Young children learn equally from real and thought experiments

Igor Bascandziev[1]

Susan Carey[1]

1. Harvard University

Abstract

As the history of science has documented, there is an important role for thought experiments in scientific progress. Yet, there is very little empirical research about whether and how children learn from thought experiments. Here, we asked that question in the context of 6-year-olds' developing theory of matter. At the outset of the study, over half of the children claimed that small pieces of matter weigh nothing at all. Children were randomly assigned to a Real (RE) and a Thought Experiment (TE) condition. The goal of each condition was to show – via demonstration in the RE and via mental simulation in the TE – that the weight of a single grain of rice can cause a card resting on a fulcrum to topple. We found that children simulated accurately in the TE, and they changed their judgments and justifications concerning the weight of small pieces *equally* from the TE and RE.

Keywords: Thought Experiment; Experiment; Weight; Matter.

Introduction

The "child as scientist" research program is often based on the observation that both children and scientists learn from data, generating the research question of how both populations do so (e.g., Gopnik & Shultz, 2007). But, as many have pointed out, both populations also often learn from testimony (e.g., Harris, 2012), and from thinking (e.g., Clement, 2009; Lombrozo, 2019). A paradigm example of learning from thinking is learning from thought experiments (TEs; Kuhn, 1977, Gendler, 2004; Neressian, 1992; Norton, 2004; see Bascandziev and Harris, 2019). Thought experiments are widely attested in episodes of theory construction in the history of science, with Galileo, Kepler, Einstein, and many others appealing to them in both published work and in the day-to-day notebooks they kept of their ongoing research.

The hypothesis that thought experimentation leads to new knowledge seems highly paradoxical. How can a process that involves **no** new data generate new knowledge? It would seem that the person engaging in a thought experiment already must know everything needed to generate its conclusion. This question is much discussed in the philosophy and history of science literature, with at least four classes of answers offered (see Brown and Fehige, 2019 for review). These include: 1) Thought Experiments (TEs) are examples of inferential reasoning. New knowledge results from inferences not yet drawn; this is equally true of all newly made deductive and inductive inferences (Norton, 2004), 2) TEs provide information that current knowledge leads to contradictions; it identifies areas of understanding requiring conceptual change (Kuhn, 1977), 3) Thought experimentation is often a form of conceptual modelling that itself may produce new data (Nersessian, 1992), and similarly 4) Sometimes TEs *do* generate new data; there is knowledge encapsulated in perceptual systems and systems of core cognition that can be used in perception based simulations that *do* generate new data not previously encountered (Gendler, 2004; Mach, 1897). These resolutions to the paradox of TEs are not mutually exclusive. Different episodes of thought experimentation may draw on distinct mechanisms, and they can also act in concert.

Despite the importance of TEs in the history of science, there is little empirical research on whether people actually *do* learn from thought experimentation. Here we ask whether even young children can do so. In the present experiment we compare learning from a thought experiment and from a real experiment with exactly the same structure.

Our participants are 6- and 7-year-olds at the very beginning of constructing a theory of matter in which the extensive concept of weight is differentiated from the intensive concept of density, and in which weight is taken to be a necessary feature of material entities that distinguishes them from non-material physically real entities, such as shadows, heat, and light (see Piaget and Inhelder, 1974; Smith, Snir, & Grosslight, 1992; Carey, 2009). At the beginning of this episode of conceptual change, young children take weight to be an accidental property of some objects. Children assert that a single grain of rice, a small piece of playdoh, or a grain of sugar weigh "nothing at all", "0 grams." They thus do not conceive of the weight of a pile of sugar or a large ball of playdoh as the sum of the weights of the individual grains of rice or the small pieces of playdoh that constitute the aggregate. This misconception results from a failure to distinguish felt weight from objective weight, a failure to differentiate weight from density, a lack of appreciation of the sensitivity of measuring devices, and a failure to distinguish physically real objects made of matter from physically real immaterial entities (see Carey, 2009, for review). These aspects of an intuitive theory of matter are constructed over the years of 6 to 12, partly in the course of elementary and junior high school science education.

