International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

# The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy

Martin Schwichow, Simon Christoph, William J. Boone & Hendrik Härtig

**To cite this article:** Martin Schwichow, Simon Christoph, William J. Boone & Hendrik Härtig (2016): The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy, International Journal of Science Education, DOI: 10.1080/09500693.2015.1137651

**To link to this article:** http://dx.doi.org/10.1080/09500693.2015.1137651

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, 2016
http://dx.doi.org/10.1080/09500693.2015.1137651

Routledge
Taylor & Francis Group

# The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy

Martin Schwichow[a], Simon Christoph[a], William J. Boone[b] and Hendrik Härtig[a]

[a]Department of Physics Education, Leibniz Institute for Science and Mathematics Education (IPN), Kiel, Germany; [b]Department of Educational Psychology, Miami University, Oxford, OH, USA

**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. 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.

**ARTICLE HISTORY**
Received 23 March 2015
Accepted 29 December 2015

**KEYWORDS**
Control-of-variables strategy; Rasch analysis; scientific reasoning; inquiry skills; experimental skills; assessment instrument



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, 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 that the choice of test instruments used to measure outcomes had a significant influence on student control-of-variables strategy (CVS) measures (Ross, 1988; Schwichow, Croker, Zimmerman, Höffler, & Härtig, in press). A reason for the incoherence of student measures across instruments might be that different instruments cover different sub-skills of the broader construct 'CVS' detailed by Chen and Klahr (1999).

**CONTACT** Dr. Martin Schwichow schwichow@ipn.uni-kiel.de Department of Physics Education, Leibniz Institute for Science and Mathematics Education (IPN), Olshausenstraße 62, Kiel 24098, Germany

Supplemental data for this article can be accessed at 10.1080/09500693.2015.1137651.

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.

## The 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' (Tschirgi, 1980) or '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. (Figure 1)

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 (1988, 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

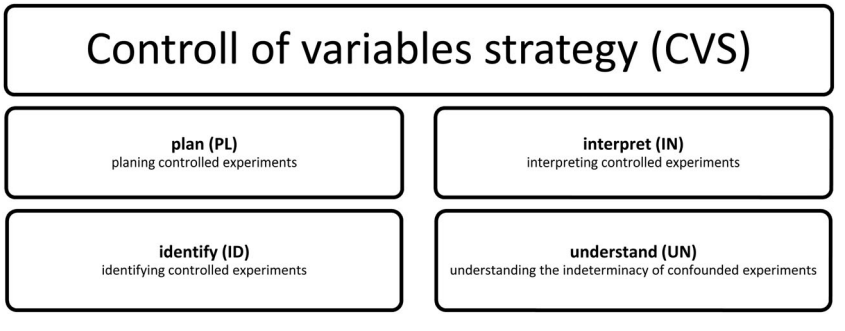



**Figure 1.** Sub-skills of the CVS construct according to Chen and Klahr (1999).

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 it is linked to broader educational goals such as inquiry skills and argumentation (Kuhn, 2005). 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-skill. To build this knowledge more extensive CVS instruments are required to evaluate the effect of instructions on student achievement regarding different CVS sub-skills.

## Literature review

Concerns about the comparability of CVS measures based on existing instruments originate from a meta-analysis (Schwichow et al., in press) 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 significantly 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. (in press) 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 are 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 et al., in press). 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 sub-skills 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 et al. (in press) 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 the 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 students 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 eighth 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 Klahr's (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, the 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 CVS 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 (ID) 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 practice students' science process skills like CVS depend 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 and 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).

## 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 1 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

**Table 1.** Overview of existing CVS multiple-choice instruments.

| Test | PL | ID | IN | UN | Format | Domain |
|---|---|---|---|---|---|---|
| Test by Staver (1984) | | | 1 | | Open response/multiple-choice | Biology |
| Test by Staver (1986) | 1 | | | | Open response/multiple-choice | Physics |
| Chewing gum test by Koerber et al. (2005) | | | 1 | | Interview | Everyday |
| Oral health test by Croker and Buchanan (2011) | 1 | | | | Interview | Everyday |
| 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 | Everyday |
| Airplane task by Bullock and Ziegler (1999) | 1 | 1 | | | Interview with card choice | Everyday |
| CVS posttest by Chen and Klahr (1999) | | 15 | | | Multiple-choice | Biology, Everyday |
| CVS tests by Kuhn and Dean (2005) | 5 | | | | Online interactive test | Geo-science, Everyday |
| CVS posttest by Dean and Kuhn (2007) | 5 | | | | Online interactive test | Geo-science |
| Meta-strategic knowledge test by Zohar and David (2008) | | | | 6 | Open-response items | Biology |

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



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 sub-skill might be present because some formats lend themselves to accessing specific CVS sub-skills. 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 sub-skill might be testing efficiency. For instance, 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 item types which have been used for specific CVS sub-skills (but not all sub-skills) the isolated impact of instrument format and utilized CVS sub-skill in existing CVS instruments is not known.

