Title

Challenging the control-of-variables strategy: How confounded comparisons can support children's science learning

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Lörch, Lucas
Bonawitz, Elizabeth
Brod, Garvin

Publication Date

2024

Peer reviewed

Challenging the control-of-variables strategy: How confounded comparisons can support children's science learning

Lucas Lörch[1,2], Elizabeth Bonawitz[3], Garvin Brod[1,2,4]

[1]DIPF Leibniz Institute for Research and Information in Education, Frankfurt/Main, Germany
[2]Center for Individual Development and Adaptive Education of Children at Risk (IDeA), Frankfurt/Main, Germany
[3]Harvard Graduate School of Education, Cambridge MA, USA
[4]Department of Psychology, Goethe University Frankfurt, Frankfurt/Main, Germany

Abstract

The control-of-variables strategy is often considered to be the superior strategy when children learn from experiments. However, by simulating Bayesian likelihoods of outcomes from a water displacement task, we show that certain confounded comparisons may support belief revision better than controlled comparisons. We tested this assumption by experimentally varying the types of comparisons that participants observed in a learning task involving balls of different sizes and materials (N = 90, age range 6- to 9-yrs). In the Size, Material, and Mixed conditions we presented controlled comparisons. In the Confounded Incongruent Condition, we presented confounded comparisons in which the larger ball was made of the heavier material. In line with our hypotheses, children in the Confounded Incongruent Condition revised their beliefs more than children in the other conditions, as indicated by higher transfer test scores. These findings suggest that confounded comparisons may in fact sometimes provide more optimal information for learning.

Keywords: control-of-variables strategy; science learning; Bayesian computational modeling; virtual simulation

Introduction

As part of inquiry-based science instruction, experiments can be an effective way to learn about scientific phenomena (Minner et al., 2010). The logic is that when learners conduct or observe experiments, or interpret their results, they can learn about the involved phenomena. In the context of such learning experiments, the control-of-variables strategy (CVS; Chen & Klahr, 1999) is generally considered the gold standard. In this strategy, only the variable of interest is varied, while all other variables are held constant. Because this strategy is central to confirm causal effects in scientific experiments, it has been assumed that it is also central to learning from experiments in an educational setting. Accordingly, research in science learning has focused mainly on investigating how to foster CVS as a skill (Schwichow et al., 2016), rather than questioning its superiority as a general learning strategy.

However, the philosophical community has argued that one can also learn from "soft interventions", i.e., interventions that change the value of one variable without necessarily de-confounding other variables (see, e.g. Eberhardt & Scheines, 2007). Cognitive Developmentalists have similarly pointed to noisier forms of interventions to support learning. For example, Gopnik and Wellman (2012) claimed that informative interventions "need not be the systematic, carefully controlled experiments of science" and that "even less controlled interventions can be extremely informative about causal structure" (Gopnik & Wellman, 2012, p.18).

To explore how learners reason about different possible causal associations, powerful tools for analyzing the informativeness in learning experiments have been developed. In particular, Bayesian computational models of learning provide a framework for considering how a learner might probabilistically evaluate competing causal beliefs in light of evidence. These models have helped to characterize and explain the causal learning of children and adults alike (e.g. see Gopnik & Tenenbaum, 2007, Gopnik & Bonawitz, 2014 for a review).

In the present study, we show that confounded interventions can be not only as informative, but even more informative, than controlled interventions. We first use Bayesian methods to provide a logical and mathematical basis for this argument. We then provide empirical support for this hypothesis by reporting results from a science learning study. The paradigm and learning context were based on a study by Theobald and Brod (2021). We used a virtual learning environment in which two balls of different size (small, medium, large) and material[1] (polystyrene, wood, iron) were shown above water containers with equal amounts of water. Elementary school children were presented with confounded or controlled comparisons and were asked to

[1] "Volume" and "density" would be the more physically accurate terms. However, as this is not a physical study, we will use the more intuitive and simple terms "size" and "material".