The current study has a pre-training, intervention, post-training design. The pre-training and post-training assessed the progress the child had made towards the earliest stages in the construction before and after the intervention. Children were randomly assigned to one of two interventions: A

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

Thought Experiment and a Real Experiment. Both experiments targeted the belief that a single grain of rice weighs nothing at all.

The full experiment probes the child's concepts of matter as predictors of their beliefs about the weight of a single grain of rice, probes the effects of the thought experiment (TE) and real experiment (RE) on concepts of matter not directly targeted in the intervention, and explores the mechanisms through which the TE generates new knowledge. Here we report only one aspect of the full study: whether the RE changed children's beliefs about the weight of a single grain of rice, and if so, whether the TE did so as well, and to a similar extent. We analyze near transfer to beliefs about small pieces of other kinds of matter (playdoh and sugar) and analyze whether the RE and TE changed the nature of justifications children gave for their judgements that small portions of matter weigh something. In all cases we assess the efficacy of the TE to change beliefs, comparing this efficacy to that of the RE. We test two hypotheses and then address several further questions in the data we present here:

H1: Children in the TE condition will be able to simulate the RE data, drawing on knowledge encapsulated within perception.

H2: Children will learn from the RE that a grain of rice weighs something.

Q1: Do children learn from the TE that a grain of rice weighs something?

Q2: If so, is learning comparable to that from the real experiment?

Q3: Does the child generalize what is learned about a single grain of rice in the RE to other kinds of matter?

Q4: Do either or both interventions lead to progress in the construction of an extensive concept of weight for all material entities, as reflected in the child's justifications?

Method

Participants

A total of 122 children were recruited. Three children discontinued participation before the post-training trials, and were excluded, leaving a final sample of 119 children (M_{Age} = 82.30 months, SD = 6.92, range = 69 – 95 months). Children were randomly assigned to two conditions: Thought Experiment and Real Experiment condition. The two groups were comparable in terms of age (M_{Age_TE} = 82.58 months and M_{Age_RE} = 82.02 months, $t(117)$ = .44, p = .657) and the distribution of boys and girls in each condition was similar (($\chi 2$(1, N = 119) = .73, p = .393). The sample was drawn from a predominantly white, non-Hispanic, middle-class population from the Boston metro area. The testing took place in a quiet room at the Harvard Laboratory for Developmental Studies.

Procedure

All children received pre- and post-training interviews, as well as a training intervention (either a real or a thought experiment). All testing was conducted in a single session.

Pre- and Post-Training Interviews. To assess the effect of the training, all children received pre- and post-training interviews that probed their beliefs about whether a single grain of rice, a small piece of playdoh, and a single grain of sugar weighed a lot, a little, or nothing at all, and were asked to justify their responses. These interviews also probed children's understanding of addition involving 0, their understanding of scales, and probed other aspects of their concepts of matter (not reported in detail here).

Real Experiment Condition. Children were introduced to a fulcrum made out of 6 popsicle sticks stuck together (see Figure 1). After the experimenter placed a card on top of the fulcrum, she put a single grain of rice on a designated spot on one side of the card, and she asked the child: "What happened? Why?" The experimenter then added grains of rice, one at a time, until the card toppled. It took an average of 9 grains of rice to topple the card resting on a 6-stick fulcrum. When the card toppled, children were asked why adding the last grain of rice toppled the card. Next, children were introduced to a 3-stick fulcrum. The procedure was repeated: after seeing a grain of rice put on one side of the card, children were asked: "What happened? Why?" It took fewer grains (an average of 4) of rice to topple the card resting on a 3-stick fulcrum. Again, children were asked why adding the last grain of rice toppled the card. Finally, children were introduced to a single stick fulcrum where only one grain of rice was sufficient to topple the card. Children were asked: "What happened? Why?" Importantly, the experimenter did not invite children to make predictions or simulate what's going to happen. Children only observed the outcomes and were asked to interpret them.