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.

## 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 three research questions:

(1) What is the evidence for validity and reliability of the new CVSI instrument?
(2) What is the pattern of item difficulty of the CVS sub-skills?
(3) What is the pattern of item difficulty for items covering different physics topics?



## Instrument development

We decided to develop the CVSI using a 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 representations 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 U.S. (National Research Council, 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 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 2.

The five items for the CVS sub-skill interpreting (IN) and the seven 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 3 shows an example of an understanding (UN) item. The full CVS inventory is available as online supplemental material.

## Data collection

The CVSI was administered to 386 seventh-, eighth- and ninth-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 were 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 3 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.

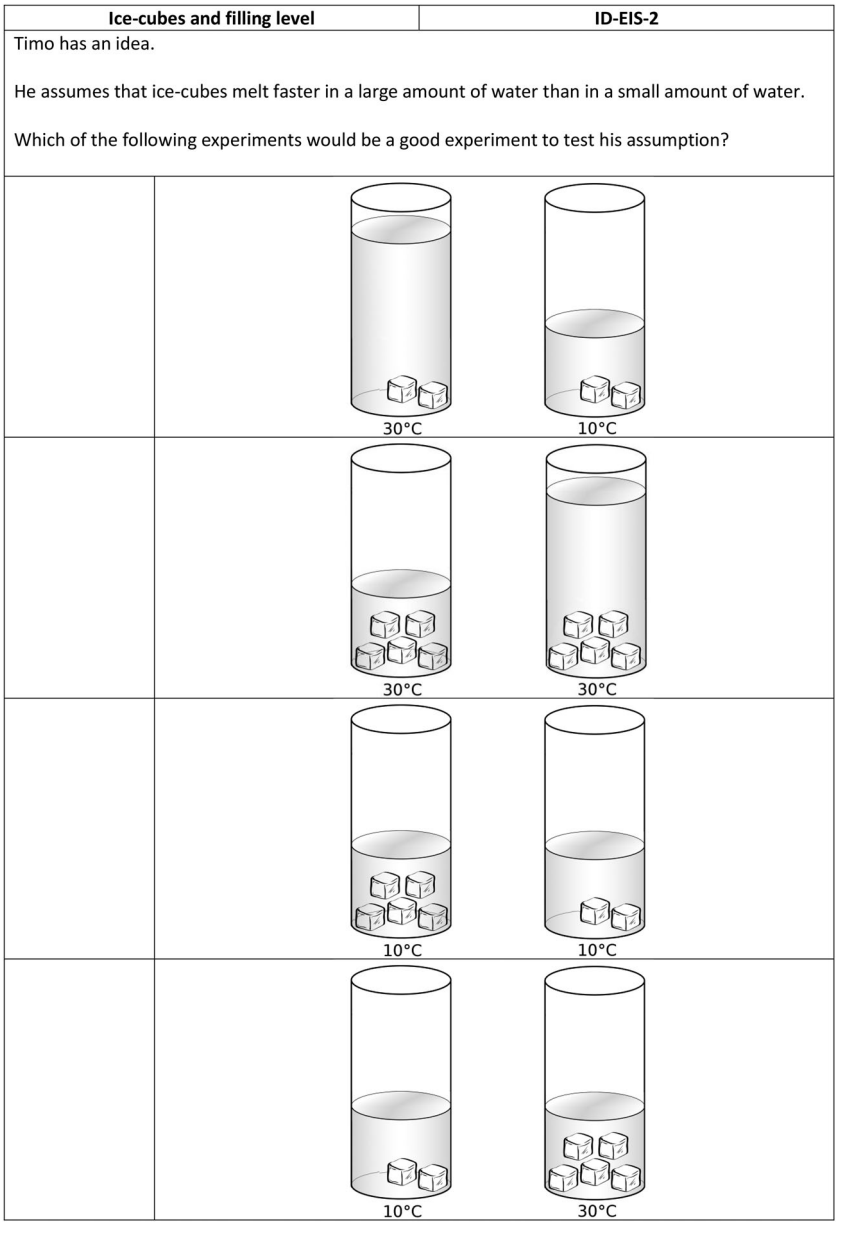



**Figure 2.** Example of an identifying (ID) item. Answer two is correct because the critical variable 'filling level' varies, while all other variables are the same between both conditions.