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

predict which of the balls displaces more water on a 5-point scale (left, rather left, equal, rather right, right). Then, participants were shown an animation of the outcome.

Learning from confounded and controlled experiments

If certain conditions are met, i.e. if all explanatory variables are independent of each other and if we have no prior assumptions about the causal relationships between these variables and the outcome, the CVS will always be superior. In natural learning situations, however, these conditions are almost never met. First, children hold specific assumptions about natural phenomena in order to explain their everyday experiences. These assumptions can sometimes be inaccurate (Confrey, 1990; Vosniadou & Ioannides, 1998). For example, in the domain of water displacement, elementary school children typically believe that the mass, rather than the volume, of an object determines how much water it displaces (Dawson & Rowell, 1984; Linn & Eylon, 2000; Piaget & Inhelder, 1978).

Second, children's reasoning takes place in the context of more abstract framework theories (Carey, 2000; Keil, 1991; Wellman & Gellman, 1992; Gopnik & Meltzoff, 1997; Vosniadou, 1994). These framework theories define which concepts or causal relationships are considered during learning. For example, children apparently have a "more causes more" heuristic. As a result, they appear to consider only positive associations between the variables involved (size, material, mass) and the amount of displaced water. That is, they may not consider the hypothesis that smaller objects displace more water than larger ones.

Third, the physical world is full of interrelated variables that are potentially relevant to a given phenomenon. For example, the mass of an object is the product of its volume and its density. Thus, if either volume or density is varied, mass will necessarily be varied as well. This means that the CVS may not be readily applicable in such a case[2]. In summary, learners' preconceptions about scientific phenomena, their framework theories, and the interrelatedness of variables uniquely characterize reasoning in the context of science learning.

How do these characteristics affect the adequacy of the CVS? In our example of water displacement, consider a learner who believes that mass determines water displacement, but is not entirely sure—it could also be volume. The learner strictly follows the CVS for deciding between mass and volume as causal variables. Therefore, in a learning experiment, this learner compares a large iron ball with a small iron ball to conduct a controlled test of the effect of mass. This person will observe that the large iron ball displaces more water and will take this as evidence supporting the hypothesis that mass determines water displacement. Because of the inherent confounding of size and mass, this seemingly controlled experiment is actually misleading.

Now imagine another learner who holds a mass misconception but does not follow the CVS. This learner compares a large but light object (such as a large polystyrene ball) with a small but heavy object (such as a small iron ball). If the learner had no framework theories, the result that the large polystyrene ball displaces more water would be ambiguous. It could indicate that volume has a positive effect or that mass has a negative effect on water displacement. However, because negative effects are not considered by the learner, the result clearly supports volume as the causal variable. The comparison, although confounding volume and density, is highly informative. These examples demonstrate that there are cases where the CVS is not optimal. This calls for methods to decide which types of comparisons are most informative in a novel reasoning situation

Bayesian likelihood as a measure of informativeness

Bayesian statistics defines principles of induction, i.e., of learning from data about the true state of the world. The main principle is Bayes' theorem: the probability of a state after seeing the data (called the posterior) is proportional to the product of the probability of that state before seeing the data (called the prior) and the probability of the data given that state (called the likelihood) (see e.g. Griffiths et al., 2001 for a review).

This logic can be translated to reasoning in learning experiments. The prior corresponds to a learner's initial knowledge, the likelihood corresponds to the information provided by a learning experiment, and the posterior corresponds to the knowledge after learning. For example, a learner's prior knowledge about water displacement can be represented by a probability distribution over the candidate causal variables (size, material, mass). If a learner believes that mass determines water displacement, then mass would have the largest probability in this distribution. If this learner is then presented a comparison of a large polystyrene ball against a small metal ball in a learning experiment, the outcome that the larger polystyrene ball displaces more water is very likely given volume is the true causal variable, but unlikely if mass or density is the true causal variable. Multiplying the learner's prior probability distribution with these likelihoods yields a posterior probability distribution and thus, a model of the learning process. Recently, this modeling approach has been applied to children's learning in water displacement tasks, demonstrating the feasibility of such an approach to capturing children's learning (Colantonio et al., 2023).

