Unterrichtswiss (2021) 49:91–115
https://doi.org/10.1007/s42010-020-00094-4

The use of scaffolding to promote preschool children's competencies of evidence-based reasoning

Ilonca Hardy · Simone Stephan-Gramberg · Astrid Jurecka

Received: 18 September 2020 / Revised: 23 December 2020 / Accepted: 29 December 2020 / Published online: 22 January 2021
© The Author(s) 2021

Abstract Scientific reasoning encompasses individuals‘ evaluation of evidence with regard to a given hypothesis. In this study, we investigated whether preschool children are able to reason with empirical evidence in the science context of elasticity. *N*= 63 preschoolers were presented with tasks following the deductive reasoning paradigm and were asked to evaluate the relevance of given events (objects) with regard to a hypothesis. In a repeated measures experimental design with three groups, we tested whether different forms of scaffolding (adaptive prompts with/without modeling of advanced reasoning) would promote children's reasoning compared to a control group without intervention. We found that adaptive prompts with modeling significantly improved children's evaluation of irrelevant events in the posttest. Further, these children's reasoning patterns scored significantly higher than those of the control group. Our results suggest that preschool children are able to reason with evidence if they are given adequate support. Specifically, the modeling of advanced reasoning functioned as a scaffold beyond the use of adaptive prompts in irrelevant event evaluations.

Keywords Scientific reasoning · Deductive reasoning · Preschoolers · Scaffolding

I. Hardy · A. Jurecka (✉)
Fachbereich Erziehungswissenschaften, Institut für Pädagogik der Elementar- und Primarstufe,
Goethe-Universität Frankfurt, Theodor-W.-Adorno Platz 6, 60323 Frankfurt, Germany
E-Mail: jurecka@em.uni-frankfurt.de

I. Hardy
E-Mail: hardy@em.uni-frankfurt.de

S. Stephan-Gramberg
Maria-Scholz Schule, Wiesbadener Str. 27, 61350 Bad Homburg, Germany
E-Mail: Simone.stephan-gramberg@schule.hessen.de

Scaffolding zur Förderung des evidenzbasierten Begründens bei Kindern im Elementarbereich

Zusammenfassung Wissenschaftliches Denken erfordert die Evaluation von empirischer Evidenz in Bezug auf formulierte Hypothesen. In dieser Studie untersuchten wir, inwiefern Kinder im Vorschulalter in der Lage sind, Evidenz im naturwissenschaftlichen Kontext von Elastizität angemessen zu evaluieren. N= 63 Vorschulkindern wurden Aufgaben nach dem deduktiven Schlussfolgerungsparadigma gestellt und sie wurden gebeten, die Relevanz von gegebenen Ereignissen (Objekten) für eine Hypothese zu beurteilen. In einem Messwiederholungsdesign mit drei experimentellen Gruppen untersuchten wir die Auswirkungen von unterschiedlichen Formaten des Scaffolding (adaptive Prompts mit/ohne Modellierung von fortgeschrittenen Denkmustern) auf das schlussfolgernde Denken im Vergleich zu einer untrainierten Kontrollgruppe. Es zeigte sich, dass das Training mit adaptiven Prompts und Modellierung die Evaluation von irrelevanten Ereignissen im Posttest signifikant verbesserte. Die Schlussfolgerungsmuster der Kinder dieser Gruppe erreichten zudem eine signifikant höhere Qualität als diejenigen der Kontrollgruppe. Unsere Ergebnisse legen nahe, dass Vorschulkinder Evidenz evaluieren können, wenn sie angemessene Unterstützung erhalten. Die Modellierung von fortgeschrittenen Schlussfolgerungen hat sich dabei bei der Beurteilung von irrelevanten Ereignissen als bedeutsame instruktionale Unterstützungsmaßnahme über die Anwendung von adaptiven Prompts hinaus erwiesen.

