerer Variablen untersucht werden kann. Sie verfolgen also das Ziel durch Experimentieren etwas herauszufinden, verändern aber trotzdem mehr als eine Variable, da sie den Einfluss mehrerer Variablen in einem Experiment untersuchen wollen. Ein weiterer Grund für die Durchführung unkontrollierter Experimente ist, dass Schülerinnen und Schüler unterschiedliche Variablenausprägungen nicht wahrnehmen und daher zwei Bedingungen vergleichen, die sich in mehr als einer unabhängigen Variable unterscheiden. Schließlich ist einigen Schülerinnen und Schülern nicht bewusst, dass in Experimenten grundsätzlich Variablenausprägungen verglichen bzw. kontrastiert werden, um Wissen zu generieren. Sie führen daher nur ein Teilexperiment bestehend aus einer Versuchsbedingung oder ein Expe-

riment mit zwei identischen Versuchsbedingungen (nicht kontrastives Experiment) durch (Siler & Klahr, 2012).

Zur erfolgreichen Förderung der VKS sollten die vorgestellten fehlerhaften Präkonzepte der Schülerinnen und Schüler im Unterricht überwunden und durch elaboriertere Konzepte ersetzt werden. Aus den vorgestellten Präkonzepten können daher Unterrichtsziele für die Förderung der VKS abgeleitet werden. Um erfolgreich die VKS anzuwenden, müssen Schülerinnen und Schüler zunächst verstehen, dass beim Experimentieren kein Effekt erzeugt, sondern eine Hypothese (Vermutung) überprüft werden soll. Eine weitere Voraussetzung für eine erfolgreiche Anwendung der VKS ist, dass die Schülerinnen und Schüler beim Experimentieren Variablenausprägungen erkennen und identifizieren können. Schlussendlich sollten sie verstehen, dass beim Experimentieren eine unabhängige Variable verändert und die Auswirkungen auf die abhängige Variable beobachtet wird und dass in einem Experiment nur der Einfluss einer Variablen untersucht werden kann. Wie die Befunde explorativer Studien zeigen, sind ohne gezielte unterrichtliche Intervention die meisten Schülerinnen und Schüler nicht in der Lage die VKS adäquat und unabhängig von ihren Vermutungen beim Experimentieren anzuwenden. Im folgenden Kapitel werden daher bisherige Befunde zu unterrichtlichen Fördermaßnahmen vorgestellt.

1.3 Befunde zur Förderung der VKS

Erste Interventionsstudien zur Wirksamkeit einer unterrichtlichen Förderung der VKS hatten das Ziel, die von Inhelder und Piaget (1958) aufgestellte Hypothese zu prüfen, dass der Erwerb der VKS nicht durch Unterricht beschleunigt werden kann. Die Befunde dieser frühen Interventionsstudien zeigen, dass bereits kurze unterrichtliche Interventionen große Effekte auf die Probandenfähigkeit zur Variablenkontrolle haben und die Hypothese somit nicht haltbar ist (Case & Fry, 1973; Siegler & Liebert, 1975). Weitere Interventionsstudien entstanden in Folge der Entwicklung von Curricula, die einen Fokus auf die Vermittlung naturwissenschaftlicher Arbeitsweisen anstelle der Vermittlung von Fachwissen legen (z. B. Science a Process Approach SAPA oder Science Curriculum Improvement Study SCIS). Aufgrund der Bedeutung der Variablenkontrolle für naturwissenschaftliche Experimente ist die VKS ein zentraler Aspekt dieser Curricula. Die im Rahmen der Entwicklung und Implementierung der Curricula durchgeführten Evaluationsstudien zeigen, dass die unterrichtliche Implementation der VKS einen positiven Effekt auf die Schülerfähigkeit bezüglich der VKS hat (siehe z. B. Bowyer & Linn, 1978).

In einer ersten Meta-Analyse wurden von Ross (1988a) die Befunde von 65 Interventionsstudien zur Vermittlung der VKS aus den Jahren 1970 bis 1988 zusammengefasst. Die berechnete mittlere Effektstärke von $d = .73$ verdeutlicht, dass eine unterrichtliche Förderung der VKS nicht nur möglich, sondern auch effektiv ist. Im Detail zeigt sich jedoch, dass die Effekte der Interventionsstudien äußerst heterogen sind und sich die Studien bezüglich der Stichproben, der verwendeten Unterrichtsmethoden und Testinstrumente unterscheiden. Mittels varianzanalytischer Methoden konnte Ross (1988a) u. a. zeigen, dass Studien, in denen Probanden ein Feedback zu den von ihnen geplanten Experimenten erhalten, signifikant größere Effekte aufweisen als Studien, in denen die Probanden kein Feedback erhalten. Außerdem konnte gezeigt werden, dass Eigenschaften der Testinstrumente einen signifikanten Einfluss auf den Effekt der Interventionsstudien haben.

Von 1988 bis 2012 (Beginn des Promotionsprojekts) wurden zahlreiche neue Interventionsstudien (42 wurden in einer Literaturrecherche identifiziert) durchgeführt. Das Ziel dieser Studien ist weniger die Überprüfung von Piagets Theorie oder der Wirksamkeit von Curricula, sondern vor allem die Identifizierung effektiver Instruktionsstrategien zur Förderung der VKS. Ferner unterscheiden sich neuere von früheren Interventionsstudien dahingehend, dass sie vermehrt virtuelle Instruktionsmaterialien (z. B. Kuhn & Dean, 2005; Lin & Lehman, 1999) sowie virtuelle Testinstrumente einsetzen (z. B. Beishuizen, Wilhelm, & Schimmel, 2004; Marschner, 2011) und häufiger Probanden im Grundschulalter benutzten (z. B. Grygier, 2008). Untersuchte und divers diskutierte Fragestellungen sind u. a. ob Methoden der direkten Instruktion oder des entdeckenden Lernens effektiver zur Förderung der VKS sind (Chen & Klahr, 1999; Dean & Kuhn, 2007) und welche Vermittlungsmethoden nachhaltigere Lerneffekte erzielen (Klahr & Nigam, 2004; Strand-Cary & Klahr, 2008). Die Befunde neuerer Interventionsstudien sind jedoch nicht eindeutig. So zeigt z. B. eine Studie von Klahr und Nigam (2004) die Überlegenheit direkter Instruktion gegenüber Unterrichtsmethoden des entdeckenden Lernens, während eine Studie von Dean und Kuhn (2007) das Gegenteil belegt. Auch bezüglich der Nachhaltigkeit von VKS Förderunterricht gibt es widersprüchliche Befunde. Einerseits gibt es Studien, die einen signifikanten Lernzuwachs in Nachtests (bis zu drei Jahre nach der Intervention) nachweisen (Chen & Klahr, 1999; Strand-Cary & Klahr, 2008) während andere Studien keinen nachhaltigen Lernerfolg feststellen (Brotherton & Preece, 1996; Sao Pedro, Gober, & Raziuddin, 2010).

2. Zielsetzung der Arbeit und Konzeption der Studien

Aufgrund der zentralen Rolle der Variablenkontrolle in den Naturwissenschaften und ihrer hohen Bildungsrelevanz ist das Ziel dieser Arbeit die Identifizierung und Erprobung effektiver Methoden zur Förderung der VKS im Physikunterricht. Wie dargelegt, berücksichtigen Schülerinnen und Schüler ohne gezielte unterrichtliche Förderung die VKS nur unzureichend beim Experimentieren. Zwar befassen sich bereits eine Vielzahl von Interventionsstudien und eine Meta-Analyse mit dem Thema, doch ist bisher nicht hinreichend geklärt, welche Methoden besonders geeignet sind. Angesichts der uneinheitlichen Befunde der vorliegenden Interventionsstudien wurde zu Beginn des Promotionsprojekts zunächst eine aktuelle Meta-Analyse durchgeführt. Ziel der Meta-Analyse ist die Identifizierung von Merkmalen effektiver Instruktionsmethoden, sowie weiterer Faktoren bezüglich der Probanden und Testinstrumente, die einen Einfluss auf den Interventionseffekt haben. Eine neue Meta-Analyse erscheint notwendig, da die vorliegende Meta-Analyse von Ross (1988a) zahlreiche aktuellere Studien nicht berücksichtigt. Diese neueren Studien unterscheiden sich systematisch von älteren Interventionsstudien, da sie andere Instruktionsmethoden, Unterrichtsmaterialien, Testinstrumente und Probanden benutzen. Mit Hilfe einer aktuellen Meta-Analyse soll daher geklärt werden, inwiefern die Befunde von Ross (1988) noch aktuell sind. Ferner soll der Einfluss weiterer Studienmerkmale auf den Interventionseffekt untersucht werden.

Ein wesentlicher Befund der Meta-Analyse ist, dass die Effekte der Interventionsstudien von dem zur Evaluation eingesetzten Testinstrument abhängen. Studien, die Multiple-Choice Testinstrumente einsetzten weisen signifikant kleinere Effektstärken auf als Studien, die offene Aufgabenformate oder praktische Experimentiertests einsetzten. Ein Vergleich von Instrumenten unterschiedlichen Formats zeigt, dass sich die Instrumente nicht nur im Aufgabenformat, sondern auch hinsichtlich der operationalisierten Teilfähigkeiten der VKS unterscheiden. Während Probanden in Multiple-Choice Tests kontrollierte und unkontrollierte Experimente unterscheiden müssen, erfordern offene Aufgaben und praktische Experimentiertests kontrollierte Experimente zu planen bzw. durchzuführen. Es ist somit fraglich, ob Testinstrumente unterschiedlichen Formats dasselbe Konstrukt erfassen und inwiefern Studien, die Testinstrumente unterschiedlichen Formats einsetzen, vergleichbar sind. In einer Anschlussstudie wurde daher zunächst ein neues Testinstrument zur Erfassung der VKS entwickelt und

pilotiert, welches mehrere relevante Teilfähigkeiten der VKS erfasst und somit mutmaßlich validere Maße der VKS-Fähigkeit generiert.

Ein weiterer und für die Unterrichtspraxis relevanter meta-analytischer Befund ist, dass Studien, die Schülerexperimente zum Üben der VKS einsetzen, geringere Effekte aufweisen als Interventionen, die keine Schülerexperimente einsetzen. Auch wenn dieser in der Meta-Analyse identifizierte Trend nicht signifikant war, ist er doch aufgrund des häufigen Einsatzes von Schülerexperimenten im Unterricht von großer praktischer Relevanz. In einer Interventionsstudie wird daher der Einsatz von Schülerexperimenten und reinen Papier-und-Bleistift Übungsaufgaben bei der Förderung der VKS verglichen. Da die Ergebnisse der Interventionsstudie - entgegen den meta-analytischen Befunden - keine geringere Effizienz von Schülerexperimenten zeigen, wird abschließend in einer unterrichtsnahen Publikation ein beispielhaftes Schülerexperiment zur Förderung der VKS samt den Ergebnissen einer unterrichtlichen Erprobung vorgestellt.

2.1 Publikation 1: Teaching the control-of-variables strategy: A meta-analysis

Die in der ersten Publikation präsentierte Meta-Analyse fasst die Befunde von 72 Interventionsstudien (im Treatment-Kontrollgruppendesign) zur Förderung der VKS zusammen. Zur Datenanalyse wurde das Verfahren der *Robusten Meta-Regression* verwendet, da zahlreiche Studien multiple Effektstärken berichten und somit eine geklusterte Datenstruktur vorliegt. Nach Ausschluss von Ausreißern wurde eine mittlere Effektstärke von $g = 0.61$ (95% CI = 0.53-0.69) bestimmt. Zusätzlich wurde in einer Analyse potentieller Moderatorvariablen der Einfluss von Studien-, Probanden-, Treatment- und Testmerkmalen auf den Studieneffekt untersucht. Es zeigt sich, dass nur die zwei Treatmentmerkmale “Induktion kognitiver Konflikte” und „Einsatz von Demonstrationsexperimenten“ einen signifikanten Einfluss auf die Probandenfähigkeit haben. Weiterhin wurde ein nicht signifikanter aber für die Praxis des naturwissenschaftlichen Unterrichts relevanter negativer Effekt von Schülerexperimenten auf die Probandenfähigkeit zur Variablenkontrolle gefunden. Darüber hinaus hängt die Effektstärke der Studien vor allem von dem Format des eingesetzten Testinstruments ab. In der Publikation werden sowohl das eigene methodische Vorgehen, als auch Implikationen für den naturwissenschaftlichen Unterricht und zukünftige Forschungsarbeiten kritisch diskutiert.

2.2 Publikation 2: The impact of sub-skills and item content on students' skills with regard to the control-of-variables-strategy (CVS)

Aufbauend auf den Befunden der Meta-Analyse wird in Publikation 2 die Abhängigkeit der VKS-Fähigkeit von dem verwendeten Testinstrument untersucht. Ein systematischer Vergleich existierender Testinstrumente offenbart, dass Instrumente unterschiedlichen Formats verschiedene Teilfähigkeiten der VKS abfragen. Der Einfluss unterschiedlicher VKS-Teilfähigkeiten auf die Schwierigkeit von VKS-Aufgaben ist bisher jedoch nicht bekannt. In Publikation 2 wird zunächst ein neu entwickelter VKS Multiple-Choice-Test vorgestellt, der die drei Teilfähigkeiten: Verständnis der Invalidität unkontrollierter Experimente, Identifizierung und Interpretation kontrollierter Experimente erfasst. Mittels Rasch-Analyse kann gezeigt werden, dass Messwerte, die mit dem neuen Instrument gewonnen wurden, reliabel und valide sind. Des Weiteren zeigt ein Vergleich der Item-Schwierigkeiten, dass Items, die das Verständnis der Invalidität unkontrollierter Experimente abfragen signifikant schwieriger sind als Items, die eine Identifizierung oder Interpretation kontrollierter Experimente fordern. Der physikalische Fachinhalt der Items (Wärme und Temperatur versus Elektrizitätslehre und Elektromagnetismus) hat jedoch keinen Einfluss auf die Item-Schwierigkeit. Folglich scheinen bestehende Testinstrumente, welche die Teilfähigkeit „Verständnis der Invalidität unkontrollierter Experimente" nicht abfragen, die Probandenfähigkeit bezüglich der VKS zu überschätzen. Die Konsequenzen der Befunde für die Interpretation von VKS Messwerten und für die Unterrichtspraxis werden diskutiert.

2.3 Publikation 3: What students learn from hands-on activities

Die dritte Publikation beschreibt eine Interventionsstudie, die den Einfluss von Schülerexperimenten und nicht experimentellen Arbeitsblättern auf das Erlernen der VKS vergleichend untersucht. Die Interventionsstudie greift damit einen in der Meta-Analyse identifizierten Trend des negativen Effekts von Schülerexperimenten auf die Lernwirksamkeit unterrichtlicher Interventionen auf. Auch wenn dieser Effekt nicht signifikant ist, ist die Frage der Wirkung von Schülerexperimenten für die Unterrichtspraxis aufgrund des regelmäßigen Einsatzes von Schülerexperimenten im Unterricht hochgradig relevant. Eine eventuell negative Wirkung von Schülerexperimenten könnte auf den im Vergleich zu alternativen Übungsaufgaben relativ hohen *Cognitive Load* von Schülerexperimenten zurückzuführen sein. In der quasi-experimentellen Studie wurden N = 161 Achtklässler, in Abhängigkeit von ihrem Abschneiden in einem Vortest (der in Publikation 2 beschriebene VKS Multiple-Choice-Test) in zwei

gleich starke Treatmentgruppen aufgeteilt. Nach einem kurzen Unterrichtseinstieg, in dem ein kognitiver Konflikt zur Einführung in die VKS benutzt wurde, übten beide Gruppen die Planung und Interpretation kontrollierter Experimente am Beispiel von Aufgaben zum Elektromagnetismus. Während eine Gruppe direkt mit Schülerexperimenten interagierte, plante die andere Gruppe Experimente nur theoretisch und interpretierte Experimente, die auf einem Foto präsentiert wurden. Die Ergebnisse des Posttests zeigen, dass sich beide Gruppen nur in Testaufgaben, die identisch oder sehr ähnlich zu den Übungsaufgaben sind, unterscheiden. Entgegen der ursprünglichen Hypothese konnte somit keine generelle negative Wirkung von Schülerexperimenten nachgewiesen werden.

2.4 Publikation 4: Förderung der Variablen-Kontroll-Strategie im Physikunterricht

Schülerexperimente zum Üben der VKS unterschieden sich von Schülerexperimenten, die zum Erlernen von Fachwissen oder der Demonstration von Phänomenen konzipiert wurden. VKS-Übungsexperimente sollten u. a. die Schülerinnen und Schüler mit einer größeren Auswahl an Variablen und Variablenausprägungen konfrontieren, so dass diese eine wirkliche Auswahl und Kontrolle von Variablen vornehmen müssen. In Publikation 4 werden daher in einem unterrichtsnahmen Beitrag die Merkmale von VKS-Übungsexperimenten, sowie ein Beispielexperiment zum Widerstand von Leitern vorgestellt. Außerdem werden die Befunde einer unterrichtlichen Erprobung berichtet, die zeigen, dass Schülerinnen und Schüler trotz der erhöhten Komplexität von VKS-Übungsexperimenten sinnvoll mit den Materialien interagieren.

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

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 science education and future research are discussed.*

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

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" (CVS; 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* (NGSS Lead States, 2013; NRC, 2012) 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 Lead States, 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 (Cloutier & Goldschmid, 1976; Linn, Clement, & Pulos, 1983), or *intervention studies*, which explore the impact of instruction on students' achievement on CVS tasks (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 (Croker & Buchanan, 2011, Schulz & Gopnik, 2004, 2004; Sodian et al., 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 (Croker & Buchanan, 2011; Koslowski, 1996; Kuhn et al., 1995; Linn et al., 1983; Tschirgi, 1980; 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; Sodian & Bullock, 2008; Zimmerman, 2000, 2007). 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 (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 (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.

3.1 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 (Linn, 1978) to asking students to discover the principles of CVS on their own (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 $d = 0.73$ 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.

3.2 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 (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.

3.3 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 (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 in-

struction (Zohar & David, 2008). In other cases, the comparison group may use the same equipment that CVS group uses, but without any CVS-related instruction (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 (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 (Klahr & Li, 2005, Raudenbush, Rowan, & Cheong, 1993) 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 (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 (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 (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 (Ford, 2005; Lawson & Wollman, 1976) or virtual experimental setups (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 which 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 (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 et al., 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 (Adey & Shayer, 1990; Lawson & Wollman, 1976). For example, Lawson & 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 & 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 *context* 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 & Wollman (1976) demonstrated the difference between good and bad experiments on bouncing balls, and Beishuizen et al. (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-administrated 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 (Rosenthal, 1979) whereas in other tests the participants have to identify the variables on their own (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, & Zimmerm, 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 (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.

3.4 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 (Scardamalia, 1976) or games such as *Mas-*

termind (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 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 the overall school achievement were reported (e.g., 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 included.
8. The participants were students without learning disabilities.
9. The study was available in English or German.

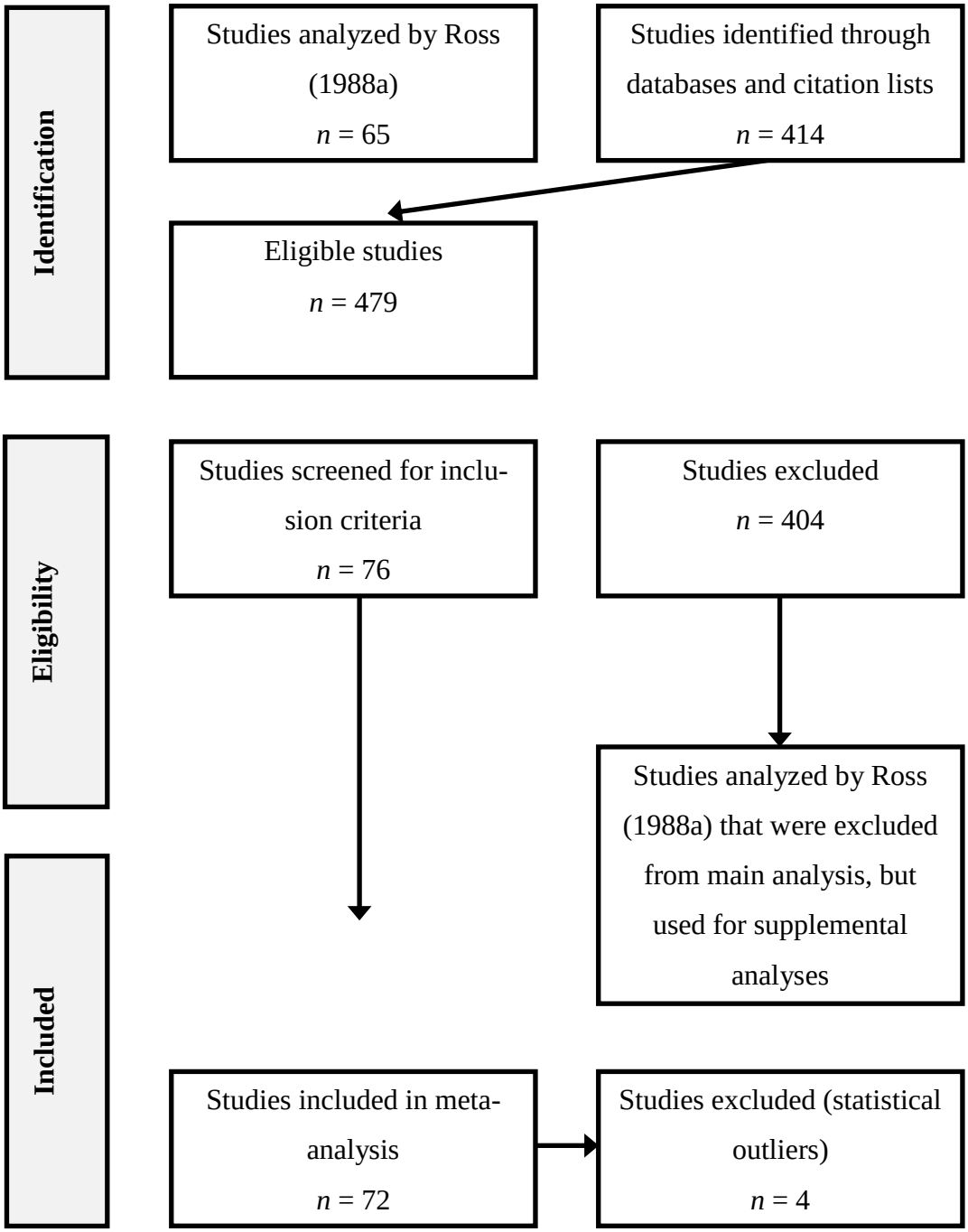


Figure 1 *Study selection flow chart. Please see text for a description of the nine inclusion criteria and the exclusion criteria for the 404 excluded studies.*

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 previous-

ly 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 Figure 1.

Study Coding and Interrater Reliability

All studies were coded by the first author. 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 are 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):

- 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 do-

ing 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 context. 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. All initial coding was done by the first author. A second rater coded a random subsample of 41 studies (10% of the 414 studies detected during the literature research) to determine the objectivity and reliability of the coding procedure. The inter-rater agreement was high (90%). 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, design).

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: $d = (M_T\text{-}M_C)/sd_p$ 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 out-

come 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). 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: $\text{var} = (n_1+n_2)/n_1n_2 + d^2/2(n_1+n_2)$ 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; Lipsey & Wilson, 2001). 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 Figure 2). The first SAMD-value divergent from the flat, gradual slope is the cut-off value (Huffcutt & Arthur, 1995). We excluded 11 (4.6%) of the 237 pairwise comparisons and 4 whole studies (see Table 1).

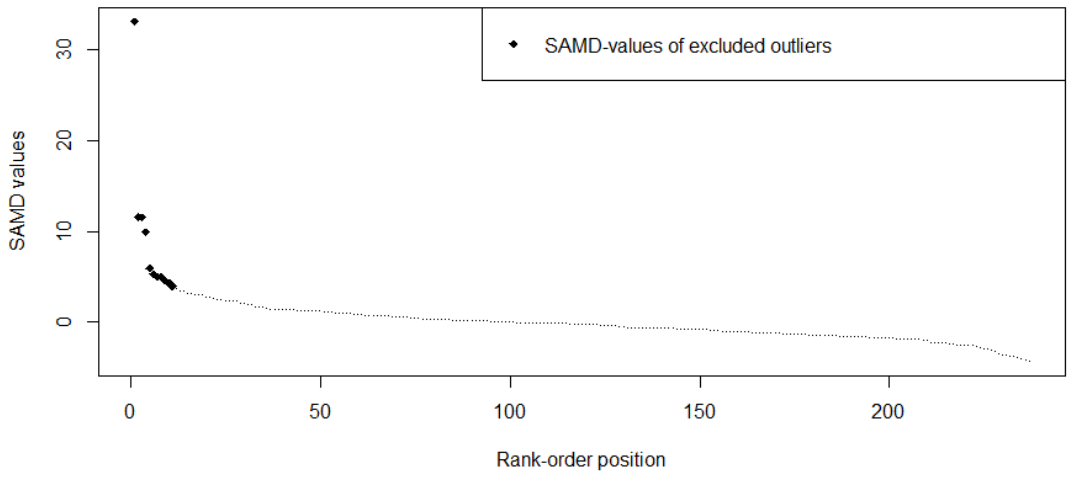


Figure 2 *Sample-adjusted meta-analytic deviancy (SAMD) values over rank-order position. The SAMD cut-off value is estimated by identifying the first value divergent from the flat gradual line of SAMD values (Huffcutt & Arthur, 1995). Outlying SAMD-values represent pair-wise comparisons with an unreasonable 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 (1988a) 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.

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, Tipton, & Johnson, 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 (ρ = 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. 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).

3.5 Results

The 72 studies included in this meta-analysis were published between 1972 and 2012 (see Figure 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' (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, 5,355 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.

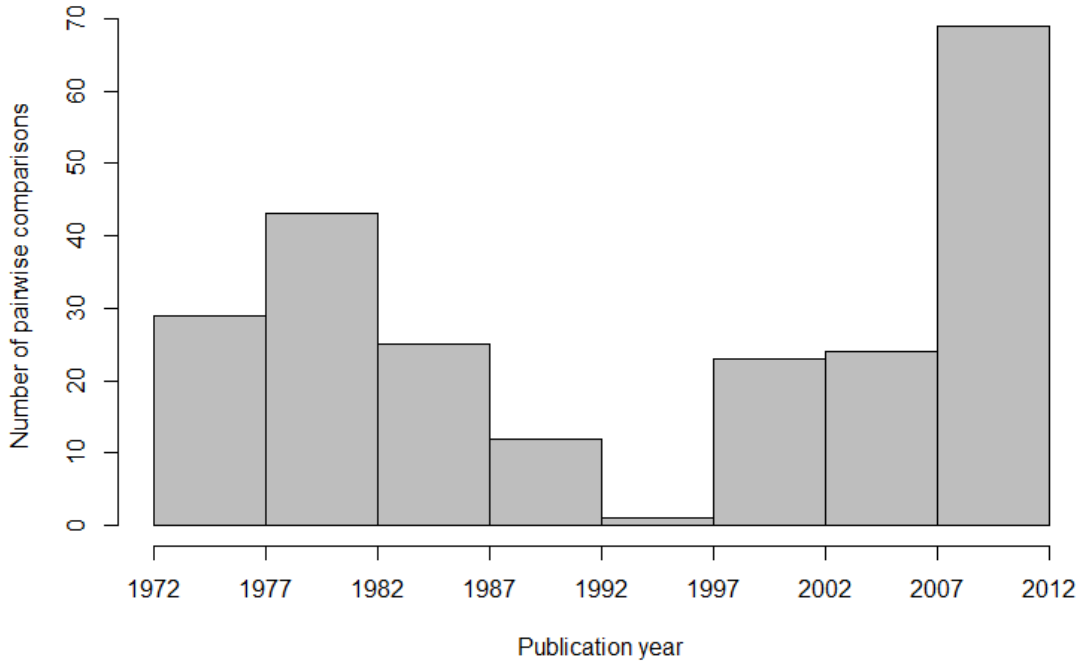


Figure 3 *Distribution of studies over publication year.*

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 Figure 4) 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. A significant QE-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).

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*	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. *Datasets are not identical due to differing inclusion criteria.

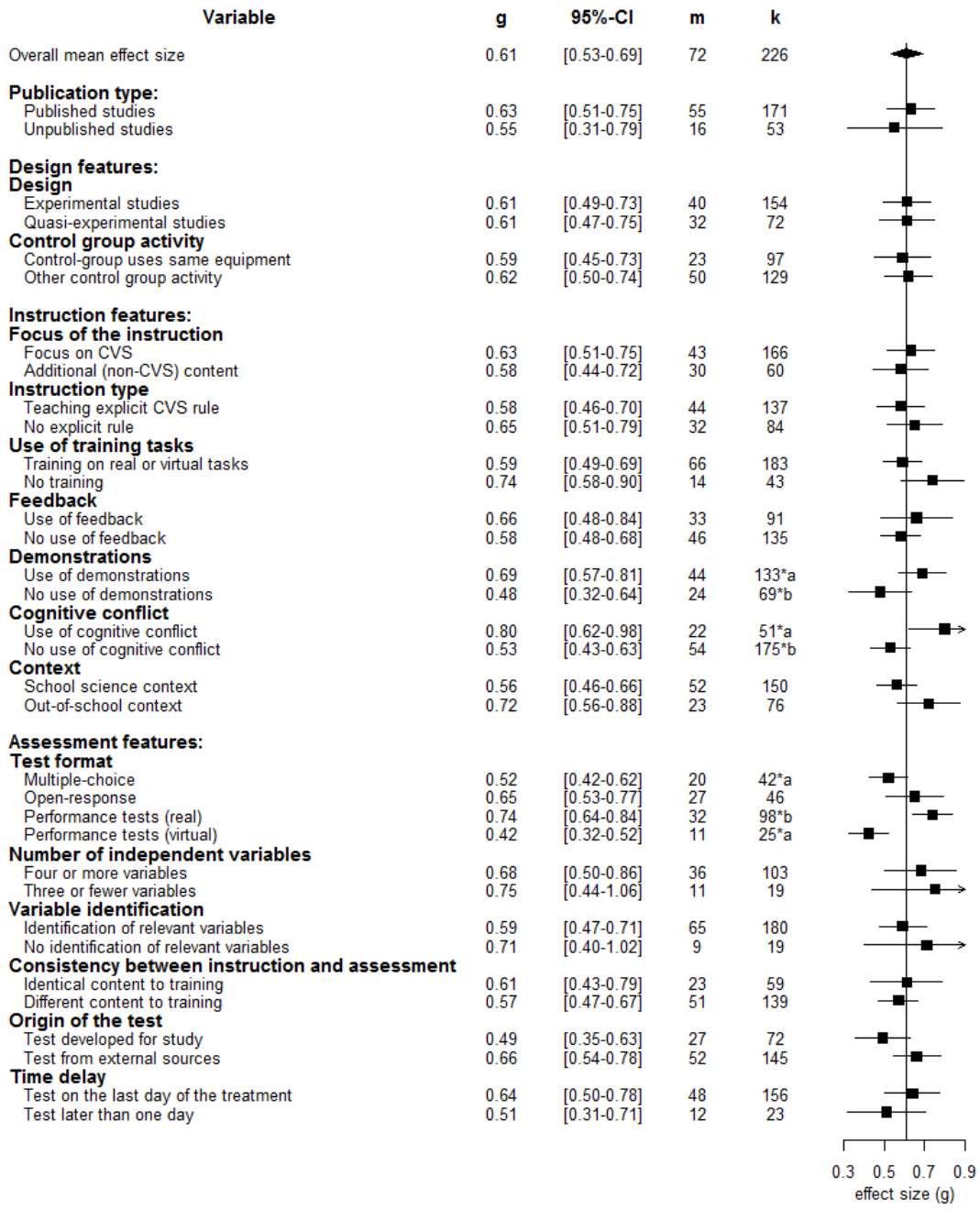


Figure 4 *Forest plot of the moderator effects (g) and 95% confidence intervals (CIs). Note. Items with subscripts a and b mark groups differing from each other at $p < 0.05$. The column m refers to the number of studies, and k represents the total number of pairwise comparisons.*

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 Figure 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 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 Figure 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.

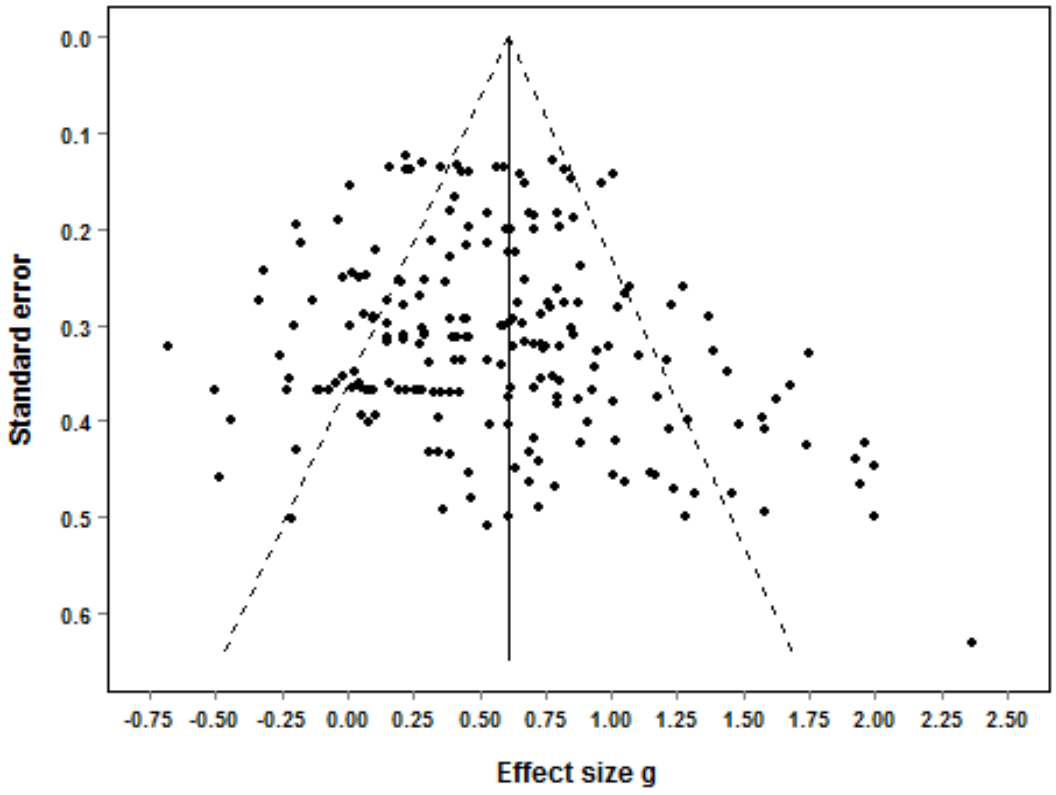


Figure 5 *Funnel plot showing the mean weighted effect size in relation to the corresponding standard error for every pair-wise comparison. The dotted line indicates the 95% confidence interval.*

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 shows 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 Figure 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 (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 (Lazarowitz & Huppert, 1993) or other content knowledge (Zimmerman, Raghavan, & Sartoris, 2003), but only three studies taught three or more additional content areas. In contrast with findings from single intervention studies (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 (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 mul-

tiple-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 to select an un-confounded 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, 1988a) 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 to 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.

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		

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.

3.6 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' (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 (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 primary 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 (NRC, 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 (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 (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 atten-

tion 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 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 per 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 (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' (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 (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 towards 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).

3.7 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) and Dean and 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. In 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 effect 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.

4. The impact of sub-skills and item content on students' skills with regard to the control-of-variables-strategy (CVS)

Abstract: *The so called control of variables strategy (CVS) incorporates the important scientific reasoning skills of designing controlled experiments and interpreting experimental outcomes. As CVS is a prominent component of science standards appropriate assessment instruments are required to measure these scientific reasoning skills and to evaluate the impact of instruction on CVS development. However, evidence from a meta-analysis raises concerns about the validity of existing CVS assessment instruments. A detailed review of existing CVS instruments suggests that they utilize different, and only a few, of the four critical CVS sub-skills in the item development. This study presents a new CVS assessment instrument (CVS-Inventory, CVSI) and investigates the validity of student measures derived from this instrument utilizing Rasch-analyses. The results indicate that the CVSI produces reliable and valid student measures with regard to CVS. Furthermore, the results show that the item difficulty depends on the CVS sub-skills utilized in item development, but not on the item content. Accordingly, previous instruments that are restricted to a few CVS sub-skills tend to over- or underestimate students' CVS skills. In addition these results indicate that students are able to use CVS as a domain general strategy in multiple content areas. Consequences for science instruction and assessment are discussed.*

Key words: Control-of variables-strategy, Rasch-analysis, Scientific reasoning, Inquiry skills, Experimental skills, Assessment instrument

4.1 Introduction

The ability to design controlled experiments and interpret experimental outcomes is a core scientific reasoning skill and a prominent object of science curricula and standards (National Research Council, 1996, 2000; NRC, 2012). Hence, appropriate assessment instruments are needed in order to 1) measure this core scientific reasoning skill and 2) evaluate the impact of specific science instruction on the development of those skills. However, two separate meta-analyses that summarized and evaluated the findings of more than 60 intervention studies found the choice of test instruments used to measure outcomes had a significant influence on student CVS measures (Ross, 1988a; Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2015). A reason for the incoherence of student measures across instruments might be that different instruments cover different sub-skills of the broader construct "control of variables strategy" [CVS] detailed by (Chen & Klahr, 1999).

The purpose of this article is to present a new assessment instrument [CVS Inventory (CVSI)] which utilizes items addressing the CVS sub-skills of identifying controlled experiments (ID), interpreting the outcome of a controlled experiments (IN) and understanding the indeterminacy of confounded experiments (UN). We analyzed a large data set collected with the CVSI and demonstrate that the difficulty of the CVSI items depends on these CVS sub-skills. Accordingly, we suggest that an over- or underestimation of student abilities with previous instrumentation may result from the restricted range of CVS which is measured by many instruments. The CVSI appears to provide a measurement scale which can be used to monitor the ability level of students concerning CVS more precisely.

4.2 The Control of Variables Strategy (CVS) in science education

In the literature skills related to controlling variables have often been referred to as "isolation of variables" (Inhelder & Piaget, 1958), "vary-one-thing-at-a-time" [VOTAT] (Tschirgi, 1980) or "control of variables strategy" [CVS]" (Chen & Klahr, 1999). According to Chen and Klahr (1999, p. 1098) "CVS is a method for creating experiments in which a single contrast is made between experimental conditions. The full strategy involves not only creating such contrasts, but also being able to distinguish between confounded and unconfounded experiments. This includes the ability to make appropriate inferences from the outcomes of unconfounded experiments as well as an understanding of the inherent indeterminacy of confounded experiments" (see Figure 6).

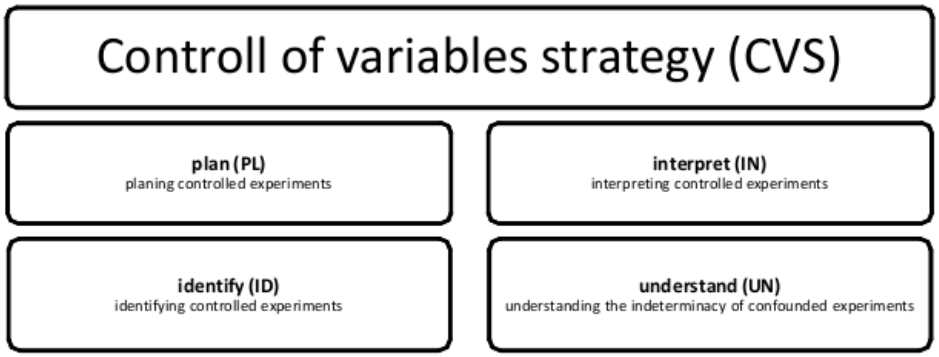


Figure 6 *Sub-skills of the CVS-Construct according to Chen and Klahr (1999).*

Most alternative definitions of CVS are imprecise because they define no (sub-) skills or performance expectations. For instance, CVS is defined as "isolation and control of variables" by Lawson (1978, p. 12) or as "(eliminating) alternative interpretations of a situation" by Millar and Driver (1987, p. 49). An exception to this lack of precision is the definition by Ross (1988a, p. 407) who summarized the sub-skills implemented in different CVS instruments. According to his definition CVS consists of the four sub-skills "distinguishing controlled and uncontrolled experiments", "remediating uncontrolled experiments", "planning controlled experiments" and "justifying experimental designs by referring to a general rule" (p. 407). However, this definition is incomplete because it lacks the sub-skills of interpreting experimental outcomes and understanding the indeterminacy of confounded experiments which are crucial for conducting scientific inquiries. Moreover, remediating uncontrolled experiments is not an independent sub-skill but instead a combination of the sub-skills identifying uncontrolled experiments and planning controlled experiments. Furthermore, justifying experimental designs using a general rule is not a process skill and thus not the focus of this study. For the purpose of this paper we utilize Chen and Klahr's (1999) definition of CVS because 1) it is the most extensive existing definition, 2) it defines crucial CVS sub-skills and 3) the definition can be used to evaluate existing CVS test instruments.

CVS has a prominent role in science standards because it is the fundamental principle which leads the investigation of causal relations by scientific experiments (Rousmaniere, 1906). Beyond that, scientific process skills like CVS are necessary for learning through inquiry as they enable students to conduct their own informative investigations. In addition, reasoning based on unconfounded evidence is important not only in science but in all argumentation about causality. Accordingly, CVS is crucial for learning scientific literacy and linked to broader educational goals such as inquiry skills and argumentation (Kuhn, 2005a). Current research

about students' CVS skills is limited as existing instruments are restricted to single CVS sub-skills. Assessments in science education and science education research require instruments that measure the complete CVS construct because students who are supposed to work independently on their own inquiries need to apply all four CVS sub-skills. Conclusions based on restricted instruments therefore give an inaccurate picture of students' actual abilities to utilize CVS. Furthermore, to introduce the complete CVS concept to students, knowledge must be obtained regarding the best instruction method for every single CVS sub-skills. To build this knowledge more extensive CVS instruments are required to evaluate the effect of instructions on student achievement regarding different CVS sub-skills.

4.3 Literature Review

Concerns about the comparability of CVS measures based on existing instruments originate from a meta-analysis (Schwichow, Croker et al., 2015) that summarize the results of 72 intervention studies designed to increase students' CVS skills. This analysis suggests that studies utilizing multiple-choice instruments to assess student CVS achievement have significant smaller effect sizes than studies utilizing other instrument formats (e.g. open response, virtual / hands-on experimental tasks). However, in a detailed comparison of these instruments Schwichow et al. (2015) found that instruments with different formats in fact measure different sub-skills of the broader CVS construct. Thus, instrument format and measured CVS sub-skill are confounded in existing CVS instruments so that the isolated effect of the utilized CVS sub-skill on CVS measures is unknown.

In addition, existing CVS instruments exhibit a range of item content from biology, chemistry and physics to everyday life content. Again, item content and instrument format is confounded in existing CVS instruments. Hands-on instruments focus particularly on physics experiments while biology content is only utilized with paper-and-pencil or virtual CVS instruments and chemistry content is rarely utilized in any instrument (Schwichow, Croker et al., 2015). In summary, existing CVS instruments differ regarding the instrument format, the utilized CVS sub-skills and the item content. Below we present an overview of past research regarding 1) the impact of instrument format upon CVS measures, 2) the impact of CVS subskills upon CVS measures and 3) the impact of item content upon CVS measures.

The Impact of Instrument Format on CVS Measures

Evidence from various research fields shows that students' performance on assessment tests is influenced by the utilized test format (e.g. open-response items versus multiple-choice items or hands-on items). It seems that differently formatted instruments require different cognitive skills and hence measures of the same construct but from differently formatted instruments are not comparable (Martinez, 1999; Shavelson, Baxter, & Pine, 1992). In their meta-analysis Schwichow, and Croker et al. (2015) found that CVS multiple-choice instruments seem to be easier than open-response or hands-on items when it comes to CVS. However, the meta-analysis compares test instruments that differ not only in format but also in content, number of independent variables and utilized CVS sub-skills. Only two studies (Staver, 1984, 1986) isolate the effect of instrument format by comparing CVS measures on instruments of different formats while holding the item content, the utilized CVS sub-skills and number of independent variables constant. In the first study by Staver (1984) 253 biology freshman students were assigned either to open-response or multiple-choice CVS items. Both item formats utilized the CVS sub-skill of interpreting experiments (IN) by asking student to interpret the outcome of a controlled experiment and to justify their interpretation. The results suggest that item format leads to a significant amount of variance in student CVS measures. The second study by Staver (1986) with 548 eight graders investigated the effect of item format and number of independent variables upon CVS measures. The study had a two (open-response versus multiple-choice item format) times four (2, 3, 4 or 5 independent variables) research design and entire science classes were assigned to one of the eight conditions. All items asked students to plan experiments (PL) by choosing materials from a list and to justify their choice either by selecting or by formulating a justification. In contrast to his first study Staver (1986) found no direct effect of the test format on student CVS measures. Instead, his results showed that items with four or five independent variables are significantly more difficult than items with two or three independent variables regardless of instrument format. Moreover, he found an interaction effect of item format and number of independent variables indicating that the number of variables has a larger impact on CVS measures in open-response than in multiple-choice items. In summary, the presented studies suggest that item format has 1) a direct effect on CVS measures and 2) an indirect effect on CVS measures moderated by further instrument features such as the number of independent variables.