Since the likelihood represents the information provided by the learning experiment, it is the key to a formal analysis of informativeness. The likelihood of an outcome in a learning

[2] To overcome this confound, one could use a container and fill it with different amounts of water. However, this brings other problems as the density of such an object is unclear.

experiment is determined in 5 steps: (1) numerically coding the features in question for both conditions shown in the experiment, (2) defining normal distributions with the feature value as the mean and the feature value*0.22 (Weber's ratio representing perceptual noise; Dehaene, 2007; Droit-Volet et al., 2008) as the standard deviation, (3) drawing a large number of values from these distributions and counting whether these values are equal or different, (4) translating these counts to the answering scale used in the learning experiment by means of linear regression for the upper and lower part of the scale. These steps are described in detail in Colantonio et al. (2023).

The method is applied to each relevant feature, i.e., size, material, and mass. The chosen code for each of these features (as defined in step 1) should be plausible in the context of the study design and the assumed knowledge of the children, and it should account for the interrelatedness of the variables. For simplicity, we chose to code the three ball sizes (small, medium, large) by the integers 1, 2, and 3, respectively, because we assumed that this would reflect children's perception of these sizes. We also decided to code the three materials (polystyrene, wood, iron) with the integers 1, 2, and 3, respectively, as we assumed that this would reflect children's knowledge of the densities of the involved materials. Moreover, we coded mass as the product of the values for volume and density, as this reflects the actual relationship between these variables.

In addition to believing that either volume, density, or mass determines water displacement, children might also believe that water displacement does not follow a systematic causal rule. Therefore, we defined the likelihood for a "random" hypothesis by dividing 1 by the number of response options (in our case there were five options: left, rather left, equal, rather right, right). That is, given the random hypothesis, all outcomes have a likelihood of 0.2. Note that while likelihoods are calculated for all response options, and while the likelihoods across options sum to 1, the most relevant likelihood for modeling the learning process is that of the actual outcome. When we use the term likelihood, we mean the likelihood of the outcome of the learning experiment. It is this outcome that children observe and that provides them with new information (Colantonio et al., 2023).

A taxonomy of comparisons

In order to decide which comparisons are most informative in a novel reasoning situation, we must first define the types of comparisons that can occur in this context. One can distinguish between controlled (only one variable varies), confounded (multiple variables vary), and identical (no variable varies) comparisons. Subcategories can then be defined for both controlled and confounded comparisons. Controlled comparisons contain a subcategory for each variable involved (only variable A varies, only variable B varies, etc.). In our case, we distinguish between comparisons in which only the size of the balls is varied (called size comparisons) and comparisons in which only the material of the balls is varied (called material comparisons). For confounded comparisons, one can distinguish whether the variables involved are contrasted or matched. In our case, we distinguish between comparisons in which one ball is both larger and made of the heavier material (called confounded congruent) and comparisons in which one ball is larger but made of the lighter material (called confounded incongruent). In cases with more than two variables, more subcategories arise. Table 1 shows an example stimulus for each of the five types.

Results of computational analyses Calculating likelihoods for these five types of comparisons provides insight into their informativeness. The likelihoods in Table 1 show that identical comparisons and confounded congruent comparisons are uninformative. The likelihood of the outcome (that the large iron ball displaces more water/that the two wooden balls displace equal amounts of water) is equally likely given the three hypotheses. Thus, nothing can be learned from these comparisons. Size comparisons are potentially misleading, even though they vary only one variable. The result that the large polystyrene ball displaces more water is equally likely under either the size hypothesis or the mass hypothesis. Thus, a learning experiment presenting such a comparison could be misinterpreted as indicating that heavier objects displace more water. Finally, both material comparisons and confounded incongruent comparisons are informative. The outcome of these comparisons is quite likely given the correct size hypothesis, but rather unlikely given the incorrect material or mass hypothesis. The likelihoods suggest that confounded incongruent comparisons are the most informative, as they allow the clearest distinction between the correct and incorrect causal variables.