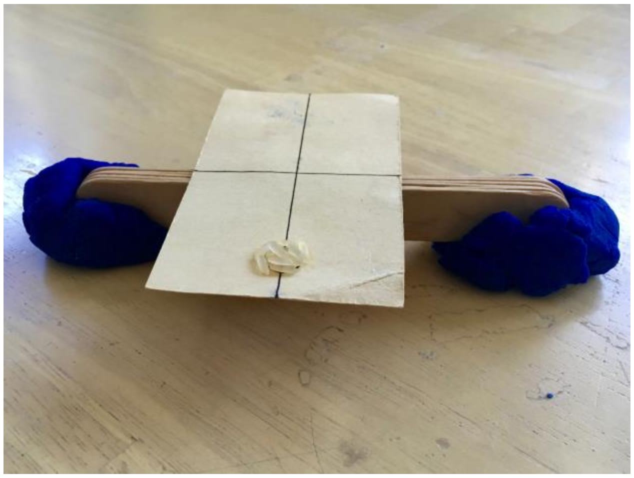

Figure 1. Picture of the card resting on a 6-stick fulcrum

Thought Experiment Condition. Children were first introduced to the fulcrum made out of 6 popsicle sticks and the card. Next, they were invited to imagine putting one grain of rice on one side of the card. After simulating what would happen and explaining why, children were invited to imagine adding more grains of rice and they were asked if the card would ever topple. After agreeing that an X number of grains would topple the card, children were asked why adding the X-th grain of rice would make the card topple. Next children were introduced to the fulcrum made out of 3 popsicle sticks. They were asked the same questions. The procedure was repeated with the fulcrum made out of 1 popsicle stick. If they predicted the card would topple, they were asked "Why?" If children did not acknowledge that a single grain of rice would be sufficient to topple the card, they were asked if there could ever be a piece of wood thin enough so that only one grain of rice would topple the card and were asked "why?" if they said yes.

Results

A preliminary analysis of children's understanding of addition showed that 58 out of 59 children in the Real Experiment, and 57 out of 60 children in the Thought Experiment condition knew that zero plus n zeros equals zero.

Could children simulate accurately in the thought experiment?

All 60 children in the Thought Experiment condition – a striking statistic for this age range – said that the card would eventually topple if grains of rice were being added on one side. This was true for the 6-stick fulcrum (children's estimate of number of grains needed to topple the card was approximately 17 (range 1 to 100), 3-stick (average estimate approximately 9, range 1 to 100), and 1-stick fulcrum (average estimate approximately 3, range 1 to 50). Thus, they simulated what would happen, and correctly predicted that the thinner the fulcrum the fewer grains of rice would be needed to topple the card. The estimates of the number of grains needed were a linear function of fulcrum width, with a slope of approximately 3. To formally test this, we conducted a simple linear regression analysis where children's estimates about the number of grains needed to topple the card was an outcome variable and the number of sticks on the fulcrum was a predictor variable. The analysis showed that the number of sticks on the fulcrum was a significant predictor of children's estimates about the number of grains needed to topple the card (p < .001). The Beta coefficient was 2.9 meaning that for each increase in fulcrum width (1, 3, 6), there was an average 2.9 increase in children's estimate about the number of grains needed to topple the card.

As mentioned above, in the real experiment, the decline also follows a linear function, with a slope of approximately 1.5. In the Real Experiment condition, the average number of grains needed to topple the card on the 6-stick fulcrum was approximately 9 (range 3 to 20), approximately 4 for the 3-stick fulcrum (range 2 to 9), and it was exactly 1 for the 1-stick fulcrum (always 1). In summary, not only did children correctly simulate that adding grains would topple the card, and that even a single grain of rice would topple the card if the fulcrum were thin enough, but their simulations were on average very accurate estimates of what would happen in the actual world in terms of the relationship between the width of the fulcrum and the number of grains needed to topple the card.

For the 1- stick fulcrum TE, 51 out of 60 children (85%) said that a single grain of rice is sufficient to topple the card. When asked if there could be a fulcrum thin enough so that only one grain of rice would topple the card, all but 2 of the 9 remaining children answered yes. These results provide strong evidence that children were able to correctly simulate in the thought experiment despite their explicitly stated belief that a single grain of rice weighs nothing at all.