The Impact of CVS Sub-Skills on CVS Measures

By utilizing Chen and Klahrs (1999) definition of CVS it is possible to classify instrumentation with respect to the inclusion or exclusion of the four critical CVS sub-skills planning controlled experiments (PL), identifying controlled experiments (ID), interpreting the outcome of a controlled experiment (IN) and understanding the indeterminacy of confounded experiments (UN). No previous study has investigated the impact of utilized CVS sub-skills on CVS measures while holding the instrument format, item content and the number of independent variables constant. However, the Munich longitudinal study (Bullock & Ziegler, 1999) compared CVS measures on different CVS sub-skills with items of varying content with the same "format". In that study 200 children of 8-12 years were interviewed on different control of variables tasks. For example, children had to suggest an experimental setup to evaluate the impact of different airplane features on fuel-efficiency (PL) before they were asked to choose an appropriate experimental design for the identical problem from the presented examples (ID). In a further example children were asked to plan experiments about variables that influence the extension of springs (PL) and to interpret (IN) the outcomes of experiments regarding identical problems presented to them afterwards. The study results suggested that independent of participants' age, planning items (PL) are the most difficult items while interpreting items (IN) are easier than identification items (Bullock, 1991; Bullock & Ziegler, 1999). No empirical study has compared understanding items (UN) to the CVS sub-skills planning (PL), identifying (IN) and interpreting (IN).

The Impact of Item Content on CVS Measures

In theory CVS is a content-independent strategy that can be applied to investigations of causal effects in science, social sciences and everyday life. However in praxis students' science process skills like CVS depends on their knowledge and preconceptions about the item content (Eberbach & Crowley, 2009; Millar & Driver, 1987). Accordingly, the relation between students' content knowledge and the item content must be kept in mind when interpreting CVS skills. With respect to item content existing CVS instruments can be classified as either "domain general" or "domain specific" instruments. Domain general instruments attempt to minimize the impact of students' content knowledge (e.g. knowledge about mechanics) on CVS measures. Such instruments use items which utilize everyday contexts and/or abstract contexts. For example, tasks present fictional experimental data that compare the impact of "color of chewing gum" on teeth. Students have to interpret these data to find out which color gum supports healthy teeth. Students' prior beliefs play no role in answering this question because

there is no reason to expect a specific gum color to foster healthy teeth (Koerber, Sodian, Thoermer, & Nett, 2005). A further example of a domain general instrument is one developed by Bullock (1991). Bullock asks students to plan experiments to test which of three variables (decoration, candle length, roof style) makes a difference in how well a candle lantern will remain illuminated in the wind. In particular, such domain general instruments have been used by developmental psychologists (e.g. Bullock, 1991) and educational researchers (Koerber et al., 2005) to investigate the scientific reasoning skills of pre- and elementary school children.

The second type of existing CVS instrumentation can be characterized as "domain specific". Such instruments explicitly use items with a scientific content to assess students' scientific reasoning ability in what are termed "realistic contexts". An example of an instrument composed of domain specific items is the work of Dillashaw and Okey (1980). Their instrument of integrated science process skills asks students, within the context of biology, to 1) choose a controlled experiment (an experiment with a single contrast) from a set of potential examples (ID) and to 2) choose a hypothesis that can be tested by a described experiment (IN). A second example of an instrument utilizing domain specific CVS items is a classroom test of scientific reasoning (Lawson, 1978). This instrument requires students to choose a controlled experiment (ID) and to interpret experimental outcomes (IN). The items of this test cover topics in the fields of physics, chemistry and biology. Predominantly domain specific CVS instruments have been utilized to measure students at the high school, college and university level.

The domain targeted by CVS instruments seems to impact conclusions about students' skills in designing and students' skills in interpreting controlled experiments. A study by Song and Black (1992) contrasting CVS tasks with everyday life and scientific content that are comparable regarding the utilized CVS sub-skills, item content and number of independent variables showed that students perform better on everyday life tasks than on scientific tasks. Studies using domain general instruments consistently suggest that very young and older students have a basic understanding of controlled experiments (Zimmerman, 2000, 2007). Studies using domain specific CVS instrumentation have suggested a range of conclusions. It seems that student CVS measures depend on whether students' beliefs conflict with the experimental outcome or the supposed experimental outcome (e.g. whether they believe that the mass of a pendulum has an impact on its period). Students use CVS more often in the case of belief

consistent outcomes (e.g. candy is bad for teeth) (Croker & Buchanan, 2011; Keating, 1990). It also seems that students tend to produce an expected effect instead of testing a hypothesis and designing experiments that produce an anticipated outcome by varying more than one variable (Penner & Klahr, 1996). A possible explanation for this finding is that students try to avoid conflicts between experimental evidence and their conceptual knowledge by adapting the evidence to their knowledge. They do not choose the alternative approach of adapting their concepts to the evidence because they cannot explain the mechanism that caused the experimental outcome (Koslowski, 1996). Taken together, studies utilizing domain specific CVS instruments show that beside students' CVS skills impacting their measures, a second key issue is the students' level of content knowledge regarding item content. Accordingly, domain general CVS instruments tend to be applied 1) to test young students with little science knowledge or 2) to produce CVS measures not contaminated with content knowledge. However, students' performance on domain general tasks is non-predictive for their ability to utilize CVS on tasks with scientific content because their performance on domain specific tasks depends on their preconceptions (Millar & Driver, 1987). Consequently, for classroom assessment a CVS domain specific approach is preferred over a CVS domain general approach because one common goal of science education is to foster students' use of process skills in scientific contexts (Pellegrino, Wilson, & Koenig, 2013).

4.4 Past CVS instrumentation

The CVS definition proposed by Chen and Klahr (1999) provides an overarching theory with which existing CVS instrumentation can be classified. Past CVS instruments have addressed some, but not all, of the CVS sub-skills. **Table 4** presents a summary of past CVS instrumentation efforts. Generally multiple-choice CVS instruments have been restricted to items which involve identifying (ID) and interpreting (IN) (e.g. test of integrated science process skills by Dillashaw & Okey, 1980). Hands-on instruments have been restricted to items which address the CVS sub-skill planning (PL) (e.g. Piagetian Interview by Inhelder & Piaget, 1958). In summary, item format and utilized CVS sub-skills are confounded in existing CVS instruments (e.g. hands on instrument to evaluate PL). This pattern of a specific item type and subskill might be present because some formats lend themselves to accessing specific CVS subskills. For example, in multiple-choice tests it is easier to utilize identification items (ID) that ask students to choose an appropriate experimental design in comparison to presenting students with test items that ask for planning a controlled experiment (PL). Another reason for the range of pairings of item format and CVS subskill might be testing efficiency. For in-

stance, to present students with identification items (ID) using a hands-on instrument is inefficient compared to the use of multiple-choice items. As a result of the mix of items types which have been used for specific CVS subskills (but not all subskills) the isolated impact of instrument format and utilized CVS sub-skill in existing CVS instruments is not known.

Table 4 Overview of existing CVS Multiple-Choice Instruments.

Test	PL	ID	IN	UN	Format	Domain
Test by Staver (1984)			1		Open respond / Multi-ple-choice	Biology
Test by Staver (1986)	1				Open respond / Multi-ple-choice	Physics
Chewing gum test by (Koerber et al., 2005)			1		Interview	Every-day
Oral health test by (Croker & Buchanan, 2011)	1				Interview	Every-day
Piagetian Interview (Inhelder & Piaget, 1958)	1				Interview with Hands-on tasks	Physics, Chemistry
Test of integrated science process skills (Dillashaw & Okey, 1980)		3	9		Multiple-Choice	Biology
Classroom test of scientific reasoning (Lawson, 1978)		3	9		Multiple-Choice	Physics, Chemistry Biology.
Lantern task by (Bullock, 1991)	1	1	1		Interview with card choice	Every-day
Airplane task by Bullock and Ziegler (1999)	1	1			Interview with card choice	Every-day
CVS posttest by (Chen & Klahr, 1999)		15			Multiple-choice	Biology, Every-day
CVS tests by Kuhn and Dean (2005)	5				Online interactive test	Geo-science, Every-day
CVS posttest by Dean and Kuhn (2007)	5				Online interactive test	Geo-science
Meta-strategic knowledge test by (Zohar & David, 2008)				6	Open response items	Biology

Note. Numbers are the total number of items which belong to a specific CVS sub-skill.

A further limitation of existing CVS instruments is that most current CVS instruments lack items that ask students to demonstrate an understanding of the indeterminacy of confounded experiments (UN). The only example of considering UN items involved a study by Zohar and David (2008). In this study students were confronted with a fictional story about a person who wanted to investigate which variables impact the speed of sail boats. The experiment designed by the character in the story is confounded and students are asked to evaluate the conclusions made by the character. In summary, the review of existing CVS instruments suggests that 1) existing CVS instruments tend to be limited to a few sub-skills of the broader CVS construct and 2) certain instrument formats are predominantly utilized to implement specific CVS sub-skills. These limitations suggest that the student measures which can be computed with existing instruments may have limited validity. For example, a restricted coverage of the CVS construct can cause an over- or underestimation of students' abilities. Another problem with existing instruments is the incomparability of measures as the result of utilizing different CVS subs-kills, formats and content. In particular, it is not clear whether format effects are caused by the "format" or by the utilization of different sub-skills because existing CVS instruments utilize different sub-skills in items of different formats.

4.5 Research questions

The aim of this study is to develop a multiple-choice instrument (CVSI) that involves relevant CVS sub-skills in the context of middle-school physics and to present evidence of the validity of student measures based on this instrument. Furthermore, we use the CVSI to answer the following two research questions:

(1)What is the pattern of item difficulty of the CVS sub-skills?

(2)What is the pattern of item difficulty for items covering different physics topics?

4.6 Instrument development

We decided to develop the CVSI using multiple-choice item format for a number of reasons. First, multiple choice instruments provide the opportunity to administer a larger number of items to respondents. This can provide the opportunity to increase the precision with which person measures can be determined (often more items administered to respondents can decrease measurement error). Second, multiple choice instruments utilizing graphical represen-

tations can minimize the impact of students' varying writing ability levels on CVS measures by avoiding the use of written responses. Third, multiple-choice instruments can facilitate quick data collection and scoring in comparison to instruments using alternative formats (see Martinez, 1999, for a review of different item formats). A drawback of using multiple-choice item format is that items measuring the CVS sub-skill of planning (PL) cannot be assessed. Following a weighing of the pros and cons of instrument format, the new instrument developed in the multiple-choice format (the CVSI) is restricted to the CVS sub-skills of identification (ID), interpretation (IN) and understanding (UN).

A standardized procedure for item development was utilized with respect to item content, the number of independent variables, and the formulation of answer options. In short, the CVSI items were developed so that they differ only regarding the utilized CVS sub-skill. The CVSI consists of 23 multiple-choice items each having one correct answer and three distractors. All items of the CVSI are embedded in middle school physics contexts of heat and temperature or electricity and electromagnetism (further referred to as electro/magnetism) because middle school is known to be the timeframe of the largest changes in science concept knowledge and an important period for the development of long-term interest and engagement in science (Ma & Wilkins, 2002). Accordingly, such middle school context instruments are particularly important because many intervention studies and surveys focus on this important time period. Both topics (heat and temperature and electro/magnetism) are a component of the middle school science curriculum in most German states. Furthermore, these two topics are also an integral part of curricula in many other countries including the US (NRC, 2012), England (Department for Education, 2014) and Singapore (Curriculum Planning & Development Division, 2007). Each of the 23 items has graphical illustrations in order to minimize the influence of reading ability on students' CVS measures.

The CVSI includes 11 items which belong to the CVS sub-skill of identifying controlled experiments (ID). Each of these items starts with a short story about a fictitious person who wants to prove a specific hypothesis about a causal relationship. Afterwards, students have to select one correct experiment from one of of four graphically presented experiments to prove or disprove the hypothesis. Only one experiment shows a controlled experiment and is therefore correct. The distractors show confounded experiments with two, three or four variables

changed. For each ID item, the order of the answer options was random. An example of an ID item is presented in Figure 7.

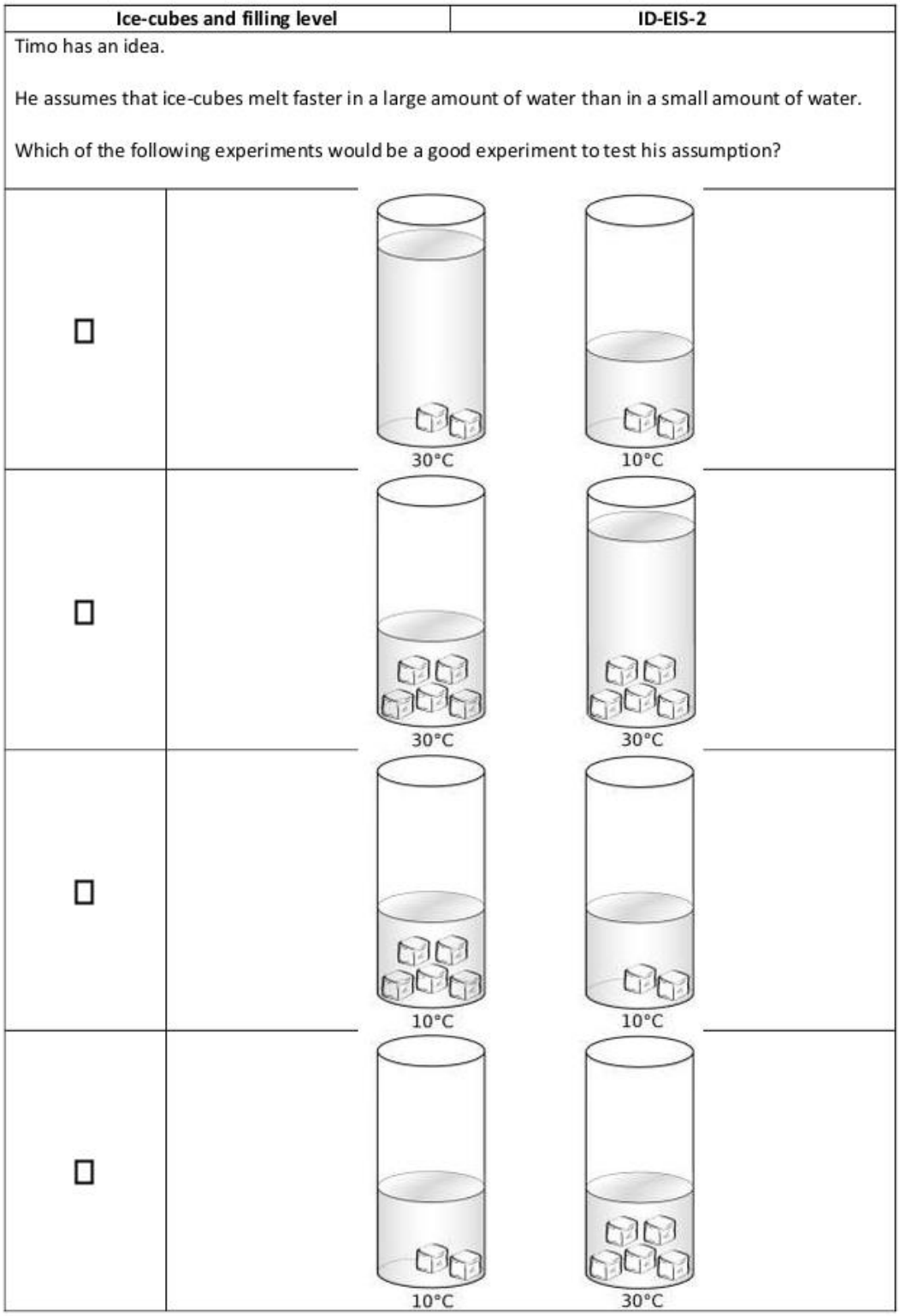


Figure 7 *Example of an identifying (ID) item. Answer two is correct because the critical variable "filling level" varies, while all other variables are the same between both conditions.*

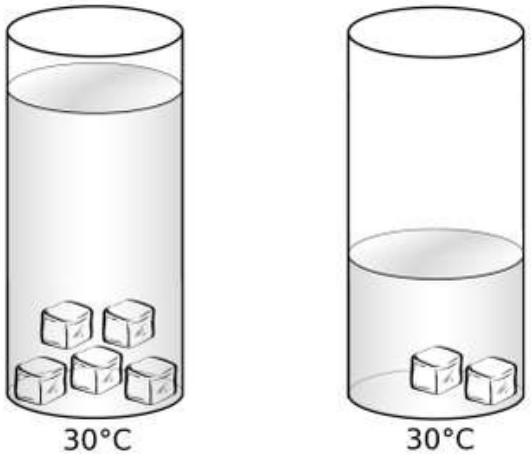
The 5 items for the CVS sub-skill interpreting (IN) and the 7 items for the sub-skill understanding (UN) have a highly similar structure. Items of both types start with a drawing that shows the outcome of an experiment. Students are then asked to interpret the presented experimental outcome. The only difference between the two item types is that interpreting (IN) items include the outcome of controlled and valid experiments. The understanding (UN) items consider the outcome of confounded and thus invalid experiments. For IN items students have to draw appropriate inferences from a controlled experiment. For UN items students have to decide that the presented experiment is confounded and the students have to recognize that they cannot draw a valid conclusion from the presented outcome. The four written reply options for the IN items and the UN items are standardized and are always presented in the same order. The response options are:

1. Variable X has an impact on the outcome of the experiment.
2. Variable Y has an impact on the outcome of the experiment.
3. Variable X and variable Y have an impact on the outcome of the experiment.
4. The experiment does not allow any valid conclusion.

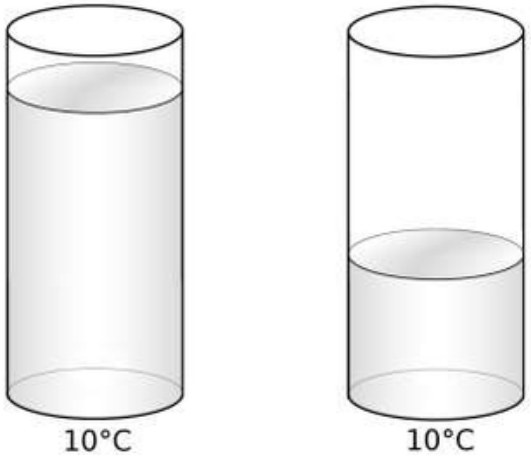
Figure 8 shows an example of an understanding (UN) item. The full CVS inventory is available as online supplemental material.

Disappearing ice-cubes	UN-EIS-2

Julia did the following experiment:



After 15 minutes…



What does her experiment show?

☐	The amount of ice-cubes influences the temperature change after adding ice-cubes.
☐	The original water temperature influences the temperature change after adding ice-cubes.
☐	The amount of ice-cubes and the original water temperature influences the temperature change after adding ice-cubes.
☐	The experiment does not allow any valid conclusion.

Figure 8 *Example of an understanding (UN) item. Answer four is correct because this experiment is confounded as more than one variable differs between the contrasted conditions.*

4.7 Data collection

The CVSI was administered to 386 7th, 8th and 9th grade students from four comprehensive schools in northern Germany. The students of these schools range from students with special education needs to students who plan to pursue a university degree and also include students who do not plan to attend a university. As the research project was confined to the research questions no demographic data was collected from students. The complete 23 item CVS inventory was answered by 215 students while the remaining 171 students completed a subset of 12 items. The shortened version of the CVSI (12 items opposed to 23 items) consists of three different booklets of 12 items each. The three booklets share at least six anchor items (Boone, Staver, & Yale, 2014). Each test booklet includes six identifying items (ID), three interpreting items (IN) and three understanding items (UN). The students were given 25 minutes to complete the entire CVS inventory and 15 minutes to complete the short version instrument.

4.8 Data Analysis

First, we present procedures that were taken to convert the nonlinear raw scale data to a linear scale by utilizing the Rasch model. All further analyses are based on Rasch measures (e.g. item difficulties). Additionally, we detail analysis steps utilized to compute Rasch item / person measures, to investigate the instrument functioning, and to conduct statistical tests.

Utilizing the Rasch Model

Raw test data of the type collected with the CVSI cannot be assumed to represent linear measures and thus must be converted to a linear scale utilizing techniques such as Rasch measurement. We utilized the Rasch model (Rasch, 1960) and Rasch analysis (Wright & Masters, 1982; Wright & Stone, 1979) to compute person and item measures which were used in further analyses to answer the research questions. The Rasch model expresses item measures (e,g, the items of the CVSI) and person measures (e.g. students taking the CVSI) on the same scale and therefore allows an evaluation of which items are typically solved by students with a specific ability level. A further benefit of the Rasch analysis is that it provides additional measures like item and person reliability and outfit values that are useful to evaluate and document aspects of instruments (e.g. CVSI) with regard to validity and reliability. Aspects of validity and reliability must be accessed in order to test whether measures are confident and to rigorously evaluate the functioning of instruments. For these reasons the application of the Rasch model is considered a required step in instrument development, instrument

revision, and outcome measure computations. The recent text Rasch Analysis in the Human Sciences (Boone et al., 2014) provides details as to the application of the model.

Computation of Scale Score Outcome Measures

The Winsteps Rasch analysis program (Linacre, 2014) was utilized for the computation of person outcome measures and item difficulties (the linear measures needed for parametric statistical tests). In this analysis we used the same probability value of 62% as used in PISA. That is, a person with the same measure as an item has a 62% probability of correctly answering the item and that person has greater than a 62% probability of correctly answering the items which have a measure below the measure of the person. Rasch measures in an initial analysis are expressed using a logit scale. Commonly the average item difficulty is defined as 0 logits. With such a definition of the zero point of a scale (which extends from negative infinity to positive infinity), item difficulty (and person measures) will be expressed with both positive and negative numbers. The pure numbers of the scale are not informative because the scale is relative. Instead, values from of the same scale have to be compared with each other. Lower logit values represent easier items (or less able students) and larger values correspond to more challenging items or more able students. All statistical analyses, as well as qualitative analyses were conducted with these logit values.

Instrument Functioning

The Rasch analysis program Winsteps (Linacre, 2014) provides numerous additional indices that can be used to further evaluate instrument functioning. In particular, we reviewed item fit, item reliability and person reliability. Moreover, we created Wright Map to study how CVSI items target the students' abilities in our sample. A Wright Map presents both item difficulties and person abilities on a single plot. More difficult items, solved by more able students are plotted in the upper part of the map while less challenging items, solved by most students are plotted at the bottom of the plot. By analyzing Wright Maps one can identify challenging and easy items and investigate whether the items of an instrument cover the ability spectrum of the sample.

Descriptive and Statistical Analyses

Following the computation of person measures and item measures, a range of statistical tests and descriptive analyses were conducted to evaluate patterns in the data. Of primary interest was the manner in which the instrument items defined the CVS trait and whether the item difficulty depends on the CVS sub-skills and item content (heat and temperature versus electro/magnetism). A three-way ANOVA with Post- hoc Bonferronie test was used to compare the mean item difficulties of identifying (ID), interpreting (IN) and understanding (UN) items. The mean item difficulties of heat and temperature and electro/magnetism items were compared using an independent t-test. For these analyses the R statistical packages were utilized.

4.9 Results

First, we present evidence for the reliability and validity of CVSI measures prior to presenting results to address the research questions.

Item Fit

A requirement of high-quality measurement is that items which are utilized to define a trait (as is done with the pool of items from the CVSI), fit the Rasch model. A common technique to explore this is through a review of MNSQ item outfit. Linacre (2002) has suggested that MNSQ values below 2.0 are not degrading for measurement and that MNSQ values of 0.5-1.5 are productive for measurement. An initial analysis suggested that no CVSI item exhibited a MNSQ Outfit value below 0.5 and that only three items exhibited MNSQ Outfit values greater than 1.5. Those items were UN.SO.1, UN.MS.2, UN.FL.2. Review of these three potentially misfitting items revealed that these three items were three of the four most difficult items of the CVSI for the sample. A review of all respondents answers to these three items and a comparison with each respondent's overall measure suggested that the misfit of the items was the result of a low number of respondents (who had low CVSI measures) having in contrast to the assumptions of the Rasch model correctly answered one or more of these items. Following the identification of these low performing respondents who very unexpectedly answered correctly, these respondents were retained in the analysis but were not utilized for the computation of item calibrations which defines the measurement scale. By this procedure the MNSQ Outfit values of the UN.SO.1, UN.MS.2 and UN.FL.2 items dropped below 1.30. The mean MNSQ Outfit of the whole set of 23 CVSI items was .95.

Item Reliability and Person Reliability

To further evaluate the functioning of the measurement scale we compute Rasch item reliability and Rasch person reliability. A person reliability of .73 *(Cronbach's α = .88)* and an item reliability of .99 resulted from the analysis. The high item reliability, in part, resulted from the very large number of respondents who answered each item. The lower but still acceptable person reliability resulted from the fact that with most testing scenarios there is a limit to the number of items which can be completed by respondents.

Item Difficulty of the CVS Sub-Skills

Figure 9a illustrates the mean item difficulty for items of the three CVS sub-skills. Items belonging to the understanding (UN) sub-skill *(mean item difficulty = 2.72, sd = 0.92)* are more difficult than identifying (ID) *(mean item difficulty = -1.24, sd = 1.10)* and interpreting (IN) items *(mean item difficulty = -1.08, sd = 0.57)*. No statistical difference was found between the difficulty of identifying (ID) and interpreting (IN) items. The small 95% confidence interval bands even in the case of low item numbers are further evidence for a similar difficulty of items of the same trait. An analysis of variance for the three CVS traits was computed to investigate whether the item difficulty depends on the CVS sub-skills. There was a significant and large effect of sub-skills on item difficulty, $F(2, 20) = 40.25$, $p < .01$, $\omega^2 = 0.77$. Post hoc Bonferroni tests show significant difference between understanding and identifying items, $p < .01$ $d = 3.82$ and between understanding and interpreting items, $p < .01$, $d = 4.76$. The difficulties of identifying and interpreting items do not differ significantly.

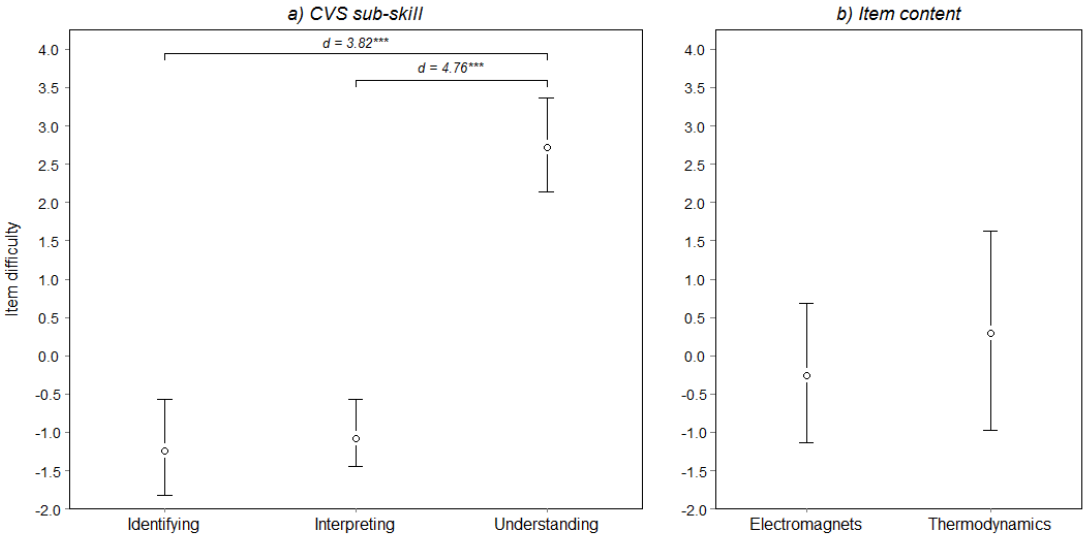


Figure 9 *Mean item difficulties and standard errors (in logits) for a) identifying (ID), interpreting (IN) and understanding (UN) items and b) for items with content from heat and temperature and electro/magnetism. Horizontal lines represent significant differences in item difficulties (p > .01). The corresponding effect size is reported using Cohen's d.*

Item Difficulty of Heat and Temperature and Electro/magnetism Items

The mean item difficulties of heat and temperature and electro/magnetism items (see Figure 9b) are similar. Our hypotheses concerning the impact of the item content on the item difficulty was that the mean item difficulties of electricity and heat items do not differ from each other. Accordingly, we should use the more conservative criterion of $p < 0.20$ to prove the truth of a null-hypothesis. An independent t-test shows that the difference between the mean item difficulty of electro/magnetism (*m = -0.26, sd = 1.70, n = 12 items*) and heat and temperature items (*m = 0.29, sd = 2.44, n = 11 items*) is non-significant (*t = 0.62, df = 17.71, p = 0.54*).

4.10 Discussion

First, we will discuss our findings regarding the reliability and validity of CVSI measures in detail and consequences. Second we will discuss the impact of CVS sub-skills and item content with respect to the item difficulty. Third we will interpret the implications of our results for science education.

Validity and Reliability of CVSI Measures

The CVSI provides reliable student measures as is evidenced by the person reliability of .73 (*Cronbach's α = .88)*. All items fit the Rasch model as the MNSQ Outfit values are below 1.3. A Wright Map (see Figure 10) was constructed to further evaluate the validity of student measures derived from the CVSI. In a Wright Map item difficulties and student measures are plotted in one figure with lower item difficulties and lower student measures at the bottom. Using a Wright Map one can see which items are typically solved by more able students because student measures and item difficulties are presented on the same scale in the same plot. The Wright Map of this study with the CVSI shows a clear pattern. All understanding items (UN) are in the upper part of the scale (more difficult items, solved by more able students), while identification (ID) items tend to be at the bottom and interpreting (IN) items in the middle. A statistical comparison of the mean item difficulties of the CVS sub-skills shows that only understanding items are significantly more difficult than interpreting and identifying items. This difference in item difficulty seems not to reflect differences in construct irrelevant item features because the understanding items and the easier interpreting items are designed to be highly similar (see instrument development). Instead, this pattern confirms findings from other studies which show that even preschool students are able to interpret and identify controlled experiments (Gopnik, Sobel, Schulz, & Glymour, 2001; Koerber et al., 2005; Piekny, Grube, & Maehler, 2014). However, striking is that three identification items (ID) are much more challenging than the remaining items that belong to this sub-skill. All three items cover content from electro/magnetism and require some content knowledge to identify variables (e.g. that car-batteries and mono-cells differ in their voltage). This might indicate that content knowledge is crucial for solving CVS tasks because students need knowledge about the variables to identify variables. Evidence from further studies about students' ability on understanding (UN) items does not exist. A further piece of evidence for the validity of CVSI measures is that the differences in item content (heat and temperature versus electro/magnetism) do not explain differences in the item difficulty as the mean item difficulty of heat and temperature and electro/magnetism items do not differ. Moreover, one can see by comparing Figure 9a and 9b that grouping items by content produces larger standard errors than grouping items by CVS sub-skills (this is a strong argument as the number of items per group is smaller when items are grouped by sub-skills compared to grouping items by content). However, it might be that we found no content effects because we utilized content that is part of the science curriculum. Accordingly, the variance in students' content knowledge regarding the item content

might be too low to detect content effects. In conclusion, our findings show that the difficulty of CVSI items depends primarily on the utilized CVS sub-skill and not on context or construct irrelevant item features.

The Wright Map shows that the current item set of the CVSI does not cover all student abilities of the sample. A gap of more than one logit appears between the most difficult identifying (ID) item (0.39) and the easiest understanding (UN) item (1.76). To improve the quality of student measures subsequent versions of the CVSI should include items that fill this gap. One possible alteration that could be made to the existing CVSI items is to make understanding (UN) items easier. This might be done by including an explicit statement with regard to which variables are confounded in the correct answer option. Thus, the item difficulty of the revised and original UN items could be compared to investigate whether students' poor skills on UN items are caused by inattention or by a misconception about the validity of uncontrolled experiments. The difficulty of the revised and original UN items should not differ if students hold a misconception about valid experimental designs. Another possible alternation is to increase the difficulty of identification (ID) and interpreting (IN) items by increasing the number of independent variables. This item revision would facilitate investigations whether students CVS skills depend on the number of variables or not? Hence, the suggested item revisions could not only lead to a better coverage of student abilities by the CVSI, but further increase our knowledge about the structure of the CVS construct. Currently, not all features that influence the difficulty of CVS items are known. All changes of item structure for the development of a new item pool should of course be evaluated using the psychometric techniques we have detailed before. In summary, the results of the Rasch analysis provide evidence that the CVSI is an instrument that provides reliable and valid student measures. A strong argument for the validity of CVSI measures is that the utilized CVS sub-skill is the only item feature that systematically influences the item difficulty. The CVSI is a new instrument that seems to offers a more complete picture of students' CVS skills. The CVSI is of relevance for science education and research because CVS is a crucial scientific reasoning skill that is a basic requirement for learning by inquiry.

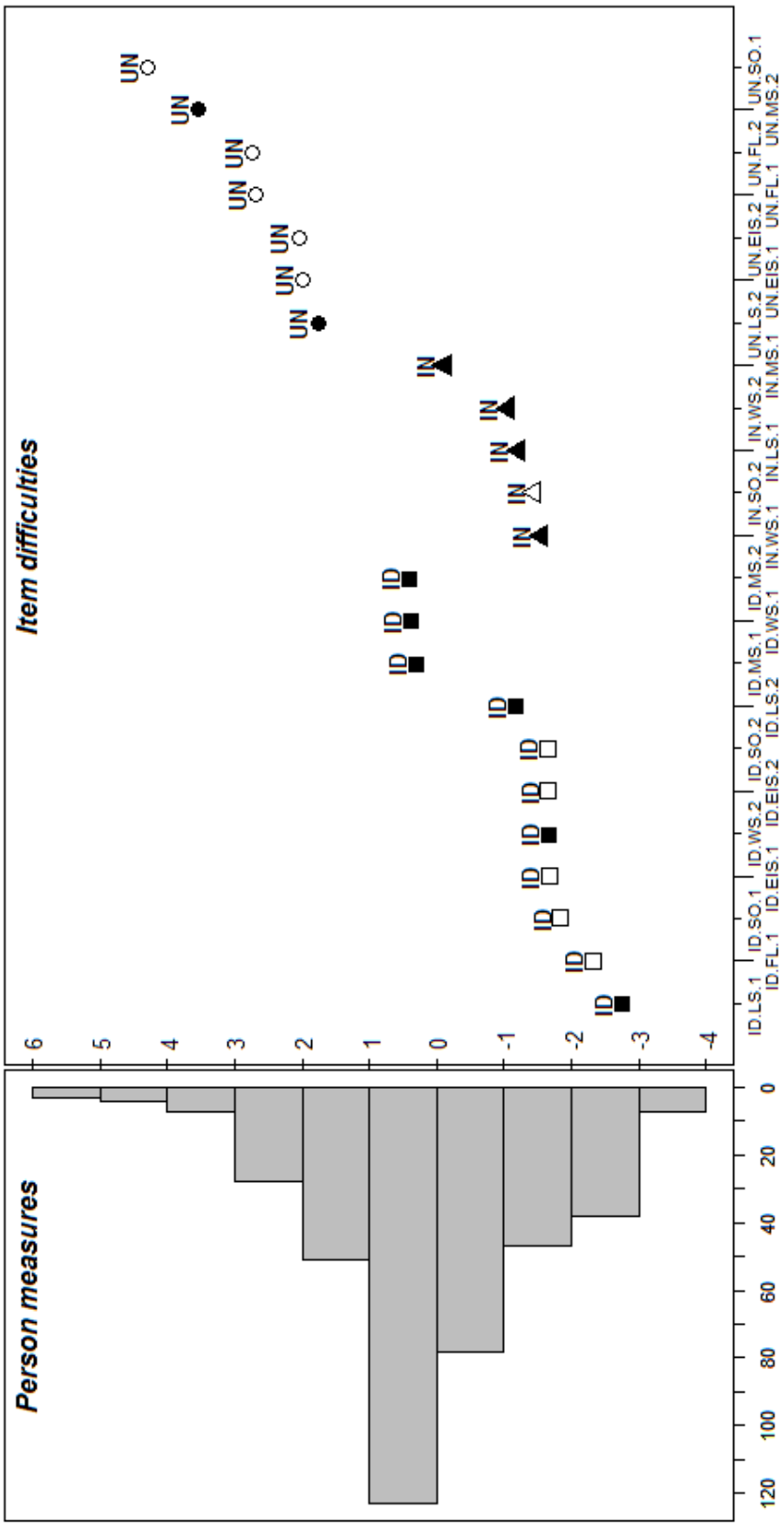


Figure 10 Wright Map of person measures and item difficulties (in logits) derived from the CVSI. Item difficulties and person measures are expressed one the same scale with easy items and less able students at the bottom and challenging items and more able students in the upper part of the scale. Squares represent ID-items, triangles represent IN-items and dots identify UN-Items. Items with content from heat and temperature have white symbols while black symbols represent items with content from electro/magnetism.

Pattern of Item Difficulty by CVS Sub-Skills and Item Content

This study allows a systematic investigation of the impact of CVS sub-skills and item content on item difficulty because additional item features such as item format or number of independent variables are constant in the new instrument. The results of our study show that understanding items (UN) are systematically more challenging than items utilizing the CVS subskill of identifying (ID) and interpreting (IN) (see Figure 9a or Figure 10). One explanation for this observed pattern of item difficulty is that understanding items (UN) ask students to think about the validity of experimental comparisons instead of identifying items (ID) and interpreting items (IN) which ask students to identify a presented contrast. To solve understanding items (UN) students have to 1) identify the confounding variables in the presented experiment and 2) think about the consequences of manipulating multiple variables. However, to solve ID items students only have to choose the "most valid" experiment among a presented selection of experiments. Similarly, to solve IN items students have to search for a contrast in the presented experiments and not necessarily look for additional contrasts. In conclusion, the correct reply to understanding items (UN) requires a more complex cognitive operation than correctly answering identification (ID) or interpretation items (IN). However, an alternative explanation why understanding items are more challenging might be that teachers of regular science classes do not utilize examples of confounded experiments. This means that students are not used to experiments with "non-results" so that they have no experience in thinking about the quality of experiments while interpreting experimental outcomes. These possibilities should be explored by future intervention studies which investigate the effect of instruction focusing on the UN sub-skill upon students' ability to solve understanding (UN) items.

One important implication of our findings is that past instruments that lacked UN items may overestimate student CVS skills. A lack of UN items in previous instruments means that interventions have not been evaluated with respect to the UN sub-skill. The lack of measuring the upper range of CVS subskills has serious implications for science education researchers. Researchers need to not only evaluate whether students can plan controlled experiments and interpret the outcome of controlled experiments, but also whether students understand that invalid conclusions derive from confounded experiments. This is particularly important because students involved in inquiries need to be aware of flawed conclusions derived from confounded experiments in order to interpret and discuss experimental data and to generate valid

knowledge from their inquiries. The understanding (UN) sub-skill is of practical importance for constructive critique of one's own and the experimental evidence of other students. In addition, students who understand that conclusions based on confounded experiments are invalid might pay more attention to possible confounding variables when planning and running own experiments.

Some of our study results contradict previous studies. Our results show that the CVS item content does not influence the item difficulty. It might be that students note similarities between items of the same CVS sub-skills because both are physics contexts. This finding provides evidence that students can use CVS as a content independent strategy to plan and interpret controlled experiments. As it appears that students can use CVS as a "content independent" strategy it seems very important that CVS would play a prominent role in current science curricula.

Implications for Science Teaching

The study outcomes regarding the impact of CVS sub-skills on item difficulty can supplement further intervention studies and science teaching. Instructions on complex concepts (like CVS) should start with more familiar (and thus easier) aspects and then follow a path of increasing difficulty to the most challenging aspects of the concept (Oser & Baeriswyl, F., J., 2001). Hence, at the beginning of CVS instruction with student teachers should focus on the identification and interpretation of controlled experiments before introducing the more challenging aspect of understanding the indeterminacy of confounded experiments. This ordered and planned teaching is much of what current research on learning progressions is based upon.

There has been some research with respect to the benefits of teaching the understanding (UN) subs-skill to students. An intervention study by (Zohar & David, 2008) that explicitly focuses on the understanding (UN) aspect of CVS shows a significant gain in students' abilities to design controlled experiments and in their understanding of the indeterminacy of confounded experiments. An unanswered question is whether this understanding skill will develop as a result of traditional CVS instructions which do not focus on the understanding aspect of CVS. It could be that to develop the understanding (UN) sub-skill students have to receive the less challenging identification (ID) and interpretation (IN) sub-skills before introducing the under-

standing (UN) sub-skill. However, students who understand the more challenging understanding aspects of CVS first may automatically develop the other aspects of CVS without explicitly instruction. To investigate these effects further studies are required that contrast both instruction sequences. Nevertheless, it seems important to introduce the understanding (UN) CVS sub-skill in order to facilitate students' inquiry skills and to show students that a first step in the interpretation of experimental results is a search for potential confounding variables.

In addition, as we found no effect of item content on item difficulty results of this study suggest that teachers can choose a content area that they wish to use to introduce students to CVS. Therefore, CVS is an ideal concept to be implemented in spiral curriculum. A repetitive practice of CVS within different contexts might be especially effective for the development of robust CVS skills. Supporting students' CVS skill development is important for science education as CVS skills are known to be related to science and school achievement in general (Adey & Shayer, 1990; Bryant, Nunes, Hillier, Gilroy, & Barros, 2013).

Limitations

A limitation of the current version of the CVSI is that it does not include items which measure students' abilities to plan controlled experiments (PL). Subsequent versions of the instrument should also include items on this sub-skill. One solution to include planning (PL) items in the CVSI would be to utilize interactive online items. A further limitation of the current version of the CVSI is that the instrument covers specific physics content. In order to explore student's ability to transfer CVS, more items within the domain of physics and other sciences are needed for lengthened or alternative versions of the CVSI. This is particularly important because CVS can be introduced within multiple disciplines. Thus an appropriate instrument for all these disciplines is required. Researchers can utilize the procedures detailed for item development of the CVSI to develop new and appropriate CVSI items.

4.11 Conclusion

This study shows that it is possible to develop a CVS multiple-choice test that includes at least three out of four relevant CVS-sub skills. The presented version of the CVSI seems to produce valid student measures concerning the CVS and includes the important CVS sub-skill of understanding the indeterminacy of confounded experiments. Because of the relevance of this skill for realistic inquiry situations we highly recommend including items covering that sub-skill in CVS instruments. The pattern of item difficulties in our data set reflects the theoretical difference between strategic and meta-strategic knowledge. The CVSI seems an ideal instrument for evaluating intervention studies on CVS because the test includes relevant sub-skills within the same contexts so that learning gains on sub-skills can be compared.

Important Findings

- The new CVSI instrument produces reliable and valid CVS measures
- The understanding (UN) CVS sub-skill is systematically more challenging for students than the CVS sub-skills identifying (ID) and interpreting (IN)
- Older instruments seem to overestimate students' CVS skills
- CVS is a domain general strategy
- Science teaching should include the understanding aspect of CVS

5. What students learn from hands-on activities

Abstract *The ability to design and interpret controlled experiments is an important scientific process skill and a common objective of science standards. Numerous intervention studies have investigated how the control-of-variables-strategy (CVS) can be introduced to students. However, a meta-analysis of 72 intervention studies found that the opportunity to train CVS skills with hands-on tasks (g =.59) did not lead to better acquisition of CVS relative to interventions without a hands-on component (g = .74). A potential reason for this finding might be that hands-on tasks are more demanding than initially assumed, as they require additional non-CVS related skills such as arranging materials, taking measurements, and recording data, all of which may lead to a high cognitive load and thus hinder learning. We conducted an intervention study in which we investigated the differential effects of hands-on and paper-and-pencil training tasks on 161 eighth-grade students' achievement. CVS was introduced to all students before they were grouped into a hands-on or a paper-and-pencil training condition. In both training conditions, students designed and interpreted experiments about which variables influence the force of electromagnets. Students in the hands-on group interacted with physical equipment while students in the paper-and-pencil group planned experiments using sketches and interpreted the outcome of experiments presented in photographs. We found no general advantage or disadvantage of hands-on tasks, as both groups did equally well on CVS and content knowledge tests. However, hands-on students outperformed paper-and-pencil students on a hands-on test identical to the training tasks, whereas the paper-and-pencil students outperformed hands-on students on a science fair poster evaluation task similar to the paper-and-pencil training. In summary, students learned task-specific procedural knowledge, but they did not acquire a deeper conceptual understanding of CVS or the content domain as a function of type of training. Implications for instruction and assessment are discussed.*

Keywords: Control-of variables-strategy, Inquiry skills, experimental skills, Hands-on learning, Cognitive load theory

The ability to design and interpret controlled experiments is an important scientific process skill and a common curricular objective of science standards (National Research Council [NRC, 2012; Next Generation Science Standards, NGSS Lead States, 2013). Accordingly, numerous intervention studies have investigated how students can be taught to design and interpret controlled experiments. Many of these studies utilize hands-on tasks to train students' experimentation skills because it has been assumed that students should benefit from hands-on experiences, as they offer authentic and direct practice with designing and interpreting controlled experiments. However, the results of a recent meta-analysis of 72 intervention studies challenge this view (Schwichow, Croker et al., 2015). Across 226 comparisons, a surprising finding emerged such that studies that did not utilize hands-on training had effect sizes (g = 0.74; n = 43 comparisons) that were not significantly different from studies utilizing hands-on training (g = 0.59; n = 183 comparisons). This finding is relevant for science education because hands-on experiences are advocated as a method that promotes successful learning, student engagement, and motivation (Haury & Rillero, 1994). In this paper, we present a comparison of highly similar hands-on and paper-and-pencil training tasks designed to teach skills in controlling variables. The purpose of this study is to isolate the unique effects of hands-on experience. An additional finding of interest was that the type of test format used to assess an intervention moderated study outcomes (Schwichow, Croker et al., 2015). To investigate potential interaction effects between training tasks and assessment format, we measured student achievement with the use of four different test instruments.