## 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.

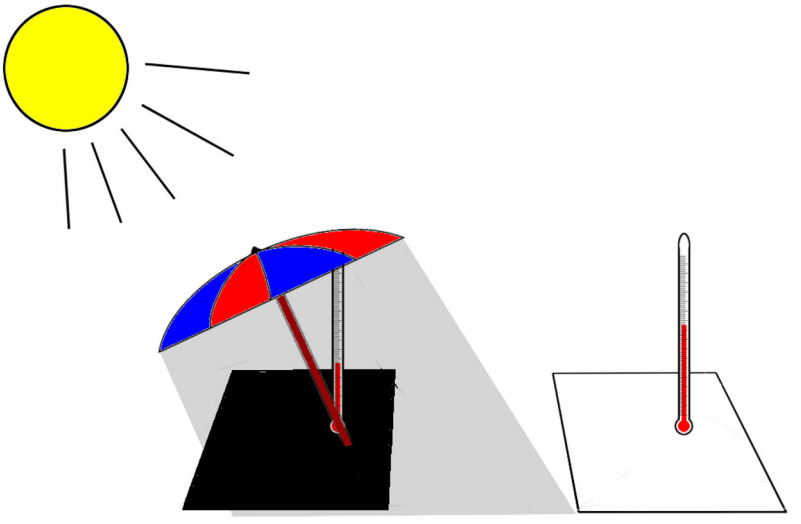
| Sun | UN-SO-1 |
|---|---|
| Toni did the following experiment:<br>  <br>What does his experiment show? | |
| | Objects in the shadow are getting less hot. |
| | The color of an object influences the object's temperature. |
| | Color and shadow influence the temperature of the object. |
| | The experiment does not allow any valid conclusion. |

**Figure 3.** 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.

### *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 indices 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 Programme for International Student Assessment (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. Lower logit values represent easier items (or less able students) and larger values correspond to more challenging items or more able students. The pure logit values of this scale are not informative because the scale is relative. However, as the scale is linear and as item and person measures are expressed on the same scale, we can compare values from the same scale (item and person measures) with each other. For example, we can identify items that are typically solved by students having a specific person measure (items that have the same or lower item measures). 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 a 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* Bonferroni 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.

## 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 an 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 and 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 person outcome 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 0.95.

### *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 $\alpha = .88$) 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 4(a) 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, s.d. = 0.92) are more difficult than identifying (ID) (mean item difficulty = −1.24, s.d. = 1.10) and

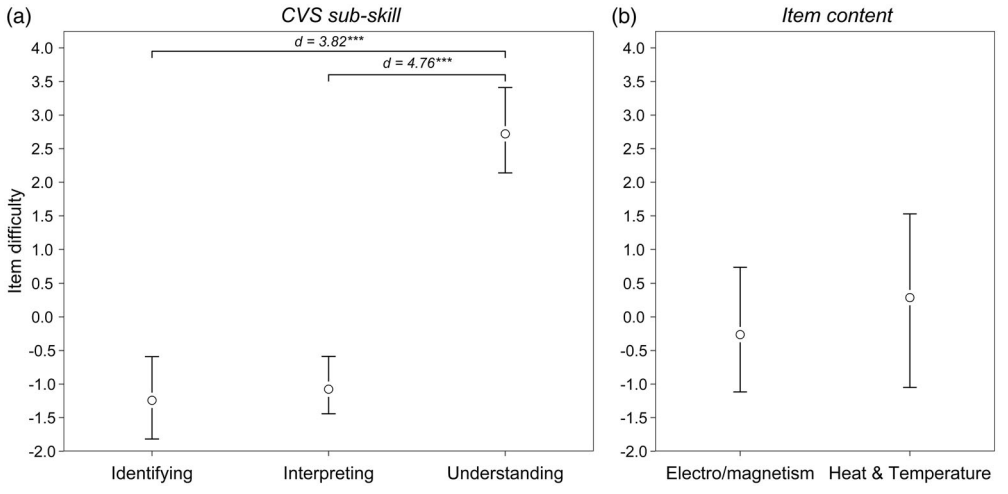



**Figure 4.** Mean item difficulties and standard errors (in logits) for (a) identifying (ID), interpreting (IN) and understanding (UN) items and (b) for items with content from heat and temperature and electro/magnetism. Horizontal lines represent significant differences in item difficulties ($p > .01$). The corresponding effect size is reported using Cohen's *d*.

interpreting (IN)) items (mean item difficulty = −1.08, s.d. = 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.