Did children learn from the experiments? Was learning comparable in the Thought Experiment and Real Experiment Conditions?

Children's Judgments. To answer the question of whether children learned from the experiments, we first investigated children's pre- and post-training judgments about a single grain of rice. Figure 2 presents the proportion of children who gave a correct judgment, namely that a grain of rice weighs something. Replicating prior findings, more than half of the children denied that a grain of rice weighs anything at all at pre-training (see Carey, 2009, for review). This was true in both the TE and the RE condition. ($\chi2$(1, N = 119) = .22, p = .64).

Children's judgments changed dramatically after the TE and the RE intervention. At post-training, approximately 80% of all children gave a correct judgment. A McNemar test showed that the change was statistically significant both in the TE and the RE condition (ps < .001). In the TE condition, of the 60 children, 7 denied weight to rice both at pre- and post-training and 23 said rice weighs something both at pre- and post-training. Only 5 children who claimed that rice weighs something at pre-training, claimed that rice weighs nothing at all at post-training. Critically, 25 children who claimed that rice weighs nothing at all at pre-training, said that rice weighs something at post-training. The pattern of results was very similar in the RE condition. Of the 59 children, 10 denied weight to rice both at pre- and post-training and 24 said rice weighs something both at pre- and post-training. Only 1 child who claimed that rice weighs something at pre-training, claimed that rice weighs nothing at all at post-training. Critically, 24 children who claimed that rice weighs nothing at all at pre-training, said that rice weighs something at post-training. As at pre-training, the proportion of children who gave a correct judgment (this time at post-training) was very similar across the two conditions ($\chi2$(1, N = 119) = .04, p = .85). Thus, children learned that a grain of rice weighs something both from the RE and the TE, which, importantly, were comparably effective.

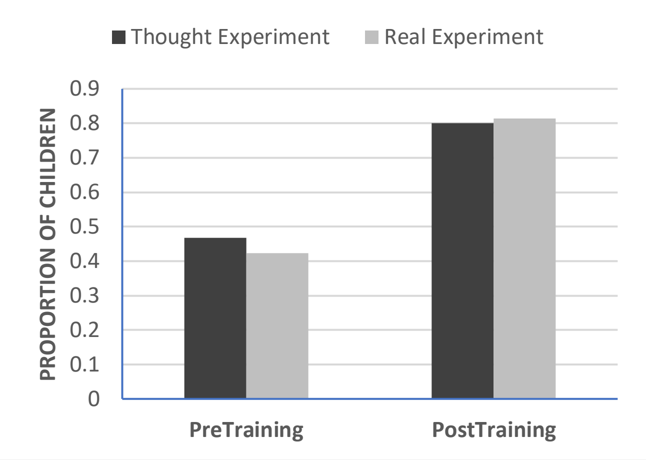

Figure 2. Proportion of children who said that a grain of rice weighs something at pre- and post-training.

Next, we asked if the RE and TE contributed to near transfer of knowledge. That is, we asked if children learned that other pieces of matter – that were not directly targeted by the intervention – also weigh something. Figure 3 represents the composite score of children's judgments on the two questions about sugar and playdough.

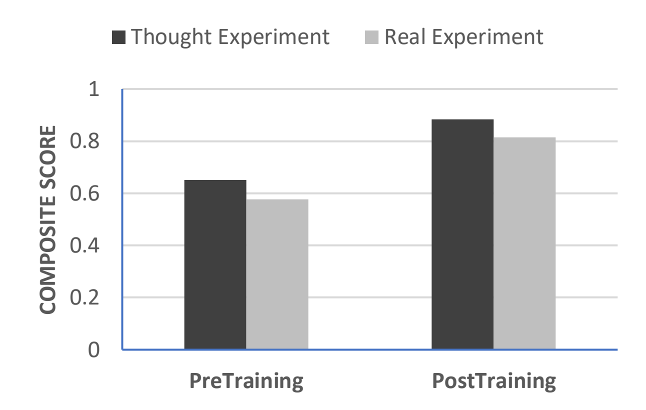

Figure 3. Composite scores on the playdough and sugar questions (possible range 0 to 2) across the two conditions at pre- and post-training