5.1 The Control-of-Variables Strategy (CVS) in Science and Science Education

A fundamental principle of scientific experiments is to isolate the causal effects of variables by contrasting conditions that differ only with the respect to the variable for which the causal status is under investigation. Thus, experiments are the primary scientific method to investigate causal relationships (Rousmaniere, 1906). The critical skills required for designing and interpreting controlled experiments are brought together under the term control-of-variables strategy (CVS). Chen and Klahr (1999) outline the four key components of CVS. Procedurally, CVS includes the ability to create experiments in which conditions differ with respect to only a single contrasting variable, as well as the ability to recognize confounded and unconfounded experiments. It is also important to understand the *logic* of CVS, which involves the ability to make appropriate inferences from the results of unconfounded experiments (e.g., that only inferences about the causal status of the variable being tested are warranted). Final-

ly, understanding CVS means that one is also aware of "the inherent indeterminacy of confounded experiments" (Chen & Klahr, 1999, p. 1098).

Many current science curricula and standards address CVS. The *Next Generation of Science Standards* (NGSS, 2013) mention CVS-related science and engineering practices like designing fair tests and interpreting evidence generated from controlled experiments at every grade level from kindergarten through grade 12. The National Research Council's *Framework for K-12 Science Education* proposes that even kindergarten students should be able to "plan and carry-out investigations […] based on fair tests" (NRC, 2012, p. 5), that middle school students should "identify independent and dependent variables and controls" (p. 55) and that high-school students should "construct and revise explanations based on valid and reliable evidence […] including students' own investigations" (p. 75). Furthermore, CVS is mentioned in the national curriculum of England (Department for Education, 2014), the national science syllabus of Singapore (Curriculum Planning & Development Division, 2007, p. 8), and the Australian curriculum (Australian Curriculum, Assessment and Reporting Authority, see e.g. year 7 science inquiry skills). In Germany, CVS is discussed as a crucial sub-skill of "experimental competence" even though CVS is not explicitly addressed by the German national science standards (Wellnitz et al., 2012). An additional reason for the prominence of CVS in science curricula is the increasing role of inquiry-based learning in science education (Abd-El-Khalick et al., 2004). Mastery of CVS is required for successful inquiry learning as it enables students to conduct their own informative investigations. Furthermore, the logical aspects of CVS are relevant for argumentation and reasoning about causality in science and everyday life, as CVS includes an understanding of the invalidity of evidence from confounded experiments (or observations) and the importance of comparing controlled conditions (Kuhn, 2005a). Taken together, CVS is crucial for developing scientific literacy and is relevant to broader educational and societal goals, such as inquiry, reasoning skills, and critical thinking.

However, without CVS instruction, students (e.g., Croker & Buchanan, 2011; Schauble, 1996) and even adults (Kuhn, 2007) have poor CVS skills (for a review see Zimmerman & Croker, 2013). Siler and Klahr (2012) have identified the various "misconceptions" that students often have about controlling variables. Typical mistakes include (a) designing experiments that vary the wrong (or "non-target") variable, (b) a tendency to vary more than one

variable, or (c) to vary no variable between the contrasted experimental conditions (i.e., to overextend the idea of "fairness" such that the two conditions are identical.

5.2 Previous Research on CVS Instruction

Due to the crucial role of CVS for developing scientific literacy and general reasoning skills, numerous intervention studies have investigated how CVS can be introduced to students. A recent meta-analysis summarizes the findings of 72 CVS intervention studies conducted between 1972 and 2012 (Schwichow, Croker et al., 2015). The analysis found a mean overall effect size of g = 0.61, indicating that CVS instruction can be effective in promoting students' CVS skills.

A detailed analysis considered the possible moderators of the overall effect size. A number of design features (e.g., quasi-experimental vs. experimental studies), instructional features (e.g., use of demonstrations), and assessment features (e.g., test format) were considered and coded for. Schwichow, and Croker et al. (2015) found that interventions that induced a *cognitive conflict* produced significantly greater gains in students' CVS abilities (g = 0.80, 95% CI = 0.62-0.98) compared to interventions not using this technique (g = 0.53, 95% CI = 0.43-0.63). The use of cognitive conflict in educational interventions has roots in Piagetian theory (Limón, 2001; McCormack, 2009) and, in a previous meta-analysis, Ross (1988a) described cognitive conflict as a strategy in which "student conceptions and expectations were overtly challenged to create disequilibrium" (p. 419). In the context of CVS instruction, a teacher draws attention to a particular (confounded) comparison and asks what conclusions can be drawn about the effect of a particular variable. Cognitive conflict is induced in students by drawing attention to a current experimental procedure or interpretation of empirical data in an attempt to get the student to notice that the comparison or conclusion is invalid or indeterminate (Adey & Shayer, 1990; Lawson & Wollman, 1976).

Furthermore, interventions that included a *demonstration* of "good experimental designs" showed greater student achievement (g = 0.69, 95% CI = 0.57-0.81) than interventions without such demonstrations (g = 0.48, 95% CI = 0.32-0.64). Moreover, the meta-analysis found a significant effect of the *test instrument* used to measure students' CVS skills. Studies that utilized multiple-choice instruments to measure student CVS skills had significantly smaller effect sizes compared to studies using either open-response or hands-on assessment instruments. For example, larger effect sizes were evident when student achievement was assessed

with real performance tests (g = 0.74, 95% CI = 0.64-0.84), compared to when either 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) were used (Schwichow, Croker et al., 2015).

As noted previously, Schwichow, and Croker et al. (2015) found a non-significant but interesting trend regarding the impact of *hands-on experience* on CVS skills. Interventions in which students practiced CVS with hands-on experimental tasks (either physical or virtual) had smaller effect sizes (g = 0.59, 95% CI = 0.49-0.69) compared to interventions in which students did not practice CVS with hands-on tasks (g = 0.74, 95% CI = 0.58-0.90). This pattern is surprising for several reasons. First, practicing to-be-learned skills is assumed to be beneficial for acquiring new concepts (Tobin, 1990). Second, hands-on experiences have been promoted as an instructional technique that supports student engagement and motivation and that leads to successful learning (Haury & Rillero, 1994). One reason for this finding might be that hands-on training tasks induce a high *cognitive load*, as these tasks require additional non-CVS-related skills, such as identifying and selecting variables, manipulating and setting up equipment, making measurements, and recording information about the experimental set-up and accompanying measurements in a physical or virtual lab book.

5.3 Cognitive Load and CVS Instruction

The basic idea behind cognitive load theory (CLT) is that the cognitive processing capacity of individuals is limited due to limitations in working memory (Sweller, 1988, 1994). CLT distinguishes between two sources of cognitive load. The *intrinsic* cognitive load originates from the complexity and structure of the concept to be learned and thus cannot be varied by instruction. For example, some concepts are inherently more difficult or complex (e.g., the concept of fractions is more complex than the concept of sums; Sweller & Chandler, 1994). In contrast, the *extraneous* cognitive load originates from the format and manner by which information is presented to the learner and thus can be decreased or increased by instruction (Sweller, 1988). In addition, cognitive load depends on the learners' prior conceptual knowledge and expertise. Learners with expertise in the area of learning might hold automated schemas that they can use to process new information. These schemas can be utilized without reducing the capacity of working memory, as they are stored in long-term memory. As a consequence, expert learners can learn successfully even from instructional materials that induce a high extraneous cognitive load, whereas novice learners might be overstrained by these same materials (Kirschner, Sweller, & Clark, 2006; Sweller, 1994). From the CLT

perspective, instruction should be designed to minimize extraneous cognitive load, particularly when students are novices in the area of instruction.

According to Siler and Klahr's (2012) analysis, to apply the basic logical and procedural aspects of CVS, one must (a) identify which potentially causal variable is being investigated, (b) understand that the two or move levels of the variable being tested must be contrasted in different testing conditions, (c) identify other potentially relevant causal variables and variable levels, (d) ensure that the same level of all possible relevant variables is held constant across different testing conditions. The intrinsic cognitive load of a CVS task at the most basic level includes a minimum of four separate steps. However, CVS tasks can become increasingly complex depending on the content and context and thus induce a high extraneous cognitive load. For example, a CVS task involving selecting or designing a controlled experiment may be more or less challenging depending on a number of contextual factors, such as the strength of prior beliefs or whether the outcome can be characterized as a good or a bad outcome (e.g., Croker & Buchanan, 2011; Schauble, Klopfer, & Raghavan, 1991; Tschirgi, 1980). Furthermore, CVS task are more complex and challenging when more variables are involved (Staver, 1986) or when students have to manipulate experimental equipment (Schwichow, Croker et al., 2015). Accordingly, CLT should be considered when designing CVS instruction, as the cognitive load of CVS tasks can vary depending on factors such as content, prior knowledge, and number of variables.

Hands-on tasks can be challenging because they add extraneous cognitive load. In addition to the logical and procedural considerations of CVS as outlined above, students must, for example, identify the target variable and make sure to contrast levels of the target variable while manipulating physical equipment, then ensure that other variables that have been identified are held constant with respect to the equipment. Additionally, students must identify which outcome variable is to be measured, make those measurements, and then record measurements for both conditions (or, at least try to remember their procedure and their results if notes are not taken). Finally, they must interpret their findings. As a further challenge they have to coordinate all these steps of the experimental process in order to produce meaningful results (Klahr, 2000). Therefore, CVS hands-on tasks require both cognitive and non-cognitive processes that are not related to CVS but necessary for solving hands-on inquiry tasks. As a consequence, the relatively high extraneous cognitive load induced by hands-on

activities might be a reason why interventions utilizing hands-on tasks have thus far not been found to produce greater learning gains relative to interventions that not utilizing hands-on tasks (Schwichow, Croker et al., 2015).

5.4 Research Comparing Hands-on to Alternative Tasks

Insight into the impact of hands-on tasks on students' learning comes from studies that compare hands-on tasks with other types of tasks (e.g., computer simulations). Numerous studies compare hands-on activities and curricula to alternatives such as textbook curricula, lectures, teachers' demonstrations or virtual experimental tasks (Schwichow, Croker et al., 2015). However, most studies contrast learning environments that differ in multiple features and thus do not isolate the unique impact of hands-on tasks on students learning (Smetana & Bell, 2012). Triona and Klahr (2003) note that previous studies examining physical vs. virtual materials often fail to control for instructional method, sequencing, and goals. For example, a study by Kiboss, Ndirangu, and Wekesa (2004) compares students following a student-centered inquiry curriculum with a traditional teacher-centered biology course. Students in both conditions were exposed to different materials but further receive a different amount of teacher guidance.

There are a few exceptions to the claim that intervention studies that compare hands-on and other types of task do not isolate the hands-on component. Two studies compared the impact of students' engagement in hands-on or virtual experimental tasks in the domains of learning physics concepts and CVS (Klahr et al., 2007); (Triona & Klahr, 2003). *Virtual training tasks* may be less complex than hands-on tasks as they do not require the manual skills required for manipulating apparatus or making measurements, and most students are familiar with how to interact with computers (Jones, Ramanau, Cross, & Healing, 2010). As a consequence, the extraneous cognitive load introduced by a virtual task should be lower than for a hands-on task. The intervention study by Triona and Klahr (2003) consisted of brief direct instruction about the concept of CVS, including demonstrations and explanations of "good" and "bad" experiments utilizing either hands-on or virtual materials. In the study by Klahr et al. (2007), students had to discover which variables influence how far a mousetrap car travels by designing either virtual or hands-on experiments. Students did not receive any explicit instruction in CVS or demonstrations of "good" experiments, and what they learned about engineering mousetrap cars was the result of the experimental context (i.e., either virtual or hands-on). In

both studies, the researchers found no differences in CVS skills or conceptual knowledge between students who worked on virtual or hands-on experiments.

Renken and Nunez (2010) were also interested in testing the assumption that first-hand interactions with physical materials are necessary for conceptual change. Across two experiments, adult and adolescent participants either read about the results of an experiment that challenged an existing physics misconception (and confirmed a correct conception), or they were provided with a set of materials in order to conduct the experiment themselves. In both experiments, participants who read about the experiment that challenged their belief had better conclusion accuracy, and were better able to generate their new knowledge to a generalization test.

Taken together, a picture is emerging where we need to question our assumptions about the advantage of hands-on over virtual materials (Triona & Klahr, 2007). Moreover, we also need to reconsider the possibility that even lower technology materials may be just as effective as hands-on materials. Renken and Nunez (2010) suggested that "at least in regard to simple physics experiments, education practitioners have prematurely dismissed non-experimental methods of teaching that resemble the read-only condition in favor of hands-on methods" (p. 807).

5.5 The Current Study

The purpose of this study was to isolate the unique impact of hands-on training on students' learning of CVS and physics content knowledge by contrasting hands-on and paper-and-pencil CVS training tasks. So far, no study has directly compared hands-on with paper-and-pencil training (Schwichow, Croker et al., 2015). Paper-and-pencil tasks have the potential to be a useful alternative to hands-on training tasks as they address the conceptual ideas underlying CVS but without the additional challenges that have been identified, such as manipulating equipment or taking measurements. In contrast to virtual training tasks (e.g., Klahr et al., 2007; Triona and Klahr, (2003) we utilized paper-and-pencil tasks because they do not require access to computers. Moreover, because paper-and-pencil tasks do not require a familiarity with computers, or a novel computer program and interface, they may further reduce extraneous cognitive load. Accordingly, we hypothesized that students who are trained on CVS with paper-and-pencil tasks will outperform students who are trained with hands-on tasks on assessments of both CVS knowledge and physics content knowledge at posttest. We theorize that this effect will be due to the lower extraneous cognitive load of paper-and-pencil tasks.

Additionally, we were interested in determining whether the type of test format used to assess the effects of a training task would reveal the need to match the type of assessment instrument with the type of training, based on the results of a meta-analysis showing that type of assessment instrument moderated study outcomes. However, no systematic attempts have been made to directly examine the type of intervention with the type of assessment. For example, a hands-on assessment has the potential to favor students receiving hands-on training, whereas a paper-and-pencil assessment has the potential to favor students receiving paper-and-pencil training. Therefore, our procedure included multiple assessment formats.

5.6 Method

Participants

The study was conducted with 161 eighth graders (ranging from 12 to 15 years of age; 54% female) from eight science classes at two comprehensive schools in a suburban area in northern Germany. The students from both schools were academically diverse, including students with special educational needs as well as students approaching a university entrance exam. Both schools enroll an equal amount of low-, medium-, and high-achieving students according to their elementary school reports. Specific demographic information regarding individual student ethnicity, socioeconomic status, and school achievement was not collected due to a lack of permission, as such information was considered to have the potential to identify participants and thus result in a loss of confidentiality.

Design

We used a 2 (training condition: hands-on vs. paper-and-pencil) times 3 (test phase: time1/pretest, time2, and time3/posttest) factorial design with training condition as a between-subjects factor and test phase as a within-subjects factor (see Table 5). The training phase was the only time that students were separated and where they received differential instruction. During the training phase, students in both conditions designed and interpreted experiments to answer identical research questions. The only difference between the training conditions is that the hands-on group (n = 82) designed and interpreted experiments with physical equipment, while the paper-and-pencil group (n = 79) designed experiments on paper and interpreted the results of experiments presented to them as photographs. Students within each class-

room were assigned to training conditions based on their CVS pretest scores. That is, we created pairs of students with equal pretest scores; one student from each pair was randomly assigned to one of the training conditions and the other student in the pair was placed in the other training condition. This procedure ensured that students in the two training conditions were equivalent with respect to their pre-intervention CVS skills.

Table 5 Study Design.

Training condition	**Time 1 (Pretest)**	**Introduction Phase**	**Time 2**	**Training Phase**	**Time 3 (Posttest)**
Hands-on (*n* = 82)	▪ Multiple-Choice CVS ▪ Content knowledge ▪ Cognitive abilities ▪ Reading abilities	Cognitive conflict	Multiple-Choice CVS	Hands-on	▪ Multiple-Choice CVS ▪ Hands-on CVS: 1) Elec.magnets 2) Light bulb ▪ Content knowledge ▪ Poster evaluation test: 1) Magnets 2) Memory
Paper-and-pencil (*n* = 79)				Paper-and-pencil	
	First unit (90 min)		Second unit (135 min)		

Note: CVS instructional phases are marked grey. The two training conditions differed only on the nature of the activity during the *Training Phase* (see text and Table 6 for additional details). Order of presentation was counterbalanced for the hands-on CVS tasks and the poster-evaluation tasks.

Procedure

The study took place during the last two weeks of the school year and was organized into two sessions of 90 and 135 minutes, which were spread over two days. The first author and an assistant provided the instruction in all classes, and all activities were carried out identically in all science classes as illustrated in Table 1. During the pretest we measured students' physics content knowledge and their initial CVS skills with a multiple-choice paper-and-pencil instrument. We did not use a CVS hands-on instrument during the pretest to avoid students receiving any CVS-specific hands-on experience prior to instruction. In addition, we measured students' general cognitive and reading abilities as control variables. The CVS instruction consisted of two phases. The CVS *introduction phase*, which employed the use of cognitive conflict, occurred at the end of the first unit and was the same for all participants.

The second unit began with the second administration of the CVS multiple-choice test, which was used to check whether the introduction of CVS via cognitive conflict was effective. For the CVS *training phase*, students were split into their assigned training condition (i.e., either the hands-on or paper-and-pencil training). For 30 minutes, the two groups worked in separate rooms on CVS training tasks. Students worked in pairs or triples (never individually) on an electromagnetism task. In the posttest phase, we measured students' physics content knowledge and their CVS skills with a third alternate form of the multiple-choice instrument. In addition, we measured students' CVS skills on a hands-on task about electromagnetism and a transfer task on light bulb brightness. The transfer task was used to investigate whether any effect of training condition was restricted to the training content (i.e., electromagnetism). Finally, we administered a paper-and-pencil poster evaluation task in which students were asked to identify the confounding variables in two experiments presented as science-fair posters (one on the topic of magnetism and the other on the topic of memory skills of boys and girls).

CVS Instruction

The CVS instruction was divided into an *introduction phase* (at the end of the first unit) and a *training phase* (at the beginning of the second unit; see Table 1). The introduction phase occurred prior to the CVS training, and neither the teacher nor the students knew which training condition students would be assigned to.

Introduction phase: Cognitive conflict. To introduce CVS, we utilized a method to induce cognitive conflict. In their meta-analysis, Schwichow, and Croker et al. (2015) identified cognitive conflict as an effective method to teach CVS. We adopted the procedure developed by Lawson and Wollman (1976), in which everyday objects are used so that no specific prior science knowledge is necessary. The teacher showed students a Ping-Pong ball and an iron ball and asked them to predict which ball would bounce higher. The teacher then dropped the balls from different heights and onto two different surfaces so that -- in contrast to the students' prediction -- the iron ball bounced higher (see Figure 11). In the group discussion that followed, the students argued that the test was unfair and discussed how to turn it in a fair test. After this, the teacher performed a controlled experiment as suggested by the students. The same procedure was repeated by asking the students to predict whether a ball bouncing on an iron surface or on a Styrofoam surface would bounce higher. This method raises a conflict between students' correct predictions (assuming a fair test) and the misleading evidence when the teacher again performed a confounded experiment. To solve this conflict, students need to

consider the idea of "fair tests" or controlled experiments. For the last part of the introduction phase, the teacher performed one additional controlled experiment and summarized why it is important to control variables and how this procedure is done, by reading a standardized explanation.

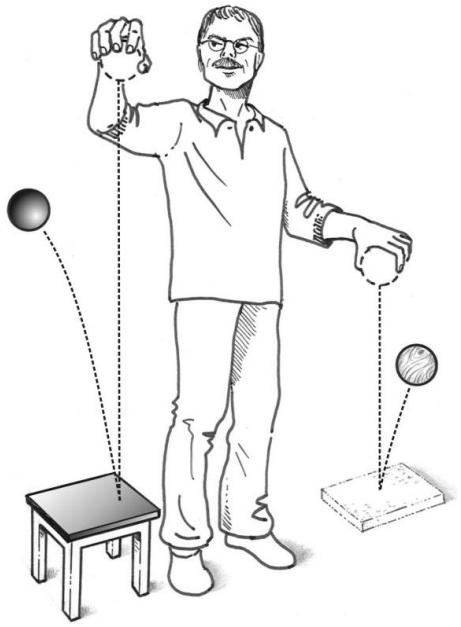


Figure 11 *Illustration of the procedure used to induce cognitive conflict (adapted from Lawson & Wollman, 1976). The teacher begins by asking students if the iron ball (left) or the Ping-Pong ball (right) will bounce higher. Most students predict the Ping-Pong ball will bounce higher. The teacher then conducts a multiply confounded (or "unfair") test by releasing the ball from different heights, and on two different surfaces (iron or Styrofoam) which were at different heights (elevated or on the floor), resulting in the iron ball bouncing higher. Students then discuss ideas about fair and unfair tests.*

Training phase. At the beginning of the training phase, students were grouped into hands-on and paper-and-pencil training conditions and separated in two rooms. This phase was the only time that students received different treatment. Students in both rooms worked in dyads on CVS tasks concerning variables that might impact the force of electromagnets (i.e., number of coils, core material and current). To control the training situation and to minimize teacher-student interaction (and more importantly, differential teacher-student interaction across the two conditions), both groups received all research questions and technical information on worksheets (see Figure 12). Furthermore, two teachers (the first author and a teaching assistant) alternated between coaching the students in the two training conditions to rule out the possibility that systematic group differences between conditions may be caused by teacher effects. Students in both training conditions worked on identical CVS tasks, all within the domain of electromagnetism (see Table 6).

The worksheets were highly similar as they included the same research questions and technical information (see Figure 12). In the first task, students were presented with an experi-

ment, which was multiply confounded. The variable of interest was the effect of the number of coils (N = 500 vs. N = 1000) on the force of electromagnets. Students were asked to identify the confounds in the proposed experiment and to suggest how the experiment could be improved. In the paper-and-pencil condition, students viewed a photo of the experimental setup (Figure 12, top panel). In the hands-on condition, physical equipment was placed in front of each student dyad. In the second and third training tasks, students were asked to plan and interpret experiments on the effects of core material (iron vs. aluminum; task 2) and current (9V vs. 1.5V; task 3) on the electromagnetic force. In both tasks, and for both conditions, students were asked to draw a sketch of the experimental set-up they had planned. No feedback was provided about their sketch. Students in the hands-on group then conducted the experiment they had planned using the physical equipment. After they sketched the setup of their proposed experiment, students in the paper-and-pencil group were given a photograph of a controlled experiment, including the resulting electromagnetic force. Students were asked to interpret their own results (hands-on condition) or the results presented in the photograph (paper-and-pencil condition) and decide whether the hypothesis under consideration was supported or unsupported by the results by selecting the corresponding multiple-choice option (see online supplemental materials for instructional materials).

In summary, the only intended difference between the two training conditions was that the hands-on group had to plan, run and interpret experiments using physical apparatus (i.e., coils, wires, and batteries) whereas the paper-and-pencil group planned identical experiments without the physical apparatus and interpreted experiments presented to them in a photograph. As a result of the extra time taken to construct experiments with apparatus, however, there was a difference in the number of tasks completed by students in the two groups. Students in the hands-on condition completed 2.19 (SD = 0.72) tasks on average while students in the paper-and-pencil condition completed 2.62 (SD = 0.56) of the three tasks, $t(160) = 4.43$, $p < .001$. A further difference between the two groups is that the students in the paper-and-pencil group always interpreted controlled experiments with unique outcomes whereas the quality of the experiments in the hands-on condition depended on how the experiments were set up by the students. However, these differences are caused by the nature of the training tasks and thus could not be controlled in our study design.

PANEL A
Task 1: Lea and Marian want to find out whether the attracting force of an electromagnet depends on the number of times the wire is wrapped around the coil. To investigate, they planned the experiment you see below.

Please look carefully at the experiment. Unfortunately, the experiment of Lea and Marian is not valid. Please write down all the problems you can find.

PANEL B

Task 1: Lea and Marian want to find out whether the attracting force of an electromagnet depends on the number of times the wire is wrapped around the coil. To investigate, they planned the experiment you see in front of you.

[*Note: The real apparatus illustrated in Panel A was present for the students in the hands-on training condition*].

Please look carefully at the experiment but do not touch the equipment yet. Unfortunately, the experiment of Lea and Marian is not valid. Please write down all the problems you can find.

Figure 12 *Worksheet used in the paper-and-pencil training condition (Panel A, top) and the worksheet used in the hands-on training condition (Panel B, bottom). Worksheets are translated from German.*

Table 6 **Sequence of Events in the Two Training Conditions.**

Training Condition	**Task 1** Coils		**Task 2** Core Material		**Task 3** Current	
Hands-on	Note that the present-ed expt. is confounded	Plan/sketch a better expt.	Plan/sketch plus run expt.	Interpret the expt.	Plan/sketch plus run expt.	Interpret the expt.
Paper-and-pencil	Note that the present-ed expt. is confounded	Plan/sketch a better expt.	Plan/sketch expt. only	Interpret a photo of an expt.	Plan/sketch expt. only	Interpret a photo of an expt.

Note: The goal was to plan (or plan and run) experiments (expt.) to test the effect of coils, core material, and current on the force of electromagnets.

Assessment Instruments

In this study we used three different CVS instruments, one cognitive abilities test, one reading abilities test, and a test of physics content knowledge. The tests are described in the same order as they were given to students.

CVS multiple-choice test. A CVS multiple-choice test was administered three times (Table 1). Three different test booklet versions were constructed so that students answered an isomorphic test version each time (test booklets were counterbalanced). All 24 items involve CVS problems in the context of heat and temperature or electricity and electromagnetism, which are two content domains that are part of the state science curriculum for middle schools in northern Germany. There were two item types. For the first type, students had to identify the controlled experiment from a selection of four experiments. For the second type, students had to interpret the outcome of controlled or confounded experiments. Each test booklets consisted of 12 items: six heat and temperature items and six electromagnetism items. Six of the 12 items involved identifying controlled experiments and six items involved the interpretation of controlled or confounded experiments. Each test booklet contained 12 items from a pool of 24 items (versions A and B had 6 items in common, and versions B and C had 6 items in common). All items presented experiments with three independent variables (two levels each) depicted with drawings to minimize the impact of reading abilities (for a detailed description and the complete item set, see Schwichow, Christoph, Boone, & Härtig, 2015).

Physics content knowledge test. The physics content test included the concepts of electric circuits, resistance, electromagnetism, heat, and temperature. The questionnaire consisted of seven multiple-choice and 11 open-response items and was used at pre- and posttest. None of the items required mathematical skills, as the focus was on knowledge and conceptual understanding of relevant physics concepts. One rater coded all open-ended responses, and a second rater coded 20% of the responses. The inter-rater agreement was moderate for three items (κ = .61- κ = .78), and very high for the remaining items (κ = .84 - κ = 1). The physics content test had an internal consistency of α = .77 and .81, at pretest and posttest, respectively.

Tests of cognitive and reading abilities. We measured students' cognitive and reading abilities in order to determine whether any effects of the intervention could be accounted for by individual differences. Students' cognitive abilities were measured by the figural analogies scale of the German-wide established cognitive abilities test for middle school children (Kognitiver Fähigkeitstest [KFT]; Heller & Perleth, 2000). All items from this scale show pairs of figures that are related according to a specific rule that students have to discover. Next, students have to choose one figure from five examples that fits with another figure according to that rule. Students had eight minutes to answer 25 items. A pilot study with a sample comparable to that of our study estimated an internal consistency of α = .86 for the figural analogies sub-scale of the KFT (Skender, 2014, p. 150). Students' reading abilities were measured using an established instrument for middle and high-school students to measure German reading abilities (Lesegeschwindigkeits-und-verständnistest (LGVT); Schneider, Schlagmüller, & Ennemoser, 2007). The instrument consists of a text of 1700 words that students have to read within 4 minutes. The text contains gaps with missing words that students must fill in by choosing the correct term from a selection of three words. The reliability LGVT measures was estimated by Schneider et al. (2007, p. 17) by calculating a test-retest reliability of r = .87. For further analyses, the sum of correct responses for each test was translated into grade and school-type specific *t*-values for each student.

CVS hands-on test. The CVS hands-on test was utilized only at posttest so that students in the paper-and-pencil condition did not receive any hands-on experience prior to the posttest. The test consisted of two tasks -- one task that was identical to the training task on electromagnetism (impact of core material and current on electromagnetic force, see section on CVS training for details) and one transfer task on the brightness of light bulbs. In the light bulb

task, students had to investigate the impact of wire material (constantan [a copper-nickel alloy] vs. copper) and battery type (*1.5V* vs. *9V*) on the brightness of two different light bulbs (*2.1W* vs. *5W*). Both tasks have three independent variables with two values each so that the number of possible experiments is identical in both tasks. In each task students had to design controlled experiments and interpret the outcomes of their experiments in order to test two hypotheses. The four hypotheses were introduced sequentially to students in a booklet with cover stories about fictitious students who want to test a suggested causal relationship. Students had to choose appropriate materials from a box to design proper experiments. One box held electromagnetism components, the other box held light bulb components. After planning the experiments, but before carrying out the experiments, a research assistant took a photograph of their experimental design. After the students conducted the experiment, they had to choose, based on their findings, whether the hypothesis under consideration was supported or unsupported. A task was scored as *correct* when (a) the corresponding photograph showed a controlled experiment, and (b) the student correctly evaluated whether the hypothesis under consideration was supported or unsupported. To control for task order, half the students started with the electromagnetism task while the other half started with the light bulb task. The inter-rater agreement on a double-coded sample of 25% of the experiment photographs (i.e., coding whether the design was controlled or confounded) was high ($\kappa = .83$ - $\kappa = .93$). In addition, we asked for students' confidence in their experimental results, as it might be the case that students in the hands-on group were more confident with their experimental results due to their experiences with conducting experiments during the training phase. Confidence was measured with a four-point Likert-scale ("very uncertain with my conclusion" to "highly confident in my conclusion") for each of the four hypotheses they tested.

Finally, we were interested in students' justification for their experimental procedure. We asked students to write short responses on how they could know for sure whether they found out something about the hypotheses. The justification questions investigate students' deeper understanding of CVS and are similar to the control questions used in Klahr and Nigam (2004)'s interview tasks. Answers were scored as correct when students referred to CVS either by a general statement such as "because everything was the same except variable *x*", or by explicitly mentioning which variables were the same and which variable was different across the contrasted experimental conditions; otherwise, answers were coded as incorrect. The inter-rater agreement on a sub-sample of 30% of double-coded justification questions for

the four justifications was very high (κ = .88 - κ = 1). In summary, the CVS hands-on task provided (a) a design and interpretation measure, (b) a confidence measure and (c) a justification measure for four experimental tasks for each student tested individually.

Poster evaluation test. As a further CVS transfer test we utilized a modified version of the poster evaluation task previously used by Klahr and Nigam (2004) and Strand-Cary and Klahr (2008). In this test, students evaluated the posters of fictitious students participating in a science fair contest. Participants were asked to help these students to make their posters better. We changed the test format from a semi-structured interview to an essay task because we had to administer the test to groups of students. Students were asked a series of questions starting with general questions about changes that should be made to improve the poster, followed by more specific questions about the experimental procedure and potential confounding variables. The questions were taken from the interview manual of Strand-Cary and Klahr (2008). To compute a poster-evaluation score, we counted the number of confounding variables mentioned by students for any sub-question. Each student evaluated two posters: one near-transfer poster about materials that are attracted by magnets and one far-transfer poster about the memory skills of boys and girls. The memory skill poster was taken from Strand-Cary and Klahr (2008) and translated into German. The magnet poster was designed for this study as a near transfer task using concepts and methods similar to those used in the training phase. Both posters presented research questions, a hypothesis, methods, results, conclusions, and contained multiply confounded experiments (each poster had 3 confounded variables). Poster order was counterbalanced. The inter-rater agreement on a sub-sample of 30% of double-coded responses to the poster task was good (κ = .70 - κ = .84).

5.7 Results

Pre-instruction Equivalence of Training Groups

The equivalence of training groups was checked by comparing their mean pretest scores on the content-knowledge test, the cognitive and reading abilities tests, and the CVS multiple-choice instrument. Single independent t-tests for all these scores showed no significant differences between training groups (all *p*-values > .50).

CVS Multiple-Choice Test

The effect of training condition was analyzed using a two-way mixed ANOVA with training condition (hands-on vs. paper-and-pencil) as a between-subjects factor and test phase (time1, time2, time3) as a within-subjects factor. The analysis yielded a significant main effect of test phase, $F(2,318) = 13.50$, $p < .001$, but no main effect of training condition, $F(1,159) = 0.1$, $p = .76$, and no interaction between test phase and training condition $F(2,318) = 2.51$, $p = .08$. A post-hoc Bonferroni test indicates that student measures increased significantly from time 1 to time 2, $p < .001$, with effect size $g = 0.24$, and from time 1 to time 3, $p < .001$, $g = 0.30$. The difference between time 2 and time 3 was non-significant, indicating that the training phase had no impact on the CVS multiple-choice instrument.

Physics Content Knowledge

A two-way mixed ANOVA with training condition (hands-on vs. paper-and-pencil) as a between-subjects factor and test phase (pretest vs. posttest) as a within-subjects factor was used to investigate the effect of training condition on physics content knowledge. There was a significant main effect of test phase, $F(1,159) = 29.82$, $p < .001$, but no main effect of training condition, $F(1,159) = 0.03$, $p = .86$, and no interaction $F(1,159) = 0.18$, $p = .67$. There was a significant difference between pretest ($M = 9.76$; $SD = 2.91$) and posttest ($M = 11.51$; $SD = 3.24$) scores on electricity and electromagnetism items, $t(320) = 5.10$, $p > .001$, but no difference between pretest ($M = 10.58$; $SD = 2.48$) and posttest ($M = 10.53$; $SD = 2.55$) scores on heat and temperature items, $t(320) = 0.18$, $p = .86$. Learning gains were limited to the content domain of the interventions.

CVS Hands-on Tests

Students' skills in designing, conducting, and interpreting controlled experiments using physical equipment were measured by (a) a task identical to the electromagnetism training task and (b) a transfer task on the brightness of light bulbs. Students were asked to test a total of four hypotheses, two for each task. In addition, students rated their confidence in their results, and we evaluated students' justifications for their experimental procedures. Students' ability to apply CVS in physical experiments was analyzed using a two-way mixed ANOVA with training condition (hands-on vs. paper-and-pencil) as a between-subjects factor and experimental task (electromagnetism vs. light bulbs) as a within-subjects factor. There was no main effect of training condition, $F(1,159) = 1.51$, $p = .22$, but there was a main effect of experimental task, $F(1,159) = 74.49$, $p < .001$, and a significant interaction between training condition and experimental task, $F(1,159) = 3.89$, $p < .05$. As seen in Figure 13, performance overall was better on the electromagnetism task, but the group who received hands-on training had a higher level of performance on the electromagnetism tasks than the group who received paper-and-pencil training. The advantage of hands-on training was not evident in the transfer task involving light bulbs. Here, students from both training conditions performed equivalently.

We also applied Strand-Cary and Klahr's (2008) "expert" criterion to compare training conditions. Students were classified as "CVS experts" when they carried out four controlled experiments of four possible experiments. After the CVS training, 10% of the hands-on students and 5% of the paper-and-pencil students were classified as CVS experts. The difference between groups was not significant, $\chi^2(1, N = 161) = 1.80$, $p = .18$.

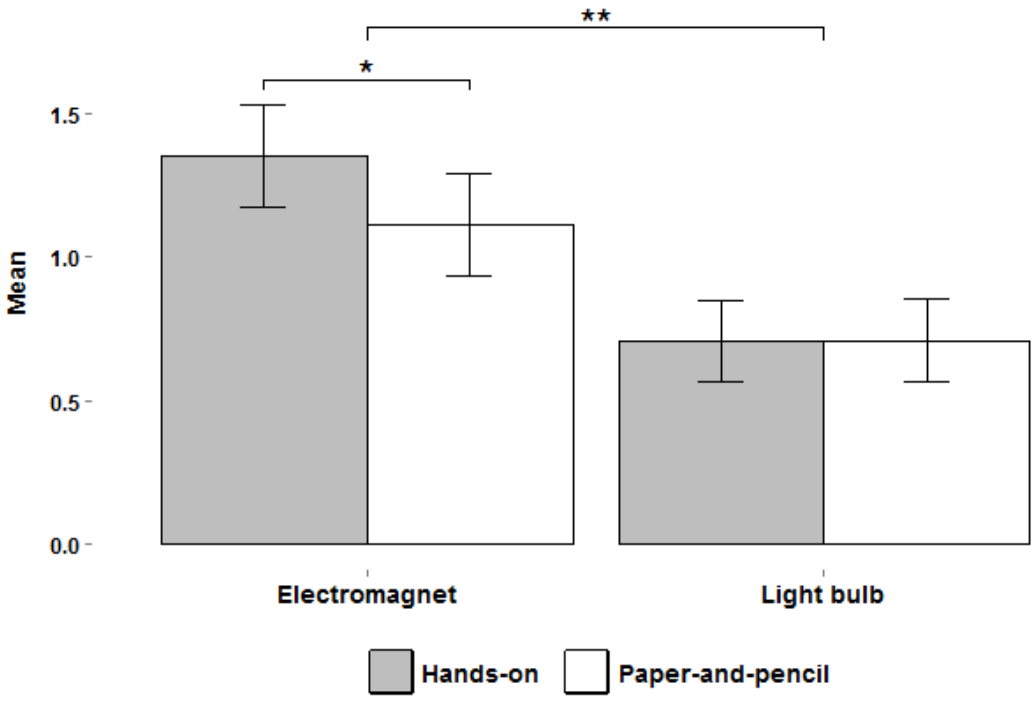


Figure 13 *Mean scores and standard errors on the CVS hands-on tests. Scores could range from 0 to 2 for each of the two tests (order was counterbalanced).*

After carrying out the experiments and answering the multiple-choice decision question (hypothesis is supported/unsupported), students answered an open-ended justification question. Students' responses were coded as correct or incorrect (i.e., one point for each correct justification for a maximum score of 4). The analyses showed no main effect of training condition, $F(1,159) = 0.10$, $p = .75$, or experimental task, $F(1,159) = 0.19$, $p = .66$, and no interaction, $F(1,159) = 1.74$, $p = .19$, indicating that training condition had no impact on students' abilities to justify their experimental procedure by mentioning CVS. On average, students correctly justified 0.99 ($SD = 1.51$) of their four hands-on experimental designs. We also examined whether training condition had an effect on students' confidence in their experimental results. There was no main effect of training condition $F(1,159) = 1.84$, $p = .18$, a significant main effect of experimental task, $F(1,159) = 14.20$, $p < .001$, but no interaction, $F(1,159) = 1.06$, $p = .30$. Training condition had no influence on students' confidence, but students were more confident in the results from their experiments with the light bulb task ($M = 2.82$; $SD = 1.27$) than from the electromagnetism task ($M = 2.47$; $SD = 1.31$; $d = 0.27$).

Poster Evaluation Task

The poster evaluation task was completed by 109 students from six intervention classes. Two classes (n = 52 students) did not complete the poster task due to logistical problems. Because students from each class were assigned to each training condition we were still able to analyze data from the subset of participants who completed the poster evaluation task. A two-way mixed ANOVA with training condition (hands-on vs. paper-and-pencil) as a between-subjects factor and poster topic (magnets vs. memory) as a within-subjects factor was used to analyze the poster evaluation task score (i.e., the number of confounding variables noted, ranging from 0 to 3 for each poster). There was a significant effect of poster topic on the number of recognized confounding variables, $F(1,107) = 84.91$, $p < .001$. Students found more confounds on the memory poster (M = 1.41; SD = 1.05) than on the magnet poster (M = 0.54; SD = 0.69; d = 0.98). There was also a main effect of training condition $F(1,107) = 8.54$, $p < .01$. Students in the paper-and-pencil condition (M = 2.37; SD = 1.53) outperformed students in the hands-on condition (M = 1.56; SD = 1.34; d = 0.56). The interaction between training condition and poster task was not statistically significant, $F(1,107) = 1.44$, $p = .29$ (see Figure 14*Figure 13*).

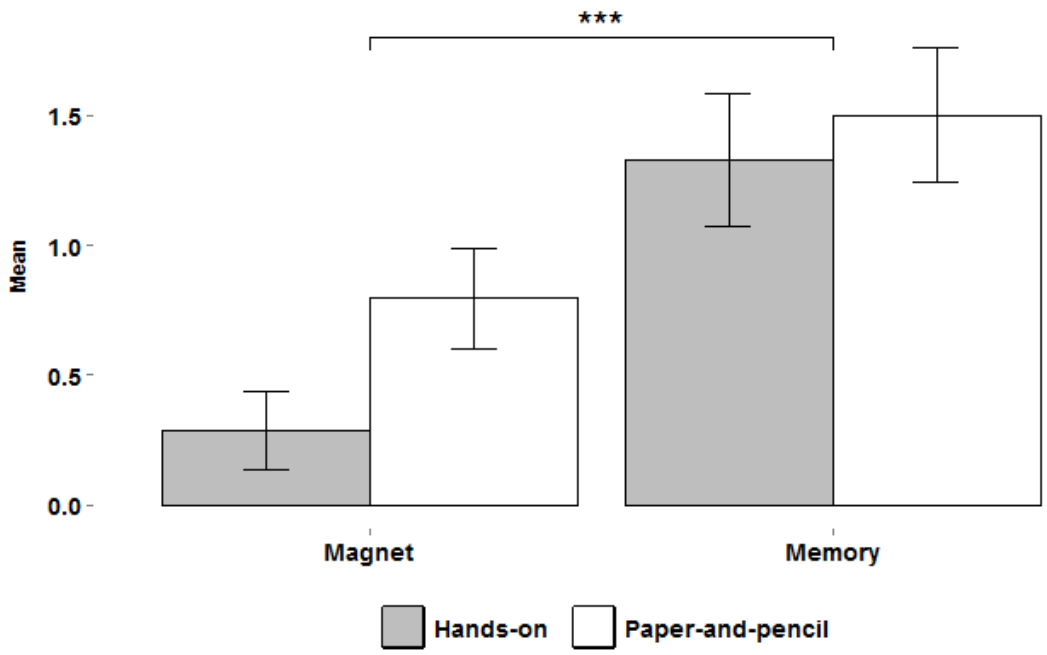


Figure 14 *Mean number of recognized confounding variables and standard errors in the poster evaluation test. Maximum score = 3 for each poster topic.*

5.8 Discussion

The aim of this study was to isolate the unique impact of hands-on experience on students' CVS achievement by contrasting the effect of hands-on and paper-and-pencil training tasks. Our hypothesis was that students who were trained in CVS with paper-and-pencil tasks would outperform students who were trained with hands-on tasks. This hypothesis was informed by cognitive load theory (CLT). Conceptually and procedurally, CVS requires at minimum four steps (Siler & Klahr, 2012) that constitute an already moderate baseline cognitive load. We reasoned that hands-on tasks, which require extra steps (e.g., manipulate equipment, make and record measurements), would add extraneous cognitive load, and would therefore be potentially less efficient for training CVS. In contrast to our hypothesis, we found no general advantage of paper-and-pencil training tasks over hands-on training tasks. Instead, we found a differential effect as a function of the type of assessment that was administered. Students in the hands-on condition outperformed students in the paper-and-pencil condition on the CVS hands-on test, whereas the students in the paper-and-pencil condition outperformed students in the hands-on condition on the poster evaluation test. However, this training advantage was only evident for items that were identical (in the case of the hands-on test) or very similar (in the case of the poster evaluation test) to the training task. On the transfer items (light bulb hands-on task and memory poster evaluation) students in both training conditions performed equally well. Interestingly, task effects were also evident. Across the board, students performed better on the hands-on electromagnet task relative to the hands-on light bulb (transfer) task, which makes sense given that electromagnetism was the content domain of the CVS training. Surprisingly, both groups had better performance in finding confounds when evaluating the science fair poster about memory (i.e., the transfer task) relative to the poster about magnets.

Also contrary to our hypothesis, the two training conditions had equivalent performance on the CVS multiple-choice test and the content knowledge test. However, this finding is in line with previous studies that found no difference between hands-on and virtual experimental tasks (Klahr et al., 2007; Triona & Klahr, 2003). Thus, the effect of training condition (hands-on vs. paper-and-pencil) was restricted to test instruments that are very similar to the training tasks. This finding is consistent with results from a meta-analysis showing that the type of test format used to assess an intervention moderated study outcomes (Schwichow, Croker et al., 2015). We utilized multiple test instruments, including those that were identical or similar to

the tasks utilized in both training conditions so that we were able to detect the differential effect of the task type. The CVS multiple-choice test results did show that our initial lessons, using the cognitive conflict procedure, did have an effect. However, this assessment was not sensitive enough to detect any additional learning gains provided by the CVS training.