Schlüsselwörter Wissenschaftliches Denken · Deduktives Schlussfolgern · Vorschulkinder · Scaffolding

Goals of early science education encompass the development of conceptual knowledge in basic domains of science as well as getting familiar with processes of scientific knowledge construction, modeling, and communication (Leuchter 2017; National Research Council (NRC) 2012; Steffensky 2017). Children involved in inquiry science activities therefore need to be equipped with basic scientific reasoning skills such as formulating hypotheses, observing, and interpreting observations (Stephan-Gramberg 2015; NRC 2012; van der Graaf et al. 2016, 2018) to participate in science preschool education. While there is growing evidence for elementary school children's capabilities of scientific reasoning (Edelsbrunner et al. 2018; Legare 2014; review by Sandoval et al. 2014), studies also show specific constraints with regard to the evaluation of evidence in the domain of science in preschool, elementary school, and secondary school age (Zimmermann 2007). In the present study, we investigated young children's ability to evaluate a given hypothesis in the domain of elasticity, employing tasks derived from the deductive reasoning paradigm.

Deductive reasoning is relevant to scientific reasoning as it involves individuals' decision-making as to the adequacy of a given hypothesis on the basis of empirical evidence (see Robisch et al. 2014). For example, in a discovery science activity, children may be asked to generate hypotheses with regard to a certain phenomenon ("things with holes will sink") and to judge whether specific observations (e.g., a solid piece of clay that sinks) will confirm or disconfirm the given hypothesis. Studies based on the deductive reasoning paradigm revealed that these types of event evaluations are difficult for elementary school students both in context-lean and in contextualized tasks (Barrouillet et al. 2008; Gauffroy and Barrouillet 2011; Troebst et al. 2011). Prior studies also showed that elementary school children are able to improve their event evaluations by means of adaptive instructional support in individual training sessions (Robisch et al. 2014) and in classroom contexts (Grimm et al. 2018). Against this background, we investigated whether even preschoolers are able to reason with evidence in contextualized tasks and if so, which specific instructional scaffolds are suitable to advance reasoning in the age group of four- to six-year-olds. In individual sessions, we confronted preschoolers with specific hypotheses concerning the bouncing behavior of balls (for example, "balls with air in them bounce") and with balls of differing qualities and bouncing behavior that would confirm, disconfirm, or else, be irrelevant to the given hypothesis. The task design was based on the paradigm of deductive reasoning advanced by Wason (1966) and Barrouillet et al. (2008). In tasks with scaffolding by adaptive prompts (van de Pol et al. 2010) as well as by modeling of correct responses (see also Vorholzer and von Aufschnaiter 2019), we investigated the effects of different formats of instructional scaffolds on children's answers.

1 Evidence-based reasoning in young children

Scientific reasoning is regarded a process of sense-making from observations of natural phenomena, experiments, and data patterns (Sandoval et al. 2014; Windschitl et al. 2008). Scientific reasoning is based on analytical processes that allow individuals to derive inferences and integrate observations with theoretical constructs, hypotheses, and beliefs—a process that has also been described as the coordination of theory and evidence (Kuhn 2002; Sodian et al. 1991). According to Duschl (2003), these processes involve transformations from (a) data to evidence, thus determining whether data are anomalous or count as valid evidence; (b) evidence to patterns, thus searching for patterns in and generating models for data; and (c) patterns to explanations, thus developing explanations on the basis of the evidence selected. Accordingly, Klahr (2000) and Klahr and Dunbar (1988) conceptualize scientific reasoning as a complex search within a "dual search model" that includes the hypothesis space and the experimental space.

Studies with mostly four- to-six-year-old preschool children show that they are able to generate hypotheses, recognize data patterns, and use data to derive inferences (e.g., Köksal-Tuncer and Sodian 2018). For example, studies by van der Graaf et al. (2016, 2018) showed that four- to six-year-old preschoolers can solve tasks of variable control and data covariation, and that variance in performance predicted children's conceptual development. Piekny and Mähler (2013) found that preschoolers were able to generate at least one hypothesis based on empirical data, and Piekny et al. (2014) showed that preschoolers evaluated evidence entirely or partially correctly from given data patterns. In addition, it was found that preschool children's corresponding prior beliefs supported correct interpretation, whereas conflicting prior beliefs hindered interpretation (Koerber et al. 2005). Importantly, four- to six-year-olds were also able to generate evidence and verbal explanations needed for the falsification of an incorrect causal claim (Köksal-Tuncer and Sodian 2018). This strategy of falsification involves deductive reasoning (Lawson 2010) since the validity of claims is judged with suitable evidence.