Table 1: Exemplary comparisons of each type together with likelihood (LH) of the outcome given the three hypotheses.

Type	Exemplary comparison	LH given size	LH given material	LH given mass
Size		0.99	0.26	0.99
Material		0.51	0.08	0.07
Confounded incongruent		0.91	0.00	0.03
Confounded congruent		0.99	0.99	1.00
Identical		0.53	0.51	0.53

Empirical exploration of modeling results

Our computationally simulated results reveal cases in which the confounded comparisons can actually provide *more* informative data than controlled comparisons. Further, our results reveal that some controlled comparisons (i.e., size comparisons) may even provide misleading information to a learner.

In the present study, we sought to explore whether children's learning outcomes follow these simulated predictions. How might confounded incongruent comparisons support children's learning about water displacement? To investigate this, participants completed three trials in the learning experiment described above. The types of comparisons presented in the learning experiment were varied between participants. In the Size, Material, and Confounded Incongruent experimental conditions, the respective comparisons were shown. In addition, there was a Mixed Condition in which material and size comparisons were subsequently shown. This condition was included because it reflects a standard application of the CVS, i.e., varying the relevant causal variables one after the other. We hypothesized that children's performance in the posttest and the transfer test would be better in the Confounded Incongruent Condition than in the Size, Material, and Mixed conditions.

Method

The present study has been preregistered on the Open Science Framework: https://osf.io/te34n.

Sample

To date, 90 participants have taken part in the study. Data collection is ongoing and will be completed when data from 168 participants has been collected. The mean age of the current sample was 8.07 years (SD = 1.04, min = 6, max = 9.97). There were 37 female and 53 male participants. Only children between the ages of six and nine years with sufficient knowledge of the German language were accepted as participants. Participants received a small thank-you gift worth approximately €5.

Procedure

Research assistants approached children or their parents at a booth in a local natural history museum and asked them to participate in a study that took approximately 15 minutes. If they agreed, they were taken to a separate room and parents signed an informed consent form. Parents were asked to wait outside or in a separate area of the testing room. The children were tested in individual sessions. They were instructed that their task was to learn about water displacement. They were told that the tasks would involve balls made of three materials (polystyrene, wood, iron). The children were given equally sized balls of these materials to feel their weight. Then a short video was played showing a person pushing a rubber ball under water and how this causes the water to rise. This ensured that the children understood that the issue was water displacement and not buoyancy.

Participants then completed the pretest (6 trials), the learning task (1 warm-up, 3 learning trials), the posttest (6 trials), and the transfer test (13 trials). In the pretest and posttest, children saw two balls printed on a sheet of paper and were asked to indicate which one would displace more water on a 5-point scale (left, rather left, equal, rather right, right). The learning task took place in a virtual learning environment in which two water containers with equal amounts of water were shown, each with a ball on top. As in the pre- and posttest, participants were asked to predict which of the balls would displace more water on a 5-point scale. After giving their answer, they saw an animation of the balls being pushed underwater and of the water rising. This animation was 5 seconds long, with the final water level shown for 2 seconds. The transfer test consisted of three subtasks. First, participants had to indicate which of two objects (e.g., pyramids, cuboids) would displace more water on a 5-point scale. Second, participants were asked to draw a line indicating how far the water would rise relative to a given object if certain objects were submerged. Finally, participants were explicitly asked why a larger and heavier ball displaces more water than a smaller and lighter one. All three tasks were completed on a sheet of paper.

Because the ability to read, write, or use a computer mouse varies greatly between the ages of 6 and 9, the experimenter's assistance was flexibly adapted to each child's individual needs. Upon completion of all tests (or premature termination of the study), the children received their thank-you gift.