Based on our findings, we argue that in the short run students learn particular item-specific procedural knowledge when engaged in hands-on or paper-and-pencil training tasks. We found no evidence that one training condition was superior for developing a deeper conceptual understanding of CVS, as the differences between training conditions were restricted to tasks very similar to the training tasks. Rather, students' test performance seems to depend on their direct experience with a task. Although these experiences with highly similar tasks might reduce students' cognitive load when they worked on corresponding test items, it is not likely that this effect reflects a pure memory effect because students did not receive any feedback or guidance on "good experimental design" during the training. Moreover, the observed effect does not seem to reflect differences in students' confidence in their task-specific skills, as we found no differences between students' confidence in their experimental findings. Instead, this effect might be similar to what Siler, Klahr, and Price (2013) found when they engaged students in CVS tasks prior to CVS training in order to prepare them for future learning. Students might develop an intuitive understanding of the materials and the nature of the task during training even when they do not receive feedback or guidance. This intuitive knowledge might reduce students' cognitive load at the assessment phase and result in better performance on the corresponding assessment tasks.

Our findings add evidence to the debate about the role of active manipulation in instruction. The argument that manipulation has to be physical or hands-on to be effective has been challenged by studies that contrasted virtual and hands-on tasks (Triona & Klahr, 2003, 2007). We found no evidence that students trained with hands-on or paper-and-pencil tasks performed better on transfer items, so the type of task used for training the procedural and logical aspects of CVS may be irrelevant for a deep understanding of CVS. Based on the findings of this study we conclude that it is not the physical or virtual manipulation that is crucial for effective learning but rather the "cognitive manipulation" (i.e., thinking about manipulating variables). Accordingly, thinking about manipulating variables and not the manipulation itself seem to be crucial for learning CVS. To induce a cognitive manipulation of variables, a teacher can uti-

lize hands-on, virtual, or paper-and-pencil tasks as long as the task requires thinking about the manipulation of variables and the consequences of those manipulation. As seen with the memory skills poster task, students from both training conditions were able to consider the confounding variables in a very different domain.

Limitations

In this study, our hands-on task functioned as a training task, rather than as a teaching demonstration or as a discovery-learning task. However, the nature of the training tasks caused additional undesired differences that may have contributed to the observed group differences. First, because the hands-on tasks were more time consuming, students completed fewer training tasks in the hands-on condition. However, the time on task was identical in both conditions, thus although hands-on students worked on fewer tasks, they spent more time on each individual task. Second, although it could be argued that our training conditions differed with respect to the amount of reading skill required, both tasks required students to follow written instructions in order to control for the amount of teacher-student interaction across training conditions). Moreover, the groups were equivalent with respect to reading scores.

An unanticipated difference between our training conditions resulted from the procedure where students in the paper-and-pencil condition interpreted experiments presented in a photo while hands-on students interpreted experiments that they designed. The photos in the paper-and-pencil condition showed unique and controlled experimental outcomes. In contrast, the quality of the experiments interpreted by hands-on students depended on their CVS and manual skills. We do not think it likely that these differences had an undue influence on our findings because (a) hands-on students outperformed paper-and-pencil students on the hands-on test involving electromagnetism, and (b) the poster evaluation test required the identification of confounding variables in the presented experiments and not the interpretation of controlled experiments. A further limitation of this study is that we tested only a restricted range of relevant outcome variables. For example, we did not test students' understanding of the nature of science, students' motivation or the long-term effects of the types of CVS training. As these variables are important goals of science education, our conclusions are clearly limited to the learning of CVS and content knowledge.

5.9 Implications for Instruction and Assessment

Our results are similar to the findings of other studies (e.g., Klahr et al., 2007; Triona & Klahr, 2003) and a meta-analysis (Smetana & Bell, 2012) comparing hands-on and virtual tasks. Similar to these studies, we found no general advantage of students' engagement in hands-on training. This finding neither supports nor refutes the utility of hands-on activities in science education. Rather, it means that the task itself is more important than the materials that students interact with. That is, curriculum developers and teachers should pay close attention to task design. According to the findings of Smetana and Bell (2012), "good" learning tasks encourage student reflection and promote cognitive dissonance.

The observed differential effect of training conditions (or, lack of difference between training conditions) on our four different assessments shows that performance on inquiry tasks strongly depends on context and item-specific features. Although CVS is often cast as a domain-general process skill, it is very rarely used as such. It is clear from the existing literature and from our study that people are rarely able to successfully deploy CVS across a wide range of contexts. We include professional scientists here as we have been on both the giving and receiving end of criticism concerning failure to adequately control variables when the study concerns a topic outside the researchers' field of expertise. In some cases, it is a challenge for students to transfer their training experience to new tasks; in other contexts, the effect of CVS training is more evident. Thus, CVS instruction should utilize training tasks that require students to practice their developing CVS skills on a variety of tasks so they can benefit from what they have learned and flexibly apply it in multiple contexts (and on multiple types of inquiry tasks that may be used as learning assessments). However, it is an open question how students can be supported to develop more general non-task-specific CVS skills. It might be that the utilization of multiple training tasks induces the development of more general CVS skills.

Our findings have obvious implications for the assessment of inquiry skills. As students' performance on inquiry tasks depend on context-specific features, assessment instruments should utilize contexts that are familiar to students in order to be fair. Because we used four different assessment instruments, our conclusions actually vary as a function of assessment type. If we consider the findings from any one assessment, our conclusions would have been quite different. We may have concluded that either of the training tasks were superior or that there were

no differences. Because we used four assessments, our conclusions are much more nuanced. As indicated by the meta-analysis (Schwichow, Croker et al., 2015), it is important to be aware that our assessment instruments are just as important as the educational interventions that we develop to improve teaching and learning.

6. Förderung der Variablen-Kontroll-Strategie im Physikunterricht

Damit Schülerinnen und Schüler lernen selbstständig zu experimentieren, sollten im Unterricht Strategien der experimentellen Erkenntnisgewinnung thematisiert werden. Eine grundlegende Strategie der experimentellen Erkenntnisgewinnung ist die Variablen-Kontroll-Strategie (VKS). Der folgende Beitrag stellt ein VKS-Übungsexperiment zum Thema Leitfähigkeit sowie Merkmale von VKS-Übungsexperimenten und Befunde einer unterrichtlichen Erprobung in einer achten Jahrgangsstufe vor.

6.1 Einleitung

Spätestens seit Einführung der bundesweit gültigen Bildungsstandards für den naturwissenschaftlichen Unterricht ist die Vermittlung von Kompetenzen zur selbstständigen Erkenntnisgewinnung durch Experimentieren ein zentrales Ziel von Physikunterricht (KMK, 2005c). Eine Förderung entsprechender Kompetenzen bedarf spezifischer Lerngelegenheiten. In der Unterrichtspraxis dominiert bislang der Einsatz von Experimenten, die primär die Vermittlung von Fachwissen zum Ziel haben. Ein Kennzeichen dieser Experimente ist, dass die Schülerinnen und Schüler eine detaillierte Experimentieranleitung abarbeiten und nicht selbstständig Experimente planen, durchführen und auswerten (Hofstein & Lunetta, 2004). Solche Experimente vermitteln einerseits ein verzerrtes Bild der naturwissenschaftlichen Erkenntnisgewinnung, andererseits schränken sie die Selbsttätigkeit und damit auch die Lerngelegenheiten der Schülerinnen und Schüler stark ein, so dass eine positive Wirkung auf die experimentelle Kompetenz kaum zu erwarten ist (Kircher, Girwidz, & Häußler, 2009). Zur Förderung experimenteller Kompetenz werden daher Experimente benötigt, die den Schülerinnen und Schülern Entscheidungsmöglichkeiten bei der Planung, Durchführung und Auswertung von Experimenten geben. In diesem Beitrag wird ein solches Schülerexperiment vorgestellt, anhand dessen das Durchführen kontrollierter Experimente gelernt bzw. geübt werden kann.

6.2 Die Variablen-Kontroll-Strategie

Um möglichst eindeutige Aussagen über Ursache-Wirkungsbeziehungen zu erlangen, sollte in Experimenten nur eine Variable verändert, die Auswirkungen auf die potentielle Wirkung beobachte und sämtliche weiteren Variablen konstant gehalten werden. Das Vergleichen kontrollierter Bedingungen ist ein Wesensmerkmal wissenschaftlicher Experimente (Weizsäcker,

1951). Wird z. B. der Einfluss der Pendellänge auf die Schwingungsdauer eines Pendels untersucht, so sollten die verglichenen Pendel sich ausschließlich in ihrer Länge und in sonst keiner Variablen unterscheiden. Würden sich die Pendel zusätzlich z. B. in der Masse unterscheiden, wäre es nicht möglich zuzuordnen, ob unterschiedliche Schwingungsdauern durch die Fadenlänge oder die Pendelmasse verursacht werden (siehe Figure 15).

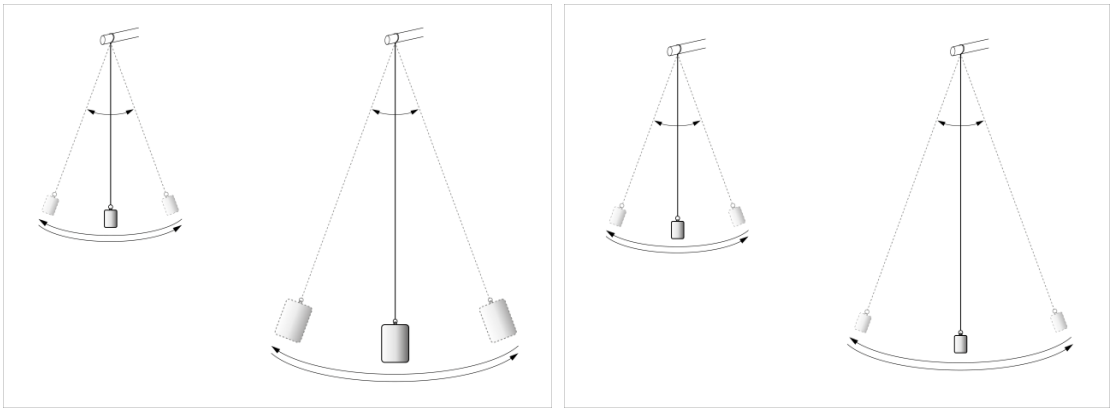


Figure 15 *Beispiel für ein unkontrolliertes (links) und kontrolliertes Experiment (rechts) zum Einfluss der Pendellänge auf die Schwingungsdauer (nach Inhelder & Piaget, 1958, S.68).*

Dieses grundlegende Prinzip der experimentellen Erkenntnisgewinnung wird als Variablen-Kontroll-Strategie (VKS) bezeichnet. Die VKS ist nicht nur bei der Planung, sondern auch bei der Durchführung und Auswertung von Experimenten von Bedeutung. So sollte bei der Durchführung der Experimente auf die Einhaltung der VKS geachtet werden. Bei der Auswertung von Experimenten ist die Suche nach Variablen, die möglicherweise zusätzlich zu der untersuchten Variable variiert wurden (konfundierende Variablen) eine wichtige Qualitätskontrolle. Aufgrund der zentralen Rolle der VKS in der wissenschaftlichen Erkenntnisgewinnung sollte sie ebenfalls ein eigenständiges Lernziel von Physikunterricht sein.

Ohne explizite Vermittlung der VKS sind vor allem jüngere und leistungsschwächere Schülerinnen und Schüler häufig nicht in der Lage, kontrollierte Experimente zu planen, durchzuführen und auszuwerten. Typische Fehler, die Schülerinnen und Schüler beim Experimentieren machen sind, dass sie (1) mehr als eine Variable, (2) keine Variable oder (3) die falsche Variable zwischen zwei kontrastierten Experimenten verändern (Siler & Klahr, 2012). Befunde von Interventionsstudien zeigen, dass Schülerinnen und Schüler, die in der VKS unterrichtet wurden, anschließend signifikant häufiger kontrollierte Experimente planen und durchführen (siehe Publikation 1). Zur Förderung experimenteller Kompetenz sollte die VKS daher explizit im Unterricht thematisiert und ihre Anwendung geübt werden. Dem Üben kommt dabei

ein besonderer Stellenwert zu, da es für eine langfristige Verfügbarkeit von Lerninhalten sowie für ein erfolgreiches Anwenden von Strategien erforderlich ist (Hänsel, 2014). Experimente, die ein sinnvolles Üben der VKS ermöglichen, unterscheiden sich allerdings von Schülerexperimenten, die eine Vermittlung von Fachwissen zum Ziel haben.

6.3 Merkmale von VKS Übungsexperimenten

In Schülerexperimenten, die zur Vermittlung von Fachwissen konzipiert werden, erhalten die Schülerinnen und Schüler häufig Versuchsmaterial, das bereits von der Lehrkraft auf das zur Durchführung eines kontrollierten Experiments notwendige Material beschränkt wurde. Die Anzahl an möglichen Versuchsaufbauten ist bei solchen Experimenten folglich gering, weil der Erwerb experimenteller Kompetenzen nicht das intendierte Lernziel ist. Die Schülerinnen und Schüler müssen bei traditionellen Schülerexperimenten nur selten Variablen auswählen bzw. nicht entscheiden, welche Versuchsbedingungen zu vergleichen sind, um aussagekräftige Ergebnisse zu bekommen (siehe z. B. Phywe o. J.). Eine Beschränkung der Versuchsmaterialien mag zur Reduktion der Aufgabenkomplexität sinnvoll sein, wenn ein Schülerexperiment vor allem mit dem Ziel der Vermittlung von Fachwissen eingesetzt wird. Allerdings sind solche Experimente als Übungsexperimente für die VKS ungeeignet, da die Schülerinnen und Schüler nicht entscheiden können, welche Versuchsbedingungen und welche Variablen bzw. Variablenausprägungen sie vergleichen wollen.

Im Unterschied zu Schülerexperimenten mit dem Fokus Fachwissen müssen bei VKS Übungsexperimenten den Schülern/innen daher mehr Materialien zur Verfügung gestellt werden, als sie für die Durchführung eines geeigneten Experiments benötigen würden. Dadurch wird gewährleistet, dass die Schülerinnen und Schüler eigenständig Variablen und Variablenausprägungen auswählen. Die Schwierigkeit von VKS Übungsexperimenten kann durch Variation der Anzahl an Variablen und Variablenausprägungen variiert werden. Je geringer die Kombinationsmöglichkeiten aus Variablen und Variablenausprägung sind, desto geringer wird die Anzahl an Experimenten, die mit den gegebenen Materialien durchgeführt werden können und desto leichter werden die Übungsexperimente (Staver, 1986).

Da der Fokus dieser Übungsexperimente ausschließlich auf der VKS liegt, sollten andere schwierigkeitserzeugende Merkmale möglichst reduziert werden. So sollten z. B. die Experimentiermaterialien so gewählt werden, dass sämtliche kontrollierten Experimente möglichst eindeutige Ergebnisse liefern. Andernfalls hängt es von der zufälligen Auswahl von Variab-

lenausprägungen bzw. dem Fachwissen der Schülerinnen und Schüler ab, ob sie eindeutige Ergebnisse erhalten. Zwar sind die Ergebnisse von wissenschaftlichen und offenen schulischen Experimenten häufig nicht eindeutig, doch dieser Aspekt der naturwissenschaftlichen Erkenntnisgewinnung soll mit VKS Übungsexperimenten nicht thematisiert werden.

6.4 Ein VKS Übungsexperiment zum Thema Leitfähigkeit

Die vorgestellten Merkmale von VKS Übungsexperimenten wurden bei der Entwicklung eines entsprechenden Schülerexperiments zum Thema Leitfähigkeit berücksichtigt. Das Thema Leitfähigkeit eignet sich im Inhaltsbereich Elektrizitätslehre zum Üben der VKS, da der Widerstand eines Leiters u. a. von den drei zu variierenden Variablen Leiterlänge, Leiterquerschnittsfläche und Leitermaterial abhängt. Folglich können die Schülerinnen und Schüler eine wirkliche Variablenkontrolle vornehmen, wenn sie den Einfluss dieser Variablen auf den Widerstand eines Leiters untersuchen. Außerdem kann durch geschickte Wahl des Leitermaterials gewährleistet werden, dass sämtliche kontrollierten Experimente auch eindeutige Ergebnisse liefern. Ein weiterer Vorteil des Themas Leitfähigkeit ist, dass es ein fester Bestandteil von Curricula für die Sekundarstufe I in zahlreichen Bundesländern ist, so dass die vorgestellten Experimente leicht in den Unterricht zu integrieren sind. Aufgrund der relativ hohen Komplexität der Experimente (es sind drei Variablen unabhängig voneinander zu untersuchen) sollten die Schülerinnen und Schüler vor dem Experimentieren allerdings eine Einführung in die VKS erhalten.

Der spezifische Widerstand eines Leiters kann durch die Formel $R = \rho \cdot {}^{l}/_{A}$ beschrieben werden. Dabei entspricht ρ dem materialabhängigen spezifischen Widerstand des Leiters, l der Leiterlänge und A dem Flächeninhalt des Leiterquerschnitts. Als Leitermaterialien wurden Konstantan (ρ = 0,5 Ω·mm²/m) und Eisen (ρ = 0,1 Ω·mm²/m) gewählt, da bei diesen Materialien schon relativ kleine Längen- bzw. Querschnittsänderungen aufgrund ihres relativ großen spezifischen Widerstands einen messbaren Effekt auf den Widerstand des Leiters haben. Bei Kupfer sind die Effekte aufgrund des kleineren spezifischen Widerstands (ρ = 0,017 Ω·mm²/m) deutlich geringer. Für jedes Material wurden Leiter in sämtlichen Kombinationen dreier Längen (15 cm, 30 cm, 45 cm) und zweier Querschnitte (0,2 mm², 0,4 mm²) angefertigt. Den Schülern/innen wurden folglich zwei (Materialien) mal drei (Längen) mal zwei (Querschnitte) d. h. 12 unterschiedliche Leiter zur Verfügung gestellt (siehe Table 7).

Table 7 Übersicht über die im Experiment zur Verfügung gestellten Leiter.

		Länge		
		15 cm	**30 cm**	**45 cm**
Querschnitts-fläche	**0.2 mm²**	Eisen / Konstantan	Eisen / Konstantan	Eisen / Konstantan
	0.4 mm²	Eisen / Konstantan	Eisen / Konstantan	Eisen / Konstantan

Unter der Annahme, dass in einem Experiment mindestens zwei unterschiedliche Leiter verglichen werden, könnten mit den Experimentiermaterialien insgesamt N = 4.083[1] verschiedene Experimente durchgeführt werden. Von diesen theoretisch möglichen Vergleichen stellen jedoch nur 16 Vergleiche (bei der Frage nach dem Einfluss der Länge) bzw. 6 Vergleiche (bei der Frage nach dem Einfluss des Leitermaterials bzw. des Querschnitts) ein kontrolliertes Experiment dar. Dieses Beispiel zeigt, wie groß und unübersichtlich die Anzahl theoretisch möglicher Experimente schon bei drei Variablen ist, und verdeutlicht, dass die Auswahl geeigneter Vergleiche keineswegs trivial ist. Erst die Beherrschung der VKS ermöglicht es den Schülern/innen, aus der großen Auswahl an möglichen Vergleichen diejenigen Vergleiche auszuwählen, die kontrollierte Experimente darstellen auf deren Grundlage sinnvolle Schlüsse gezogen werden können.

Für den Einsatz im Unterricht wurden die 12 Leiter (siehe Tabelle 1) zu einer Experimentierbox zusammengestellt (siehe Figure 16). Jede Form der Vorsortierung der Leiter wurde vermieden, da das Sortieren der Leiter nach ihren Eigenschaften (sprich nach den relevanten Variablen) ein wichtiger Teilschritt bei der Planung eines Experiments ist. Der Leiterwiderstand soll von den Schülern/innen mit einer einfachen Strom-Spannungsmessung bestimmt werden, so dass neben der Experimentierbox jeweils eine Spannungsquelle, ein Amperemeter, ein Voltmeter und ein Taschenrechner erforderlich sind. Mit den beschriebenen Materialien können Experimente zum Einfluss der drei Variablen Leiterlänge, -querschnitt und -material auf den Widerstand des Leiters durchgeführt werden. Der Messbereich des Amperemeters sollte zwischen den Messungen nicht variiert werden, da ein veränderter Innenwiderstand des Amperemeters die Spannungs- und Strommessung beeinflusst. Ferner ist die Spannung nur über

[1] Die Anzahl möglicher Vergleiche N berechnet sich wie folgt: $N = \sum_{r=2}^{n} \frac{n!}{(n-r)!\cdot r!}$. Dabei entspricht n dem Umfang des größtmöglichen Vergleichs (Bei diesem Beispiel ist n = 2 Materialien x 3 Längen x 2 Querschnitte = 12).

dem Leiter (spannungsrichtige Widerstandsmessung) und nicht über der Spannungsquelle oder über dem Leiter und dem Amperemeter zu messen. Der Stromfluss durch das Voltmeter ist aufgrund des großen Innenwiderstands des Voltmeters im Vergleich zum Widerstand des Leiters genauso wie der Einfluss der Kontaktstecker zu vernachlässigen. Bei der Durchführung der Experimente sollte darauf geachtet werden, dass die Strom- und Spannungsmessung relativ zügig vorgenommen werden, da die Schaltung einen Kurzschluss darstellt und es zu einer stärkeren Erwärmung der Leiter kommen kann. Aus diesem Grund sollten die Schülerinnen und Schüler auch darauf hingewiesen werden, den Leiter nicht zu berühren und keine Spannungen oberhalb von 2 V zu wählen. Die Informationen zum Messvorgang sollten den Schülern/innen explizit mitgeteilt werden, um ein erfolgreiches und sicheres Messen zu gewährleisten. Im Gegensatz zu Experimenten mit dem Fokus auf Fachwissen erhalten die Schülerinnen und Schüler jedoch keine detaillierte Anleitung, wie kontrollierte Experimente zu den gegebenen Forschungsfragen durchzuführen sind. Die Schülerinnen und Schüler müssen somit selbstständig entscheiden welche Variablenausprägungen sie wählen und vergleichen.

Figure 16 *Übersicht über die verwendeten Experimentiermaterialien.*

Zur Anleitung und als Hilfestellung für die Schülerinnen und Schüler wurde ein Arbeitsblatt (siehe Anhang Publikation 4) entwickelt. Auf dem Arbeitsblatt werden zunächst das fachliche Ziel des Experiments sowie das Vorgehen bei der Planung kontrollierter Experimente und bei der Bestimmung von Widerständen wiederholt. Anschließend erhalten die Schülerinnen und Schüler jeweils ein vorgefertigtes Protokoll (siehe Anhang Publikation 4) für die Untersuchung der drei Variablen Leiterlänge, -querschnitt und -material. Neben den entsprechenden inhaltlichen „Forschungsfragen“ enthalten die Protokolle Dokumentationstabellen, Platz für eine graphische Auswertung, vorgefertigte Antwortsätze und eine Reflexionsfrage. Durch die Reflexionsfrage sollen die Schülerinnen und Schüler zum Nachdenken über ihr experimentelles Vorgehen angeregt werden und auf die Bedeutung der Variablenkontrolle für die Güte der

experimentellen Befunde eingehen. Zwar wird auf dem Arbeitsblatt ein allgemeiner Hinweis zum adäquaten Vorgehen beim Experimentieren gegeben, jedoch enthalten weder die Protokolle noch das Arbeitsblatt eine konkrete Anleitung (Rezept), wie ein kontrolliertes Experiment zu den entsprechenden Forschungsfragen durchzuführen ist. Die relativ geschlossene Protokollform mit vorgegebenen Antwortsätzen wurde gewählt, da die Komplexität der Experimente durch die offene Aufgabenstellung und die große Anzahl möglicher Experimente bereits relativ groß ist.

6.5 Unterrichtliche Erprobung der Übungsexperimente

Die VKS Übungsexperimente wurden im Rahmen einer Doppelstunde mit N = 46 Schülern/innen der achten Jahrgangsstufe eines Gymnasiums erprobt. Im vorangegangenen Unterricht wurden das Messen von Stromstärke und Spannung und der Ohm'sche Widerstand bereits behandelt, so dass die Schülerinnen und Schüler das notwendige Fachwissen zur Bestimmung von Widerständen besaßen. Zu Beginn der Doppelstunde erfolgte ein kurzer Unterrichtseinstieg mittels eines gelenkten Unterrichtsgesprächs, in dem die Merkmale kontrollierter (aussagekräftiger) Experimente besprochen wurden. Die verbleibenden etwa 65 Minuten haben Schülerinnen und Schüler in Kleingruppen von bis zu vier Personen experimentiert und die Arbeitsblätter bearbeitet. Am Ende der Doppelstunde wurden die Arbeitsblätter zur Evaluation der entwickelten Übungsexperimente eingesammelt. Ziel der Evaluation ist es, herauszufinden, inwiefern Schülerinnen und Schüler in der Lage sind, aus einer Vielzahl von möglichen Experimenten geeignete (kontrollierte) Experimente auszuwählen. Es wurde kein Prä-Post Vergleich durchgeführt, da es ausschließlich um eine Erprobung des Übungsmaterials ging.

Zu diesem Zweck wurde bei der Auswertung der Arbeitsblätter zunächst untersucht, ob und wie viele kontrollierte und unkontrollierte Experimente (Vergleiche) die Schülerinnen und Schüler in ihren Protokollen aufgezeichnet haben. Als ein kontrolliertes Experiment wurde ein Vergleich von mindestens zwei Teilexperimenten gewertet, die sich nur in der untersuchten Variable unterscheiden. Bei der Frage nach dem Einfluss der Leiterlänge auf den Widerstand stellen z. B. nur solche Vergleiche ein kontrolliertes Experiment dar, die Leiter unterschiedlicher Länge aber gleichen Querschnitts und Materials vergleichen. Darüber hinaus wurde bewertet, ob die Schülerinnen und Schüler fachlich korrekte Schlussfolgerungen aus den Ergebnissen gezogen haben und ob sie bei der Beantwortung der Reflexionsfrage auf die VKS eingegangen sind. Die Ergebnisse sind in den Abbildungen 17-19 (Figure 17-19) darge-

stellt. Es zeigt sich, dass die meisten Schülerinnen und Schüler kontrollierte Experimente durchgeführt haben und nur wenige Schülerexperimente unkontrolliert sind. Die Wahrscheinlichkeit, dass Schülerinnen und Schüler aus der Vielzahl von möglichen Experimenten zufällig ein kontrolliertes Experiment auswählen, liegt bei nur 0,1 % für die Fragen zum Leitermaterial und -querschnitt bzw. 0,3 % für die Frage zur Leiterlänge. Folglich scheinen die meisten Schülerinnen und Schüler bewusst kontrollierte Experimente durchgeführt zu haben. Des Weiteren hat die überwiegende Mehrheit der Schülerinnen und Schüler auch die entsprechende fachlich korrekte Schlussfolgerung gezogen. Bei der Beantwortung der Reflexionsfrage hingegen haben nur wenige Schülerinnen und Schüler entweder explizit benannt welche Variablen sie konstant gehalten haben bzw. argumentiert, dass ihre Befunde gültig sind, da sie auf kontrollierten Experimenten beruhen.

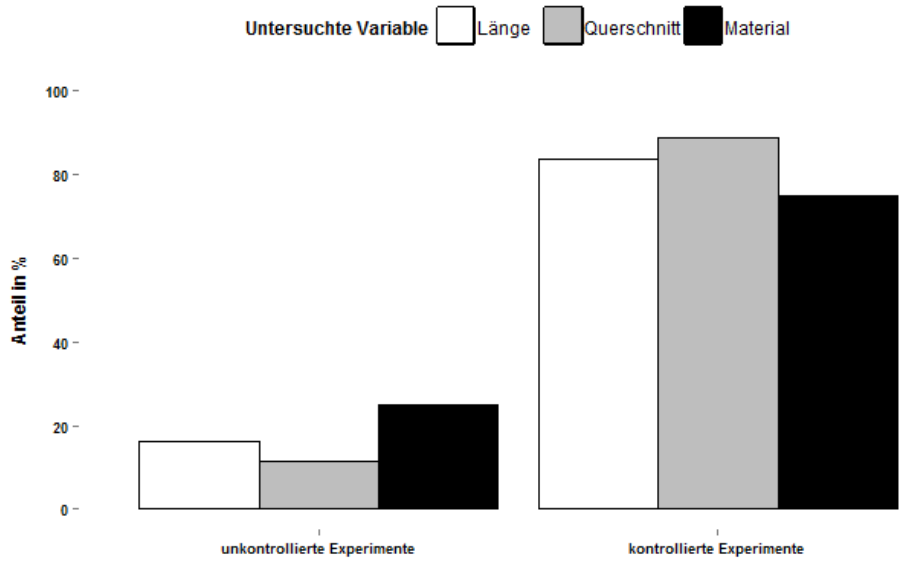


Figure 17 *Anteil an den Schülerexperimenten, die kontrollierte bzw. unkontrollierte Experimente darstellen, aufgeteilt nach der untersuchten Variable.*

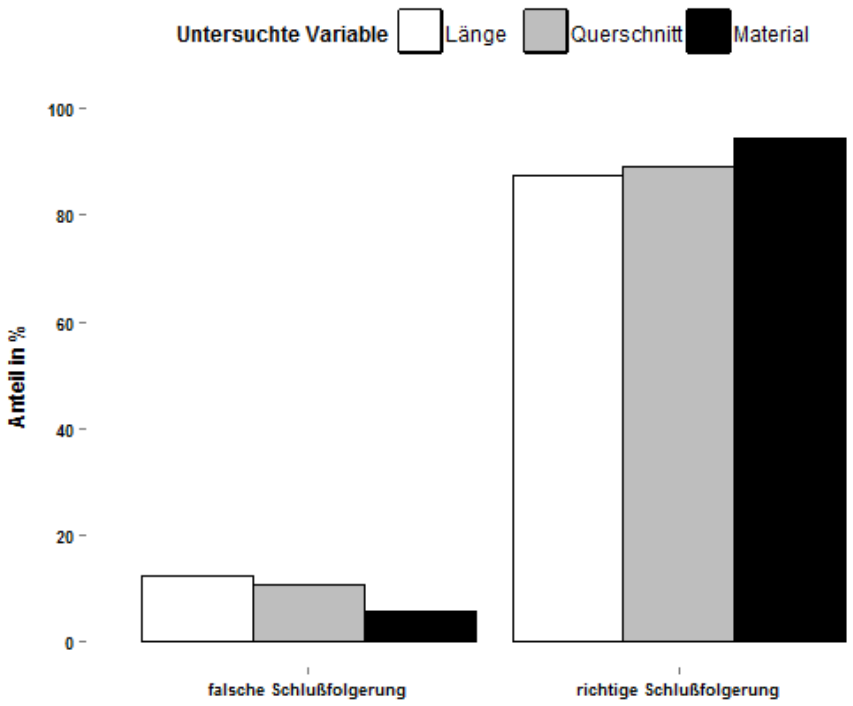


Figure 18 *Anteil der fachlich richtigen Schlussfolgerungen dargestellt für alle drei untersuchten Variablen.*

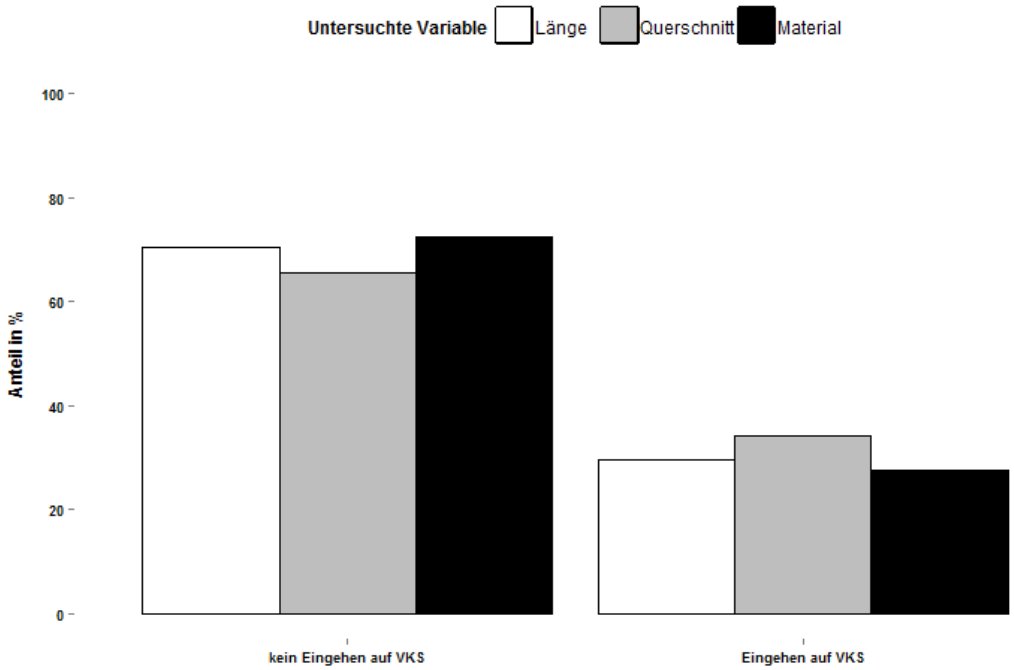


Figure 19 *Schülerantworten auf die Reflexionsfrage. Anteil der Schülerinnen und Schüler die korrekt auf die VKS in ihren Antworten eingehen.*

Die Befunde der Unterrichtserprobung zeigen, dass die meisten Schülerinnen und Schüler in der Lage sind aus der großen Anzahl an theoretisch möglichen Vergleichen diejenigen auszuwählen, die kontrollierte Experimente darstellen. Folglich sind die entwickelten Übungsexperimente nicht zu komplex und für den Einsatz im Unterricht geeignet. Es scheint für die Schülerinnen und Schüler jedoch deutlich schwieriger zu sein, die Relevanz der VKS für die Gültigkeit der experimentellen Befunde zu benennen. Da gerade dieser Reflexionsprozess jedoch für die Kontrolle eigener Befunde und für den wissenschaftlichen Argumentationsprozess von großer Bedeutung ist, sollte dieser Aspekt im Unterricht explizit behandelt werden.

6.6 Fazit

Das vorgestellte VKS Übungsexperiment konfrontiert die Schülerinnen und Schüler im Vergleich zu Experimenten, die zur Vermittlung von Fachwissen konzipiert wurden, mit wesentlich mehr Experimentiermaterialien. Trotz der für die Schülerinnen und Schüler ungewohnten Anforderung scheinen die vorgestellten Experimente für den Einsatz im Unterricht geeignet, da die meisten Schülerinnen und Schüler der Erprobungsklassen bewusst und nicht zufällig kontrollierte Experimente durchgeführt haben. Dies spricht auch dafür, dass die vorgestellten Merkmale von VKS Übungsexperimenten auf andere physikalische bzw. allgemeine naturwissenschaftliche Kontexte übertragen werden können, um weitere VKS Übungsexperimente zu konstruieren. Entsprechende Übungsexperimente wurden bereits zu den Themen Schmelzen von Eis, Elektromagnetismus, physikalisches Pendel und Strahlungswärme erstellt. Die Materiallisten und Protokollbögen können bei den Autoren angefragt werden.

Es zeigt sich allerdings auch, dass ein Großteil der Schülerinnen und Schüler bei der Reflexion über ihr experimentelles Vorgehen nicht auf die VKS eingeht. Dies ist problematisch, da eine kritische Reflexion darüber ob Befunde auf kontrollierten Experimenten beruhen eine zentrale Fähigkeit für das Arbeiten mit eigenen bzw. fremden experimentellen Befunden ist. Darüber hinaus fördert eine adäquate Reflexion des eigenen experimentellen Vorgehens ein nachhaltiges Lernen (Smetana & Bell, 2012). Wenn dieses im Rahmen der Bearbeitung der Arbeitsblätter nicht erfolgt, so ist eine Reflexion im Unterrichtsgespräch denkbar. Um einen solchen Reflexionsprozess zu initiieren, könnten z. B. in einer Tabelle die Längen, Querschnitte und Materialien der von den Schülern/innen verglichenen Leiter gesammelt und verglichen werden. Bei alternativen Experimenten, deren Aufbau auf einem Foto gut nachvollziehbar ist, sind auch Fotos der Schülerexperimente zur Initiierung eines Reflexionsprozesses denkbar. Fotos bieten zudem für die Lehrkraft eine Möglichkeit, die experimentellen Fähig-

keiten der Schülerinnen und Schüler zu evaluieren, typische Fehler zu erkennen oder ein gruppenspezifisches Feedback zu geben. Die vorgestellten Übungsexperimente eignen sich zum Einsatz im Physikunterricht und können ein entscheidender Schritt zur Förderung der selbstständigen Erkenntnisgewinnung durch Experimentieren sein. Schlussendlich ist das Beherrschen der VKS nur ein erster Schritt zum erfolgreichen selbstständigen Experimentieren im Unterricht.

7. Diskussion

In diesem Kapitel werden die zentralen Ergebnisse der vier vorgestellten Studien zusammengefasst und ihr Bezug zu dem übergeordneten Ziel der Arbeit dargestellt. Anschließend werden Implikationen dieser Ergebnisse für zukünftige Forschung und die Praxis des Physikunterrichts diskutiert.

7.1 Zusammenfassung und Diskussion der Studienergebnisse

Ziel dieser Forschungsarbeit ist, effektive Methoden zur Förderung der VKS im Physikunterricht zu identifizieren. Aufgrund der Vielzahl existierender Interventionsstudien und der uneinheitlichen Befunde dieser Studien, wurde das Forschungsprojekt mit einer Meta-Analyse begonnen. Die in der Meta-Analyse (Publikation 1) bestimmte mittlere Effektstärke (Vergleich zwischen Treatment- und Kontrollgruppen) g = 0.61 (95% CI = 0.53-0.69) verdeutlicht, dass eine Förderung der VKS möglich und darüber hinaus effektiv ist. Sie ist aufgrund der Exklusion von Ausreißern kleiner als die von Ross (1988a) bestimmte Effektstärke (g = 0.73; 95% CI = 0.54-0.92) und folglich eine konservativere und realistischere Schätzung von Interventionseffekten.

Die Meta-Analyse liefert zudem aufgrund der systematischen Literaturrecherche ein umfassendes Bild existierender Interventionsstudien zur Förderung der VKS. Durch Vergleich der Effektstärken unterschiedlicher Studien (Treatment-Kontrollgruppen-Vergleiche) kann der Einfluss von Unterrichts- (z. B. Nutzung von Demonstrationsexperimenten), Probanden- (z. B. Probandenalter) und Testmerkmalen (z. B. Item-Format) auf den Effekt von Interventionsstudien untersucht werden. Die meta-analytischen Befunde zum Einfluss dieser Merkmale auf die Studieneffekte sind in Figure 4 zusammengefasst. Für die Zielsetzung dieser Forschungsarbeit sind besonders die Ergebnisse zu effektiven bzw. ineffektiven Unterrichtsmethoden von Interesse. Sie zeigen u. a., dass sich direkte Instruktion und entdeckendes Lernen in ihrer

Wirksamkeit nicht unterscheiden. Die kontrovers geführte Debatte um die unterschiedliche Wirksamkeit beider Unterrichtsmethoden (Kuhn, 2005b versus Klahr, 2005) scheint somit von geringer praktischer Relevanz.

Stattdessen haben andere, bisher kaum systematisch betrachtete Unterrichtsmerkmale einen Einfluss auf die Wirksamkeit von Interventionen. Studien, die kognitive Konflikte und/oder Demonstrationsexperimente einsetzen, erzielen signifikant größere Effekte als Studien, die diese Methoden nicht einsetzen. Diese Ergebnisse legen nahe, dass zur Förderung der VKS eine rezeptartige Anleitung zum Durchführen kontrollierter Experimente weniger effektiv ist als implizierte Methoden, die aufzeigen bei welchen Fragestellungen und warum die VKS anzuwenden ist. Erklärt werden kann dieser Befund durch ein spätestens im Grundschulalter vorhandenes elementares Verständnis der VKS. Bereits Probanden im Vorschul- und Grundschulalter lösen erfolgreich alltagsnahe VKS-Aufgaben, die nicht im Konflikt mit ihrem Vorwissen stehen (domain general tasks) (Croker & Buchanan, 2011; Gopnik & Schulz, 2004; Piekny et al., 2014; Sodian et al., 1991). Auch die Ergebnisse der Poster-Evaluations-Aufgabe in Publikation 3 (Interventionsstudie) belegen, dass Probanden, die kein Verständnis der VKS in physikalischen Kontexten zeigen, in alltagsnahen Kontexten adäquat mit der VKS argumentieren. Effektive Interventionsstudien zeichnet aus, dass den Probanden durch kognitive Konflikte oder Demonstrationsexperimente gezeigt wird, weshalb und bei welchen Fragestellungen die ihnen aus Alltagskontexten bekannte VKS auch in naturwissenschaftlichen Kontexten anzuwenden ist. Sowohl Demonstrationsexperimente als auch kognitive Konflikte eignen sich dazu, da sie die Aufmerksamkeit der Schülerinnen und Schüler auf das Vorgehen und Argumentieren beim Experimentieren lenken. Aufgrund des bereits vorhandenen elementaren Verständnisses der VKS sollte zur unterrichtlichen Förderung demnach weniger Merkmale kontrollierter Experimente vermittelt werden. Stattdessen sollte die Aufmerksamkeit der Schülerinnen und Schüler auf die VKS gelenkt und aufgezeigt werden, dass die ihnen (implizit) bekannte VKS analog zu Alltagskontexten auch bei naturwissenschaftlichen Fragestellungen anzuwenden ist.

Meta-Analysen liefern einen Überblick über den bisherigen Forschungsstand indem sie die Generalisierbarkeit einzelner Studienbefunde überprüfen. Darüber hinaus können in Meta-Analysen Merkmale (z. B. Unterrichtsmerkmalen) untersucht werden, die bisher nicht innerhalb einer, sondern nur zwischen Studien variiert wurden. Eine kausale Interpretation me-

ta-analytischer Befunde ist jedoch nicht möglich, da sich Primärstudien im Allgemeinen in mehr als einem Merkmal systematisch unterschieden, sodass keine Kontrolle alternativer Ursache-Wirkungszusammenhänge erfolgt (Borenstein et al., 2010, S. 379ff). Die Ergebnisse der Meta-Analyse (Publikation 1) zeigen z. B. ein überwiegend gemeinsames Auftreten der Unterrichtsmerkmale „Induktion eines kognitiven Konflikts“ und „Nutzung von Demonstrationsexperimenten“. Darum ist nicht unterscheidbar, welches der beiden Merkmale oder ob die Kombination beider Merkmale einen Einfluss auf den Interventionseffekt hat. Der kausale Einfluss meta-analytischer Befunde, die für zukünftige Forschung und für die Unterrichtspraxis besonders relevant sind, wurde daher in experimentellen (bzw. quasi-experimentellen) Studien überprüft.

Die erste Anschlussstudie (Publikation 2) untersucht die in der Meta-Analyse identifizierte Abhängigkeit der Studienbefunde von den eingesetzten Testinstrumenten. Dieser Befund ist für zukünftige Forschung von Relevanz, da er nahe legt verschiedene Instrumente einzusetzen, um herauszufinden, ob Studienbefunde auf einzelne Testinstrumente beschränkt oder übertragbar sind. Wie in der Meta-Analyse dargestellt, ist bislang nicht geklärt, ob die Operationalisierung unterschiedlicher VKS-Teilfähigkeiten den beobachteten Effekt verursacht. In Publikation 2 wird ein neu entwickeltes Testinstrument (Control-of-Variables-Strategy Inventory – CVSI) vorgestellt, das mehrere relevante Teilfähigkeiten der VKS erfasst. Die Befunde einer Pilotierungsstudie zeigen, dass die Aufgabenschwierigkeit von den operationalisierten VKS-Teilfähigkeiten abhängt. Aufgaben, die testen ob Schülerinnen und Schüler die fehlende Aussagekraft unkontrollierter Experimente erkennen, sind signifikant schwieriger als Aufgaben, die eine Identifizierung oder Interpretation kontrollierter Experimente fordern. Folglich kann die in der Meta-Analyse identifizierte Abhängigkeit der Studienbefunde von den eingesetzten Testinstrumenten auf die Operationalisierung unterschiedlicher Teilfähigkeiten zurückgeführt werden. Der Befund zeigt außerdem, dass Aufgaben, die eine Reflexion über Experimente erfordern, schwieriger sind, als Aufgaben, die eine Anwendung der VKS erfordern. Dieser Unterschied zwischen kognitiven und metakognitiven Aufgabenanforderungen (Zohar, 2012) spiegelt sich auch in den Ergebnissen der Unterrichtserprobung (Publikation 4) wider. Während es Schülerinnen und Schülern kein Problem bereitet kontrollierte Experimente aus der großen Anzahl theoretisch möglicher Experimente auszuwählen, gehen nur wenige Probanden in der Reflexionsfrage auf die VKS ein. Diese Ergebnisse deuten erneut an, dass Schülerinnen und Schüler bereits vor dem Unterricht ein elementares Verständnis der VKS

haben, bzw. mit geringem Aufwand ein solches Verständnis erlangen können. Nur wenige Probanden scheinen hingegen ein vollständiges Verständnis der VKS zu entwickeln, welches ein explizites Referieren auf die VKS bzw. ein Verständnis der Konsequenzen unkontrollierter Experimente einschließt.