2 The role of deductive reasoning in evidence evaluation

2.1 Development of deductive reasoning

When individuals formulate hypotheses, they typically use conditional statements, for example "when an object is light, it will float." Accordingly, the interpretation of these conditional statements is considered a type of deductive reasoning, i.e., reasoning about truths. In contrast to reasoning about possibilities, reasoning about truths requires the individual to evaluate the correctness of a given assumption on the basis of evidence. In order to evaluate a given assumption, individuals need to distinguish between confirming, disconfirming, and irrelevant events (Barrouillet et al. 2008). This competency may thus be a decisive factor promoting conceptual development, as the validation or rejection of hypotheses based on empirical evidence is a vital aspect of science knowledge revision and extension (see Grimm et al. 2018; Robisch et al. 2014). Several studies in science contexts showed that especially the evaluation of irrelevant and disconfirming evidence poses a challenge to students (for a review see Zimmermann 2007; Sandoval et al. 2014). In addition, Chinn and Brewer (1998) found that adult learners' evaluation of disconfirming evidence, or anomalous data, followed a taxonomy of reactions, ranging from ignoring and rejecting evidence to modifying one's own theory. In their study, only few learners interpreted disconfirming evidence adequately, and learners rarely used empirical information to induce conceptual change. Similarly, research on deductive reasoning about truths with context-lean tasks demonstrates that individuals have difficulties with the interpretation of disconfirming and irrelevant events. Here, studies showed that adults failed to use falsification strategies when evaluating events, but succeeded with a higher probability when reasoning tasks were embedded in (everyday) contexts (Wason and Shapiro 1971).

In the paradigm of deductive reasoning, conditional statements are defined with antecedent (p) and consequent (q) (Barrouillet et al. 2008; Lawson 2010; Troebst et al. 2011). For example, the question of "Why do things float in water?" may be answered with the respective hypothesis "things that are light, will float." which may be framed as a conditional statement with (p) light things and (q) float. The combination of p and q results in four possible events: p (present/absent) and q (present/absent), see also Table 1 for examples. Reasoning about truths requires the evaluation of a conditional statement (i.e., a hypothesis) on the basis of a given event in combination of p and q. The developmental model of Barrouillet et al. (2008) describes three specific reasoning patterns in elementary and secondary school students' event evaluations. Third-graders' reasoning about truths showed predominantly conjunctive reasoning patterns, where the connection between antecedent and consequent is constructed conjunctively, evaluating only instances of p *and* q as confirming. Sixth-graders predominantly showed biconditional reasoning patterns, taking the antecedent p as the single valid precondition for the consequence q. Ninth-graders showed conditional reasoning patterns, in which the combination of p/q is correctly interpreted as confirming, p/–q as disconfirming, and –p/q and –p/–q as irrelevant. This reasoning pattern was also commonly found with adults. Whether this model also holds for preschoolers in the domain of science has not yet been investigated. Yet, several assumptions speak to the general possibility of its application in younger age: While there are mean differences in the predominant reasoning patterns between third and sixth grade, there is also a variance of patterns found across all age groups, ranging from conjunctive to conditional reasoning. Furthermore, there are several studies indicating that children as young as four years of age are able to reason conditionally when given appropriate input and tasks (de Chantal et al. 2019) in context-lean tasks as well as in contextualized tasks (Köksal-Tuncer and Sodian 2018).

2.2 Models and correlates of deductive reasoning

Developmental patterns of deductive reasoning are frequently explained by the Dual Process Mental Models Theory (Gauffroy and Barrouillet 2009; Johnson-Laird and Byrne 2002). According to this theory, individuals construct mental models when confronted with the evaluation of a conditional statement, representing all states that are compatible with a given assumption[1]. Two processes are involved in the construction of mental models. First, an initial representation is constructed that represents the states explicitly mentioned in the conditional statement and that confirms these states (p/q). In a second step, two more models are derived based on analytical reasoning (e.g. Gauffroy and Barrouillet 2009). These mental models delineate the states of affairs that are compatible with the assumption (the so-called "fleshing out" of the implicit initial representation) with regard to –p/q and –p/–q. In the end, ideally, three mental models of states of affairs result that are compatible with the given assumption. Events that are incompatible with a given assumption and that are not represented in a mental model (p/–q) would then lead to a falsifying event evaluation. The developmental sequence of conjunctive, bi-conditional, and conditional reasoning may well be described as the increasing success in fleshing out mental models besides the initial mental model of p/q, therefore leading to an increase in correct evaluations of the irrelevant events.