To answer if children's judgments changed between pre- and post-training, and whether they changed more in one of the two conditions, we conducted a repeated measures ANOVA, which examined the effects of condition (Thought Experiment vs. Real Experiment) and training session (Pre-Training vs. Post-Training) on the Composite Weight of Playdough and Sugar variable. This analysis revealed a significant effect of Training session, Wilks' Lambda = .91 $F(1, 117) = 11.11$, $p = .001$, $\eta 2 = .087$. There was no significant effect of condition ($p = .56$), nor a significant interaction ($p = .98$). This result confirms that children's improvement in this study was not restricted to their understanding of the weight of rice only (i.e., the substance used in the thought and the real experiment). Children also showed near transfer to substances other than rice, namely playdough and sugar. Importantly, the changes in judgements between pre- and post-training about the weight of a single grain of rice, as well as small portions of other kinds of stuff, was essentially identical whether the intervention involved real data or merely simulated data.

Children's Justifications. Children's correct and incorrect judgments (analyzed in the section above) provide only partial insight into children's reasoning. For example, children might have judged that a small piece of matter weighs a little bit by merely guessing, or by associating the word "little" with the small piece of matter and the phrase "little bit" (i.e., everything about the piece is little) without having any understanding of weight as an extensive property of matter. Other children might have relied on their understanding of weight when making the judgment and used evidence that even small pieces of matter weigh something (e.g., holding a piece of matter in the hand may produce a sensation of feeling it in the hand, which is interpreted as weighing something). Finally, *some* children might have given correct judgments because they have a more abstract understanding of weight, such that all pieces of matter weigh something, and that the weight of any material entity is a function of the weights of arbitrarily small parts of it (an extensive concept of weight, as a necessary property of material entities). According to this abstract understanding, even though our perceptual system cannot detect the weight of a small piece of matter, we know that it weighs something. We coded children's justifications for correct judgements according to a coding scheme that differentiated the different types of reasoning outlined above. Children who gave no justification or some irrelevant justification received a score of 1. Children who gave justification that a piece of matter weighs a little because it is small/ little received a score of 2. Children who appealed to evidence that a piece of matter weighs something (e.g., it feels something in the hand, or it topples sensitive scales) received a score of 3. Finally, children who offered a generalization that everything that is material weighs something received a score of 4. The important distinction is between 1 and 2, on the one hand, where children are at most justifying their choice of "a little bit" rather than "a lot", and 3 and 4, on the other, where they are explaining why they said this tiny piece of matter weighs something.

We first investigated qualitatively the type of answers that children gave. Table 1 presents percentages of children who gave justifications scored 1 or 2 on the one hand, and justifications scored 3 or 4 on the other hand out of all correct judgments. As is evident in Table 1, at pre-training most children gave justifications that fell in the 1 or 2 scoring category: irrelevant justifications or justifications for their judgment of "a little" vs "a lot", and fewer children tried to explain their judgment that the small piece weighed *something* (Categories 3 and 4). This was equally true in both the TE and RE conditions.

Table 1. Proportion of children who gave a level 1 or 2 and a level 3 or 4 justification

	Thought Experiment						**Real Experiment**					
	RICE		PLAYDOUGH		SUGAR		RICE		PLAYDOUGH		SUGAR	
	1-2	**3-4**	**1-2**	**3-4**	**1-2**	**3-4**	**1-2**	**3-4**	**1-2**	**3-4**	**1-2**	**3-4**
Pre-Train	75%	25%	81%	19%	55%	45%	76%	24%	71%	29%	56%	44%
Post-Train	44%	56%	59%	41%	50%	50%	33%	67%	58%	42%	50%	50%
Increase 3-4		31 pp		22 pp		5 pp		43 pp		13 pp		6 pp

However, that pattern of results changed dramatically at post-training. In both conditions, many more children provided Category 3 and 4 justifications at post-training compared to pre-training. Indeed, for the playdough and rice pieces, the justifications for attributing weight at all increased between 13 and 43 percentage points between pre- and post-training, whereas they increased by 5 and 6 percentage points for the sugar question. This suggests that both the Real and the Thought Experiment contributed to focusing attention on the fact that small pieces of matter weigh something. The overall change toward justifications that small pieces of matter weighed a little bit coded 3 or 4 across all three types of matter was similar between the Real and the Thought Experiments.