Design & Material

There were four between-participants conditions (Size, Material, Confounded Incongruent, Mixed) that defined which types of comparisons were presented in the learning task. In the Size Condition and in the Material Condition, comparisons were presented in which only the size or only the material of the two balls was varied, respectively. In the Confounded Incongruent Condition, both the material and the mass of the two balls were varied in such a way that the larger ball was made of the lighter material. In the Mixed Condition, only the mass and only the material of the two balls was varied in subsequent trials, reflecting a standard application of the CVS.

The second of the three learning trials was an uninformative (confounded congruent) comparison. This trial was introduced to prevent children from drawing conclusions from the fact that the same one variable was varied throughout the task. In all four conditions, the same uninformative comparison (small polystyrene ball vs. large wooden ball) was used. Thus, the experimental manipulations only affected the first and third learning trials. The order of the first and third learning trials was counterbalanced.

The pretest contained two comparisons of each type involved in the learning task (material, size, confounded incongruent). The posttest contained the same items, but with the sides of the balls reversed and the order of the items changed. In the warm-up trial, an uninformative (confounded congruent) comparison was shown to allow children to become familiar with the virtual learning environment and its user interface. In the Mixed Condition, one item from the Material Condition and one item from the Size Condition was shown.

Results

Figure 1 shows the mean posttest and transfer test accuracies in the four conditions. To statistically test for differences between conditions, we fitted mixed logit regression models with the package lme4 (Bates et al., 2015). The dependent variable in these models was the accuracy of solving a posttest or transfer test item (0/1). The predictors were a categorical variable indicating the condition and a z-standardized decimal variable indicating children's age in years. The models contained by-participant random intercepts. The model for the posttest explained a significant amount of variance in the data ($\chi^2(4) = 14.33$; $p < .01$). The parameter estimates for this model are shown in the left columns of Table 2. The reference category is the Confounded Incongruent Condition. That is, the intercept indicates the log odds of solving a posttest item in that condition for a child with mean age (8.07 years). The parameter estimates show that accuracy on the posttest was significantly reduced when participants learned from size comparisons compared to confounded incongruent comparisons.

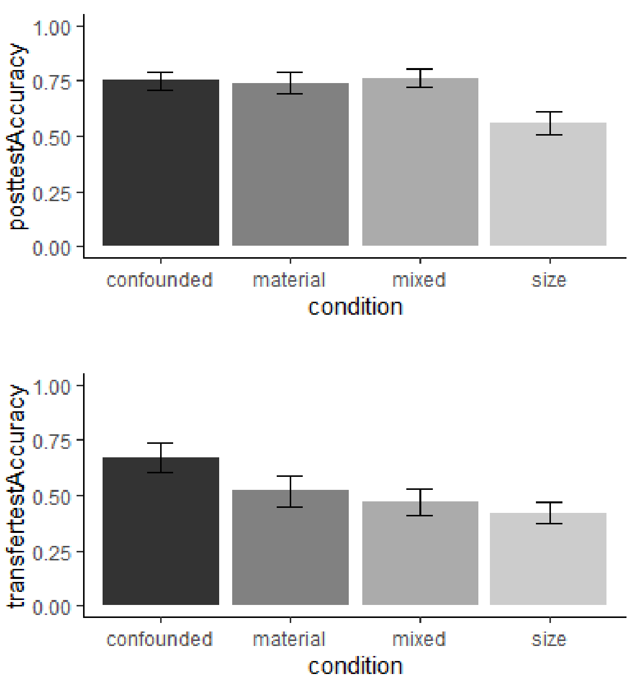

Figure 1: Means and standard errors of the posttest (top) and transfer test accuracy (bottom) in the four conditions.

The model for the transfer test also explained a significant amount of variance in the data ($\chi^2(4) = 10.77$; $p < .05$). The parameter estimates (Table 2, right columns) indicate that transfer test accuracy was significantly reduced when children learned from material, mixed, and size comparisons compared to confounded incongruent comparisons. Age had a positive effect on transfer test performance. In summary, confounded incongruent comparisons were just as helpful for learning as material and mixed comparisons as measured by posttest scores, and significantly better for learning across all conditions, as measured by transfer of knowledge.