Publikation 3 stellt eine quasi-experimentelle Studie vor, welche die Wirkung von Schülerexperimenten und reinen Papier-und-Bleistift Übungsaufgaben auf den Erwerb der VKS kontrastiert. Die Studie basiert auf dem in der Meta-Analyse identifizierten Trend einer negativen Wirkung von Schülerexperimenten auf die Fähigkeit zur Variablenkontrolle. Aufgrund der häufigen Nutzung von Schülerexperimenten im Physikunterricht (Tesch, 2005) ist die Frage nach ihrer Lernwirksamkeit von hoher praktischer Relevanz. Die in der Studie kontrastierten VKS Übungen wurden so gestaltet, dass sie sich hinsichtlich der anzuwendenden kognitiven Operationen (Teilfähigkeiten der VKS), der zu kontrollierenden Variablen und dem Kontext und Inhalt der Aufgaben nicht unterscheiden. Ziel der Studie ist die unique Wirkung der Interaktion mit Experimentiermaterialien zu isolieren.

Die Ergebnisse zeigen keine generelle Überlegenheit oder negative Effekte von Schülerexperimenten. Die Wirkung der unterschiedlichen Lernumgebungen (Schülerexperimente versus Arbeitsblätter) hängt, wie bereits in der Meta-Analyse, von dem verwendeten Testinstrument ab. Die Probanden schneiden besser in Testinstrumenten ab, die vergleichbare Inhalte und Formate zu ihrer jeweiligen Lernumgebung aufweisen. Dieser Vorteil überträgt sich aber nicht auf Instrumente gleichen Formats mit unterschiedlichen Inhalten oder Instrumenten gleichen Inhalts aber unterschiedlichen Formats, sodass er auf eine spezifische Kombination von Inhalt und Format beschränkt ist. Ferner zeigen die Ergebnisse, dass eine aktive Manipulation von Variablen weder zwingend notwendig, noch der vermeintliche höhere *cognitive load* von Schülerexperimenten hinderlich für das Erlernen der VKS ist. Stattdessen scheinen für die Förderung der VKS die gedankliche Manipulation von Variablen (kognitive Variablenmanipulation) und eine Reflexion der Konsequenzen entscheidend. Die Ergebnisse der Interventionsstudie ermöglichen daher nicht nur Rückschlüsse auf geeignete Methoden zur Förderung der VKS, sondern auch auf die Struktur der VKS. Sie untermauern, dass die VKS rein kognitive Fähigkeiten umfasst, da das Üben manueller Fertigkeiten keinen Einfluss auf die VKS hat. Zwar ist anzunehmen, dass manuell anspruchsvolle Experimente ein Üben des

Messvorgangs oder der Variablenmanipulation erfordern, doch sind diese kontextspezifischen Fertigkeiten nicht der allgemeinen VKS zuzuordnen.

Auch wenn das Forschungsprojekt neue Erkenntnisse zur unterrichtlichen Förderung der VKS liefert, unterliegt es Einschränkungen. Beispielsweise wurde der Einfluss des Alters auf die Wirksamkeit unterschiedlicher Unterrichtsmethoden weder in der Meta-Analyse noch in den Anschlussstudien betrachtet. Zwar wurde in der Meta-Analyse kein Effekt des Probandenalters auf die Wirksamkeit von Interventionen gefunden, doch heißt dies nicht, dass Probanden jeden Alters von identischem Unterricht in gleicher Weise profitieren. Stattdessen sind die in Interventionsstudien eingesetzten Unterrichtsmethoden vermutlich an das Probandenalter angepasst. Zur Untersuchung der Altersabhängigkeit sind daher Längsschnittstudien erforderlich, welche die Wirkung identischer Unterrichtsmethoden auf Probanden unterschiedlichen Alters unter Verwendung identischer Testinstrumente untersuchen. Des Weiteren wurde in der Interventionsstudie (Publikation 3) der Einfluss des Unterrichtseinstiegs mit einem kognitiven Konflikt und Demonstrationsexperimenten nicht untersucht. Um zu zeigen inwiefern der Unterrichtseinstieg für die berichteten Effekte notwendig oder ausreichend ist, wären weitere Kontrollgruppen erforderlich. Die beiden vorhandenen Treatmentgruppen müssten dazu mit einer Gruppe, die keinerlei CVS Treatment erhält und einer Gruppe, die nur den Unterrichtseinstieg (Kognitiver Konflikt und Demonstrationsexperiment) erhält, verglichen werden.

Trotz dieser Limitationen liefert die Forschungsarbeit neue Impulse in einem bereits elaborierten Forschungsfeld und für die Unterrichtspraxis. Möglich war dies durch die Kombination einer Meta-Analyse und eng auf ihre Ergebnisse abgestimmte Anschlussstudien. Das Vorgehen zeigt exemplarisch die Vorteile der Kombination von Meta-Analysen und Primär-Studien.

7.2 Implikationen für zukünftige Forschung

Die gefundene Abhängigkeit der Studienergebnisse von den eingesetzten Testinstrumenten (Publikationen 1, 2, 3) hat weitreichende Implikationen für zukünftige Forschungsarbeiten. Eine Abhängigkeit der Studienergebnisse von den Testinstrumenten lässt in der fachdidaktischen, wie in der naturwissenschaftlichen Forschung (siehe S. 3) Zweifel an der Existenz des zugrundeliegenden Konstrukts bzw. an der Eignung der eingesetzten Messinstrumente aufkommen (Messick, 1989). Bezogen auf die VKS ist fraglich, ob diese überhaupt ein den unterschiedlichen Operationalisierungen zugrunde liegendes Konstrukt ist, da Messinstrumente unterschiedlichen Formats uneinheitliche Befunde über die Probandenfähigkeit zur Variablenkontrolle liefern. Dennoch erscheint es aus zweierlei Gründen sinnvoll die VKS weiterhin als ein einheitliches, aus vier Teilfähigkeiten bestehendes Konstrukt anzusehen. Erstens legen theoretische Überlegungen (siehe Einleitung) nahe, die VKS als ein domänenunabhängiges und aus mehreren Teilfähigkeiten bestehendes Konstrukt zu betrachten. Zweitens kann die Frage ob die VKS ein einheitliches Konstrukt ist auf Grundlage bisheriger Studien nicht beantwortet werden, da diese mit einem eingeschränkten Probandenkreis durchgeführt wurden. Sowohl die in der Meta-Analyse integrierten Studien, als auch die vorgestellten Anschlussstudien, wurden mit experimentell unerfahrenen Probanden (Novizen) durchgeführt. Aus der Forschung zum Problemlösen von Experten und Novizen ist bekannt, dass Novizen Konzepte nur kontextgebunden nutzen und ein hoher Grad an Expertise erforderlich ist, um Konzepte auf neue Problemstellungen zu übertragen (Feltovich, Glaser, & Chi, 1981). Die beobachtete Anhängigkeit der Testergebnisse von dem verwendeten Testformat kann daher durchaus auf die Beschränkung bisheriger Studien auf Novizen zurückzuführen sein. Auch wenn Probanden im Grundschulalter bereits in alltagsnahen Kontexten kontrollierte Experimente planen und interpretieren können (Croker & Buchanan, 2011; Gopnik & Schulz, 2004; Piekny & Maehler, 2013), bedarf es einer größeren Expertise sämtliche Teilfähigkeiten der VKS in unterschiedlichen Kontexten anzuwenden.

Inwiefern Experten die einzelnen Teilfähigkeiten der VKS gleichermaßen beherrschen oder diese domänenübergreifend einsetzen ist jedoch eine offene Frage. Eine Klärung dieser Frage wäre z. B. mit einer Längsschnittstudie möglich, in der die Entwicklung der einzelnen VKS-Teilfähigkeiten während des Schulbesuchs erfasst wird. Würde sich mit zunehmendem Alter (zunehmender Expertise infolge der Beschulung) die Schwierigkeit von Aufgaben, die unterschiedliche VKS-Teilfähigkeiten erfassen, angleichen, so spräche dies für die Betrachtung der

VKS als ein einheitliches Konstrukt. Die vollständige Beherrschung der VKS würde allerdings einen hohen Grad an Expertise voraussetzen. Aus den Ergebnissen einer solchen Längsschnittstudie könnte zudem ein mögliches Kompetenzentwicklungsmodell abgeleitet werden, das bei der Förderung einfacher VKS-Teilfähigkeiten ansetzt und sukzessiv schwerere Teilfähigkeiten der VKS einführt.

Die vorgestellten Befunde legen nahe, in zukünftigen Studien Instrumente einzusetzen, die sämtliche Teilfähigkeiten der VKS erfassen. Nur so können Unterrichtsmethoden identifiziert werden, die möglichst viele VKS-Teilfähigkeiten fördern. Die Instrumente, die bisher in Interventionsstudien eingesetzt wurden, sind überwiegend auf die Planung bzw. Identifizierung kontrollierter Experimente beschränkt (siehe Table 4). Über Methoden zur Förderung der übrigen VKS-Teilfähigkeiten ist daher wenig bekannt. In Publikation 2 wird ein Test (CVSI), der diese Anforderungen erfüllt, vorgestellt und ein Verfahren zur Entwicklung von VKS-Testaufgaben beschrieben. Der vorgestellte Test (CVSI) ist jedoch unvollständig, da die Fähigkeit kontrollierte Experimente zu planen nicht abgefragt wird. In einer bereits entwickelten erweiterten online Version des Tests wird diese Schwäche behoben und sämtliche VKS-Teilfähigkeiten erfasst. Der Einsatz von Instrumenten wie des CVSI kann dazu beitragen, die benötigten Erkenntnisse zur Förderung bisher kaum betrachteter VKS-Teilfähigkeiten zu gewinnen.

Neben diesen praktischen Implikationen für weiterführende Forschungsarbeiten, haben die Befunde des vorgestellten Forschungsprojekts auch lerntheoretische Implikationen. Die Lerntheorie des *Conceptual Change* geht davon aus, dass die „vorunterrichtlichen Konzepte der Lerner ersetzt werden müssen, um wissenschaftliche Konzepte zu erlernen“ (Duit & Treagust, 2003, S. 673). Diese Aussage trifft auf das Konzept der VKS nicht zu, da Schülerinnen und Schüler bereits vor Besuch naturwissenschaftlichen Unterrichts die VKS in gewissen Kontexten adäquat nutzen (siehe u. a. vorheriges Kapitel). Damit sie lernen, die VKS auch in naturwissenschaftlichen Kontexten adäquat anzuwenden, müssen sie sich folglich kein neues Konzept aneignen, sondern ein ihnen bekanntes Konzept auf neue Kontexte übertragen. Unterrichtsmethoden, die eine solche Übertragung z. B. durch Induktion eines kognitiven Konflikts anregen, sind daher besonders effektiv. Unterstützt wird diese Annahme durch Studienbefunde, die zeigen, dass die Vermittlung metastrategischen Wissens darüber, warum und in welchen Kontexten die VKS anzuwenden ist, besonders effektiv sind (Zohar & David, 2008; Zo-

har & Peled, 2008) während eine direkte „rezeptartige“ Vermittlung der VKS sich nicht positiv auf das Erlernen der VKS auswirkt (siehe Publikation 1).

Um die Entwicklung der VKS zu beschreiben scheint daher die *Knowledge in Pieces* Lerntheorie von diSessa (1988) geeigneter. Diese Theorie geht davon aus, dass zum Lernen neuer Konzepte vorunterrichtliche Schülervorstellungen nicht radikal, sondern zunächst kontextbezogen durch elaboriertere Konzepte ersetzt werden. Mit zunehmender Expertise abstrahieren Lerner das auf wenige Kontexte beschränkte Konzept und wenden es in zunehmend mehr Kontexten an. Die Übertragung eines Konzepts auf sämtliche relevanten Kontexte erfordert allerdings eine hohe Expertise und ist das Ergebnis eines längeren Lernprozesses. Bezogen auf die VKS bedeutet dies, dass die Schülerinnen und Schüler die Anwendung der VKS in möglichst verschiedenen Kontexten üben müssen, um die beabsichtigte Verallgemeinerung und Abstraktion der VKS zu erlangen. Dabei sollten sowohl die Kontexte als auch die notwendigen VKS-Teilfähigkeiten variiert werden.

Eine offene Frage ist, wie Schülerinnen und Schüler ihre vorunterrichtliche Fähigkeit zur Variablenkontrolle erwerben. Auch wenn diese Frage bisher ungeklärt ist, scheint eine Voraussetzung für die Entwicklung der VKS zu sein, dass Probanden die Abhängigkeit zwischen Vorstellungen (Theorien) und Beobachtungen (Evidenzen) kennen. Die Erkenntnis, dass Vorstellungen keine allgemeine Realität widerspiegeln, sondern von persönlichen Beobachtungen und Erfahrungen abhängen, wird als „Theory-of-Mind“ bezeichnet. Sie ist eine Voraussetzung dafür, dass Probanden sich mit der Aussagekraft experimenteller Daten beschäftigen und die Validität kontrollierter Experimente erkennen. Bislang ist allerdings nicht geklärt wie genau sich das frühe Verständnis der VKS entwickelt. Diese Frage ist jedoch eine entwicklungspsychologische und keine fachdidaktische Fragestellung, da sie sich auf vorunterrichtliche Entwicklungen bezieht (Kuhn & Pearsall, 2000; Sodian, 2005).

Eine aus fachdidaktischer Sicht interessante weiterführende Fragestellung ist, wie die VKS zur Entwicklung experimenteller Kompetenz im weiteren Sinne beiträgt. Auch wenn die VKS theoretisch eine zentrale Kompetenz sowohl im weiteren als auch im engeren Sinne des Experimentierens ist, fehlen bislang empirische Befunde dazu inwiefern die eine Beherrschung der VKS prädiktiv für erfolgreiches Experimentieren ist. Denkbar ist, dass die VKS zwar eine notwendige aber aufgrund der Vielzahl erforderlicher Fähigkeiten keine hinreichende Bedin-

gung für erfolgreiches Experimentieren ist. Außerdem ist es aufgrund einer unterschiedlichen Relevanz der VKS je nach experimenteller Fragestellung möglich, dass der Einfluss der VKS auf die experimentelle Kompetenz von der Fragestellung abhängt. Da das langfristige Ziel einer unterrichtlichen Förderung der VKS die Förderung experimenteller Kompetenz ist, sind diese Fragen nicht nur von theoretischer, sondern auch von praktischer Relevanz.

7.3 Implikationen für die Förderung der VKS im Physikunterricht

Aus den Befunden des Forschungsprojekts lassen sich Implikationen für die Förderung der VKS im Physikunterricht ableiten. Wie im vorherigen Kapitel dargelegt, sollte Unterricht zur Förderung der VKS weniger darauf eingehen wie kontrollierte Experimente zu planen und durchzuführen sind. Stattdessen sollte vermittelt werden bei welchen Fragestellungen die VKS anzuwenden ist und warum unkontrollierte Experimente keine kausalen Schlussfolgerungen ermöglichen. Die Ergebnisse der Publikationen 2 und 4 zeigen, dass gerade diese meta-kognitiven Teilfähigkeiten für Schülerinnen und Schüler besonders anspruchsvoll sind und daher einer gezielten unterrichtlichen Förderung bedürfen. Die meisten Schülerinnen und Schüler beherrschen bereits vor dem Unterricht die Planung und Interpretation kontrollierter Experimente, zu mindestens in alltagsnahen Kontexten. Daher scheint für eine Förderung der VKS weniger die Vermittlung dieser Fähigkeiten, sondern vielmehr die Kenntnis, dass Experimente zur Erkenntnisgewinnung dienen (siehe Kapitel 1.2), entscheidend. Damit Schülerinnen und Schüler die ihnen in Alltagskontexten bekannte VKS auf naturwissenschaftliche Kontexte übertragen, sollten sie wissen, dass durch Experimentieren keine Effekte erzeugt, sondern etwas herausgefunden werden soll. Der Grundgedanke einer unterrichtlichen Förderung der VKS sollte sein, Schülerinnen und Schülern aufzuzeigen, dass in naturwissenschaftlichen Experimenten, analog zu Alltagskontexten, eine Kontrolle alternativer Ursache-Wirkungsbeziehungen vorzunehmen ist. Dies kann z. B. durch die Induktion kognitiver Konflikte oder durch Demonstrationsexperimente erfolgen (Publikation 1 und 3).

Beim Experimentieren im engeren Sinne müssen nicht nur kontrollierte Experimente geplant, sondern auch konfundierende Variablen erkannt und die Ergebnisse von Experimenten interpretiert werden. Um Experimente im weiteren Sinne, wie bei der Entwicklung neuer Messinstrumente nachzuvollziehen, sind neben praktischen gerade die meta-kognitiven Teilfähigkeiten der VKS von besonderer Bedeutung. Folglich sind sämtliche VKS-Teilfähigkeiten zentral für die Entwicklung experimenteller Kompetenz, sowohl im Sinne der engeren als auch im Sinne der weiteren Definition des Experimentierens. Da bisherige Interventionsstudien bis auf

wenige Ausnahmen (Zohar & David, 2008; Zohar & Peled, 2008) keine meta-kognitiven Teilfähigkeiten betrachten, ist das Wissen über effektive Unterrichtsmethoden zur Förderung dieser Teilfähigkeiten beschränkt.

Eine thematische Ausweitung des Unterrichts unter Einbeziehung sämtlicher VKS-Teilfähigkeiten kann nicht nur experimentelle Kompetenz, sondern auch die Ausbildung eines realistischen Bilds des Wesens der Naturwissenschaften (engl. Nature of Science, NOS) fördern. Gerade das Nachdenken über die Methodenwahl verdeutlicht, dass wissenschaftliche Methoden auch Gegenstand von Diskursen sind und je nach Fragestellungen, Datengrundlage und angewendetem Forschungsparadigma angepasst werden. Aus diesem Grund sollten Schülerinnen und Schüler ein Verständnis der fehlenden Aussagekraft konfundierter Experimente haben. Ein solches Verständnis ermöglicht ihnen wissenschaftliche Diskurse über kausale Zusammenhänge, sowie das Vorgehen bei nicht experimentellen Methoden nachzuvollziehen (z. B. Ausschluss alternativer Ursachen durch Erhebung von Kontrollvariablen). Ziel einer unterrichtlichen Förderung sollte daher sein, dass Schülerinnen und Schüler die VKS nicht nur als „die experimentelle Methode“ anwenden können, sondern sie als ein universelles Konzept zur Untersuchung und Diskussion kausaler Zusammenhänge verstehen. Beschränkt sich Unterricht auf die Vermittlung eines idealisierten Vorgehens zum Planen kontrollierter Experimente, kann dies bei den Schülerinnen und Schülern den Eindruck erwecken, dass Wissen in den Naturwissenschaften ausschließlich durch das Abarbeiten einer genau vorgegebenen Methode gewonnen wird. Diese Vorstellung von naturwissenschaftlicher Forschung wäre unrealistisch, da kreative und individuelle Methoden für den wissenschaftlichen Fortschritt unabdingbar sind (Lederman, Antink, & Bartos, 2014). Inwiefern die erweiterte Einführung der VKS diese positiven Effekte hat, ist jedoch bisher nicht empirisch geprüft und daher ein möglicher Anschlusspunkt für weiterführende Studien.

Die Ergebnisse dieser Arbeit geben erste Hinweise, wie eine Förderung sämtlicher VKS-Teilfähigkeiten im Sinne eines Kompetenzentwicklungsmodells möglich wäre. Die in Publikation 2 berichteten Befunde zur Schwierigkeit der einzelnen VKS-Teilfähigkeiten legen nahe, dass eine Kompetenzentwicklung bei der Identifizierung und Interpretation kontrollierter Experimente beginnt und erst später ein Verständnis der fehlenden Aussagekraft unkontrollierter Experimente entwickelt wird. Diese Befunde könnten als Ausgangspunkt für die Entwicklung eines Kompetenzentwicklungsmodells zur Förderung der VKS dienen. Demnach

sollte eine unterrichtliche Förderung der VKS bei den einfacheren früh entwickelten Teilfähigkeiten (Interpretation und Identifikation) beginnen und schwierigere Teilfähigkeiten (Verständnis der fehlenden Aussagekraft unkontrollierter Experimente) erst später thematisieren (Neumann, Viering, Boone, & Fischer, 2013). Eine Förderung der VKS kann, wie mehrere Interventionsstudien zeigen (Chen & Klahr, 1999; Grygier, 2008; Strand-Cary & Klahr, 2008), bereits im Grundschulalter begonnen werden, wenn alltagsnahe Kontexte gewählt werden. Vor der Entwicklung eines solchen Kompetenzentwicklungsmodells sollte jedoch in einer Längsschnittstudie geklärt werden, ob die Kompetenzentwicklung wie angenommen stattfindet. Nach Möglichkeit sollte dabei auch die bisher nicht berücksichtigte Teilfähigkeit der Planung kontrollierter Experimente einbezogen werden.

Die in Publikation 4 vorgestellten VKS-Übungsexperimente können von Lehrkräften zur Förderung der VKS im Physikunterricht übernommen werden. Ferner können Lehrkräfte basierend auf den vorgestellten Merkmalen solcher Experimente weitere Übungsexperimente für andere naturwissenschaftliche Fächer entwickeln. Für die Unterrichtspraxis sind diese Merkmale von besonderer Relevanz, da sich VKS-Übungsexperimente deutlich von Schülerexperimenten zum Fachwissenserwerb unterscheiden (siehe Publikation 4). Die meisten Interventionsstudien und Unterrichtsmaterialien wurden bisher in physikalischen Kontexten durchgeführt (Publikation 1). Diese eignen sich im Vergleich zu chemischen oder biologischen Kontexten besonders zur Förderung der VKS, da sie mit geringerem zeitlichen Aufwand durchzuführen sind und viele Variablen (z. B. Leitermaterial) anschaulich und leicht zu operationalisieren sind. Die VKS sollte jedoch in allen naturwissenschaftlichen Fächern behandelt werden, um zu erreichen, dass Schülerinnen und Schüler die VKS auf neue Kontexte übertragen. Aus demselben Grund wäre auch eine explizite Thematisierung der VKS in gesellschaftswissenschaftlichen Fächern wünschenswert. Durch eine solche umfangreiche Implementierung der VKS können einerseits fachspezifische Kompetenzen (z. B. experimentelle Kompetenz) gefördert, aber andererseits auch die im Sinne der Allgemeinbildung relevante Übertragbarkeit der VKS auf neue Kontexte (auch nicht naturwissenschaftliche Kontexte) ermöglicht werden. Dass ein solcher Transfer prinzipiell möglich ist, zeigen die Befunde von Interventionsstudien, die eine Übertragung der VKS auf nicht naturwissenschaftliche Kontexte nachweisen (Dean & Kuhn, 2007; Strand-Cary & Klahr, 2008). Bisher fehlen jedoch geeignete Unterrichtsmethoden und vor allem Übungsaufgaben, um die breiten Anwendungsmöglichkeiten der VKS im Unterricht zu verdeutlichen.

Literaturverzeichnis

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., & Tuan, H.-l. (2004). Inquiry in science education: International perspectives. *Sci. Ed*, *88*(3), 397–419.

Adey, P., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school students. *J. Res. Sci. Teach.*, *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.

Australian Curriculum, Assessment and Reporting Authority (ACARA). *Australian curriculum*. Retrieved from http://www.australiancurriculum.edu.au/science/curriculum/f-10?layout=1#level5 (16.10.2015).

Beishuizen, J., Wilhelm, P., & Schimmel, M. (2004). Computer-supported inquiry learning: Effects of training and practice. *Computers & Education*, *42*(4), 389–402.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). *Introduction to meta-analysis* (Reprinted). Chichester: Wiley.

Bortz, J., & Döring, N. (2002). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (3., überarb. Aufl). *Springer-Lehrbuch*. Berlin [u.a.]: Springer.

Bowyer, J. B., & Linn, M. C. (1978). Effectiveness of the science curriculum improvement study in teaching scientific literacy. *J. Res. Sci. Teach.*, *15*(3), 209–219.

Brotherton, P. N., & Preece, P. F. W. (1996). Teaching science process skills. *International Journal of Science Education*, *18*(1), 65–74.

Bryant, P., Nunes, T., Hillier, J., Gilroy, C., & Barros, R. (2013). The importance of being able to deal with variables in learning science. *Int J of Sci and Math Educ*, *13*(1), 145–163.

Bullock, M. (1991). *Scientific reasoning in elementary school: Developmental and individual differences.: Paper presented at SRCD*. Seattle, WA. Retrieved from http://www.eric.ed.gov/PDFS/ED350149.pdf (16.10.2015).

Bullock, M., & Ziegler, A. (1999). Scientifc 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.

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, *21*(1), 13–19.

Case, R., & Fry, C. (1973). Evaluation of an attempt to teach scientific inquiry and criticism in a working class high school. *J. Res. Sci. Teach.*, *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.

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*.

Curriculum Planning & Development Division. (2007). *Science syllabus lower secondary: Express/normal (academic)*. Singapur: Ministry of Education.

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. *Sci. Ed.*, *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.

Department for Education. (2014). *The national curriculum in England: Key stages 3 and 4 framework document.*

Dewey, J. (2002). *Logik : Die Theorie der Forschung [Logic: The theory of inquiry]* (1st ed.). Frankfurt am Main: Suhrkamp.

Dillashaw, G. F., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Sci. Ed.*, *64*(5), 601–608.

diSessa, A. (1988). Knowledge in Pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the Computer Age* (pp. 49–70). Hillsdale, NJ: Erlbaum.

Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, *25*(6), 671–688.

Eberbach, C., & Crowley, K. (2009). From every day to scientific observation: How children learn to observe the biologist's world. *Review of Educational Research*, *79*(1), 39–68.

Feltovich, P. J., Glaser, R., & Chi. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.

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.

Franklin, A. D. (1981). What makes a ’good’ experiment? *British Journal for the Philosophy of Science*, *32*(4), 367–374.

Gilbert, J. K., Osborne, R. J., & Fensham, P. J. (1982). Children's science and its consequences for teaching. *Sci. Ed.*, *66*(4), 623–633.

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.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, *8*(8), 371–377.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Dev Psychol*, *37*(5), 620–629.

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.

Hacking, I. (1996). *Einführung in die Philosophie der Naturwissenschaften*. *Universal-Bibliothek: Vol. 9942*. Stuttgart: Reclam.

Hänsel, M. (2014). Intelligentes Üben in den Naturwissenschaften. *Der mathematische und naturwissenschaftliche Unterricht (MNU)*, *67*(5), 288–291.

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. *Res. Synth. Method, 1*(1), 39–65.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen [Cognitive abilities test of students from 4th to 12th grade]*. Göttingen: Hogrefe.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Sci. Ed., 88*(1), 28–54.

Höttecke, D., & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung – Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift für Didaktik der Naturwissenschaften*, 1–13.

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, Calif: 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.

Jones, C., Ramanau, R., Cross, S., & Healing, G. (2010). Net generation or Digital Natives: Is there a distinct new generation entering university? *Computers & Education*, *54*(3), 722–732.

Keating, D. P. (1990). Adolescent thinking. In S. S. Feldman & G. R. Elliott (Eds.), *At the threshold. The developing adolescent* (pp. 54–89). Cambridge, Mass: Harvard University Press.

Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *J. Res. Sci. Teach., 40*(9), 898–921.

Kiboss, J. K., Ndirangu, M., & Wekesa, E. W. (2004). Effectiveness of a Computer-Mediated Simulations Program in School Biology on Pupils' Learning Outcomes in Cell Theory. *Journal of Science Education and Technology*, *13*(2), 207–213.

Kircher, E., Girwidz, R., & Häußler, P. (2009). *Physikdidaktik* (2nd ed.). Heidelberg [u.a.]: Springer.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Klafki, W. (1996). *Neue Studien zur Bildungstheorie und Didaktik: Zeitgemäße Allgemeinbildung und kritisch-konstruktive Didaktik. Reihe Pädagogik*. Weinheim [u.a.]: Beltz.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, Mass: MIT Press.

Klahr, D. (2005). Early science instruction: Addressing fundamental issues. *Psychological Science, 16*(11), 871–873.

Klahr, D. (2009). To everything 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. *J. Res. Sci. Teach.*, *44*(1), 183–203.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, *333*(6045), 971–975.

KMK. (2005a). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss: Beschluss vom 16.12. 2004*. München: Wolters Kluwer Deutschland GmbH.

KMK. (2005b). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss: Beschluss vom 16.12. 2004*. München: Wolters Kluwer Deutschland GmbH.

KMK. (2005c). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss: Beschluss vom 16.12. 2004*. München: Wolters Kluwer Deutschland GmbH.

Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology*, *64*(3), 141–152.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning* (1st ed.). *Learning, development, and conceptual change*. Cambridge Mass. u.a: MIT Pr.

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? *Psychol Sci*, *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., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development*, *1*(1), 113–129.

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. (1978). The development and validation of a classroom test of formal reasoning. *J. Res. Sci. Teach.*, *15*(1), 11–24.

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. *J. Res. Sci. Teach.*, *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.

Lederman, N. G., Antink, A., & Bartos, S. (2014). Nature of science, scientific inquiry, and socioscientific issues arising from genetics: A pathway to developing a scientifically literate citizenry. *Sci & Educ*, *23*(2), 285–302.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy. In K. A. Renninger, I. E. Sigel, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology* (6th ed., Vol. 4). Hoboken: John Wiley & Sons.

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. *J. Res. Sci. Teach*, *36*(7), 837–858.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2014). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com (16.10.2015).

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). *J. Res. Sci. Teach.*, *20*(8), 755–770.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, Calif: 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.

Ma, X., & Wilkins, J. (2002). The development of science achievement in middle and high school. Individual differences and school effects. *Evaluation review*, *26*(4), 395–417.

Marschner, J. (2011). *Adaptives Feedback zur Unterstützung des selbstregulierten Lernens durch Experimentieren [Supporting self-regulated learning from experiments with adaptive feedback]* (Dissertation). Universität Duisburg-Essen, Essen.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218.

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? *Instr Sci*, *41*(3), 621–634.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *American Council on Education series on higher education. Educational measurement.* Phoenix: Oryx Press; American Council on Education.

Millar, R., & Driver, R. (1987). Beyond Processes. *Studies in Science Education*, *14*(1), 33–62.

Morris, B. J., Croker, S., Masnick, A., & Zimmerm, C. (2012). The emergence of scientific reasoning. In H. Kloos (Ed.), *Current topics in children's learning and cognition* (pp. 61–82). InTech.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, D.C: National Academy Press.

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *J. Res. Sci. Teach.*, *50*(2), 162–188.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, D.C.: The National Academies Press.

NRC. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, D.C: The National Academies.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational and Behavioral Statistics*, *8*(2), 157–159.

Oser, F. K., & Baeriswyl, F., J. (2001). Choreographies of teaching: Bridging instruction to learning. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th ed., pp. 1032–1065). Washingon, D.C: American educational research association.

Padilla, M. J., Okey, J. R., & Garrard, K. (1984). The effects of instruction on integrated science process skill achievement. *J. Res. Sci. Teach.*, *21*(3), 277–287.

Pellegrino, J. W., Wilson, M. R., & Koenig, J. A. (2013). *Developing Assessments for the Next Generation Science Standards*.

Penner, D. E., & Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: A study with sinking objects. *Child Development*, *67*(6), 2709–2727.

Peterson, K. (1977). *An experimental evaluation of a science inquiry training program for high school students* (Dissertation). university of California, Berkeley.

Phywe. *Der Widerstand von Drähten - Abhängigkeit von Länge und Querschnitt*. Retrieved from http://repository.phywe.de/files/versuchsanleitungen/p1372500/d/13725_01.pdf (16.10.2015).

Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education*, 334–354.

Piekny, J., & Maehler, C. (2013). Scientific reasoning in early and middle childhood: The development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *Br J Dev Psychol*, *31*(2), 153–179.

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. *Sci. Ed.*, *67*(1), 85–98.

Rasch, G. (1960). *Probalistic Models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

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. *Appl. Cognit. Psychol.*, *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.

Rousmaniere, F. H. (1906). A definition of experimentation. *The Journal of Philosophy, Psychology and Scientific Methods*, *3*(25), 673–680.

Rutherford, F. J., & Ahlgren, A. (1990). *Science for all Americans*. New York: Oxford University Press.

Samarapungavan, A. (1992). Children's judgments in theory choice tasks: Scientific rationality in childhood. *Cognition*, *45*(1), 1–32.

Sao Pedro, M. A., Gober, J. D., & Raziuddin, J. J. (2010). Comparing Pedagogical Approaches for the Acquisition and Long-Term Robustness of the Control of Variables Strategy. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)* (pp. 1024–1031). Chicago: International Society of the Learning Sciences.

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. *J. Res. Sci. Teach.*, *28*(9), 859–882.

Schneider, W., Schlagmüller, M., & Ennemoser, M. (2007). *LGVT 6-12 Lesegeschwindigkeits- und -verständnistest für die Klassen 6-12 [Reading speed and reading understanding test for grad 6 to 12]*. Göttingen: Hogrefe.

Schulz, A., & Wirtz, M. (2012). Analyse kausaler Zusammenhänge als Ziel des Experimentierens. In W. Rieß, M. Wirtz, B. Barzel, & A. Schulz (Eds.), *Experimentieren im mathematisch-*

naturwissenschaftlichen Unterricht. Schüler lernen wissenschaftlich denken und arbeiten (pp. 39–56). Münster, New York, NY, München, Berlin: Waxmann.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Dev Psychol*, *40*(2), 162–176.

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2015). The impact of sub-skills and item content on students' skills with regard to the control-of-variables-strategy (CVS). *Submitted for publication*.

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2015). *Teaching the Control-of-Variables Strategy: A Meta-Analysis*: submitted for publication.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, *21*(4), p 22-27.

Shayer, M., & Adey, P. S. (1992). Accelerating the development of formal thinking in middle and high school students II: Postproject effects on science achievement. *J. Res. Sci. Teach.*, *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.

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology*, *11*(3), 401–402.

Siler, S. A., & Klahr, D. (2012). Detecting, classifying and remediating: 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.

Siler, S. A., Klahr, D., & Price, N. (2013). Investigating the mechanisms of learning from a constrained preparation for future learning activity. *Instr Sci*, *41*(1), 191–216.

Simonyi, K. (2001). *Kulturgeschichte der Physik: Von den Anfängen bis heute* (3rd ed.). Frankfurt am Main: H. Deutsch.

Sirin, S. R. (2004). *The relationship between socioeconomic status and school outcomes: Meta analytic review of research 1990-2000* (Dissertation). Boston College, Boston.

Skender, S. (2014). *Reliabilitäts- und Validitätsuntersuchungen zum Kognitiven Fähigkeitstest KFT 4-12+ R an einer Stichprobe von Fünft- und Siebtklässlern der Willy-Brandt-Gesamtschule München: Reliability an validity of the kft test*. Rostock: university of Rostock.

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. (2005). Theory of mind-the case for conceptual development. In W. Schneider, R. Schumann-Hengsteler, & B. Sodian (Eds.), *Young children's cognitive development. Interrelationships among executive functioning, working memory, verbal ability, and theory of mind* (pp. 95–132). Mahwah, N.J.: L. Erlbaum Associates.

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. In M. Prenzel (Ed.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (pp. 147–160). Münster [u.a.]: Waxmann.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*(4), 753–766.

Song, J., & Black, P. J. (1992). The effects of concept requirements and task contexts on pupils' performance in control of variables. *International Journal of Science Education*, *14*(1), 83–93.

Staver, J. R. (1984). Effects of method and format on subjects' responses to a control of variables reasoning problem. *J. Res. Sci. Teach.*, *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. *J. Res. Sci. Teach.*, *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.

Strawitz, B. M. (1984). Cognitive style and the acquisition and transfer of the ability to control variables. *J. Res. Sci. Teach.*, *21*(2), 133–141.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Sweller, J., & Chandler, P. (1994). Why Some Material Is Difficult to Learn. *Cognition and Instruction*, *12*(3), 185–233.

Tesch, M. (2005). *Das Experiment im Physikunterricht: Didaktische Konzepte und Ergebnisse einer Videostudie [The experiment in physic classes results of a video study]*. Berlin: Logos-Verlag.

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.

Tobin, K. (1990). Research on science laboratory activities: In pursuit of better questions and answers to improve learning. *School Science and Mathematics*, *90*(5), 403–418.

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.

Triona, L. M., & Klahr, D. (2007). Hands-on science: Does it matter what students' hands are on? *The Science Education Review*, *6*, 126–130.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*(11), 1–10.

Weizsäcker, C. F. von. (1951). *Zum Weltbild der Physik* (5th ed.). Stuttgart: Hirzel.

Wellnitz, N., Fischer, H. E., Kauertz, A., Mayer, J., Neumann, I., Pant Hans Arne, & Walpuski, M. (2012). Evaluation der Bildungsstandards - eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung [Evaluation of National Educational Standards – an interdisciplinary test design for the competence area acquirement of knowledge]. *Zeitschrift für Didaktik der Naturwissenschaften (ZfDN)*, *18*, 261–291.

Wollman, W. T., & Chen, B. (1982). Effects of structured social interaction on learning to control variables: A classroom training study. *Sci. Ed.*, *66*(5), 717–730.

Woodward, J. (2003). Experimentation, causal inference and instrumental realism. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 87–118). Pittsburgh, Pa.: University of Pittsburgh Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press/University of Chicago.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design. Rasch Measurement*. Chicago: MESA Press/University of Chicago.

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. J. Feist & M. E. Gorman (Eds.), *Handbook of the psychology of science* (pp. 49–70). New York: Springer Publishing Company.

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. (2012). Explicit teaching of metastrategic knowledge: Definitions, students' learning, and teachers' professional development. In A. Zohar & Y. Dori (Eds.), *Contemporary trends and issues in science education: v. 40. Metacognition in science education. Trends in current research* (pp. 197–223). Dordrecht, New York: Springer.

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

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang Publikation 1

Appendix A: List of studies included in the meta-analysis

Note. Studies marked with an R are studies included in Ross (1988a) original meta-analysis and in this meta-analysis. Studies marked with an * are studies meeting the inclusion criteria but either the entire study was excluded because of outlying effect sizes, or include one or more pairwise comparisons that were excluded because of outlying effect sizes. Studies marked with an R2 are studies that did not meeting our inclusion criteria but which were used to re-analyze Ross's (1988a) sample.

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.

R2 Armstrong, A. (1983). *Earth science instruction as a factor in enhancing the development of formal reasoning patterns with transitional subjets* (Dissertation). University of New York, Albany.

Babai, R., & Levit-Dori, T. (2009). Several CASE lessons can improve students' control of variables reasoning scheme ability. *Journal of Science Education and Technology*, *18*(5), 439–446.

Beishuizen, J., Wilhelm, P., & Schimmel, M. (2004). Computer-supported inquiry learning: effects of training and practice. *Computers & Education*, *42*(4), 389–402.

Bitner, B. L. (1990). *Thinking processes model: Effect on logical reasoning abilities of students in grade six through twelve*. Retrieved from http://www.eric.ed.gov/PDFS/ED322009.pdf

Black, R. W. (1980). *An assessment of the effects of the use of the Gagne teaching model on cognitive performance and development in the Piagetian interpretation.* Temple University.

R2 Bluhm, W. J. (1979). The effects of science process skill instruction on preservice elementary teachers' knowledge of, ability to use, and ability to sequence science process skills *Journal of Research in Science Teaching*, *16*(5), 427–432.

R Bowyer, J., Chen, B., & Their, H. D. (1978). A free-choice environment: Learning without instruction. *Science Education*, *62*(1), 95–107.

R Bredderman, T. A. (1973). The effects of training on the development of the ability to control variables. *Journal of Research in Science Teaching*, *10*(3), 189–200.

R Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology*, *6*(4), 544–574.

R*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.

R 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.

R de Ribaupierre, A. (1975). *Mental space and formal operations* (Dissertation). University of Toronto.

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.

[R2] Denson, D. (1986). *The relationship betweeen cognitive styles, methods of instruction, knowledge and process skills of college chemistry students* (Dissertation). University of Southern Mississippi, Hattiesburg.

[R2] Doty, L. C. (1985). *A study comparing the influence of inquiry and traditional science instruction methods on science achievement, attitudes toward science, and integrated process skills in ninth grade students and the relationship between sex, race, past performance in science, intelligence and achievement* (Dissertation). University of Southern Mississippi, Hattiesburg.

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.

Hall, J. F. (1972). *The use of history of science case studies with first year education students to teach skills involved in scientific thinking* (Dissertation). Oregon State University.

[R] Howe, A. C., & Mierzwa, J. (1977). Promoting the development of logical thinking in the classroom. *Journal of Research in Science Teaching*, *14*(5), 467–472.

Huber, M. (2010). *Einführung in die experimentelle Methode. Ein Vergleich zwischen »Guided Scientific Inquiry« und »Direkter Instruktion in der Jahrgangsstufe 8 der Realschule [Introduction into the experimental method. A comparison of guided scientific inquiry and direct instruction in eight grade classes]* (Zulassungsarbeit zur ersten Staatsprüfung für das Lehramt an Realschulen in Bayern nach der LPO I). Universität Regensburg, Regensburg.

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.

Hushman, C. (2011). *Examing the influence of instuctional strategy on student learning and self-efficacy in science* (Dissertation). University of New Mexico, Albuquerque, New Mexico.

Iqbal, H. M., & Shayer, M. (2000). Accelerating the development of formal thinking in Pakistan secondary school students: Achievement effects and professional development issues. *Journal of Research in Science Teaching*, *37*(3), 259–274.

Janoschek, K. (2009). *Empirische Studie zum kumulativen Kompetenzaufbau des Experimentierens mit lebenden Tieren (Asseln) [Empirical studies about the cumulative acquisition of competences when experimenting with living woodlouse]* (Diplomarbeit). Universität Wien, Wien.

[R2] Jaus, H. H. (1975). The effects of integrated science process skill instruction on changing teacher achievement and planning practices. *Journal of Research in Science Teaching*, *12*(4), 439–447.

Kallio, E. (1998). *Training of students' scientific reasoning skills*. Dissertation (Jyväskylä studies in education, psychology and social research No. 139). Jyväskylä, Finland. Retrieved from https://jyx.jyu.fi/dspace/bitstream/handle/123456789/13385/9513912922.pdf?sequence=1

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., & 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.

[R] Klausmeier, H. J., & Sipple, T. S. (1980). *Learning and teaching concepts: A strategy for testing applications of theory*. New York: Academic Press.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction, 18*(4), 495–523.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*(11), 866–870.

[R]*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.

[R2] Leising, R. (1986). *Investigation of the relationship between personability type and selected teaching strategies in developing students' science process ability, logical thinking ability and science achievement* (Dissertation). University of Michigan.

[R] Lewis, N. (1986). *A study of the effects of concrete experiences on the problem-solving ability of tenth-grade students* (Dissertation). University of Michigan, Hattiesburg.

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.

[R] Linn, M. C. (1978). Influence of cognitive style and training on tasks requiring the separation of variables schema. *Child Development*, *49*, 874–877.

[R] Linn, M. C. (1980a). Free-choice experiences: How do they help children learn? *Science Education*, *64*(2), 237–248.

[R] Linn, M. C. (1980b). Teaching students to control variables: Some investigations using free choice experiences. In S. Modgil & C. Modgil (Eds.), *Toward a theory of psychological development within a Piagetian framework* (pp. pp. 673-697). London: National foundation for Educational Research.

[R] Linn, M. C., Chen, B., & Thier, H. D. (1976). Personalization in science: Preliminary investigation at the middle school level. *Instructional Science*, *5*(3), 227–251.

[R2] Linn, M. C., Chen, B., & Thier, H. D. (1977). Teaching children to control variables: Investigation of a free-choice environment. *Journal of Research in Science Teaching*, *14*(3), 249–255.

Linn, M. C., Clement, C., Pulos, S., & Sullivan, P. (1989). Scientific reasoning during adolescence: The influence of instruction in science knowledge and reasoning strategies. *Journal of Research in Science Teaching*, *26*(2), 171–187.

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.

Lucian, W. (2012). *Direct instruction vs task structuring Are they equally effective to teach the control of variables strategy?* (Masterarbeit). University of Twente, Department of Instructional Technology, Enschede.

Lugeder, J. (2010). *Experimentieren in der 5. Jahrgangsstufe der Realschule. Ein Vergleich zwischen eigenverantwortlicher Planung mit Hilfe von Inquiry Boards und Kochrezept [Experimenting with fifth grade students. A comparison of experiments planned by students with the support of inquiry boards and cookbook experiments]* (Zulassungsarbeit zur ersten Staatsprüfung für das Lehramt an Realschulen in Bayern nach der LPO I). Universität Regensburg, Regensburg.

Marschner, J. (2011). *Adaptives Feedback zur Unterstützung des selbstregulierten Lernens durch Experimentieren [Supporting self-regulated learning from experiments with adaptive feedback]* (Dissertation). Universität Duisburg-Essen, Essen.

Matlen, B. J., & Klahr, D. (2010). Sequential effects of high and low guidance on children's early science learning. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceed-*

ings of the 9th International Conference of the Learning Sciences (ICLS 2010) (pp. 1019–1023). Chicago: International Society of the Learning Sciences.

[R] Mazzei, A. L., Fogli-Muciaccia, M. T., & Picciarelli, V. (1986). Effect of a Piagetian mini-laboratory on the acquisition of typical formal operational schemata. *European Journal of Science Education, 8*(1), 87–93.