Results of Barrouillet et al. (2008) imply that children in elementary school fail to flesh out additional mental models, regarding events of p/–q, –p/q, –p/–q, as falsifying (conjunctive reasoning pattern). In late childhood and early adolescence, an additional mental model is constructed, leading to event evaluations of p/–q and –p/q as falsifying and event evaluations of–p/–q as irrelevant to the assumption (biconditional reasoning pattern). Here, the antecedent p is regarded as the exclusive precondition for the occurrence of the consequent q. In late adolescence and early

[1] We use the term assumption in the context of theories of conditional reasoning. We use the term hypothesis with regard to context-specific conditional reasoning in science.

adulthood, individuals are typically capable to flesh out both implicit models and to evaluate events accordingly (conditional reasoning pattern). The antecedent is interpreted as one precondition among others, i.e. as a sufficient precondition, for the occurrence of the consequent.

Beyond describing the developmental sequence of reasoning patterns, research has revealed two individual difference variables relevant to deductive reasoning in childhood: working memory capacity and inhibition (see Handley et al. 2004). Working memory refers to individuals' capacity for the temporary storage of information, affecting performance of other cognitive tasks (Baddeley 1983). Working memory capacity is related to individuals' success in fleshing out mental models, predicting primary school children's difficulties in event evaluations beyond p/q (Barrouillet and Camos 2001; Barrouillet and Lecas 1999). Besides working memory, inhibitory control has been shown to be associated with reasoning in contexts disconfirming prior beliefs (Handley et al. 2004). The construct of inhibition refers to individuals' ability to suppress goal-irrelevant stimuli and responses (Tiego et al. 2018). Gropen et al. (2011) claim that inhibition is relevant to children's scientific reasoning in contextualized science tasks as it allows children to inhibit premature evidence evaluation and enhances analytical reasoning. Accordingly, for preschool children, van der Graaf et al. (2018) found that executive functions and grammatical abilities predicted scientific thinking in a longitudinal study (see also van der Graaf et al. 2016).

2.3 Fostering young children's deductive reasoning

There is empirical evidence that young children's deductive reasoning in context-lean tasks may be fostered by interventions and trainings. For example, studies employing the Cognitive Training for Children (CTC) showed an increase in deductive reasoning patterns of elementary school children (e.g., Meiser and Klauer 2000; see also for meta-analytic results Hager and Hasselhorn 1998). Further, Barkl et al. (2012) employed CTC in an intervention study and found effects on inductive and deductive reasoning tasks, but no domain-specific effects on performance in mathematics with elementary school children. English (1997) investigated the impact of different tasks environments prior to deductive reasoning tasks in an experimental design with three intervention groups of 10- to 12-year-olds. She varied whether children received an explication of options (definite, indefinite) of deductive reasoning tasks, adaptive prompts based on children's individual performance that intended to attune the children to options of uncertainty, or both. Results showed that only the group that received adaptive prompts and the explication of answers showed superior performance in subsequent deductive reasoning tasks. Recently, de Chantal and Markovits (2017) found that task contexts in which preschool children were asked to generate alternatives and show divergent (flexible) reasoning were associated with advanced deductive reasoning patterns. Moreover, studies in the context of scientific reasoning deliver evidence on training formats. Based on results of English (1997), elementary school children's deductive reasoning was investigated in different task environments using adaptive prompts, with significant effects on children's correct event evaluations in the context of elasticity (Robisch et al. 2014; Troebst et al.

2011). However, the effects of instructional support to promote deductive reasoning in science have not yet been investigated for the age group of preschool children.