Table 2: Parameter estimates of the mixed logit regression model predicting posttest and transfer test accuracy by condition and age.

Parameter	Posttest		Transfer test	
	Estimate	*SE*	Estimate	*SE*
Intercept	1.12***	0.25	1.00***	0.30
Material	0.07	0.35	-0.84*	0.41
Mixed	0.11	0.34	-0.84*	0.41
Size	-0.78*	0.34	-1.05*	0.42
Age	0.21	0.12	0.19	0.14

Note. * $p < .05$; ** $p < .01$; *** $p < .001$; The reference is the Confounded Incongruent Condition. Age is in years and z-standardized.

Discussion

When learning from experiments, the control-of-variables strategy (CVS; Chen & Klahr, 1999) is generally considered to be most informative. However, because children have certain preconceptions about scientific phenomena (Vosniadou & Ioannides, 1998) and because physical variables can be related, the CVS is not optimal in many reasoning situations. One way to formally analyze the informativeness of different types of comparisons is to perform simulations of the Bayesian likelihood. We performed such simulations for a learning task on water displacement. The results suggested that confounded incongruent comparisons may be most informative and that controlled size comparisons may be potentially misleading. This assumption was empirically supported with data from 90 elementary school children.

Logit regression models for the transfer test showed that accuracy was increased in the Confounded Incongruent Condition compared to the other three conditions, as expected. For the posttest, we found that performance was lowest in the Size Condition, while it was similar in the other conditions. This result was not as expected, as we hypothesized that the posttest performance would be most accurate in the Confounded Incongruent Condition. Nevertheless, this result is highly plausible. As discussed above, size comparisons in particular are potentially misleading. If children initially believed that mass determines water displacement, the size comparisons may have supported this inaccurate belief. Besides, we believe that the transfer test is more meaningful than the posttest because it indicates the actual change of concepts.

Our findings have important practical implications. Educational psychology has put a lot of effort into studying the ability to use the CVS and how to promote it (Schwichow et al., 2016). However, our results suggest an additional relevant skill. That is, the ability to analyze a reasoning situation in order to figure out what kinds of comparisons can actually be helpful. Since children already have difficulty understanding the logic of controlled experiments, the fact that these controlled experiments are sometimes misleading may be even more difficult to grasp. However, some pilot studies in our lab in which children created their own comparisons suggest that there may be an intuitive understanding of the informativeness of confounded comparisons. In our next study, we plan to contribute to this issue by investigating the intuitive or deliberate application of confounded comparisons in learning experiments.

In addition, our work raises an interesting theoretical question: what is the difference between learning and scientific inquiry? In discussing this question, it is important to note that science has two modes of inquiry, i.e., the exploratory mode and the confirmatory mode. These modes have different goals, i.e., to discover new causal patterns and to test expected causal patterns, respectively. We believe that the exploratory mode of scientific inquiry is more akin to learning. In this mode of inquiry, it is appropriate to confound variables in order to find new effects, because any effect will subsequently be backed by further confirmatory analyses. However, the confirmatory mode of scientific inquiry substantially differs from learning. In this mode of inquiry, any confound is a problem because it creates ambiguity. Controlled experiments maximize certainty about the causal effects of a variable and hence are the method of choice in the confirmatory mode of inquiry.

In summary, the ultimate goal of scientists is accuracy, certainty, and consideration of all possible effects (even unexpected and unlikely ones). For learners, the goal may be more practical—to be able to correctly predict certain outcomes in order to feel in control of the environment. Further theoretical and empirical work is needed to clarify the implications of these issues for human reasoning.

Conclusion

Our study challenges the CVS as a superior strategy in learning experiments by showing, both computationally and empirically, that confounded comparisons can support learning just as well (if not better) than controlled ones. We hope that our findings and the methods involved inspire researchers to delve deeper into the intricacies of reasoning in science learning.