McCormack, L. (2009). *Cognitive acceleration across the primary-second level transition* (Dissertation). Dublin City University, Dublin.

[R2] McKinnon, J. W. (1971). Are colleges concerned with intellectual development? *American Journal of Physics, 39*(9), 1047.

[R] Novak, J. D., Lawson, A. E., Blake, A. J. D., & Nordland, F. H. (1975). Training effects and generalization of the ability to control variables in high school biology students. *Science Education, 59*(3), 387–396.

[R2] 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.

[R]Peterson, K. (1977). *An experimental evaluation of a science inquiry training program for high school students* (Dissertation). University of California, Berkeley.

[R2] Purser, R. K., & Renner, J. W. (1983). Results of two tenth-grade biology teaching procedures. *Science Education, 67*(1), 85–98.

[R2] Rivers, R. H., & Vockell, E. (1987). Computer simulations to stimulate scientific problem solving. *Journal of Research in Science Teaching, 24*(5), 403–415.

Rösch, F., Rieß, W., & Nerb, J. (2012). Förderung "experimenteller Problemlösefähigkeit" im problemorientierten Ökologieunterricht der 6. Klasse - Teilprojekt 2 [Supporting experimental problem solving skills in problem orientated ecology classes]. In W. Rieß, M. Wirtz, B. Barzel, & A. Schulz (Eds.), *Experimentieren im mathematisch-naturwissenschaftlichen Unterricht. Schüler lernen wissenschaftlich denken und arbeiten* (pp. 183–198). Münster, New York, NY, München, Berlin: Waxmann.

*Rosenthal, D. (1979). The acquisition of formal operations: The effect of two training procedures. *Journal of Genetic Psychology, 134*, 125–140.

[R]*Ross, J. A. (1986). Cows moo softly: Acquiring and retrieving a formal operations schema. *European Journal of Science Education, 8*(4), 389–397.

[R]*Ross, J. A. (1988). Improving social-environmental studies problem solving through cooperative learning. *American Educational Research Journal, 25*(4), 573–591.

[R] Ross, J. A. (1990). Learning to control variables: Main effects and aptitude treatment interactions of two rule-governed approaches to instruction. *Journal of Research in Science Teaching, 27*(6), 523–539.

[R2] Ross, J. A., & Maynes, F. J. (1983a). Development of a test of experimental problem-solving skills. *Journal of Research in Science Teaching, 20*(1), 63–75.

[R2] Ross, J. A., & Maynes, F. J. (1983b). Experimental problem solving: An instructional improvement field experiment. *Journal of Research in Science Teaching, 20*(6), 543–556.

Ross, R. J., Hubbell, C., Ross, C. G., & Thompson, M. B. (1976). The training and transfer of formal thinking tasks in college students. *Genetic Psychology Monographs, 93*(2), 171–187.

[R] Rowell, J. A., & Dawson, C. J. (1984). Controlling variables: Testing a programme for teaching a general solution strategy. *Research in Science & Technological Education, 2*(1), 37–46.

[R] Rowell, J. A., & Dawson, C. J. (1985). Equilibration, conflict and instruction: A new class-oriented perspective. *European Journal of Science Education, 7*(4), 331–344.

Sao Pedro, M. A., Gober, J. D., & Baker, R. S. J. (2012). *Assessing the learning and transfer of data collection inquiry skills using educational data mining on students' log files*. Retrieved from http://users.wpi.edu/~rsbaker/SaoPedroetal_AERA2012_FINAL.pdf

Sao Pedro, M. A., Gober, J. D., Heffman, N. T., & Beck, J. E. (2009). Comparing pedagogical approaches for teaching the control of variables strategy. In N. Taatgen & vanRijn H. (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1294–1299). Austin.

Sao Pedro, M. A., Gober, J. D., & Raziuddin, J. J. (2010). Comparing pedagogical approaches for the acquisition and long-term robustness of the control of variables strategy. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)* (pp. 1024–1031). Chicago: International Society of the Learning Sciences.

[R2] Saunders, W. L., & Shepardson, D. (1987). A comparison of concrete and formal science instruction upon science achievement and reasoning ability of sixth grade students. *Journal of Research in Science Teaching*, *24*(1), 39–51.

[R2] Schneider, L. S., & Renner, J. W. (1980). Concrete and formal teaching. *Journal of Research in Science Teaching*, *17*(6), 503–517.

[R] Shaw, T. J. (1983). The effect of a process-oriented science curriculum upon problem-solving ability. *Science Education*, *67*(5), 615–623.

[R] Sneider, C., Kurlich, K., Pulos, S., & Friedman, A. (1984). Learning to control variables with model rockets: A neo-piagetian study of learning in field settings. *Science Education*, *68*(4), 465–486.

Sodian, B., Jonen, A., Thoermer, C., & Kircher, E. (2006). Die Natur der Naturwissenschaften verstehen [Understanding the nature of science]: Implementierung wissenschaftstheoretischen Unterrichts in der Grundschule. In M. Prenzel (Ed.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (pp. 147–159). Münster [u.a.]: Waxmann.

Sodian, B., Thoermer, C., Kircher, E., Grygier, P., & Günther, J. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule [Teaching epistemological understanding in elemantary schools]. *Zeitschrift für Pädagogik*, 192–206.

[R] Stone, C. A., & Day, M. C. (1978). Levels of availability of a formal operational strategy. *Child Development*, *49*(4), 1054.

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, *23*(4), 488–511.

[R]*Strawitz, b. M. (1984). Cognitive style and the acquisition and transfer of the ability to control variables. *Journal of Research in Science Teaching*, *21*(2), 133–141.

[R2]Strawitz, b. M. (1993). The effects of review on science process skill acquisition. *Journal of Science Teacher Education*, *4*(2), 54–57.

[R2] Strawitz, b. M., & Malone, M. R. (1987). Preservice teachers' acquisition and retention of integrated science process skills: A comparison of teacher-directed and self-instructional strategies. *Journal of Research in Science Teaching*, *24*(1), 53–60.

[R2] Thomas, W. E., & Grouws, D. A. (1984). Inducing cognitive growth in concrete-operational college students. *School Science and Mathematics*, *84*(3), 233–243

[R]*Tomlinson-Keasey, C. (1972). Formal operations in females from eleven to fifty-four years of age. *Developmental Psychology*, *6*(2), 364.

[R2] Wilson, A. H. (1987). Teaching use of formal thought for improved chemistry achievement. *International Journal of Science Education*, *9*(2), 197–202.

[R] 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.

Zhang, J., Chen, Q., Sun, Y., & Reid, D. J. (2004). Triple scheme of learning support design for scientific discovery learning based on computer simulation: Experimental research. *Journal of Computer Assisted Learning*, *20*(4), 269–282.

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.

*Zion, M., Michalsky, T., & Mevarech, Z. R. (2005). The effects of metacognitive instruction embedded within an asynchronous learning network on scientific inquiry skills. *International Journal of Science Education, 27*(8), 957–983.

Zohar, A., & Aharon-Kravetsky, S., (2005). Exploring the effects of cognitive conflict and direct teaching for students of different academic levels. *Journal of Research in Science Teaching, 42(7), 829-855.*

*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.

Appendix B: Effect sizes and key characteristics of studies included in the meta-analysis

Note. Studies are ordered by the mean effect size of the pairwise comparisons extracted from the same study. Within one study the pairwise comparisons are ordered by their effect sizes from the small to large. Pairwise comparisons marked with an R are included in Ross (1988a) original meta-analysis. Pairwise comparisons marked with an * are pairwise comparisons meeting the inclusion criteria but excluded because of outlying effect sizes.

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Lugeder, 2010	-0.2	43	10.7	90	No	Yes	Yes	No	No	MC
Lugeder, 2010	0.01	43	10.7	90	No	Yes	Yes	No	No	MC
Marschner et al, 2012	-0.34	52	14.62	90	Yes	Yes	Yes	No	No	Virtual
Marschner et al, 2012	0.15	45	14.62	90	Yes	Yes	Yes	No	No	Virtual
Marschner et al, 2012	0.21	50	14.62	90	Yes	Yes	Yes	No	No	Virtual
Huber, 2010	-0.13	52	13.61	135	Yes	Yes	No	No	No	Open
Huber, 2010	0.15	52	13.61	135	Yes	Yes	No	No	No	Open
Beishuizen et al, 2004	-0.02	62	11.32	60	No	Yes	Yes	No	No	Virtual
Beishuizen et al, 2004	0.04	62	11.32	60	No	Yes	Yes	No	No	Virtual
Sao Pedro et al, 2010	-0.25	35	12.64	180	No	Yes	Yes	No	Yes	MC
Sao Pedro et al, 2010	-0.25	35	12.64	180	No	Yes	Yes	No	Yes	MC
Sao Pedro et al, 2010	-0.02	31	12.64	180	No	Yes	Yes	No	Yes	MC
Sao Pedro et al, 2010	0.29	41	12.64	180	No	Yes	Yes	No	Yes	MC
Sao Pedro et al, 2010	0.29	41	12.64	180	No	Yes	Yes	No	Yes	MC
Sao Pedro et al, 2010	0.46	20	12.64	180	No	Yes	Yes	No	Yes	Open
Linn, 1980^{R}	-0.68	39	11.67	840	No	Yes	Yes	No	NA	Real
Linn, 1980^{R}	0.15	39	11.67	840	No	Yes	Yes	No	NA	Real
Linn, 1980^{R}	0.21	40	11.67	120	No	No	Yes	No	NA	Real
Linn, 1980^{R}	0.71	40	11.67	120	No	No	Yes	No	NA	Real
Hall, 1972	0.02	31	19.5	1440	Yes	Yes	No	No	No	MC
Hall, 1972	0.22	260	19.5	1440	Yes	Yes	No	No	No	MC
Goossens et al, 1987	-0.23	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	-0.12	28	13.72	45	No	No	No	No	No	NA

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Goossens et al, 1987	-0.11	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	-0.07	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.02	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.05	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.07	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.1	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.19	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.19	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.22	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.26	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.33	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.35	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.39	28	13.72	45	No	No	No	No	No	NA
Goossens et al, 1987	0.42	28	13.72	45	No	No	No	No	No	NA
Adey & Shayer, 1990	0.01	164	12	NA	Yes	Yes	No	Yes	Yes	Real
Adey & Shayer, 1990	0.09	48	12	NA	Yes	Yes	No	Yes	Yes	Real
Adey & Shayer, 1990	0.29	234	12.4	NA	Yes	Yes	No	Yes	Yes	Real
McCormack, 2009	0.2	64	11.67	900	Yes	Yes	Yes	Yes	Yes	Open
Zohar & Aharon-Kravetsky, 2005	-0.32	67	14.45	180	No	Yes	Yes	No	NA	Virtual
Zohar & Aharon-Kravetsky, 2005	0.76	54	14.45	180	No	Yes	Yes	No	NA	Virtual
Rösch et al, 2012	0.16	220	11.9	585	Yes	Yes	Yes	No	Yes	Open

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Rösch et al, 2012	0.22	213	11.9	585	Yes	Yes	Yes	No	Yes	Open
Rösch et al, 2012	0.24	211	11.9	585	Yes	Yes	Yes	No	Yes	MC
Rösch et al, 2012	0.36	218	11.9	585	Yes	Yes	Yes	No	Yes	MC
Huppert et al, 2002	0.07	64	15.54	NA	Yes	Yes	No	No	NA	MC
Huppert et al, 2002	0.31	41	15.54	NA	Yes	Yes	No	No	NA	MC
Huppert et al, 2002	0.39	76	15.54	NA	Yes	Yes	No	No	NA	MC
Kuhn et al, 2000	0.28	42	12.5	750	No	Yes	No	No	Yes	Virtual
Kuhn et al, 2000	0.28	42	12.5	750	No	Yes	No	No	Yes	Virtual
Day & Stone, 1982	-0.44	24	11.8	40	No	Yes	No	No	No	Real
Day & Stone, 1982	0.05	24	13.9	40	No	Yes	No	No	No	Real
Day & Stone, 1982	0.11	24	11.8	40	No	Yes	No	No	Yes	Real
Day & Stone, 1982	0.34	24	11.8	40	No	Yes	No	No	No	Real
Day & Stone, 1982	0.61	24	13.9	40	No	Yes	No	No	No	Real
Day & Stone, 1982	1.02	24	13.9	40	No	Yes	No	No	No	Real
Sao Pedro et al, 2009	-0.17	85	13.12	180	No	Yes	Yes	No	No	MC
Sao Pedro et al, 2009	0.32	88	13.12	180	No	Yes	Yes	No	Yes	MC
Sao Pedro et al, 2009	0.45	85	13.12	180	No	Yes	Yes	No	No	virtual
Sao Pedro et al, 2009	0.53	88	13.12	180	No	Yes	Yes	No	Yes	Virtual
Chen & Klahr, 1999	-0.21	14	7.97	80	No	Yes	No	No	Yes	Real
Chen & Klahr, 1999	0.08	23	9.81	80	No	Yes	No	No	Yes	Real
Chen & Klahr, 1999	0.31	20	9.08	80	No	Yes	No	No	Yes	Real
Chen & Klahr, 1999	0.36	15	8	80	No	Yes	Yes	No	Yes	Real
Chen & Klahr, 1999	0.63	19	9.08	80	No	Yes	Yes	No	Yes	Real

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Chen & Klahr, 1999	0.88	23	9.81	80	No	Yes	Yes	No	Yes	Real
Novak et al, 1975[R]	0.02	65	15.4	200	No	Yes	Yes	No	Yes	Real
Novak et al, 1975[R]	0.67	65	15.4	200	No	Yes	Yes	No	Yes	Real
de Rib-aupierre, 1975[R]	0.16	29	12.64	90	Yes	Yes	No	No	Yes	Real
de Rib-aupierre, 1975[R]	0.54	24	15.16	90	No	Yes	No	No	Yes	Real
Greenbowe et al, 1981	-0.22	30	10.7	45	No	Yes	Yes	No	No	Real
Greenbowe et al, 1981	0.92	31	10.7	45	No	Yes	Yes	No	No	Real
Linn, 1980[R]	0.15	38	12.64	420	No	Yes	Yes	No	Yes	Real
Linn, 1980[R]	0.21	39	12.64	120	No	No	Yes	No	Yes	Real
Linn, 1980[R]	0.27	38	12.64	420	No	Yes	Yes	No	Yes	Real
Linn, 1980[R]	0.3	40	12.64	120	No	Yes	Yes	Yes	Yes	Real
Linn, 1980[R]	0.42	40	12.64	120	No	Yes	Yes	Yes	Yes	Real
Linn, 1980[R]	0.45	40	12.64	840	No	Yes	Yes	Yes	Yes	Real
Linn, 1980[R]	0.46	40	12.64	840	No	Yes	Yes	Yes	Yes	Real
Linn, 1980[R]	0.63	39	12.64	120	No	No	Yes	No	Yes	Real
Dean & Kuhn, 2007	-0.5	29	9.5	720	No	Yes	Yes	No	No	Virtual
Dean & Kuhn, 2007	-0.04	29	9.5	720	No	Yes	Yes	No	No	Virtual
Dean & Kuhn, 2007	0.75	44	9.5	45	No	No	Yes	No	No	Open
Dean & Kuhn, 2007	0.79	29	9.5	495	No	Yes	Yes	No	No	Virtual
Dean & Kuhn, 2007	0.87	29	9.5	495	No	Yes	Yes	No	No	Virtual
Janoschek, 2009	0.39	129	12.5	240	Yes	Yes	Yes	Yes	Yes	MC
Sao Pedro et al, 2010	0.4	147	13	90	Yes	Yes	No	No	No	Virtual
Linn et al, 1989	0.2	61	14.5	40	No	Yes	Yes	No	No	NA
Linn et al, 1989	0.29	61	14.5	40	No	Yes	Yes	No	No	NA
Linn et al, 1989	0.37	61	14.5	40	No	Yes	Yes	No	No	Real
Linn et al, 1989	0.79	61	14.5	40	No	Yes	Yes	No	No	Real
Klausmeier & Sipple, 1980[R]	-0.2	106	10.7	360	No	Yes	Yes	No	No	MC
Klausmeier & Sipple, 1980[R]	0.46	106	10.7	175	No	No	Yes	No	No	MC
Klausmeier & Sipple, 1980[R]	0.69	127	10.7	NA	No	Yes	Yes	No	No	MC

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Klausmeier & Sipple, 1980[R]	0.79	127	10.7	175	No	No	Yes	No	No	MC
Sodian et al, 2002	0.4	35	10.5	450	Yes	Yes	Yes	Yes	Yes	Open
Sodian et al, 2002	0.43	35	10.5	450	Yes	Yes	Yes	Yes	Yes	MC
Sodian et al, 2002	0.58	35	10.5	450	Yes	Yes	Yes	Yes	Yes	Open
Ross et al, 1976	0.05	29	21.8	60	No	No	Yes	No	Yes	Real
Ross et al, 1976	0.09	28	21.8	60	No	Yes	No	No	Yes	Real
Ross et al, 1976	0.25	28	21.8	60	No	Yes	No	Yes	Yes	Real
Ross et al, 1976	0.61	28	21.8	60	No	Yes	No	Yes	Yes	Real
Ross et al, 1976	0.61	28	21.8	60	No	Yes	No	No	Yes	Real
Ross et al, 1976	1.29	29	21.8	60	No	No	Yes	No	Yes	Real
Linn et al, 1976[R]	0.27	54	10.7	NA	Yes	Yes	Yes	No	NA	Real
Linn et al, 1976[R]	0.6	103	10.7	810	Yes	Yes	Yes	No	No	Open
Linn et al, 1976[R]	0.62	103	10.7	810	Yes	Yes	Yes	No	No	Real
Lin & Lehman, 1999	0.06	46	22	2100	Yes	Yes	No	No	NA	Open
Lin & Lehman, 1999	0.09	45	22	2100	Yes	Yes	No	No	NA	Open
Lin & Lehman, 1999	0.39	46	22	2100	Yes	Yes	No	No	NA	Open
Lin & Lehman, 1999	0.58	45	22	2100	Yes	Yes	No	No	NA	Open
Lin & Lehman, 1999	0.59	45	22	2100	Yes	Yes	No	No	NA	Open
Lin & Lehman, 1999	1.39	45	22	2100	Yes	Yes	No	No	NA	Open
Zimmerman et al, 2003	0.53	14	11.67	NA	Yes	Yes	NA	No	NA	Real
Ross, 1990[R]	-0.04	121	10.61	360	No	Yes	Yes	No	Yes	Open
Ross, 1990[R]	0.53	150	10.61	360	No	Yes	No	No	Yes	Open
Ross, 1990[R]	0.8	121	10.61	360	No	Yes	Yes	No	Yes	Open
Ross, 1990[R]	0.86	150	10.61	360	No	Yes	No	No	Yes	Open
Stone & Day, 1978[R]	0.28	28	10.7	60	No	Yes	Yes	No	Yes	Real
Stone & Day, 1978[R]	0.61	28	12.64	60	No	Yes	Yes	No	Yes	Real
Stone & Day, 1978[R]	0.8	28	8.76	60	No	Yes	Yes	No	Yes	Real
Keselman, 2003	0.45	46	11.67	420	Yes	Yes	Yes	No	Yes	Virtual

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Keselman, 2003	0.45	46	11.67	420	Yes	Yes	Yes	No	Yes	Open
Keselman, 2003	0.45	46	11.67	420	Yes	Yes	No	No	No	Virtual
Keselman, 2003	0.62	46	11.67	420	Yes	Yes	Yes	No	Yes	Virtual
Keselman, 2003	0.66	46	11.67	420	Yes	Yes	No	No	No	Open
Keselman, 2003	0.85	46	11.67	420	Yes	Yes	No	No	No	Virtual
Rowell & Dawson, 1984[R]	0.61	83	13.61	150	No	Yes	Yes	No	Yes	Open
Strand-Cary & Klahr, 2008	-0.48	18	8.8	45	No	Yes	Yes	Yes	Yes	Open
Strand-Cary & Klahr, 2008	-0.19	20	9.7	45	No	Yes	Yes	Yes	Yes	Open
Strand-Cary & Klahr, 2008	0.39	20	9.7	45	No	Yes	Yes	Yes	Yes	Real
Strand-Cary & Klahr, 2008	0.71	23	11	45	No	Yes	Yes	Yes	Yes	Open
Strand-Cary & Klahr, 2008	1.28	18	8.8	45	No	Yes	Yes	Yes	Yes	Real
Strand-Cary & Klahr, 2008	2	23	11	45	No	Yes	Yes	Yes	Yes	Real
Zhang et al, 2004	0.62	30	13	40	Yes	Yes	Yes	No	Yes	Open
Sodian et al, 2006	0.62	48	9.73	1260	Yes	Yes	Yes	No	Yes	Open
Shaw, 1983[R]	0.64	83	11.67	NA	Yes	Yes	No	No	Yes	MC
Lorch et al, 2010	0.42	233	9.73	40	No	No	Yes	Yes	Yes	MC
Lorch et al, 2010	0.45	207	9.73	40	No	No	Yes	Yes	Yes	MC
Lorch et al, 2010	0.56	224	9.73	70	No	Yes	Yes	Yes	Yes	MC
Lorch et al, 2010	0.59	224	9.73	70	No	Yes	Yes	Yes	Yes	Real
Lorch et al, 2010	0.82	222	9.73	70	No	Yes	Yes	Yes	Yes	Real
Lorch et al, 2010	1.01	222	9.73	70	No	Yes	Yes	Yes	Yes	MC
Lucian, 2012	0.53	35	10.66	35	Yes	Yes	Yes	No	Yes	Virtual
Lucian, 2012	0.78	33	10.66	25	Yes	Yes	No	No	No	Virtual
Lazarowitz & Huppert, 1993	0.67	181	15.54	540	Yes	Yes	NA	No	Yes	MC

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Novak et al, 1978[R]	0.11	80	11.67	1080	No	Yes	No	No	No	Open
Novak et al, 1978[R]	1.27	70	11.67	1080	No	Yes	No	No	No	Real
Grygier, 2008	0.64	53	9.73	810	Yes	Yes	No	No	Yes	Open
Grygier, 2008	0.77	53	9.73	810	Yes	Yes	No	No	Yes	MC
Babai & Levit-Dori, 2009	0.71	120	14.58	270	Yes	Yes	No	Yes	No	Open
Klahr & Ni-gam, 2004	0.71	104	9.5	40	No	Yes	Yes	No	Yes	Real
Bredderman, 1973[R]	0.69	18	11.19	640	No	Yes	No	Yes	Yes	Real
Bredderman, 1973[R]	0.79	18	11.19	160	No	Yes	Yes	No	No	Real
Ford, 2005	0.75	39	11.67	360	No	Yes	No	No	Yes	MC
Bitner, 1990	0.77	270	NA	NA	Yes	No	No	Yes	NA	Open
Rowell & Dawson, 1985[R]	0.73	52	13.61	210	No	No	No	Yes	NA	Open
Rowell & Dawson, 1985[R]	0.82	56	13.61	210	No	No	No	Yes	NA	Open
Linn, 1978[R]	0.4	40	12	NA	No	Yes	Yes	No	Yes	Real
Linn, 1978[R]	1.21	40	12	90	No	No	Yes	No	Yes	Real
Zion et al, 2005	0.43	208	16.3	NA	Yes	No	NA	No	NA	MC
Zion et al, 2005	0.65	208	16.3	NA	Yes	No	NA	No	NA	Open
Zion et al, 2005	0.85	199	16.3	NA	Yes	No	NA	No	NA	MC
***Zion et al, 2005**	1.36	199	16.3	NA	Yes	No	No	No	NA	Open
Sneider et al, 1984[R]	0.86	44	11	480	No	Yes	Yes	No	Yes	Open
Wollman & Chen, 1982[R]	0.89	83	10.7	540	No	Yes	No	Yes	Yes	Real
Case, 1974[R]	0.61	16	5.96	57	No	Yes	No	Yes	Yes	Real
Case, 1974[R]	0.61	16	5.96	57	No	Yes	No	Yes	Yes	Real
Case, 1974[R]	0.93	36	8.02	57	No	Yes	No	Yes	Yes	Real
Case, 1974[R]	1.63	36	8.02	57	No	Yes	No	Yes	Yes	Real
Iqbal & Shayer, 2000	0.97	201	11.5	NA	Yes	Yes	Yes	Yes	NA	Open
Hushman, 2011	0.67	40	9.5	60	No	Yes	Yes	No	Yes	MC
Hushman, 2011	0.73	40	9.5	60	No	Yes	Yes	No	Yes	MC
Hushman, 2011	0.8	40	9.5	60	No	Yes	Yes	No	Yes	MC
Hushman, 2011	0.95	40	9.5	60	No	Yes	Yes	No	Yes	MC
Hushman, 2011	1.1	40	9.5	60	No	Yes	Yes	No	Yes	Real
Hushman, 2011	1.68	40	9.5	60	No	Yes	Yes	No	Yes	Real
Mazzei et al, 1986[R]	0.99	43	14	180	Yes	Yes	No	No	No	Open

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Matlen & Klahr, 2010	0.35	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	0.35	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	0.35	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	0.35	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	0.72	20	9.03	100	No	Yes	Yes	No	Yes	Open
Matlen & Klahr, 2010	1.01	20	9.03	100	No	Yes	Yes	No	Yes	Open
Matlen & Klahr, 2010	1.24	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.24	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.24	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.24	20	9.03	100	No	Yes	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.31	20	9.03	100	No	No	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.31	20	9.03	100	No	No	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.31	20	9.03	100	No	No	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.31	20	9.03	100	No	No	Yes	No	Yes	Real
Matlen & Klahr, 2010	1.58	20	9.03	100	No	No	Yes	No	Yes	Open
Black, 1980	0.47	16	12.15	80	No	Yes	Yes	No	No	Real
Black, 1980	0.47	16	12.15	80	No	Yes	Yes	No	No	MC
Black, 1980	0.72	16	12.15	160	No	Yes	Yes	No	No	MC
Black, 1980	2.36	16	12.15	160	No	Yes	Yes	No	No	Real
Kallio, 1998	1.05	63	23.35	1080	No	No	Yes	Yes	Yes	Open
Kallio, 1998	1.07	67	23.35	1080	No	No	Yes	Yes	Yes	Open
Zohar & Peled, 2008	0.69	21	10.2	240	No	Yes	Yes	No	Yes	Virtual
Zohar & Peled, 2008	1.15	21	10.2	240	No	Yes	Yes	No	Yes	Real
Zohar & Peled, 2008	1.16	21	10.2	210	No	Yes	Yes	No	Yes	Virtual
Zohar & Peled, 2008	1.46	21	10.2	240	No	Yes	Yes	No	Yes	Real
Dejonckheere et al, 2011	0.71	30	11	250	Yes	No	No	No	Yes	Real

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Dejonckheere et al, 2011	1.58	30	11	250	Yes	No	No	No	Yes	Real
Strawitz, 1984[R]	0.87	56	11.3	75	No	Yes	No	No	No	Real
Strawitz, 1984[R]	1.02	56	11.3	75	No	Yes	No	No	No	Real
***Strawitz, 1984[R]**	2.16	56	11.3	75	No	Yes	No	No	No	Real
Howe & Mierzwa, 1977[R]	1.37	62	13.3	270	No	Yes	Yes	Yes	Yes	Real
Danner & Day, 1977[R]	1.44	40	17.5	60	No	Yes	Yes	No	Yes	Real
Lawson & Wollman, 1976[R]	0.73	32	10.5	120	No	Yes	Yes	Yes	Yes	Real
Lawson & Wollman, 1976[R]	0.8	32	10.5	120	No	Yes	Yes	Yes	Yes	Open
Lawson & Wollman, 1976[R]	1.18	32	12.6	120	No	Yes	Yes	Yes	Yes	Open
Lawson & Wollman, 1976[R]	1.57	32	12.6	120	No	Yes	Yes	Yes	Yes	Real
Lawson & Wollman, 1976[R]	1.97	32	12.6	120	No	Yes	Yes	Yes	Yes	Real
***Lawson & Wollman, 1976[R]**	2.55	32	10.5	120	No	Yes	Yes	Yes	Yes	Real
Rosenthal, 1979	0.91	26	11.92	135	No	Yes	No	No	Yes	Real
Rosenthal, 1979	1.01	30	11.92	135	No	Yes	No	No	Yes	Real
Rosenthal, 1979	1.22	27	11.92	135	No	Yes	Yes	Yes	Yes	Real
Rosenthal, 1979	1.48	30	11.92	135	No	Yes	No	No	Yes	Real
Rosenthal, 1979	1.93	29	11.92	135	No	Yes	Yes	Yes	Yes	Real
Rosenthal, 1979	1.94	26	11.92	135	No	Yes	No	No	Yes	Real
Rosenthal, 1979	2	29	11.92	135	No	Yes	Yes	Yes	Yes	Real
***Rosenthal, 1979**	2.47	27	11.92	135	No	Yes	Yes	Yes	Yes	Real

Study	g	Sample size	Students' Age [years]	Treatment duration [min]	Additional (non-CVS) content	Use of real or virtual training tasks	Teaching explicit CVS rule	Use of cognitive conflict	Use of demon-strations	Test format
Tomlinson-Keasey, 1972[R]	1.05	30	11.9	60	Yes	Yes	No	Yes	No	Real
***Tomlinson-Keasey, 1972[R]**	2.23	30	19.7	60	Yes	Yes	No	Yes	No	Real
Kuhn & Dean, 2005	1.74	30	11.5	540	No	Yes	No	No	No	Virtual
Lewis, 1986[R]	1.75	50	15.54	675	Yes	Yes	No	Yes	Yes	MC
***Peterson, 1977[R]**	2.16	50	NA	360	Yes	Yes	Yes	Yes	Yes	Real
***Ross, 1986[R]**	2.35	153	10.5	225	No	Yes	Yes	No	Yes	Open
Zohar & David, 2008	1.23	60	13.5	540	Yes	Yes	Yes	Yes	Yes	Open
***Zohar & David, 2008**	3.83	59	13.5	540	Yes	Yes	Yes	Yes	Yes	Open
***Case & Fry, 1973[R]**	2.98	30	14.58	480	No	Yes	No	No	Yes	Open
***Ross, 1988[R]**	2.45	186	9.73	NA	Yes	No	Yes	No	No	Open
***Ross, 1988[R]**	5.97	168	9.73	NA	Yes	No	Yes	No	NA	Open

Anhang Publikation 2

Variablenkontrolltest

Bitte trage zunächst den siebenstelligen Code ein, den du von uns erhalten hast:

Dein Code:

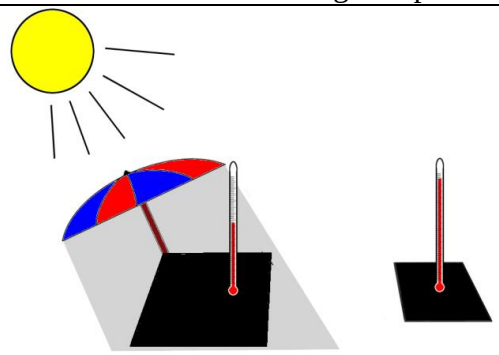
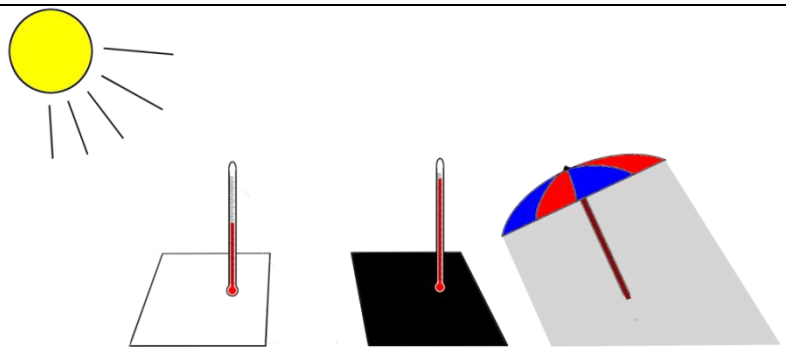
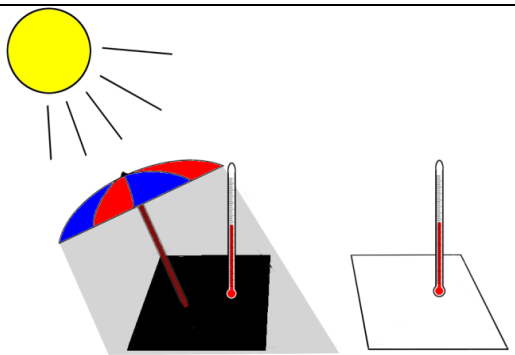
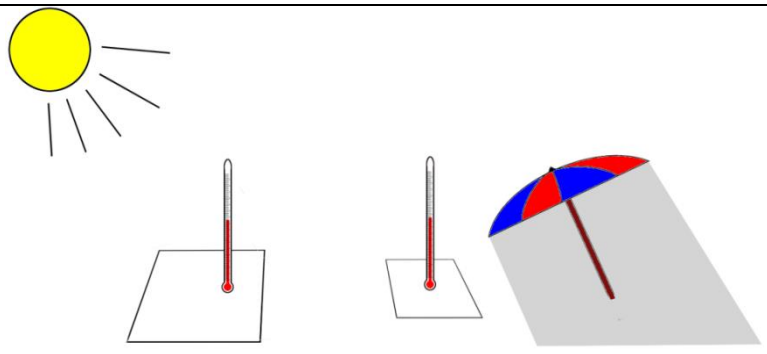
Auf der Rückseite wird dir beschrieben, wie du das Experimentierheft ausfüllen sollst.

Passende Kleidung	**ID-SO-1**

Anna möchte sich ein T-Shirt kaufen. Da die Sonne scheint und es sehr warm ist, möchte sie ein T-Shirt kaufen, in dem ihr möglichst wenig warm wird.

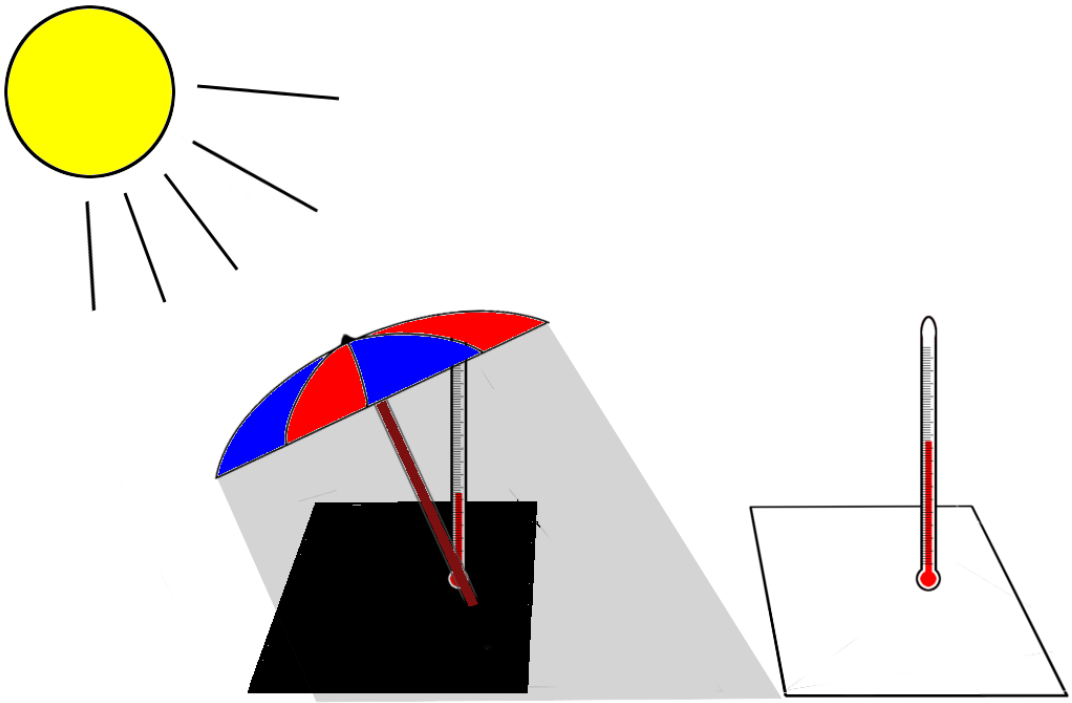
Sie vermutet, dass ihr wärmer wird, wenn sie schwarze statt weißer Kleidung trägt.

Mit welchem Experiment kann sie ihre Vermutung überprüfen?

☐	
☐	
☐	
☐	

Wirkung der Sonne	IN-SO-1

Toni hat folgendes Experiment durchgeführt:



Was zeigt dieses Experiment?

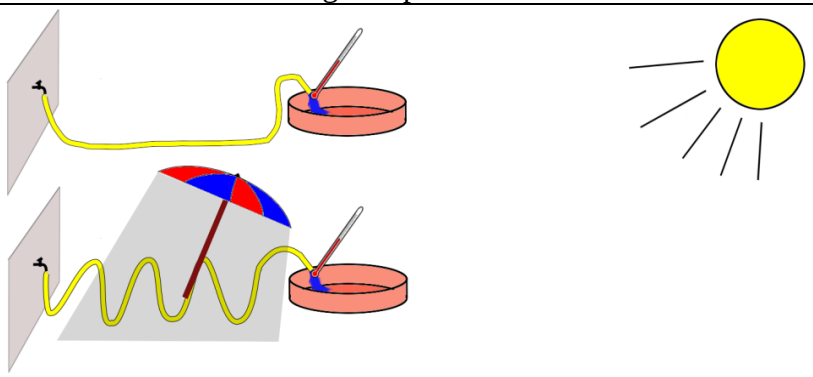
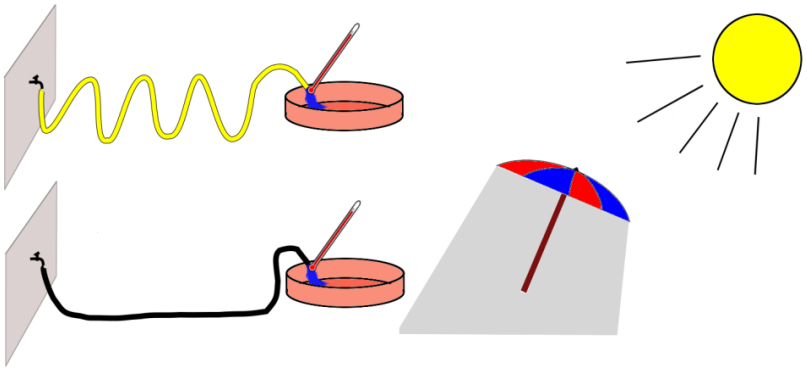
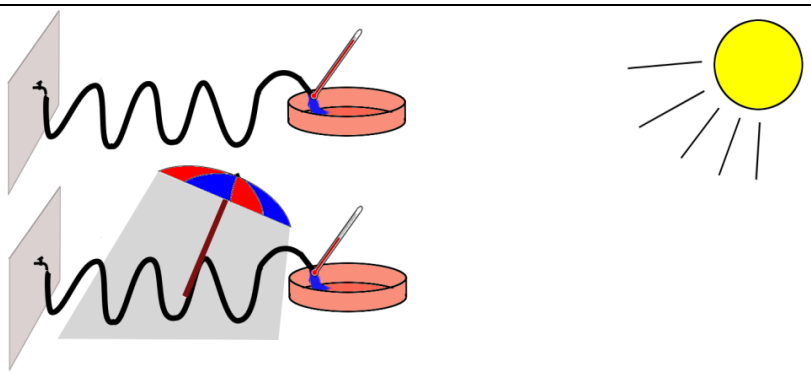
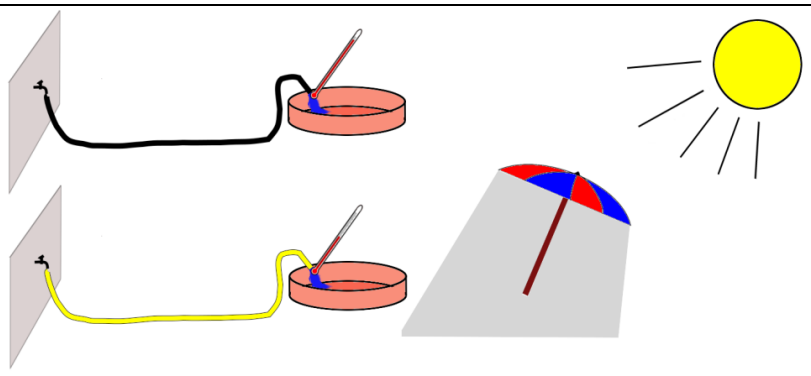
☐	Der Schatten hat einen Einfluss auf die gemessene Temperatur.
☐	Die Farbe der verwendeten Platten hat einen Einfluss auf die gemessene Temperatur.
☐	Sowohl der Schatten als auch die Farbe der Platten haben einen Einfluss auf die gemessene Temperatur.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Planschbecken	ID-SO-2

Sebastian möchte warmes Wasser in sein Planschbecken einfüllen.

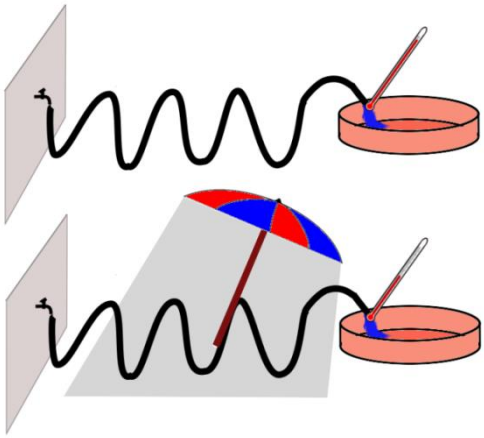
Er vermutet, dass das Wasser auf dem Weg zum Planschbecken in einem schwarzen Schlauch stärker erwärmt wird, als in einem gelben.

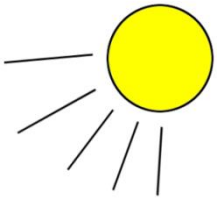
Mit welchem Experiment kann er seine Vermutung überprüfen?

☐	
☐	
☐	
☐	

Warmes Wasser	IN-SO-2

Emanuel hat folgendes Experiment durchgeführt:





Was zeigt dieses Experiment?

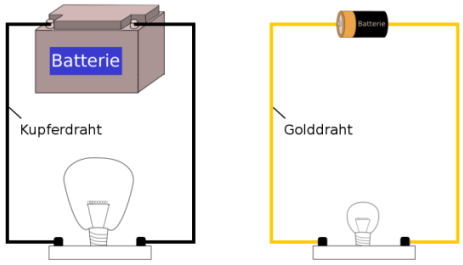
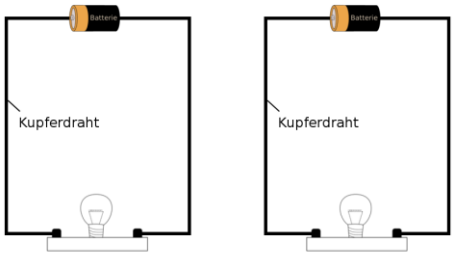
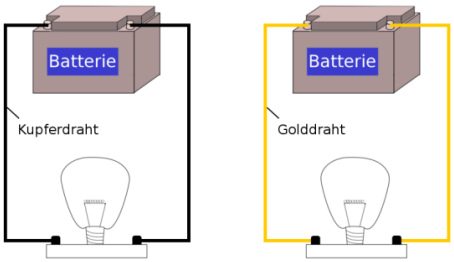
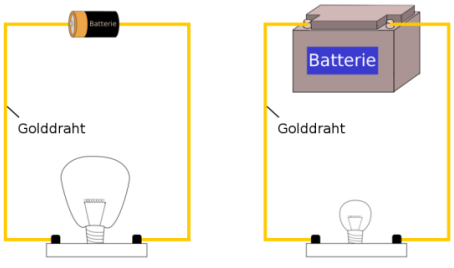
☐	Die Länge des Schlauchs hat einen Einfluss auf die Temperatur des Wassers.
☐	Schatten auf dem Schlauch hat einen Einfluss auf die Temperatur des Wassers.
☐	Sowohl der Schatten als auch die Länge des Schlauches haben einen Einfluss auf die Temperatur des Wassers.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Welcher Draht?	ID-LS-1

Peter möchte überprüfen, ob das Material eines Leiters einen Einfluss auf dessen elektrische Eigenschaften hat.

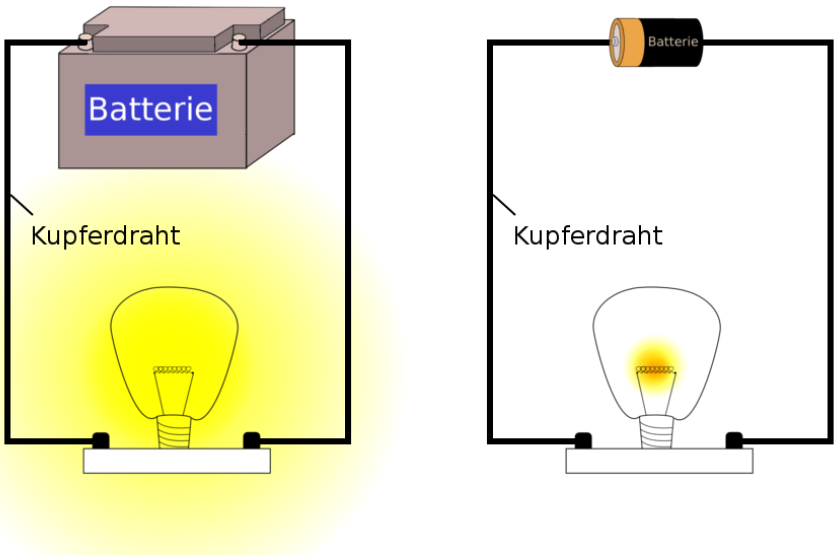
Er vermutet, dass Glühbirnen heller leuchten, wenn man als Leiter statt Kupferdraht Golddraht verwendet.