3 Scaffolding young children's evidence-based reasoning

Different means of instructional support have been described in the context of early science education (Gerde et al. 2013; Muhonen et al. 2016; Samarapungavan et al. 2011). Meta-analyses provide evidence for the impact of instructional support[2] on students' learning in different age groups in inquiry-based science learning environments (Furtak et al. 2012; Lazonder and Harmsen 2016). Specifically, Lazonder and Harmsen (2016) distinguish between prompts, heuristics, scaffolds, and explanations, and they revealed moderating, yet overall unspecific effects on student outcomes. Vorholzer and von Aufschnaiter (2019) broaden this perspective and propose three dimensions of instructional guidance—the degree of autonomy, the degree of conceptual information, and the cognitive domain, including their interplay. With respect to the present study, the following distinction taken by Vorholzer and von Aufschnaiter (2019) is relevant: Explicit guidance concerns a type of student support that makes explicit reasoning and explains strategies of inquiry. Implicit guidance, in contrast, provides prompts and hints to students on how to apply conceptual knowledge (on strategies), while this conceptual knowledge has to be discovered by students themselves. Similarly, Lazonder and Harmsen (2016) differentiate between explicit and implicit strategies, with the constructs of explanations and scaffolds as rather explicit formats, and the constructs of prompts and heuristics as rather implicit formats. Interestingly, these authors also consider the level of individual student understanding when defining means of instructional guidance. It is especially this perspective on the adaptivity of guidance that shows the conceptual overlap with the construct of scaffolding.

Scaffolding is commonly described as experts' use of adaptive prompts, hints, and explanations to solve a task that learners otherwise would not be able to perform (Wood et al. 1976). According to Hermkes et al. (2018), Puntambekar and Hübscher (2005), and van de Pol et al. (2010), the main characteristics of scaffolding include the contingency (or adaptivity) of instructional support, the transfer of responsibility (fading), and the use of diagnostic strategies. Thus, one important characteristic of scaffolding is its adaptivity, i.e., support that is contingent on a student's current (diagnosed) level of understanding (e.g., Hardy et al. 2019). In a basic distinction, Brush and Saye (2002) further differentiate between hard and soft scaffolds. Whereas hard scaffolds such a visualizations serve to support all learners by providing models of reasoning, soft scaffolds are employed adaptively based on a student's current level of understanding. Soft scaffolds may therefore be described as adaptive prompts that are faded out when a student's current understanding has reached a certain level necessary to solve a task independently (Robisch et al. 2014).

In addition to basic characteristics, Pea (2004) and Reiser (2004) point to relevant cognitive functions of scaffolding. For one, scaffolds serve to focus learners' atten-

[2] We use the term instructional support to include guidance.

tion onto essential aspects of a task. This is achieved by reducing degrees of freedom in task-based problem-solving with hints and prompts that mark relevant features of a task, and with structuring and sequencing of content. Second, scaffolds serve to model advanced reasoning processes and solution strategies (Collins et al. 1987; Pea 2004; van de Pol et al. 2010; Vorholzer and von Aufschnaiter 2019). Specifically, modeling may be employed in scientific reasoning contexts to show expert use of evidence and claims (Legare 2014; McNeill and Krajcik 2008). Typically, in direct instruction, teacher modeling of an advanced strategy will be followed by an individual learning phase during which students apply the respective strategy (Klahr and Nigam 2004). Recently, van der Graaf et al. (2019) showed the benefits of combining direct instruction and adaptive verbal support for promoting children's scientific reasoning in a science discovery context. In our training study, modeling precedes adaptive prompts that are employed contingent on a child's level of understanding during independent task solutions. Thus, children are first provided with a model of advanced reasoning on which further sense-making with similar tasks may be based.

4 Background of the study

Based on prior studies revealing gains in elementary school students' event evaluations in a training context (Robisch et al. 2014) and in a classroom context (Grimm et al. 2018), we investigated whether these effects will also be found for preschool children. Specifically, we were interested in whether scaffolding involving adaptive prompting and modeling is suitable for preschool children with presumably conjunctive or inconsistent reasoning patterns in which the relation between p and q is regarded as a simple "and." We focused on prompting as an implicit format of instructional support and combined it with preceding modeling. Based on Robisch et al. (2014), we presume that instructional support with adaptive prompting will show a positive training effect on children's event evaluations. Additionally, explicit formats of support have proven to be essential for scientific reasoning of young children in inquiry science (van der Graaf et al. 2019) and in context-lean tasks (English 1997). On a cognitive level, both modeling and prompting, when employed in a context of evidence-based reasoning, are intended to support children in correct event evaluations with regard to a given hypothesis, thus leading to the construction of fleshed-out mental models incorporating antecedent (p) and consequent (q). Following Robisch et al. (2014), we employ elements to focus children's attention onto p and q, and we employ modeling to demonstrate correct solutions and reasoning processes with regard to the connection between a given hypothesis and its respective evaluation with confirming, disconfirming, or irrelevant evidence. Modeling and prompting are therefore intended to reduce the cognitive load when processing given information, enhancing working memory capacity. They also are intended to enhance a child's inhibition of premature responses.