Another important result that's evident in Table 1 is that at pre-training, the percentage of justifications that were explaining why the small portion weighed *something* (3 or 4) was higher for the sugar question (i.e., around 45%) than for the rice and playdough questions (between 20% and 30%).

This difference was also seen in the judgments themselves. Collapsing across the TE and RE conditions, both at pre-training (mean 16.5% correct) and post-training (27% correct), children were less likely to judge that a single grain of sugar weighed a little bit than to judge that a grain of rice (44% at pre-training and 81% at post-training) or small piece of playdoh did (46% at pre-training and 58% at post-training). This is probably due to the fact that the single grain of sugar was much smaller than the other two pieces. Thus, it is likely that the few children who judged a single grain of sugar to weigh something at pre-training were already farther along in their construction of concepts of matter and weight, reflected both in their judgements and in the fact that their justifications reflected attempts to explain why this tiny piece of sugar weighed something.

The above analyses concerned justifications for correct judgments only, but of course there were more correct judgments after the TE and RE. In a final analysis of whether these experiments changed the quality of explanations overall, we investigated the percentage of justifications that received a score of 3 and a score of 4 out of *all* children (i.e., including children whose justifications were not scored because they gave incorrect judgments). Table 2 presents those data.

Table 2. Proportion of children (out of the full sample) who gave a level 3 and a level 4 justification

	Thought Experiment						**Real Experiment**					
	RICE		PLAYDOUGH		SUGAR		RICE		PLAYDOUGH		SUGAR	
Score	**3**	**4**	**3**	**4**	**3**	**4**	**3**	**4**	**3**	**4**	**3**	**4**
Pre-Train	3%	8%	3%	5%	0%	8%	0%	10%	2%	10%	0%	7%
Post-Train	33%	12%	8%	17%	2%	12%	42%	12%	8%	13%	0%	11%
Increase	30 pp	4 pp	5 pp	12 pp	2 pp	4 pp	42 pp	2 pp	6 pp	3 pp	0 pp	4 pp

The inspection of Table 2 reveals that the biggest change between pre- and post-training was on the question about rice and for the category of justifications that received a score of 3. The percentage of children whose justification received a score of 3 at post-training was much higher than at pre-training. It was 42 percentage points higher in the RE and it was 30 percentage points higher in the TE (*p*s < .001). This is not surprising for the Real Experiment condition, because children witnessed a grain of rice toppling the card, and so they explained their subsequent judgment that a piece of rice weighs something by appealing to the fact that it caused the card to topple. What is surprising is that children drew the same kind of conclusion in the Thought Experiment as well, even though they relied on their simulation of the event only.

The other important result in Table 2 is that there was very little absolute improvement in Level 4 justifications, those that explained the weight of small pieces of matter in terms of an extensive concept of weight of all material entities. There were *a few* children who made such arguments, both at pre-training and post-training. There was a 7 percentage points increase on average between pre-training and post-training in the TE condition, and 3 percentage points increase overall in the RE condition. This difference was not significant.

In sum, *neither* the TE nor the RE led to significant progress toward an extensive concept of weight of material entities. Both equally promoted an empirical generalization that small pieces of matter weigh something (reflected both in a large increase in correct judgments and in level 3 justifications).

Discussion

A long and important tradition in developmental psychology, starting with Piaget, has studied the process of construction of knowledge where the construction is seen as an inductive bottom-up process, from observed data to theories that explain the data. However, learning occurs even in the absence of new observations (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Lombrozo, 2019). More relevant to the present paper is the possibility that the thought experiments described in their writings by Galileo, Newton, Maxwell, Einstein, Heisenberg, and many others, played an important role in scientific progress and learning (Gendler, 1998; Kuhn, 1977; Nersessian, 1992; Popper, 1959). That is, thought experiments may play a productive role in the scientist's theory building, as well as a pedagogical role for communicating the major tenets of a theory (in the case of constructive thought experiments), or for refuting a particular theory (in the case of destructive thought experiments) (Brown, 1986; Popper, 1959).