Mit welchem Experiment kann er seine Vermutung überprüfen?

☐	
☐	
☐	
☐	

Helles Licht	IN-LS-1

Anna hat folgendes Experiment durchgeführt:



Was zeigt dieses Experiment?

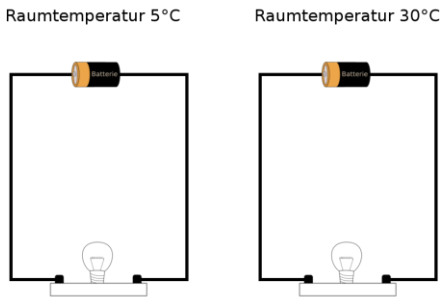
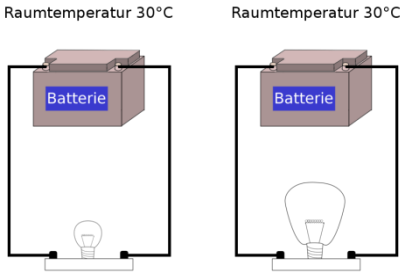
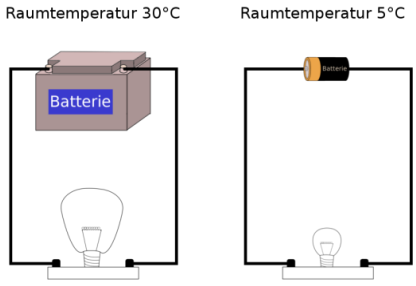
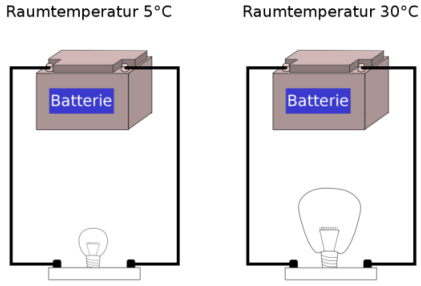
☐	Die Größe der Batterie hat einen Einfluss auf die Helligkeit der Lampe.
☐	Das Leitermaterial hat einen Einfluss auf die Helligkeit der Lampe.
☐	Sowohl das Leitermaterial als auch die Größe der Batterie haben einen Einfluss auf die Helligkeit der Lampe.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Licht einer Glühlampe	ID-LS-2

Nico möchte herausfinden, ob in einem kalten Raum eine Glühlampe heller leuchtet, als in einem warmen Raum.

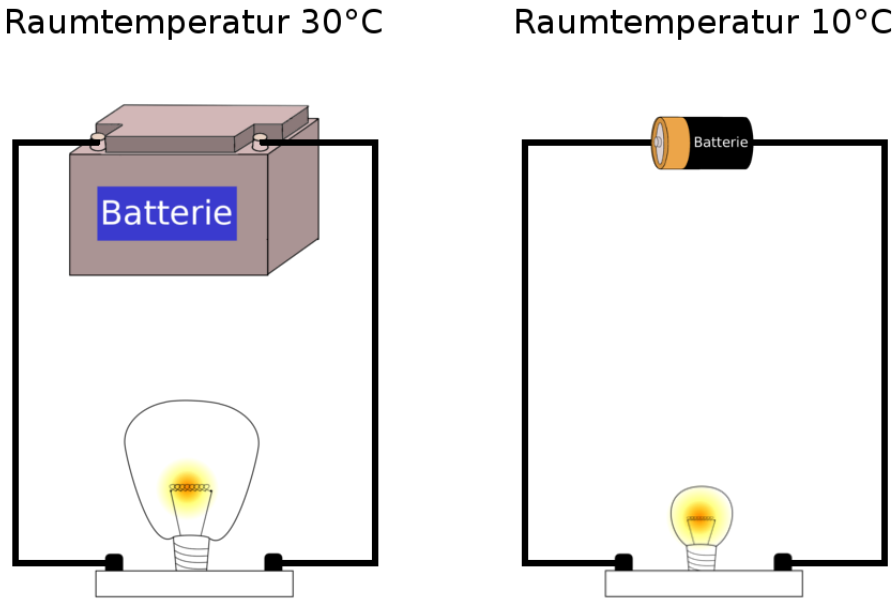
Er vermutet, dass Glühlampen heller leuchten, wenn man sie in einer kälteren Umgebung leuchten lässt.

Mit welchem Experiment kann er seine Vermutung überprüfen?

☐	
☐	
☐	
☐	

Helles Licht	IN-LS-2

Toni hat folgendes Experiment durchgeführt:



Was zeigt dieses Experiment?

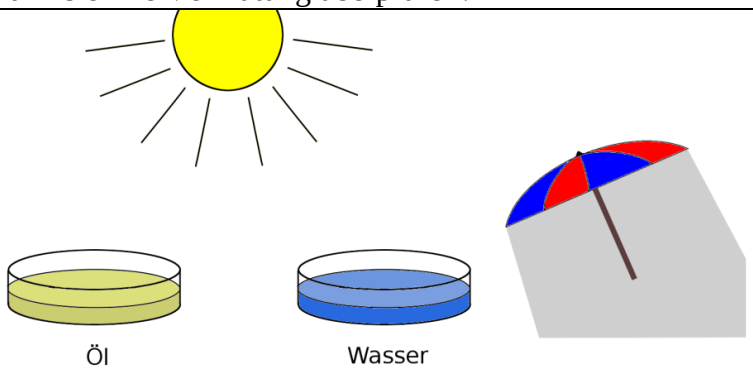
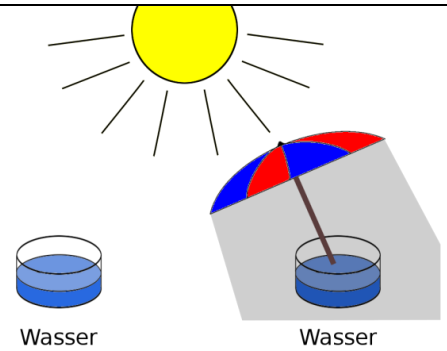
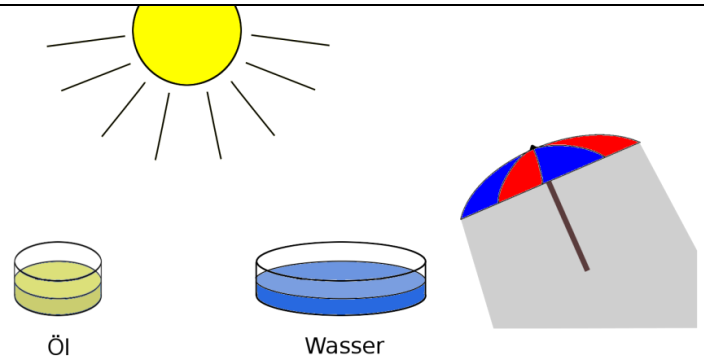
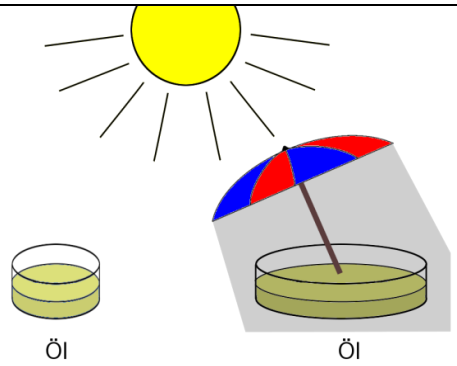
☐	Die Größe der Batterie hat einen Einfluss auf die Helligkeit der Lampe.
☐	Die Raumtemperatur hat einen Einfluss auf die Helligkeit der Lampe.
☐	Sowohl die Raumtemperatur als auch die Größe der Batterie haben einen Einfluss auf die Helligkeit der Lampe.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Verdunstung von Öl	ID-FL-1

Nach dem Braten füllt Mara immer in die kalte Pfanne mit dem übriggebliebenen Öl etwas Wasser, damit die Pfanne beim Abwasch leichter zu reinigen ist. Wartet sie mit dem Abwasch jedoch zu lange, so befinden sich erneut nur noch Öl und kein Wasser mehr in der Pfanne.

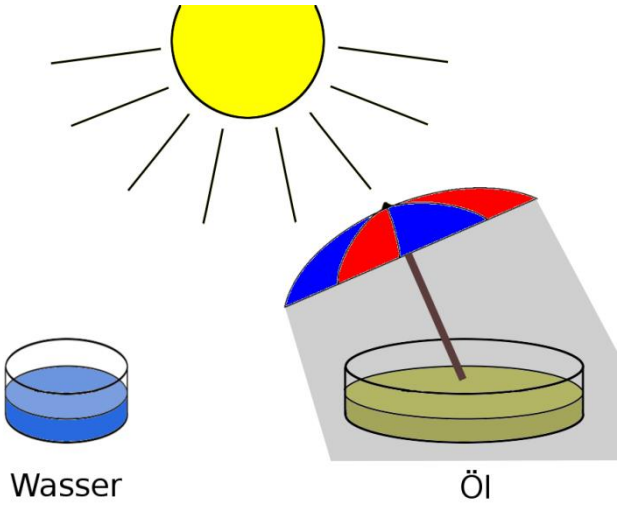
Sie vermutet, dass Waser schneller verdunstet als Öl.

Mit welchem Experiment kann sie ihre Vermutung überprüfen?

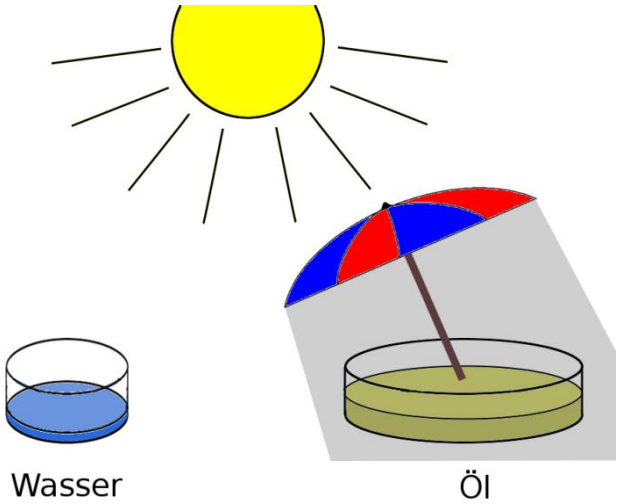
☐	
☐	
☐	
☐	

Wasser verschwindet	IN-FL-1

Michael hat folgendes Experiment durchgeführt:



drei Stunden später…

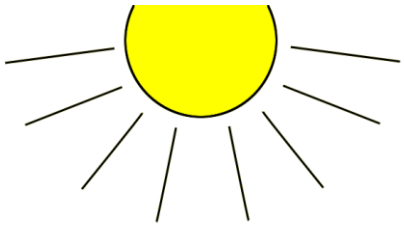


Was zeigt dieses Experiment?

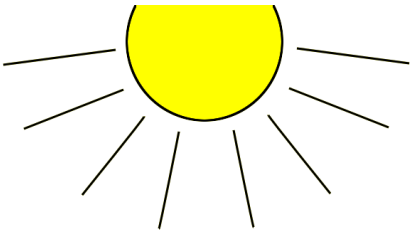
☐	Die Größe des Gefäßes hat einen Einfluss auf die Verdunstung.
☐	Der Schatten hat einen Einfluss auf die Verdunstung.
☐	Die Flüssigkeit hat einen Einfluss auf die Verdunstung.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Wasser verschwindet	IN-FL-2

Michael hat folgendes Experiment durchgeführt:



6 Stunden später…



Was zeigt dieses Experiment?

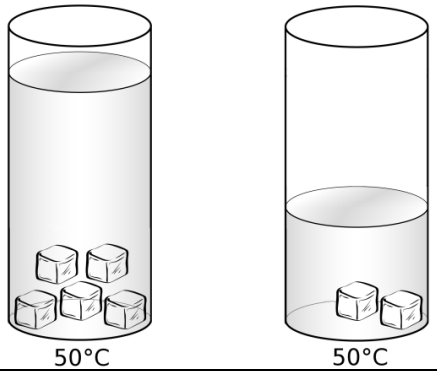
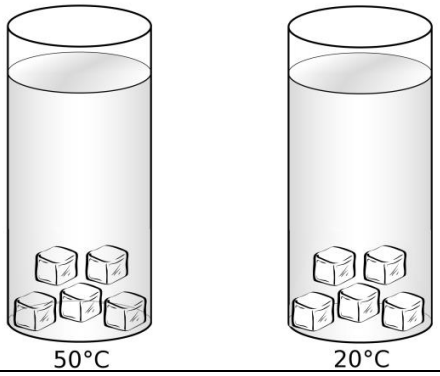
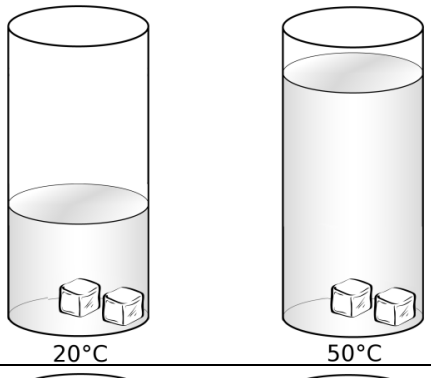
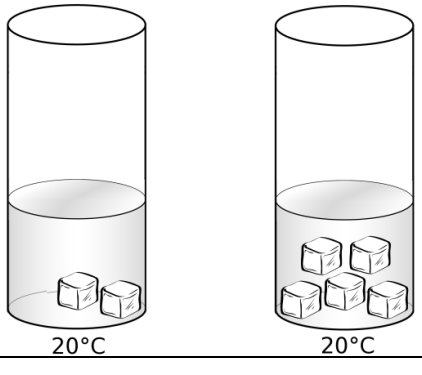
☐	Die Oberfläche des Gefäßes hat einen Einfluss auf die Verdunstung.
☐	Die Füllhöhe des Gefäßes hat einen Einfluss auf die Verdunstung.
☐	Sowohl die Füllhöhe als auch die Oberfläche des Gefäßes haben einen Einfluss auf die Verdunstung.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Eis in warmem Wasser	ID-EIS-1

Sina möchte ihr warmes Wasser aus dem Supermarkt mit Hilfe von Eiswürfeln abkühlen.

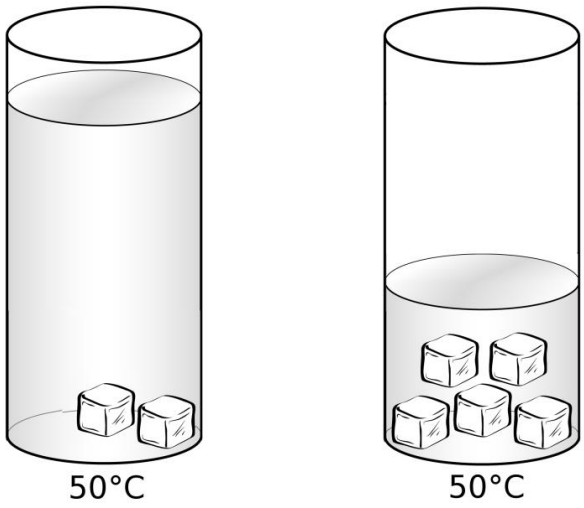
Sie vermutet, dass Eis schneller schmilzt, wenn es in warmes statt in kaltes Wasser gegeben wird.

Mit welchem Experiment kann sie ihre Vermutung überprüfen?

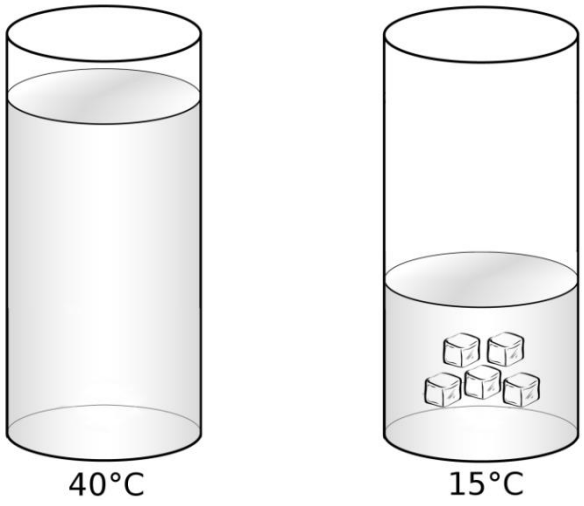
☐	
☐	
☐	
☐	

Eis verschwindet	IN-EIS-1

Andreas hat folgendes Experiment durchgeführt:



sieben Minuten später…



Was zeigt dieses Experiment?

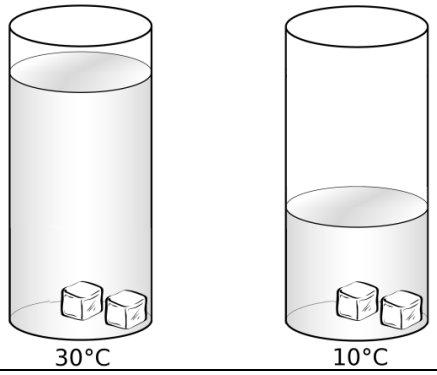
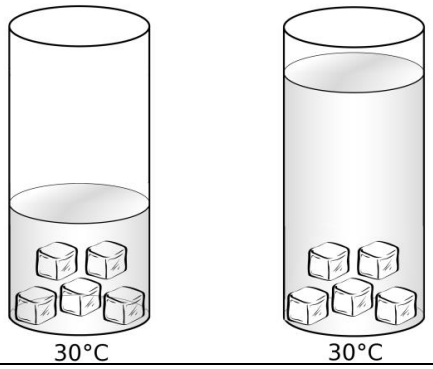
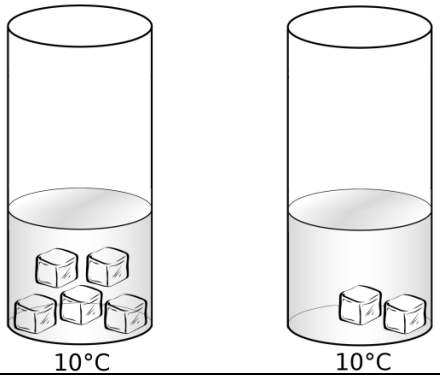
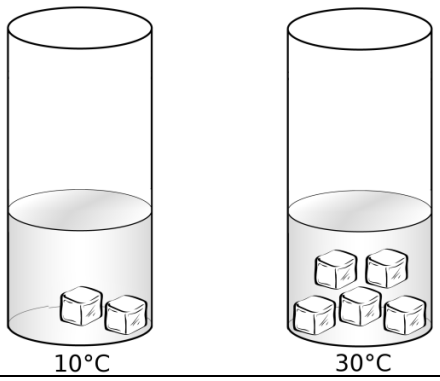
☐	Die Größe des Gefäßes hat einen Einfluss auf das Schmelzen des Eises.
☐	Die Temperatur des Wassers hat einen Einfluss auf das Schmelzen des Eises.
☐	Sowohl die Temperatur des Wassers als auch die Größe des Gefäßes haben einen Einfluss auf das Schmelzen des Eises.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Eis und Wasserpegel	**ID-EIS-2**

Timo hat eine Idee.

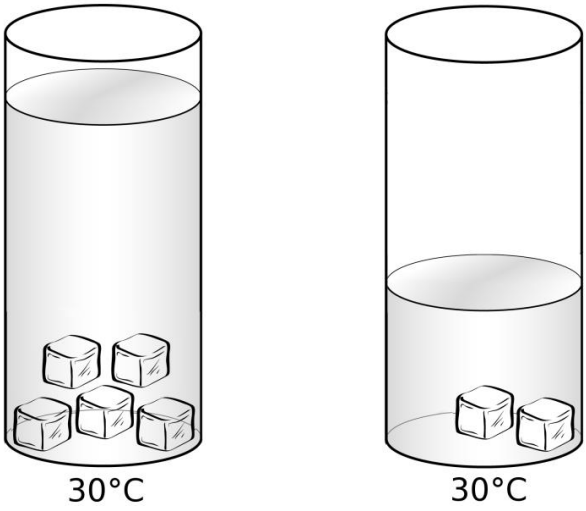
Er vermutet, dass Eis schneller schmilzt, wenn es in ein volles Wasserglas anstelle eines halb vollen Wasserglases gegeben wird.

Mit welchem Experiment kann er seine Vermutung überprüfen?

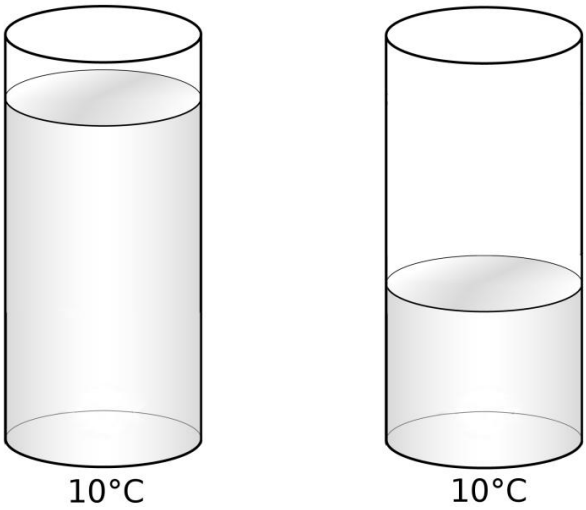
☐	
☐	
☐	
☐	

Eis verschwindet	IN-EIS-2

Jule hat folgendes Experiment durchgeführt:



15 Minuten später…



Was zeigt dieses Experiment?

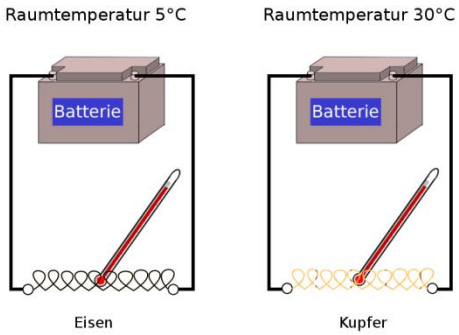
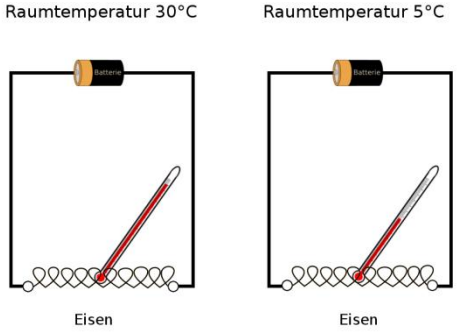
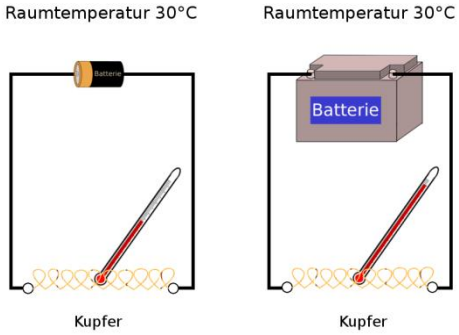
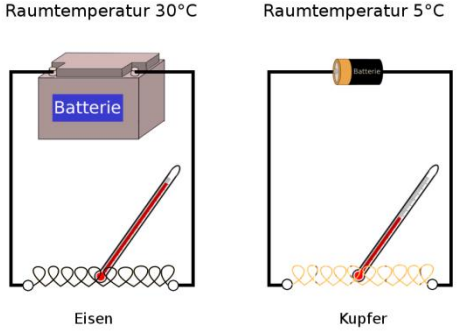
☐	Die Menge des Eises hat einen Einfluss auf die Temperaturänderung.
☐	Die Anfangstemperatur des Wassers hat einen Einfluss auf die Temperaturänderung.
☐	Sowohl die Anfangstemperatur des Wassers als auch die Menge des Eises haben einen Einfluss auf die Temperaturänderung.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Heißer Draht	ID-WS-1

Lennart ist an der Funktionsweise eines Toasters interessiert.

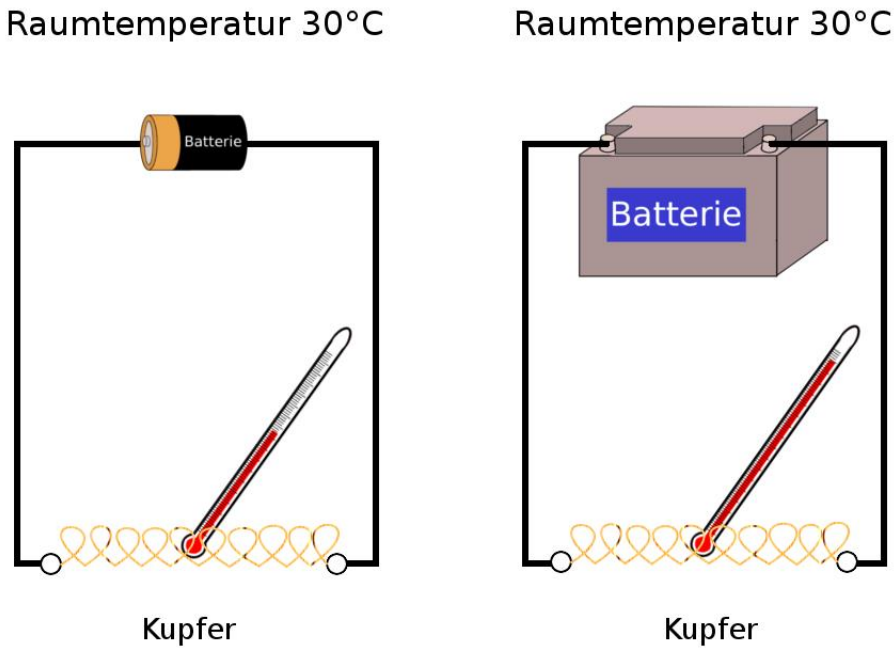
Er vermutet, dass dünne Drähte heißer werden, wenn sie von einem größeren Strom durchflossen werden.

Mit welchem Experiment kann er seine Vermutung überprüfen?

☐	
☐	
☐	
☐	

Warmer Strom	IN-WS-1

Florian hat folgendes Experiment durchgeführt:



Was zeigt dieses Experiment?

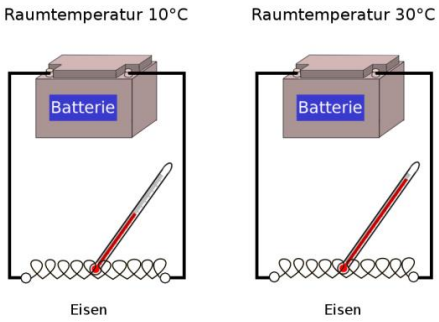
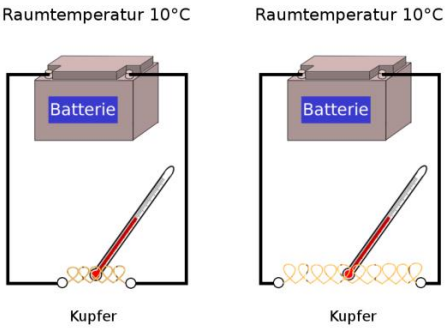
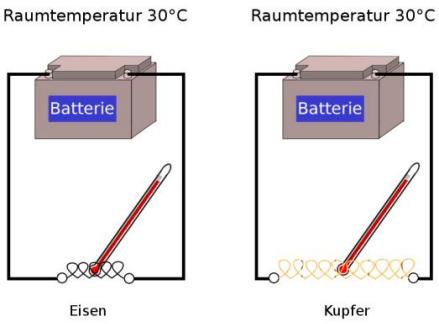
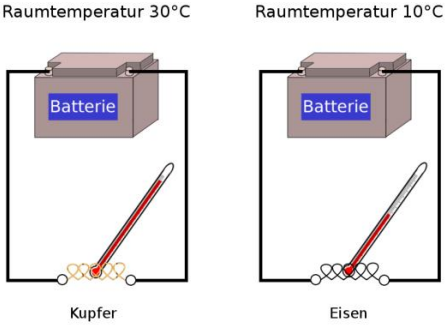
☐	Die Größe der Batterie hat einen Einfluss auf die gemessene Temperatur.
☐	Das Material des Drahts hat einen Einfluss auf die gemessene Temperatur.
☐	Sowohl die Größe der Batterie als auch das Material des Drahts haben einen Einfluss auf die gemessene Temperatur.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Lange Leitung	ID-WS-2

Saskia hat eine Idee.

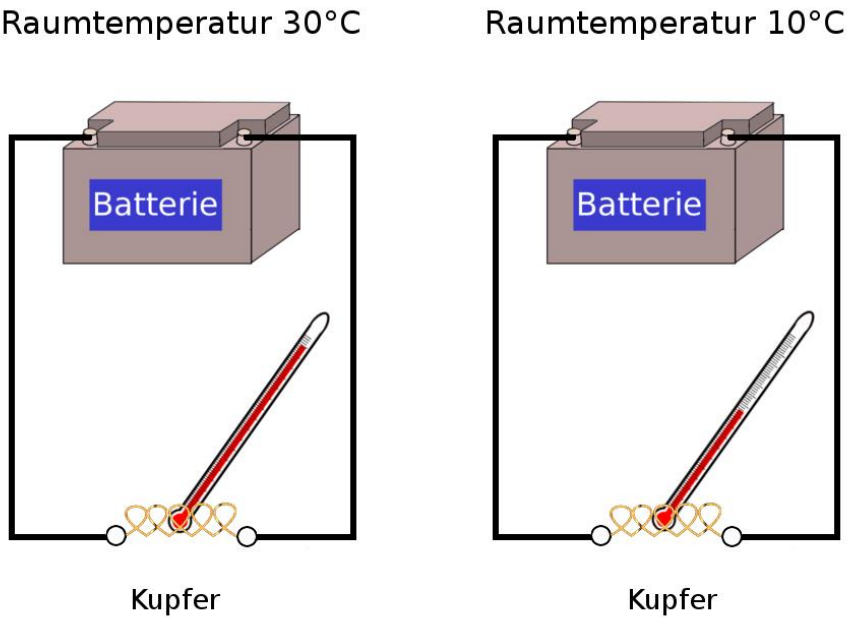
Sie vermutet, dass lange Drähte weniger heiß werden als kurze, wenn sie vom gleichen Strom durchflossen werden.

Mit welchem Experiment kann sie ihre Vermutung überprüfen?

☐	
☐	
☐	
☐	

Wärmetransport	IN-WS-2

Helge hat folgendes Experiment durchgeführt:



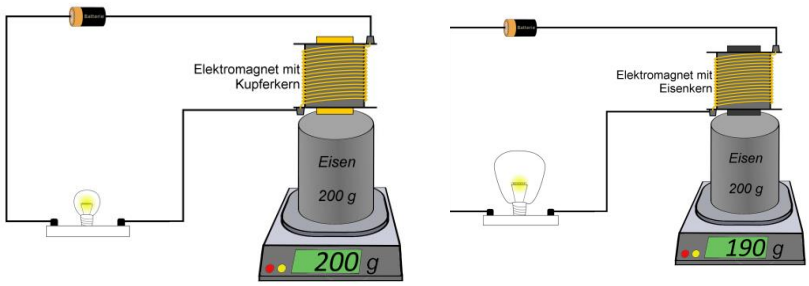
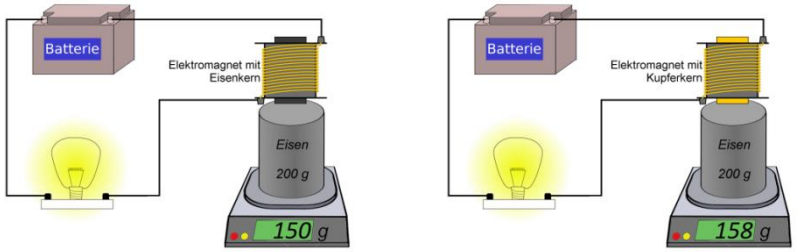
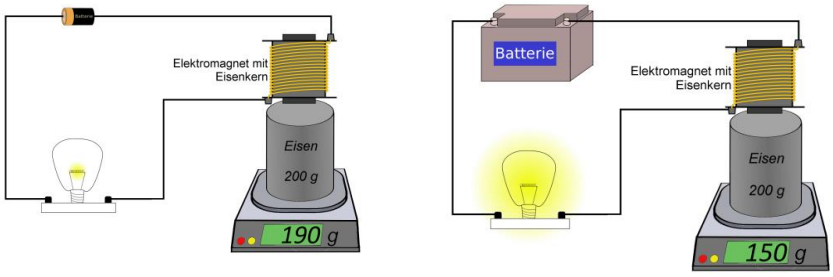
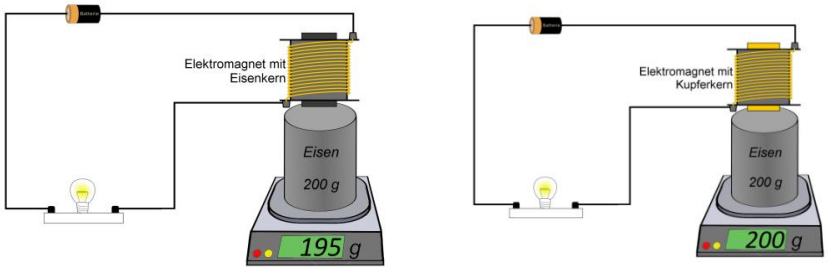
Was zeigt dieses Experiment?

☐	Das Material des Drahts hat einen Einfluss auf die gemessene Temperatur.
☐	Die Raumtemperatur hat einen Einfluss auf die gemessene Temperatur.
☐	Sowohl die Raumtemperatur als auch das Material des Drahts haben einen Einfluss auf die gemessene Temperatur.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Starker Strom	ID-MS-1

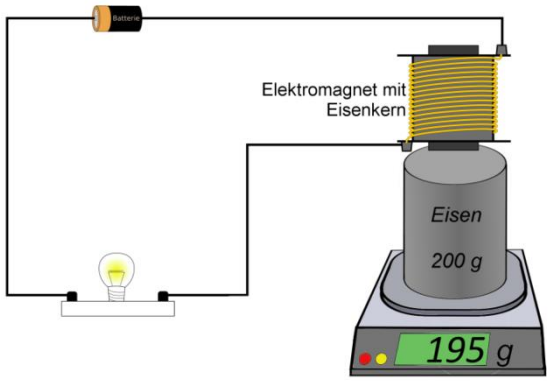
Tina vermutet, dass das magnetische Feld eines Elektromagneten stärker ist, je größer der Strom ist, der durch ihn fließt.

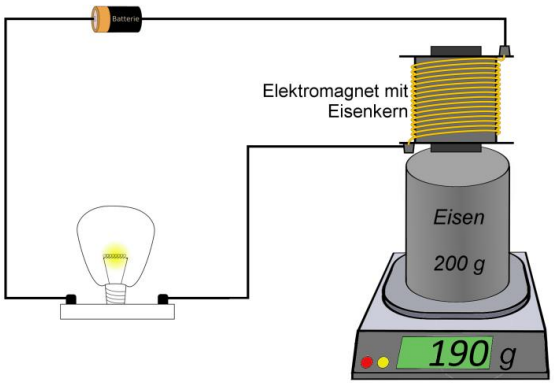
Mit welchem Experiment kann sie ihre Vermutung überprüfen?

☐	
☐	
☐	
☐	

Magnete mit Einfluss?	IN-MS-1

Patrick hat folgendes Experiment durchgeführt:





Was zeigt dieses Experiment?

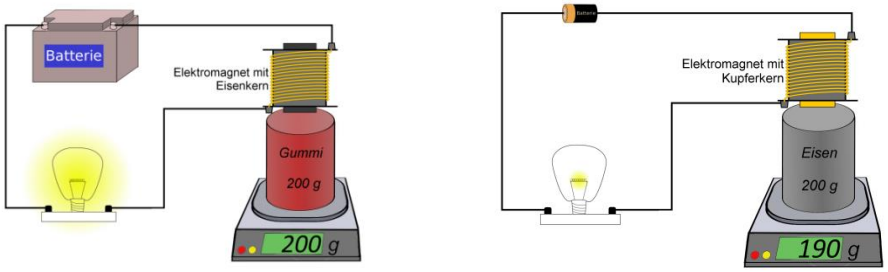
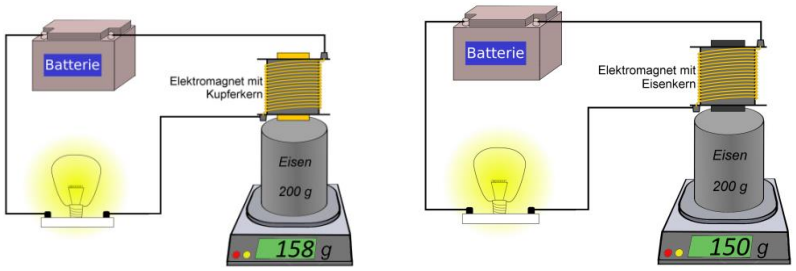
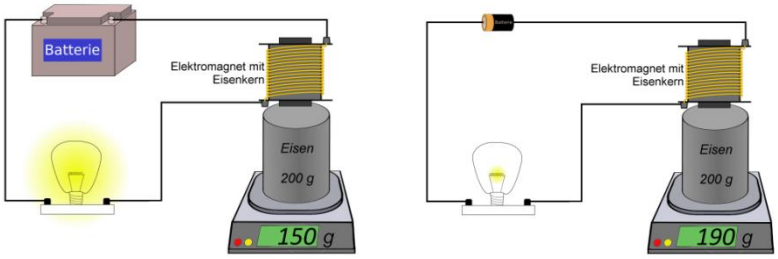
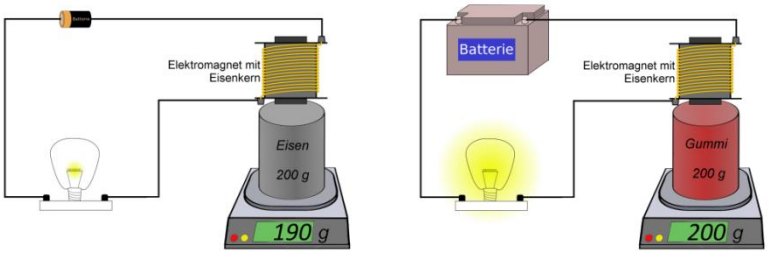
☐	Die Wahl der Glühlampen hat einen Einfluss auf das Magnetfeld des Elektromagneten.
☐	Die Größe der Batterie hat einen Einfluss auf das Magnetfeld des Elektromagneten.
☐	Sowohl die Größe der Batterie als auch die Wahl der Glühlampen haben einen Einfluss auf das Magnetfeld des Elektromagneten.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Gutes Material	ID-MS-2

Isa hat eine Idee.

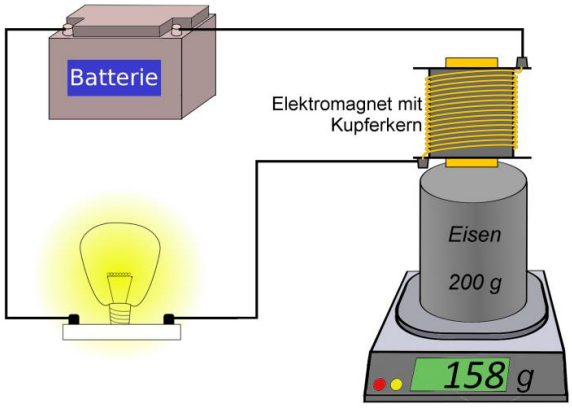
Sie vermutet, dass die Stärke des Magnetfeldes eines Elektromagneten von seinem Kernmaterial abhängt.

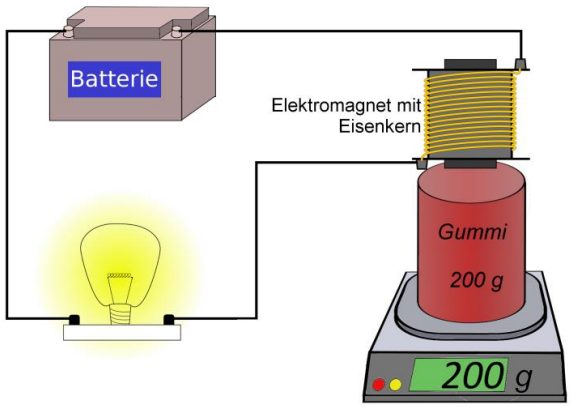
Mit welchem Experiment kann sie ihre Vermutung überprüfen?

☐	
☐	
☐	
☐	

Magnet und Strom	IN-MS-2

Lukas hat folgendes Experiment durchgeführt:





Was zeigt dieses Experiment?

☐	Das Material des Kerns hat einen Einfluss auf das Magnetfeld des Elektromagneten.
☐	Das Material des Massestücks hat einen Einfluss auf das Magnetfeld des Elektromagneten.
☐	Sowohl das Material des Massestücks als auch das des Kerns haben einen Einfluss auf das Magnetfeld des Elektromagneten.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

Mit diesem Test soll dein Wissen über gute und aufschlussreiche Experimente abgefragt werden. Bitte gib dir Mühe und versuche alle Frage richtig zu beantworten. Bei allen Fragen musst du die richtige Antwort ankreuzen. Es ist immer **nur eine** Antwort richtig!

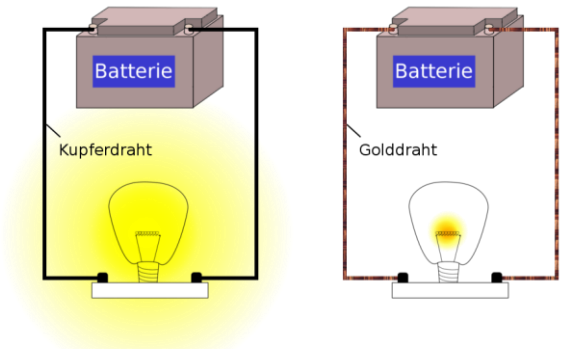
Beispiel:		
Bearbeitest du gerade ein Testheft?	☒ Ja	☐ Nein

Solltest du versehentlich eine falsche Antwort gewählt haben, so male das Kästchen mit der falschen Antwort aus und kreuze die richtige Antwort an.

Beispiel:		
Bearbeitest du gerade ein Testheft?	☒ Ja	■ Nein

In einigen Testfragen wirst du wie in folgendem Beispiel dazu befragt, was ein dargestelltes Experiment **zeigt**.

BEISPIEL : Petra hat folgendes Experiment durchgeführt:



Was zeigt dieses Experiment?

☐	Die Größe der Batterie hat einen Einfluss auf die Helligkeit der Lampe.
☐	Das Leitermaterial hat einen Einfluss auf die Helligkeit der Lampe.
☐	Sowohl das Leitermaterial als auch die Größe der Batterie haben einen Einfluss auf die Helligkeit der Lampe.
☐	Das Experiment lässt keine sichere Schlussfolgerung zu.

In diesem Fall ist die zweite Antwort die richtige, da in den beiden Versuchen nur das Leitermaterial verändert wurde und die Lampe im zweiten Fall weniger hell leuchtet. Die Größe der verwendeten Batterie hätte auch einen Einfluss auf die Helligkeit der Lampe, sie wurde jedoch im gezeigten Experiment nicht variiert und ist somit nicht anzukreuzen.

Wenn du fertig bist, bleibe bitte ruhig auf deinen Platz sitzen. Du darfst den Raum erst verlassen, wenn der Testleiter es erlaubt.

Viel Erfolg und vielen Dank für deine Mitarbeit!

Anhang Publikation 3

Anmerkung. Die folgenden Arbeitsblätter und Poster sind die in der Studie eingesetzten Materialien in deutscher Sprache. Der Publikation wurden Übersetzungen ins Englische beigefügt.

Worksheets utilized in the CVS paper-and-pencil training

Physik Elektrizitätslehre	**Kraft eines Elektromagneten**	Code:

Elektromagnete ziehen wie Dauermagnete bestimmte Metalle an. Elektromagnete haben jedoch keine dauerhaft anziehende Wirkung. Sie wirken nur anziehend, wenn sie in einen geschlossenen Stromkreis eingebaut sind. Elektromagnete können daher angeschaltet oder ausgeschaltet werden. Dies ist für viele technische Anwendungen hilfreich. So werden starke Elektromagnete z.B. auf Schrottplätzen genutzt, um Schrott anzuheben und an anderer Stelle abzulegen. Aber auch in Elektromotoren und Lautsprechern findet man Elektromagnete

Abbildung 1: Schrott wir mithilfe eines Elektromagneten verladen.

Elektromagnete bestehen aus einem Leiter der zu einer Spule gewickelt ist. In der Mitte der Spule befindet sich meistens ein Kern aus Eisen.

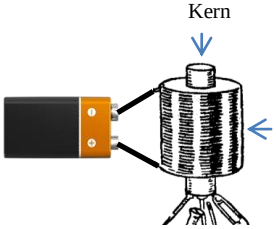


Abbildung 2: An eine Batterie angeschlossener Elektromagnet

In der letzten NAWI-Stunde haben wir besprochen, wie aussagekräftige Experimente geplant werden.