Acknowledgements

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (BR 5736/2-2) to GB and by grants from the McDonnell Foundation and the Jacobs Foundation to EB.

We want to thank Joseph Colantonio for the sharing of his Python code, the Individualized Interventions Lab and the CoCoDev Research Group for their helpful feedback, and our research assistants Mina Biegel, Antonia Schneider, Jonah Stettner, Merle Fuchs, Alina Rieb and Brigitte Röblitz for their help with data collection.

References

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

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

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. https://doi.org/10.1111/1467-8624.00081

Colantonio, J., Bascandziev, I., Theobald, M., Brod, G., & Bonawitz, E. (2023). Priors, Progressions, and Predictions in Science Learning: Theory-Based Bayesian Models of Children's Revising Beliefs of Water Displacement. *IEEE Transactions on Cognitive and Developmental Systems*, *15*(3), 1487–1500. https://doi.org/10.1109/TCDS.2022.3220963

Confrey, J. (1990). Chapter 1: A Review of the Research on Student Conceptions in Mathematics, Science, and Programming. *Review of Research in Education*, *16*(1), 3–56. https://doi.org/10.3102/0091732X016001003

Dawson, C., & Rowell, J. (1984). Displacement of water: Weight or volume? An examination of two conflict based teaching strategies. *Research in Science Education*, *14*(1), 69–77. https://doi.org/10.1007/BF02356792

Dehaene, S. (2007). Symbols and quantities in parietal Cortex: Elements of a mathematical theory of number representation and manipulation. *Attention and Performance: Sensorimotor Foundations of Higher Cognition*, *22*. https://doi.org/10.1093/acprof:oso/9780199231447.003.0024

Droit-Volet, S., Clément, A., & Fayol, M. (2008). Time, Number and Length: Similarities and Differences in Discrimination in Adults and Children. *Quarterly Journal of Experimental Psychology*, *61*(12), 1827–1846. https://doi.org/10.1080/17470210701743643

Eberhardt, F., & Scheines, R. (2007). Interventions and causal inference. *Philosophy of science*, *74*(5), 981-995. https://doi.org/10.1086/525638

Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley interdisciplinary reviews: cognitive science*, *6*(2), 75-86.

Gopnik, A., & Meltzoff, A. N. (1997). Words, thoughts, and theories. MIT Press.

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development.

Developmental Science, 10(3), 281–287. https://doi.org/10.1111/j.1467-7687.2007.00584.x

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2001). Bayesian Models of Cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (1st ed., pp. 59–100). Cambridge University Press. https://doi.org/10.1017/CBO9780511816772.006

Keil, F. C. (1991). The emergence of theoretical beliefs as constraints on concepts. The epigenesis of mind: Essays on biology and cognition, 237-256.

Linn, M. C., & Eylon, B.-S. (2000). Knowledge Integration and Displaced Volume. *Journal of Science Education and Technology*, *9*(4), 287–310. https://doi.org/10.1023/A:1009451808539

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, *47*(4), 474–496. https://doi.org/10.1002/tea.20347

Piaget, J., & Inhelder, B. (1978). *Le développement des quantités physiques chez l'enfant: Conservation et atomisme* (4 éd). Delachaux et Niestlé.

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. https://doi.org/10.1016/j.dr.2015.12.001

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive Theories as Grammars for Causal Inference. In A. Gopnik & L. Schulz (Eds.), *Causal Learning* (1st ed., pp. 301–322). Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195176803.003.0020

Theobald, M., & Brod, G. (2021). Tackling Scientific Misconceptions: The Element of Surprise. *Child Development*, *92*(5), 2128–2141. https://doi.org/10.1111/cdev.13582

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual review of psychology*, 43(1), 337-375.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, *4*(1), 45–69. https://doi.org/10.1016/0959-4752(94)90018-3

Vosniadou, S., & Ioannides, C. (1998). From conceptual development to science education: A psychological point of view. *International Journal of Science Education*, *20*(10), 1213–1230. https://doi.org/10.1080/0950069980201004