There is very little empirical work in psychology investigating whether and how learners, including children, can benefit from thought experiments. In the present study, we began addressing this gap by asking whether and how children can learn from thought experiments in the context of their developing theory of matter. We found that not only can children accurately simulate in a thought experiment, but they can learn from working through a thought experiment. Indeed, the progress made by children in the thought experiment condition was indistinguishable from the progress made by children in the real experiment condition.

What did children learn from the thought experiment and what was the mechanism by which they learned it? Children learned an empirical generalization that a grain of rice (and other similar pieces of matter) weigh something. This generalization was drawn from the outcome of a mental simulation that a single grain of rice could topple a card that is resting on a thin fulcrum. It is important to note that the thought (and the real) experiments had elements of an extreme case analysis (Zeitsman & Clement, 1997), where the aspect of ratcheting likely helped children to clearly “see” the consequences of adding grains of rice one by one, and also to “see” the consequences of reducing the thickness of the fulcrum. This type of learning is in line with the ‘experimentalism’ view according to which thought experiments are sometimes a limiting case of ordinary experiments, and they rely on perceptual-motor intuitions about the world (Aronowitz & Lombrozo, 2020; Gendler, 2004; Mach, 1897; Sorensen, 1992). In other words, conducting the thought experiment is like the process of conducting a real experiment, in the sense that the experimenter “collects” new data by running a simulation or by manipulating a model (Gendler, 2004; Nersessian, 2018). Conversely, the present study did not produce evidence in support of the “argument” view (Norton, 2004) or the “conflict” view (Kuhn, 1977). Although children knew the premises, none of them in the present study spelled out an argument along the lines of saying that if a pile of rice weighs something, then a single grain of rice must also weigh something, because we know that zero plus n zeros equals zero. Similarly, there was no evidence in the present study consistent with the view that thought experiments can reveal “a crisis” in one’s theory and can shed light on the aspects that require a theory change (Kuhn, 1977). Namely, none of the children noticed or commented on the conflict between the outcome of the thought experiment and their initial belief that a grain of rice weighs nothing at all even though children were probed to compare their pre- to their post-training judgments at the end of the interview. Of course, this may be due to children’s limited metacognitive abilities (Flavell, 1979), which may prevent them from explicitly noticing and representing contradictions and inconsistencies (Markman, 1977, Ruffman, 1999). However, recent studies with adults engaging in thought experimentation show a similar failure to spontaneously recognize conflicts among beliefs (Bascandziev, 2020). In summary, in this case study, the thought experiment led to learning based on “collecting data” from mental simulation, which was very similar to the learning that resulted in the real experiment condition.

One important result in the present study is that most children did not reach a level-4 justifications at post-training. In other words, on average, children did not advance to having an extensive concept of weight where the weight of any given material object is the sum of arbitrarily small pieces that are part of the object. This is not surprising under the view that conceptual change involves much more than merely acquiring new data from a single experiment. Furthermore, this is consistent with the conclusion that children did not engage in logical argumentation that would have allowed them to conclude that all material things must weigh something, nor did they engage in an extreme case analysis which would have allowed them to conclude that no matter how small the piece is, there can always be a fulcrum thin enough so that the piece would topple the card. Of course, this may be due to children’s age and how far they have gone in their construction of a theory of matter. Slightly older children with more advanced, but incomplete, theory of matter might have engaged in logical argumentation and extreme case analysis if given the same kind of thought experiments. Future research should explore this possibility. Another avenue for future research is to explore which aspects of children’s theory of matter (e.g., understanding of material/ nonmaterial distinction, understanding of space and occupying space) might have allowed them to benefit from the thought experiments (explored in this study, but not reported here), as well as whether broader curricular interventions based on thought experimentation alone can bring about broader conceptual change.