When assessing young children's event evaluations, a distinction between single event evaluations and reasoning patterns is relevant. Single event evaluations point to young children's potential difficulties and success in evaluations of specific combinations of p and q (Troebst et al. 2011). Reasoning patterns assess a child's overall and potentially consistent level of the four event evaluations of p/q, p/–q, –p/q, and –p/–q with regard to a given hypothesis across different tasks (Barrouillet et al. 2008). In the assessment of reasoning patterns, typically consistency scores are assigned based on reasoning across a set of tasks. We suggest that especially in difficult domains and with novel tasks, scores based on children's highest level of reasoning rather than consistency of reasoning may be used to describe potential benefits of cognitive trainings. Based on these considerations, we investigated the following research question:

What are the effects of different forms of scaffolding (adaptive prompts with/without modeling) on preschoolers' evidence evaluations in the domain of elasticity?

Hypothesis 1a We expect that both forms of scaffolding are superior to a control group in children's gains from pre- to posttest in evaluations of single events of p/q, p/–q, –p/q and –p/–q. We expect that scaffolding using adaptive prompts with modeling is superior to adaptive prompts as it provides children with a strategy for event evaluations.

Hypothesis 1b We expect that both forms of scaffolding are superior to a control group with respect to children's use of advanced reasoning patterns (consistency of patterns, most advanced reasoning patterns). We expect that scaffolding using prompts with modeling is superior to adaptive prompts only.

5 Method

5.1 Sample

A total of 63 children with German as a first language and a mean age of 5.9 (*SD* = 0.4) from five preschool institutions participated in this study[3]. Of the 63 children, 33 were female and 30 were male. Informed consent of parent guardians was collected for all participants. All children of the sample came from middle class families (assessed by teacher estimation) from suburban areas of a major German city.

5.2 Design

A 2 (Time) × 3 (Group) experimental design was employed.[4] In pilot studies (Stephan-Gramberg 2015), inhibition proved to be a strong predictor for scientific reasoning. The participating children were parallelized based on a measure for

[3] The minimum sample size for this study was computed based on results and effect sizes of a similar study with elementary school children (Robisch et al. 2014), resulting in N of 42 to detect interaction effects of group by event.

[4] This study was conducted within the research project "Förderung von Modellbildungs- und Falsifikationsprozessen beim naturwissenschaftlichen Lernen im Elementar- und Primarbereich", funded by the German Research Foundation with the funding codes of HA 3205/5-1 and MO 942/5-1.

inhibition by Jansen et al. (1999) and randomly assigned to one of three groups: A training group with adaptive prompts (PR; $N = 21$, mean age 5.9 ($SD = 0.37$)), a training group with adaptive prompts and modeling (PR + M; $N = 21$, mean age 5.8 ($SD = 0.40$)), and a control group without training (CG; $N = 21$, mean age 5.9 ($SD = 0.40$)).