### *Item difficulty of heat and temperature and electro/magnetism items*

The mean item difficulties of heat and temperature and electro/magnetism items (see Figure 4(b)) 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 $p < .20$ to prove the truth of a null hypothesis. An independent *t*-test shows that the difference between the mean item difficulty of electro/magnetism ($m = -0.26$, s.d. = 1.70, $n = 12$ items) and heat and temperature items ($m = 0.29$, s.d. = 2.44, $n = 11$ items) is non-significant ($t = 0.62$, df = 17.71, $p = .54$).

## 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 $\alpha = .88$). All items fit the Rasch model as the MNSQ Outfit values are below 1.3. A Wright Map (see Figure 5) 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 Figures 4(a) and 2(b) 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

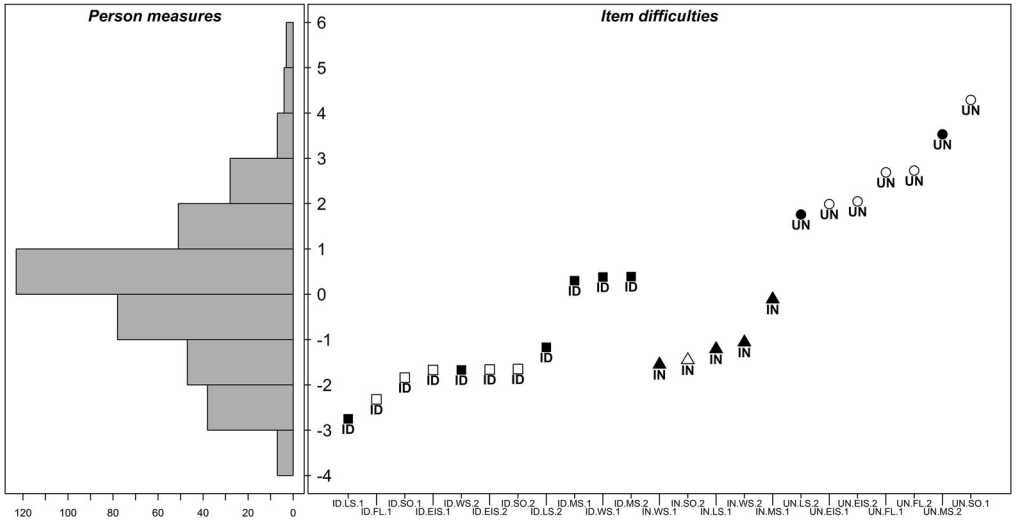



**Figure 5.** Wright Map of person measures and item difficulties (in logits) derived from the CVSI. Item difficulties and person measures are expressed on the same scale with easy items and less able students at the bottom and challenging items and more able students in the upper part of the scale. 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.

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 alternative is to increase the difficulty of identification items (ID) by asking students to identify the confounded experiments. This might be more challenging as they need to acquire similar thinking patterns to understanding items (UN) but still less demanding than understanding items (UN) as the controlled experiment prompts students to realize the differences between confounded and controlled experiments. Furthermore, the difficulty of both identification (ID) and interpreting items (IN) could be increased 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. Moreover, the item difficulty of all sub-skills might depend on students' conceptual knowledge to understand which variables might influence the experimental outcome. For example, it might be very challenging for students who lack the concept of wave optics to understand, why the orientation of a birefractive crystal influences the phenomenon of double refraction and to solve a CVS task concerning this complex content. By including items with more (including, for example, typical misconceptions) and less complex content (e.g. everyday-live content) in the CVSI, we might be able to analyze the interaction between students' conceptual understanding and their use of CVS. 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. The sufficient psychometric characteristics of the current item set and the systematic structure of the CVSI make it an ideal instrument to study further task features like item content (e.g. scientific versus everyday-live) and number of independent variables on students' CVS skills. 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 offer 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.

### *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 held constant in the new instrument. The results

of our study show that understanding items (UN) are systematically more challenging than items utilizing the CVS sub-skill of identifying (ID) and interpreting (IN) (see Figure 4(a) or 5). 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 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.