Info-Kasten

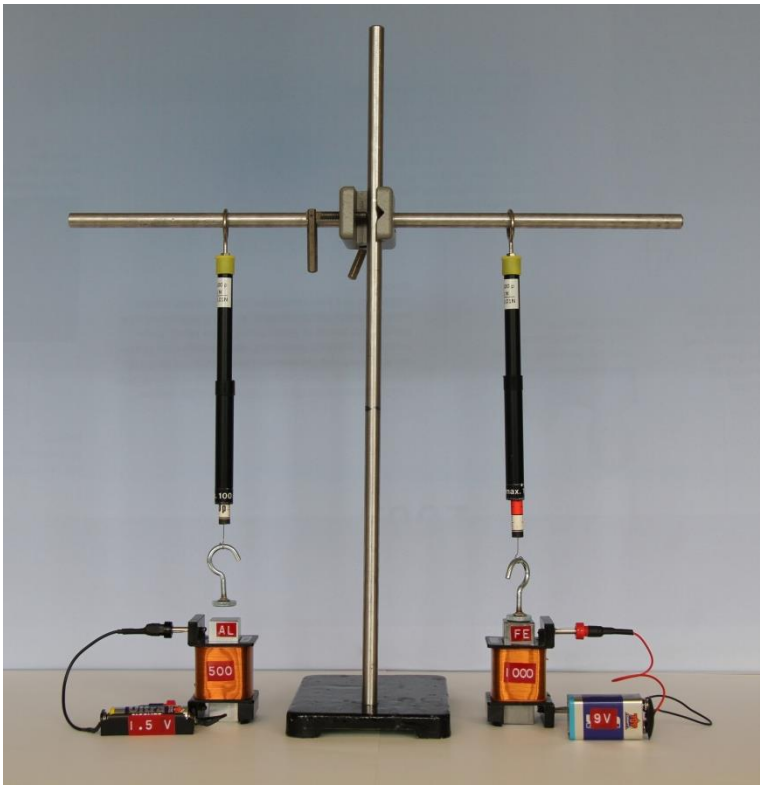
In aussagekräftigen Experimenten werden mindestens zwei Versuche miteinander verglichen. Dabei ist darauf zu achten, dass sich nur die zu untersuchende Variable zwischen den beiden Versuchsaufbauten unterschiedet. So könnt ihr sicher sein, dass Veränderungen in eurer abhängigen Variable auf die Veränderung eurer untersuchten Variablen zurückzuführen sind.

Heute sollt ihr euer Wissen über aussagekräftige Experimente nutzen, um herauszufinden von welchen Variablen die anziehende Kraft eines Elektromagneten abhängt.

Auf den folgenden Arbeitsblättern sollt ihr Experiment von anderen Schülerinnen und Schülern beurteilen und selber Experimente planen und auswerten. Dazu dürft ihr, müsst aber nicht sämtliche Materialien von der Materialiste verwenden. Blättert bitte erst auf die nächste Seite um, wenn ihr die vorherige Seite fertig bearbeitet habt.

Viel Erfolg!

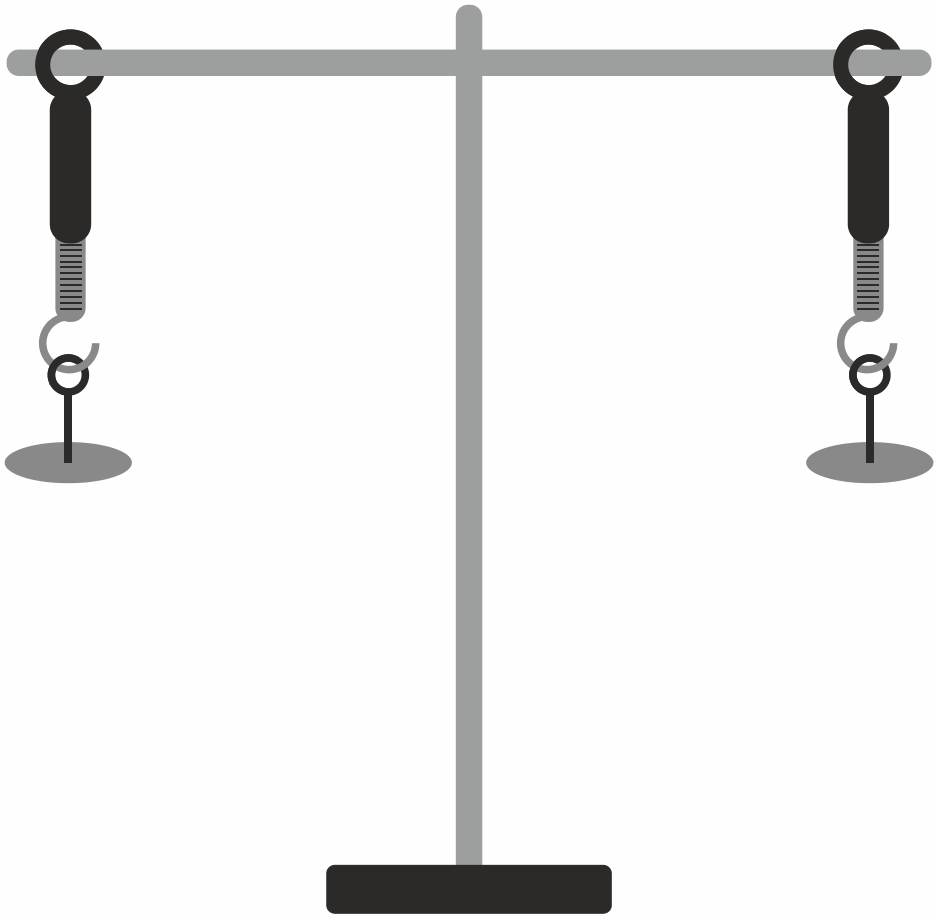
Aufgabe 1:
Lea und Marian wollen herausfinden, ob die anziehende Kraft eines Elektromagneten davon abhängt, wie oft der Draht um den Kern gewickelt ist. Sie haben dazu folgendes Experiment geplant:



Schaut euch das Experiment bitte genau an. Leider ist ihr Experiment nicht aussagekräftig. Schreibt bitte auf was die Probleme dieses Experiments sind.

Was würdet ihr verändern, damit das Experiment aussagekräftig wird? Bitte zeichnet eurer Experimentplanung in die untere Zeichnung ein und beschriftet alle Materialien.

Unser Verbesserungsvorschlag:

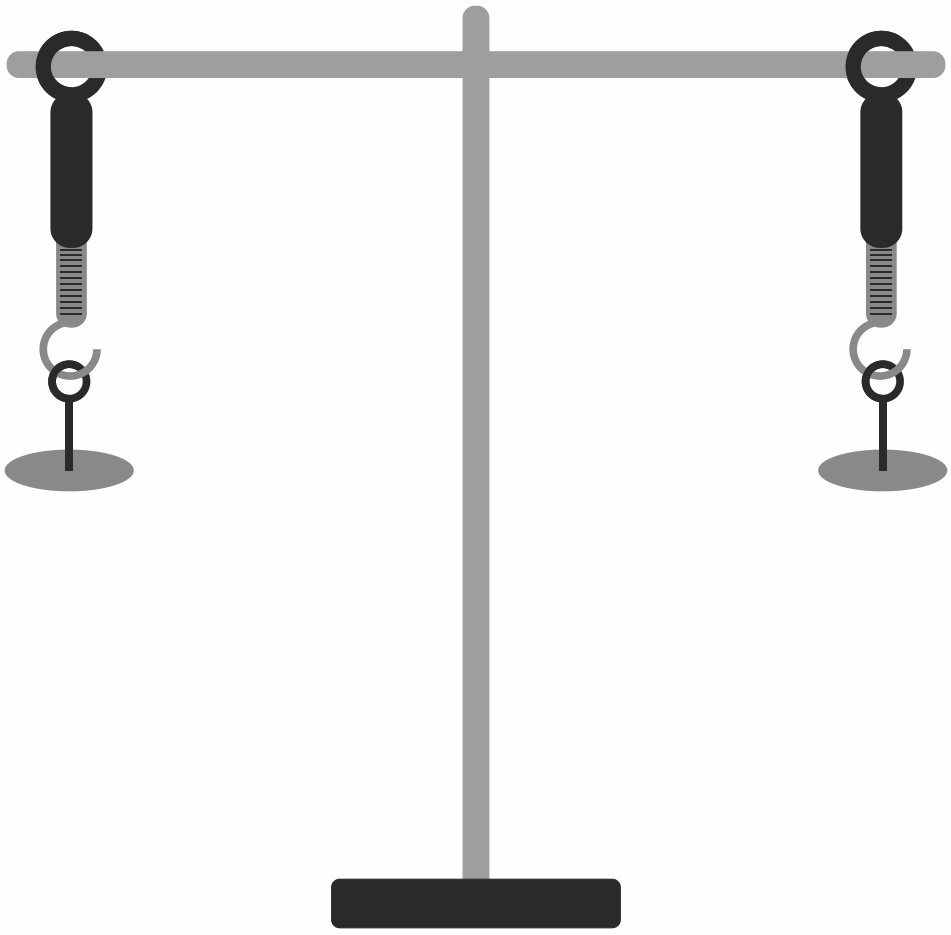


Aufgabe 2:
Marian und Lea wollen wissen, von welchen Variablen die Anziehungskraft eines Elektromagneten abhängt. Sie sammeln ihre Ideen.

Lea sagt: „*Mir ist aufgefallen, dass bei große Elektromagneten der Leiter um einen Metallkern gewickelt ist. Ich frage mich, ob das Material aus dem der Kern ist einen Einfluss auf die Anziehungskraft des Magneten hat.“*

Jetzt seid ihr gefragt. Bitte plant ein Experiment, mit dem ihr herausfinden könnt, ob Leas Vermutung richtig ist. Bitte zeichnet eurer Experimentplanung in die untere Zeichnung ein und beschriftet alle Materialien. Blättert bitte erst danach auf die nächste Seite.

Unser Experiment:



Lea und Marian haben folgendes Experiment durchgeführt:

Was haben Lea und Marian herausgefunden?

☐	Leas Vermutung ist **richtig**. Die Stärke eines Elektromagneten hängt von dem Kernmaterial ab.
☐	Leas Vermutung ist **falsch**. Die Stärke eines Elektromagneten hängt **nicht** von dem Kernmaterial ab.

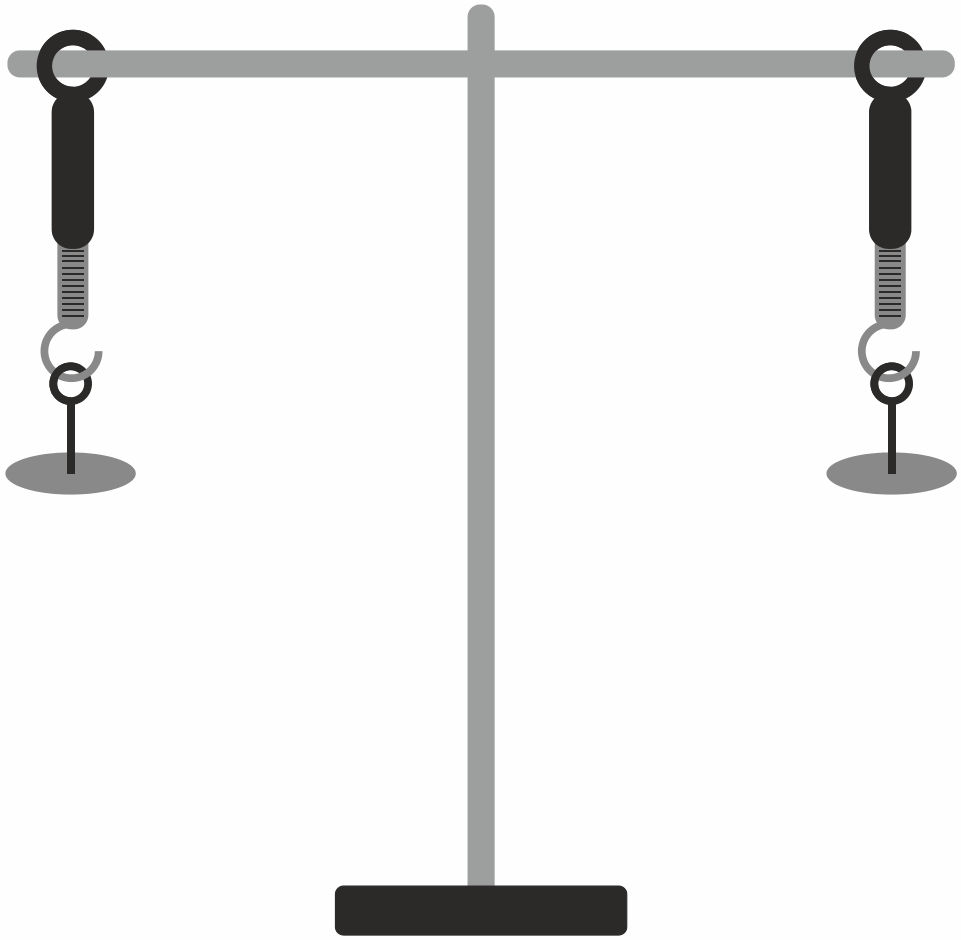
Denkt bitte noch einmal über das Experiment nach. Warum können Marian und Lea ganz sicher sein, dass sie etwas über den Einfluss des Kernmaterials auf die Anziehungskraft des Elektromagneten herausgefunden haben?

Aufgabe 3:
Marian und Lea überlegen, ob ihnen mehr Variablen einfallen, die sie untersuchen könnten.

Marian sagt: *„Ich frage mich, ob die magnetische Wirkung des Stroms, wie die Wärmewirkung des Stroms von der Stromstärke abhängt? Ich vermute, dass die Kraft eines Elektromagneten umso größer ist, je größer der Strom ist, der durch den Leiter fließt“*

Wie könnt ihr Marians Vermutungen durch ein Experiment überprüfen? Bitte zeichnet eurer Experimentplanung in die untere Zeichnung ein und beschriftet alle Materialien. Blättert bitte erst danach auf die nächste Seite.

Unser Experiment:



Lea und Marian haben folgendes Experiment durchgeführt:

Was haben Lea und Marian herausgefunden?

☐	Marians Vermutung ist **richtig**. Ein Elektromagnet ist umso stärker je größer der Strom ist, der durch den Leiter fließt.
☐	Marians Vermutung ist **falsch**. Ein Elektromagnet ist nicht stärker wenn ein größer der Strom durch den Leiter fließt.

Denkt bitte noch einmal über eure Experimente nach. Warum können Lea und Marian sich ganz sicher sein, dass sie etwas über den Einfluss der Stromstärke auf die Anziehungskraft eines Elektromagneten herausgefunden haben?

Worksheets utilized in the CVS hands-on training

Physik Elektrizitätslehre	**Kraft eines Elektromagneten**	Code:

Elektromagnete ziehen wie Dauermagnete bestimmte Metalle an. Elektromagnete haben jedoch keine dauerhaft anziehende Wirkung. Sie wirken nur anziehend, wenn sie in einen geschlossenen Stromkreis eingebaut sind. Elektromagnete können daher angeschaltet oder ausgeschaltet werden. Dies ist für viele technische Anwendungen hilfreich. So werden starke Elektromagnete z.B. auf Schrottplätzen genutzt, um Schrott anzuheben und an anderer Stelle abzulegen. Aber auch in Elektromotoren und Lautsprechern findet man Elektromagnete.

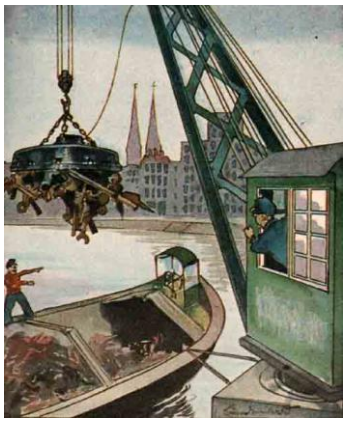


Abbildung 2: Schrott wir mithilfe eines Elektromagneten verladen.

Elektromagnete bestehen aus einem Leiter der zu einer Spule gewickelt ist. In der Mitte der Spule befindet sich meistens ein Kern aus Eisen.

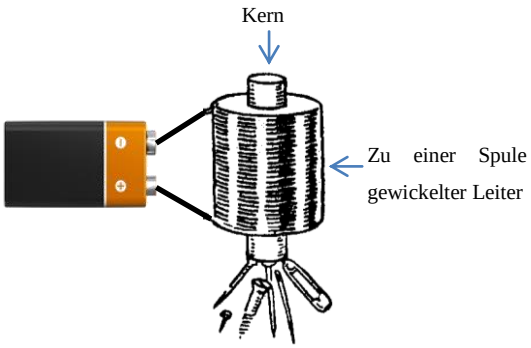


Abbildung 3: An eine Batterie angeschlossener Elektromagnet

In der letzten NAWI-Stunde haben wir besprochen, wie aussagekräftige Experimente geplant werden.

Info-Kasten

In aussagekräftigen Experimenten werden mindestens zwei Versuche miteinander verglichen. Dabei ist darauf zu achten, dass sich nur die zu untersuchende Variable zwischen den beiden Versuchsaufbauten unterschiedet. So könnt ihr sicher sein, dass Veränderungen in eurer abhängigen Variable auf die Veränderung eurer untersuchten Variablen zurückzuführen sind.

Heute sollt ihr euer Wissen über aussagekräftige Experimente nutzen, um herauszufinden von welchen Variablen die anziehende Kraft eines Elektromagneten abhängt.

Auf den folgenden Arbeitsblättern sollt ihr Experimente von anderen Schülerinnen und Schülern beurteilen und selber Experimente planen und auswerten. Dazu dürft ihr, müsst aber nicht sämtliche Materialien von der Materialiste verwenden. Blättert bitte erst auf die nächste Seite um, wenn ihr die vorherige Seite fertig bearbeitet habt.

Viel Erfolg!

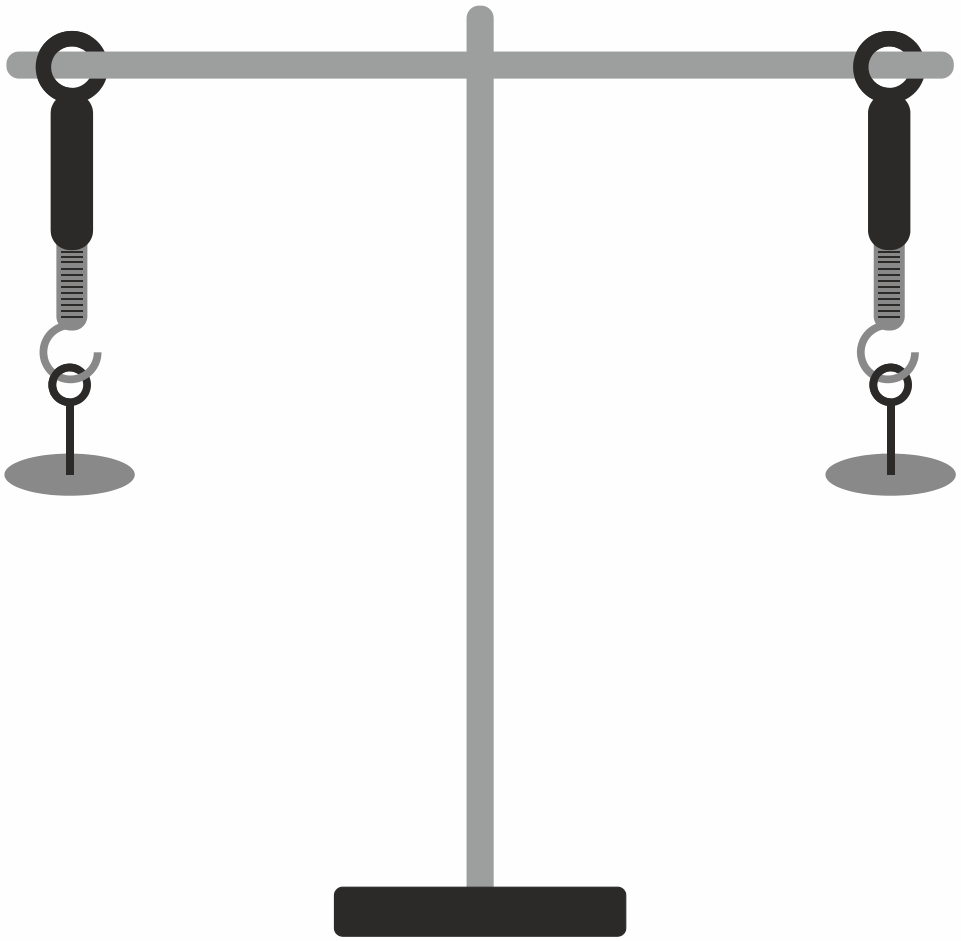
Aufgabe 1:
Lea und Marian wollen herausfinden, ob die anziehende Kraft eines Elektromagneten davon abhängt, wie oft der Draht um den Kern gewickelt ist. Sie haben dazu das vor euch aufgebaute Experiment geplant.

Schaut euch das Experiment bitte in ruhe an und verändert es zuerst einmal nicht. Leider ist ihr Experiment nicht aussagekräftig. Schreibt bitte auf was die Probleme dieses Experiments sind.

[Anmerkung: Zur Bearbeitung der ersten Aufgabe wurde den Schülerinnen und Schüler der hands-on Trainingsbedingung ein Versuchsaufbau identisch zu dem auf dem Foto der Paper-and-Penicl Arbeitsblätter präsentiert]

Was würdet ihr verändern, damit das Experiment aussagekräftig wird? Bitte zeichnet eurer Experimentplanung in die untere Zeichnung ein und beschriftet alle Materialien.

Unser Verbesserungsvorschlag:

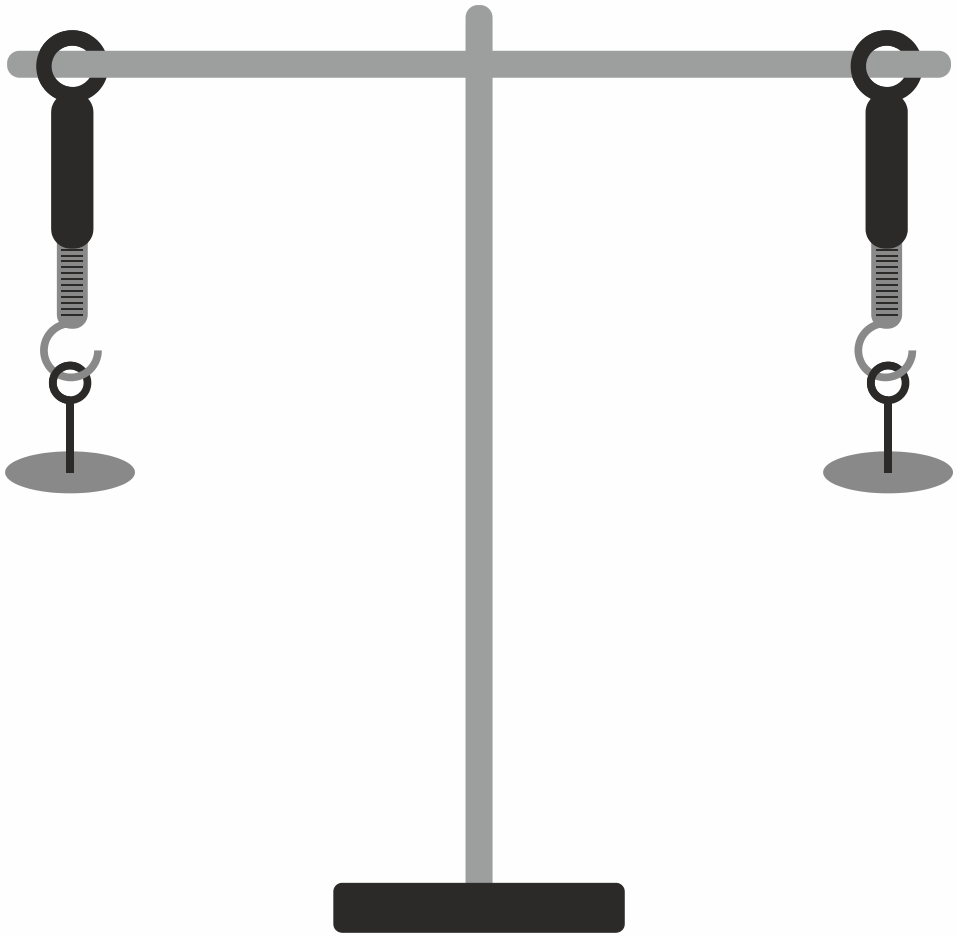


Aufgabe 2:
Marian und Lea wollen wissen, von welchen Variablen die Anziehungskraft eines Elektromagneten abhängt. Sie sammeln ihre Ideen.

Lea sagt: „*Mir ist aufgefallen, dass bei große Elektromagneten der Leiter um einen Metallkern gewickelt ist. Ich frage mich, ob das Material aus dem der Kern ist einen Einfluss auf die Anziehungskraft des Magneten hat.“*

Jetzt seid ihr gefragt. Bitte plant ein Experiment, mit dem ihr herausfinden könnt, ob Leas Vermutung richtig ist. Bitte zeichnet eurer Experimentplanung in die untere Zeichnung ein und beschriftet alle Materialien.

Unser Experiment:



Nachdem ihr das Experiment durchgeführt habt, notiert bitte alle **Messwerte und Variablen**.

Unser Messwerte und Variablen:

Was habt ihr herausgefunden?

☐	Leas Vermutung ist **richtig**. Die Stärke eines Elektromagneten hängt von dem Kernmaterial ab.
☐	Leas Vermutung ist **falsch**. Die Stärke eines Elektromagneten hängt **nicht** von dem Kernmaterial ab.

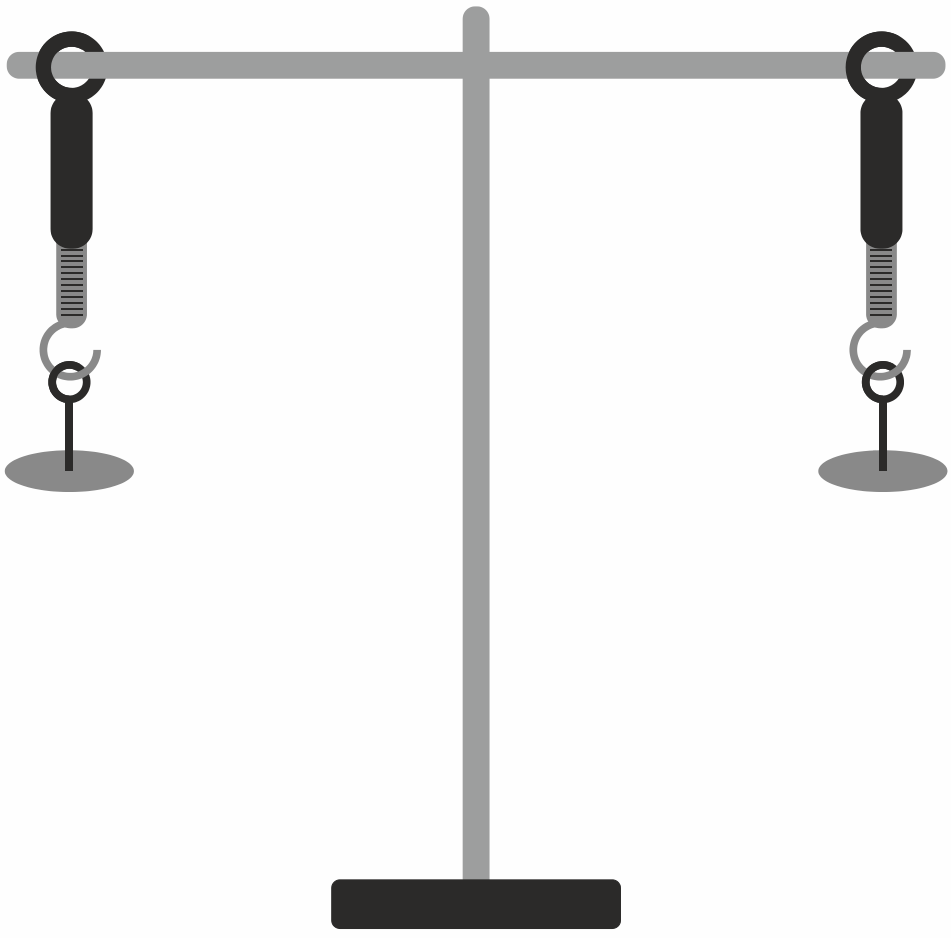
Denkt bitte noch einmal über das Experiment nach. Warum können Marian und Lea ganz sicher sein, dass sie etwas über den Einfluss des Kernmaterials auf die Anziehungskraft des Elektromagneten herausgefunden haben?

Aufgabe 3:
Marian und Lea überlegen, ob ihnen mehr Variablen einfallen, die sie untersuchen könnten.

Marian sagt: *„Ich frage mich, ob die magnetische Wirkung des Stroms, wie die Wärmewirkung des Stroms von der Stromstärke abhängt? Ich vermute, dass die Kraft eines Elektromagneten umso größer ist, je größer der Strom ist, der durch den Leiter fließt“*

Wie könnt ihr Marians Vermutungen durch ein Experiment überprüfen? Bitte zeichnet eurer Experimentplanung in die untere Zeichnung ein und beschriftet alle Materialien. Nachdem ihr das Experiment durchgeführt habt, notiert bitte alle **Messwerte und Variablen**.

Unser Experiment:



Unser Messwerte und Variablen:

Was habt ihr heraus-gefunden?

☐	Marians Vermutung ist **richtig**. Ein Elektromagnet ist umso stärker je größer der Strom ist, der durch den Leiter fließt.
☐	Marians Vermutung ist **falsch**. Ein Elektromagnet ist nicht stärker wenn ein größer der Strom durch den Leiter fließt.

Denkt bitte noch einmal über eure Experimente nach. Warum können Marian und Lea ganz sicher sein, dass sie etwas über den Einfluss der Stromstärke auf die Anziehungskraft eines Elektromagneten herausgefunden haben?

Poster of the poster evaluation task

Poster A

Wer hat das bessere Gedächtnis? Jungs oder Mädchen?

Vermutung

Ich vermute Mädchen haben ein besseres Gedächtnis als Jungs, da Mädchen sich Dinge genauer anschauen.

Teilnehmer

- 19 Mädchen aus der 8. Klasse
- 17 Jungs aus der 6. Klasse

Das habe ich gemacht:

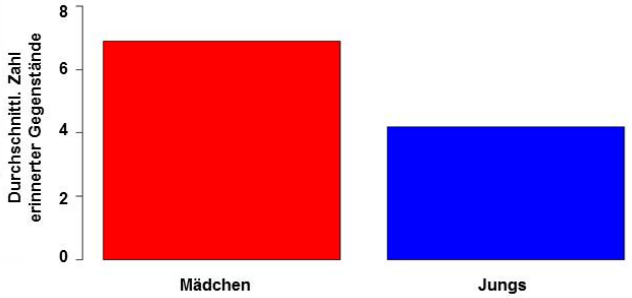
1. Ich habe den Schülern 8 Gegenstände auf einem Poster gezeigt und die Namen laut vorgelesen.
2. Die Schüler sollten sich die Namen aller Gegenstände aufschreiben. Sie hatten genau 20 sec. für jeden Gegenstand.
3. Anschließend habe ich den Schülern ihre Zettel weggenommen und das Poster umgedreht.
4. Dann sollten die Schüler alle Gegenstände aufschreiben an die sie sich erinnern.
5. Zum Schluss habe ich die richtigen Namen gezählt.

Schlussfolgerung

Mädchen haben ein besseres Gedächtnis als Jungs.

Gegenstände

Ergebnisse:



Poster B

Hat der Eisenanteil in Spänen einen Einfluss auf die anziehende Kraft eines Magneten?

Vermutung

Ich vermute, dass die Kraft eines Magneten auf eine Mischung aus Holz- und Eisenspänen größer ist, wenn mehr Eisenspäne in der Mischung sind. Ich vermute das, weil ich weiß, dass Magneten Eisen anziehen.

Material

3 kleine Eimer
Holzspäne & Eisenspäne
1 Laborwaage
1 Kraftmesser mit Stativ
1 Magnet

Das habe ich gemacht:

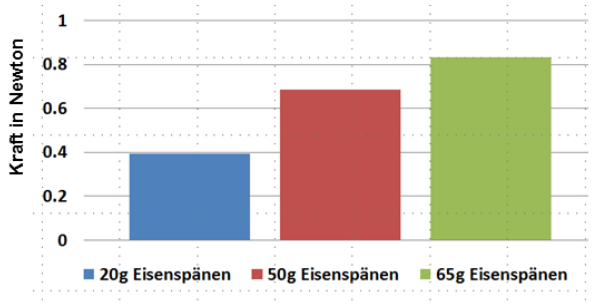
1. Ich fülle drei Behälter mit 20g, 50g und 65g Eisenspänen.
2. Dann fülle ich die Behälter mit je 20g Holzspänen auf .
3. Ich hänge den Kraftmesser auf und lege den Magneten darunter.
4. Dann hänge ich die Behälter nacheinander an den Kraftmesser und lese die Kraft ab.
5. Die Messergebnisse stelle ich in einem Diagramm dar.

Schlussfolgerung

Die Späne-Mischung mit dem wenigsten Eisen wird am stärksten angezogen.

Gegenstände

Ergebnisse:



Anhang Publikation 4

Im Rahmen der unterrichtlichen Erprobung eingesetzte Arbeitsblätter

Physik **Elektrizitätslehre** 	**Widerstand eines Leiters** 	Name: Datum:

Nicht alle Leiter haben den gleichen Widerstand. Heute sollt ihr herausfinden, ob das Leitermaterial, die Länge des Leiters und der Leiterdurchmesser einen Einfluss auf den Widerstand eines Leiters haben. Dazu sollt ihr geeignete Experimente planen, durchführen und auswerten. Ihr dürft, müsst aber nicht, sämtliche Materialien in der Experimentierbox verwenden. Bedenkt bei der Planung eurer Experimente bitte, dass bei einem aussagekräftigen Experiment immer nur diejenige Variable verändert werden darf, deren Einfluss gerade untersucht wird.

Denkt bitte bei allen Aufgaben daran, Tabellen mit euren Messwerten anzulegen. Ihr solltet neben den Messwerten und den berechneten Widerständen auch alle Eigenschaften der untersuchten Leiter notieren.

Aufgabe 1: Hat die Länge eines Leiters einen Einfluss auf dessen Widerstand?

Ergebnisstabelle 1:

Bitte stellt eure Ergebnisse in einer Graphik dar. Tragt auf der y-Achse den Widerstand und auf der X-Achse die Länge des Leiters auf.

Was habt ihr herausgefunden?

☐	Der Widerstand eines Leiters hängt von seiner Länge ab.
☐	Der Widerstand eines Leiters hängt **nicht** von seiner Länge ab.

Bitte vervollständigt den folgenden Satz, wenn ihr einen Zusammenhang zwischen Länge des Leiters und seinem Widerstand gefunden habt.

Je länger ein Leiter ist, desto …

Denkt nochmal über eure Experimente nach. Warum könnt ihr euch ganz sicher sein, dass ihr etwas über den Einfluss der Leiterlänge auf den Widerstand des Leiters herausgefunden habt?

Aufgabe 2: Hat der Durchmesser eines Leiters einen Einfluss auf dessen Widerstand?

Ergebnisstabelle 2:

Bitte stellt eure Ergebnisse in einer Graphik dar. Tragt auf der y-Achse den Widerstand und auf der X-Achse den Durchmesser eines Leiters auf.

Was habt ihr herausgefunden?

☐	Der Widerstand eines Leiters hängt von seinem Durchmesser ab.
☐	Der Widerstand eines Leiters hängt **nicht** von seinem Durchmesser ab.

Bitte vervollständigt den folgenden Satz, wenn ihr einen Zusammenhang zwischen dem Durchmesser eines Leiters und seinem Widerstand gefunden habt.

Je größer der Durchmesser eines Leiters ist, desto …

Denkt nochmal über eure Experimente nach. Warum könnt ihr euch ganz sicher sein, dass ihr etwas über den Einfluss des Durchmessers auf den Widerstand des Leiters herausgefunden habt?

Aufgabe 3: Hat das Leitermaterial einen Einfluss auf den Widerstand eines Leiters?

Ergebnisstabelle 3:

Was habt ihr herausgefunden?

☐	Der Widerstand eines Leiters hängt von dem Leitermaterial ab.
☐	Der Widerstand eines Leiters hängt **nicht** von dem Leitermaterial ab.

Denkt nochmal über eure Experimente nach. Warum könnt ihr euch ganz sicher sein, dass ihr etwas über den Einfluss des Leitermaterials auf den Widerstand von Leitern herausgefunden habt?

Lebenslauf

PERSÖNLICHE DATEN

Name:	Martin Geert Schwichow
Geburtsdatum:	21.06.1986
Geburtsort:	Henstedt-Ulzburg
Staatsangehörigkeit:	deutsch
Heimatort:	Heist

BILDUNGSWEG

Jan. 2012 – heute	Promotionsstudium in Physikdidaktik am Leibniz Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel
Nov. 2011	Erstes Staatsexamen des Landes Hessens für das Lehramt an Gymnasien für die Fächer Physik und Geographie
Okt. 2006 – Nov. 2011	Studium der Physik und Geographie für das Lehramt an Gymnasien, Philipps-Universität Marburg
2005	Allgemeine Hochschulreife
1996-2005	Integrierte Gesamtschule Wedel
1992-1996	Grundschule Heist

TÄTIGKEITEN UND AUSLANDSAUFENTHALTE

Jan. 2012 - heute	Wissenschaftlicher Mitarbeiter in der Abteilung Physikdidaktik am Leibniz Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel
Sep. 2013 – Dez. 2013	Visiting scholar im Psychologie Department der Illinois State University Normal, USA
2005 – 2006	Zivildienst in der Heilpädagogischen Kindertagesstädte der Lebenshilfe Wedel

AUSZEICHNUNGEN UND FÖRDERUNGEN

2012	Auszeichnung der Marburger Geographischen Gesellschaft für die beste Abschlussarbeit des Jahres 2011
2014	DAAD Stipendium zur Teilnahme an der NARST Konferenz 2014 / Pittsburgh, USA
2015	DAAD Stipendium zur Teilnahme an der ESERA Konferenz 2015 / Helsinki, Finnland
2015	International Young Scholar Award der Jacobs Fundation zur Teilnahme an der CDS Konferenz 2015 / Columbus, USA

PUBLIKATIONEN

Schwichow, M (2012). Welche Bedeutung hat das Vorkommen oder Fehlen von Endemiten auf flachen tropischen Inseln für die Klimaforschung und Biogeographie. Jahrbuch der Marburger Geographischen Gesellschaft 2011.
Marburger Geographische Gesellschaft. Marburg.

Schwichow, M., Croker, S.; Zimmerman, C., Höffler, T., & Härtig, H. (revised and resubmitted). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*.

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (revised and resubmitted). The impact of sub-skills and item content on students' skills with regard to the control-of-variables-strategy (CVS). *International Journal of Science Education*.

Schwichow, M., Croker, S.; Zimmerman, C., & Härtig, H. (submitted). What students learn from hands-on activities. *Journal of Research in Science Teaching*.

Schwichow, M., Christoph, S., & Härtig, H. (angenommen). Förderung der Variablen-Kontroll-Strategie im Physikunterricht. *Der mathematische und naturwissenschaftliche Unterricht (MNU).*

Schwichow, M., & Kohnen, N. (angenommen). Das Waldschattenspiel. Nutzung eines kooperativen Brettspiels im Optikunterricht. *Unterricht Physik*.

KONFERENZBEITRÄGE UND EINGELADENE VORTRÄGE

Schwichow, M., & Härtig, H. (2015). What students learn from hands-on experimental tasks. Paper presented at ESERA 2015. Helsinki, Finland.

Schwichow, M., & Härtig, H. (2015). Was lernen Schüler beim Experimentieren?. Vortrag auf der Frühjahrstagung der Deutschen Physikalischen Gesellschaft Fachverband Didaktik der Physik. 09. – 13. 03.2015. Wuppertal.

Schwichow, M., & Härtig, H., & Höffler, T. (2015). Einfluss der Lerngelegenheit auf den Erwerb experimenteller Kompetenz. In: Bernholt (Hg.) - Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. [Jahrestagung der Gesellschaft für Didaktik der Chemie und Physik, 2014, Bremen]. S. 151-153, IPN, Kiel.

Schwichow, M., Croker, S.; Zimmerman, C., Höffler, T., & Härtig, H. (2014). Teaching the control of variables strategy: A research-synthesis. Paper presented at NARST 2014. Pittsburgh, USA.

Schwichow, M., Höffler, T., & Härtig, H. (2014).Merkmale einer effektiven Vermittlung experimentellen Strategiewissens In: Bernholt (Hg.) – Naturwissenschaftliche Bildung zwischen Science und Fachunterricht.[Jahrestagung der Gesellschaft für Didaktik der Chemie und Physik, 2013, München]. S. 195-197, IPN, Kiel.

Schwichow, M. (Nov. 2013). How should we teach inquiry skills in science? Results of a meta-analysis of intervention studies on the control-of-variables strategy. Presented at the Brown Bag Speaker Series of the psychological department at Illinois State University, Normal, USA.

Schwichow, M. (19.04.2013). Messung und Förderung experimenteller Kompetenz. Vortrag im Studienseminar Physik des Landesinstituts für Lehrerbildung in Hamburg.

Schwichow, M., & Härtig, H. (2013). Überprüfung eines Modells zur Entwicklung experimenteller Kompetenz. In: Bernholt (Hg.) – Inquiry-based Learning – Forschendes Lernen.[Jahrestagung der Gesellschaft für Didaktik der Chemie und Physik, 2012, Hannover]. S. 593-596, IPN, Kiel.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation – abgesehen von der Beratung durch meine Betreuer – nach Inhalt und Form selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden. Sie wurde weder im Ganzen noch in Teilen an einer anderen Stelle im Rahmen eines Promotionsverfahrens vorgelegt. Teile dieser Arbeit wurden bereits in Form von Zeitschriftenartikeln oder Tagungsbandbeiträgen publiziert bzw. sind zur Publikation eingereicht.

Kiel, den

Martin Schwichow

Danksagung

In den vergangenen fast vier Jahre ist diese Arbeit entstanden. Viele Menschen haben mich in dieser Zeit unterstützt und so ihren Beitrag zu dieser Arbeit geleistet. All diesen Menschen möchte ich an dieser Stelle meinen herzlichsten Dank aussprechen.

Ein besonderer Dank gilt meiner Familie. Liebe Mama, lieber Papa, liebe Oma, lieber Opa, lieber Johannes – auf eure Unterstützung konnte und kann ich mich immer verlassen. Euer größter Beitrag ist jedoch, dass ihr stets an mich geglaubt habt und mir beigebracht habt an mich zu glauben. Euer Mut, euer Tatendrang, euer Humor und vor allem eure Zuversicht sind die Grundzutaten dieser Arbeit.

Mein lieber Onkel Nachum, du hast mich schon früh für das Denken begeistert und mir gezeigt, dass auch ich denken kann. Ich hoffe man erkennt, dass in dieser Arbeit Gedanken stecken.

Liebe Lise, danke für die zahlreichen praktischen Hilfestellungen und die anregenden fachlichen Diskussionen. Aber vor allem danke, für die Ablenkung und Erholung von der Arbeit.

Der Grundstein für diese Arbeit wurde an keiner Universität, sondern von den Lehrerinnen und Lehrern der IGS Wedel gelegt. Hättet ihr mir nicht unzählige Chancen gegeben und Geduld mit mir erwiesen, hätte ich niemals auch nur das Abitur erlangt. Danke dafür und für das Interesse das ihr in mir geweckt hat.

Liebe Manja, du hast mich nach Kiel geholt und ein großes Stück zu meiner Professionalisierung beigetragen. Danke, dass du deine Erfahrungen mit mir geteilt hast.

Lieber Hendrik, dir gebührt für so viele Dinge Dank. Besonders bedanken möchte ich mich für unsere kritischen und offenen Diskussionen. Du hast mich gefördert indem du mich gefordert und motiviert hast. Ich habe viel in unseren Gesprächen und von dir gelernt. Es hat mir stets Spaß gemacht mit dir zu arbeiten.

Frau Prof. Dr. Beate Sodian, Ihnen möchte ich herzlich dafür danken, dass sie sich spontan bereit erklärt haben ein Zweigutachten für meine Arbeit anzufertigen.

Dear Corinne and dear Steve, thank you so much for hosting this strange German and introducing me to American culture. Thank you for your enormous help with our publications and all the support that you offered. It is always a pleasure to work with you.

Lieber Simon, liebe Lina, wenn ihr in dieser Arbeit blättert werdet ihr deutlich euren Beitrag erkennen. Ihr habt mir durch eure verlässliche und fachlich stets einwandfreie Arbeit sehr geholfen. Ohne eure Hilfe hätte ich niemals sämtliche Studien durchführen können. Mir hat die Zusammenarbeit mit euch stehst Freude bereitet und ich habe einiges durch eure kritischen Fragen und Anregungen gelernt. Vielen Dank dafür.

Liebe Ulrike, vielen Dank für das Korrekturlesen meiner Manuskripte und natürlich die netten und interessanten Pausenunterhaltungen.

Liebe Physikabteilung, Danke für die konstruktiven Diskussionen und vor allem für die Dinge, die das Arbeiten in einer Gruppe schöner machen.

Liebe Kolleginnen und Kollegen aus dem IPN, Danke für eure Unterstützung bei der Literaturbeschaffung, der Anfertigung der Experimentierboxen, und bei all den kleineren Verwaltungsherausforderungen.