5.3 Dependent measures

5.3.1 Reasoning task: elasticity

We employed three tasks based on the truth-testing paradigm (Barrouillet et al. 2008; Wason 1966) as pretest and posttest. In each task, a hypothesis and four corresponding event evaluations are presented, see Table 1 for objects used in the pretest and the posttest. In the posttest, the same hypotheses as in the pretest were employed, but different objects were used for the event evaluations. The children were first presented with a fictive person's hypothesis about an objects' bouncing behavior (e.g., "Tim believes: round things bounce") and four corresponding events with the conditions of p/q, p/–q, –p/q, and –p/–q. Thus, in each task, children were successively presented with four objects to be evaluated—one object confirming the given hypothesis, one object disconfirming it, and two objects that were irrelevant to the given hypothesis. With the presentation of each of the four objects, the experimenter stated its quality (e.g., round) and its behavior (e.g., bounces), for example: "Look, it is round and it does not bounce." The children were then asked to try out themselves that the given object indeed bounced or did not bounce. The experimenter then asked the children with each object, while simultaneously pointing at one of three smileys: "You have three options to answer. Does this show that Tim's assumption[5] is correct, does it show that it is incorrect, or does it have nothing to do with Tim's assumption?" The children indicated whether the object was confirming (happy-looking smiley), disconfirming (sad-looking smiley), or irrelevant (thoughtful-looking smiley) to the given hypothesis by putting a token on the respective smiley. The children did not receive any feedback on their answers. Event evaluations (p/q, p/–q, –q/p, –q/–p) were randomly alternated in the pretest and posttest as well as in training sessions to avoid order effects. For single events of p/q, p/–q, –q/p, –q/–p, answers were scored with 1 for a correct answer and 0 for an incorrect answer, following the conjunctive reasoning paradigm (see Table 1). Across the three pretest and posttest tasks, sum scores were computed with a range from 0–3, respectively.

5.3.2 Coding of reasoning patterns

According to Barrouillet et al. (2008), the four trials within a truth testing task may also be combined into different reasoning patterns, see Table 2. In a conjunctive pattern, children are able to correctly evaluate the objects confirming a hypothesis and those disconfirming a hypothesis. They link antecedent with consequent by using

[5] The term assumption is used in adults' statements to refer to the more technical term of hypothesis.

Table 1 Hypotheses and events presented in pretest, posttest, and trainings

Assumption	Events	Affirming [p/q]	Disconfirming [p/–q]	Irrelevant [–p/q]	Irrelevant [–p/–q]
Pretest					
Round objects bounce	Ball	x	–	–	–
	Dough ball	–	x	–	–
	Rubber egg	–	–	x	–
	Washcloth	–	–	–	x
Soft objects bounce	Felt ball	x	–	–	–
	Grain pillow	–	x	–	–
	Wooden ball	–	–	x	–
	Booklet	–	–	–	x
Heavy objects bounce	Boccia ball	x	–	–	–
	Balloon filled with sand	–	x	–	–
	Styrofoam ball	–	–	x	–
	Cotton ball	–	–	–	x
Posttest					
Soft objects bounce	Woolen ball	x	–	–	–
	Juggling ball	–	x	–	–
	Golf ball	–	–	x	–
	Dry clay ball	–	–	–	x
Round objects bounce	Rubber ball	x	–	–	–
	Dough ball	–	x	–	–
	Egg-shaped object	–	–	x	–
	Balloon without air	–	–	–	x
Heavy objects bounce	Led ball	x	–	–	–
	Stone	–	x	–	–
	Balloon filled with air	–	–	x	–
	Bag with feathers	–	–	–	x
Training 1					
Hard objects bounce	Marble	x	–	–	–
	Ball of clay	–	x	–	–
	Softball	–	–	x	–
	Juggling ball	–	–	–	x
Training 2					
Objects filled with air bounce	Play ball	x	–	–	–
	Waterball	–	x	–	–
	Cube	–	–	x	–
	Cut balloon	–	–	–	x

Table 2 Coded reasoning patterns by time (pretest, posttest)

Trial	Equivalence	Consequence	Conjunctive	Biconditional	Conditional
[p/q]	+	+	+	+	+
[p/–q]	–	–	–	–	–
[–p/q]	–	+	–	–	○
[–p/–q]	+	–	–	○	○

+ confirming, o irrelevant, – disconfirming

"and" as a conjunction. A biconditional pattern involves children who are additionally able to correctly recognize an object of–p/q as irrelevant to a given hypothesis. A conditional pattern encompasses children who are able to correctly recognize both objects that are irrelevant according to a hypothesis. On this highest level of reasoning, subjects see the antecedent as a possible reason for the consequent. In a prior study with primary school children and a similar task context, two additional patterns were identified (e.g., Troebst et al. 2011). In an equivalence pattern, children answer in a rule-orientated manner, taking the assumptions as instantiations of a rule to be observed. In a consequence pattern, children solely evaluate an object based on the consequent, i.e., its bouncing behavior, disregarding the antecedent. If a child did not show an identifiable pattern across the four events, we coded the pattern as inconsistent. The six patterns were rank-ordered, with scores of 0 (inconsistent), 1 (conjunctive, equivalence, consequence), 2 (biconditional), and 3 (conditional). We computed two scores from the answer patterns: The score of consistent reasoning pattern reflects children's response consistency across three tasks. A score ranging from 0–3 is assigned if the respective reasoning pattern occurred in at least two of the three tasks (pretest, posttest). The score of maximum reasoning pattern reflects the most advanced (maximally reached) reasoning pattern across three tasks (pretest, posttest), and a respective score ranging from 0–3.