The present study investigated the near-term effects of a real experiment and a thought experiment, and it showed that there were no differences between the two types of intervention. That leaves the question about the long-term effects of interventions (thought and real experiments) open for future investigation. One possibility is that thought experimentation might have an advantage over real experiments. First, thought experiments typically use idealized scenarios where the attention of the learner is pointed to the relevant variables that need interpreting. Conversely, real experiments may sometimes involve manipulatives and materials that may distract children’s attention from the actual problem (Fisher, Godwin, & Seltman, 2014). Another reason why thought experiments may be more effective over the long run is because they may instill a habit of thought in the learner, which is to check one’s beliefs continuously and persistently against one’s vast database of information via thought experimentation. Future research should explore both the mechanisms by which learning occurs, as well as the educational implications and benefits that can stem from thought experimentation.

References

Aronowitz, S., & Lombrozo, T. (2020). Learning through simulation. *Philosophers’ Imprint*.

Bascandziev, I. (2020). Inconsistencies among beliefs as a basis for learning via thought experiments. In B. C. Armstrong, S. Denison, M. Mack, & Y. Xu (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Toronto, CA: Cognitive Science Society

Bascandziev, I., & Harris (2019). Can children benefit from thought experiments? In P. Godfrey - Smith & A. Levy

(Eds.). *The Scientific Imagination.* New York: Oxford University Press.

Brown, J. R. (1986). Thought experiments since the scientific revolution. *International Studies in the Philosophy of Science, 1,* 1-15.

Brown, J. R., & Fehige, Y. (2019). Thought experiments. In Edward N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy*

Carey, S. (2009). *The origin of concepts.* New York: Oxford University Press.

Clement, J. J. (2009). The role of imagistic simulation in scientific thought experiments, *Topics in Cognitive Science, 1,* 686-710.

Fisher, A. V., Godwin, K. E., & Seltman, H. (2014). Visual environment, attention allocation, and learning in young children: When too much of a good thing may be bad. *Psychological Science, 25,* 1362–1370.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911.

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

Gendler S. T. (1998). Galileo and the indispensability of scientific thought experiments. *The British Journal of the Philosophy of Science, 49,* 397-424.

Gendler, S. T. (2004). Thought experiments re-thought and reperceived. *Philosophy of Science, 71,* 1152-1163.

Gopnik, A. & Schulz, L. E. (Eds.). (2007). *Causal Learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others.* Cambridge, MA: Harvard University Press.

Kuhn, T. S. (1977). A function for thought experiments. In P. N. Johnson-Laird and P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 274-292). New York, NY: Cambridge University Press.

Lombrozo, T. (2019). Learning by thinking in science and in everyday life. In A. Levy & P. Godfrey-Smith (Ed.), *The scientific imagination* (pp. 230-249). Oxford University Press.

Mach, E. (1897/1976). On Thought Experiments. In T. McCormack & P. Foulkes (Transl.), *Knowledge and Error* (pp.134-147). Dordrecht.

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development, 50*(3), 643–655.

Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Ed.), *Cognitive models of science* (pp.3-44). Minneapolis: University of Minnesota Press.

Nersessian, N. J. (2018). Cognitive Science, Mental Models, and Thought Experiments", in Michael T. Stuart et al. (eds.), *The Routledge Companion to Thought Experiments*, London and New York: Routledge, 309–326.

Norton, J. D. (2004). Why Thought Experiments Do Not Transcend Empiricism, in C. Hitchcock (ed.), *Contemporary Debates in the Philosophy of Science*, Oxford: Blackwell, 44–66.

Piaget, J., & Inhelder, B. (1974). *The child's construction of quantities.* London: Routledge & Kegan Paul.

Popper, K. (1959). *The logic of scientific discovery.* New York, NY: Basic Books.

Ruffman, T. (1999). Children's understanding of logical inconsistency. *Child Development, 70*(4), 872–886.

Smith, C., Snir, J., & Grosslight, L. (1992). Using conceptual models to facilitate conceptual change: The case of weight-density differentiation. *Cognition and Instruction, 9*(3), 221–283.

Sorensen, R. A., (1992). *Thought Experiments*, Oxford: Oxford University Press.