Although identification (ID) and interpretation items (IN) have theoretically different requirements (see above), we found no empirical differences between their mean item difficulties. We might be unable to detect differences as our sample (seventh, eighth and ninth graders) masters both sub-skills very well. Nevertheless with respect to testing and teaching CVS, it would be of interest to get more insights about the structure of CVS. Further studies should include more heterogeneous samples (e.g. students of different ages) and confirmatory factor analyses to test the structure of the CVS construct.

One important implication of our findings is that past instruments that lacked UN items may overestimate students' 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 sub-skills 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, 2001). Of course teachers need to refer to the invalidity of confounded experiments when introducing CVS as only controlled experiments allow valid conclusions. Based on the current findings we suggest that they should compare controlled to confounded experiments (similar to the identification (ID) sub-skill) rather than directly discussing confounded experiments (similar to the understanding (UN) sub-skill). 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) sub-skill to students. An intervention study by Zohar and 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 instruction 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 understanding (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.

## 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.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Notes on contributors

***Martin Schwichow*** owns a PhD in physics ecation. His main research interests are the development of scientific reasoning and inquiry skills.

***Simon Christoph*** has a masters degree in physics and mathematics education and is currently working as a high school teacher.

***Prof. William J. Boone*** is a specialist in the use of Rasch psychometric techniques to design and evaluate tests/surveys. He also specializes in the computation of Outcome Measures.

***Prof. Hendrik Härtig*** works in the field of physics education. His main research interests are the role of language in science teaching and students' performance in scientific inquiry.

## ORCID

*Martin Schwichow* http://orcid.org/0000-0001-9694-7183

## References

Adey, P., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school students. *Journal of Research in Science Teaching*, *27*(3), 267–285.

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

Bryant, P., Nunes, T., Hillier, J., Gilroy, C., & Barros, R. (2013). The importance of being able to deal with variables in learning science. *International Journal of Science and Mathematics Education*, *13*(*1*), 145–163.

Bullock, M. (1991). *Scientific reasoning in elementary school: Developmental and individual differences.* Paper presented at SRCD, Seattle, WA. Retrieved June 2015, from http://www.eric.ed.gov/PDFS/ED350149.pdf

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12. Findings from the Munich longitudinal study* (pp. 38–54). Cambridge: Cambridge University Press.

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.

Croker, S., & Buchanan, H. (2011). Scientific reasoning in a real-world context: The effect of prior belief and outcome on children's hypothesis-testing strategies. *British Journal of Developmental Psychology*, *29*, 409–424.

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

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91* (3), 384–397.

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

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

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

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. *Developmental Psychology*, *37*(5), 620–629.

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.

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

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. Learning, development, and conceptual change* (1st ed.). Cambridge, MA: MIT Press.

Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.

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

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*(1), 11–24.

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, OR: Winsteps.com. Retrieved June 2015, from http://www.winsteps.com/

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.
Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34* (4), 207–218.
Millar, R., & Driver, R. (1987). Beyond processes. *Studies in Science Education*, *14*(1), 33–62.
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, DC: National Academy Press.
National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies.
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). Washington, DC: American Educational Research Association.
Pellegrino, J. W., Wilson, M. R., & Koenig, J. A. (2013). *Developing assessments for the next generation science standards*. Washington, DC: The National Academies Press.
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.
Piekny, J., Grube, D., & Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education*, *36*(2), 334–354.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
Ross, J. A. (1988). Controlling variables: A meta-analysis of training studies. *Review of Educational Research*, *58*(4), 405–437.
Rousmaniere, F. H. (1906). A definition of experimentation. *The Journal of Philosophy, Psychology and Scientific Methods*, *3*(25), 673–680.
Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (in press). Teaching the control-of-variables strategy: A meta analysis. *Developmental Review*. doi:10.1016/j.dr.2015.12.001
Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, *21*(4), 22–27.
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. *Journal of Research in Science Teaching*, *21*(5), 517–526.
Staver, J. R. (1986). The effects of problem format, number of independent variables, and their interaction on student performance on a control of variables reasoning problem. *Journal of Research in Science Teaching*, *23*(6), 533–542.
Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development 51* (11), 1–10.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press/University of Chicago.
Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*. Chicago, IL: 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.
Zohar, A., & David, A. B. (2008). Explicit teaching of meta-strategic knowledge in authentic classroom situations. *Metacognition Learning*, *3*(1), 59–82.