5.4 Control measures

5.4.1 Inhibition

We employed the test of inhibition by Jansen et al. (1999). The test of inhibition consists of 28 pictures of vegetables and fruit, presented in seven rows of four pictures each. Each vegetable or fruit is painted in an incorrect colour. Children were asked to tell the correct colour for each picture as quickly as possible. For every correct answer, a child received one point, with a total score ranging from 0–24. In addition, the total time needed for task completion was taken in seconds, with recorded scores ranging from 41–158.

5.4.2 Domain-general scientific reasoning

The test consists of three picture stories with each four pictures of everyday contexts (e.g., a boy climbing on a tree and falling down) to assess children's ability to coordinate theory and evidence. The content of the three stories was adopted from items of a written multiple-choice test of scientific reasoning test in elementary school (Koerber et al. 2015). In each picture story, the children were asked four questions assessing the generation of a plausible story frame (a), the recognition of a given story frame on the respective picture (b), the generation of plausible evidence for a given story frame (c), and the recognition of given evidence on the respective picture (d). For example, it was asked in what way one could see on the picture that the boy had fallen down the tree (c) or what could have led to the boy falling down the tree (a). To test the child's coordination of theory and evidence, the four pictures of each picture story were shown successively, testing generation before recognition. For the full testing material see Stephan-Gramberg (2015). The answers were scored dichotomously with a total score ranging from 0–12 and an internal consistency of $\alpha = 0.69$. All answers were double-coded ($\kappa = 0.76$).

6 Procedure and training

6.1 General procedure

A pre-posttest-design was implemented with two training sessions on consecutive days. All tests as well as the training phases took place within a one-to-one situation with one trainer and one child. On day 1, the children of all groups (PR, PR + M, CG) took the pretest of reasoning (elasticity), and the inhibition test. On day 2, the test of domain-general scientific reasoning was administered to all children. Additionally, the children assigned to one of the two experimental groups (PR, PR + M) participated in training 1. On day 3, the children of the two experimental groups participated in training 2. Children of the control group did not participate in the training sessions 1 and 2. On day 4, all children (PR, PR + M, CG) took the posttest of reasoning (elasticity) as well as a transfer reasoning test in a different domain.[6] For information on further tests see Stephan-Gramberg (2015). The tests and training sessions were conducted by one of the authors as well as trained university students.

6.2 Training days 1 and 2

The training was based on the same reasoning tasks as were used in the pretest and the posttest, but with different hypotheses. On day 1, the children of PR and PR + M worked on the hypothesis "hard objects bounce" and four respective events to be evaluated. On day 2, the children of PR and PR + M worked on the hypothesis of "objects filled with air bounce" and respective events, see Table 1. In both PR and PR + M, a visual representation was employed as a static scaffold to focus children's attention on p and q. The visual representation consisted of two trays on which both

[6] As the transfer test was employed in PR and PR + M to assess the effects of scaffolding in a comparison of training sessions and a transfer domain, a comparison between the three experimental groups including CG is not suitable.

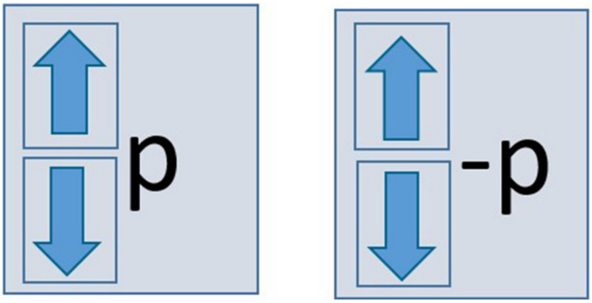

Fig. 1 Schematic Representation of p and q

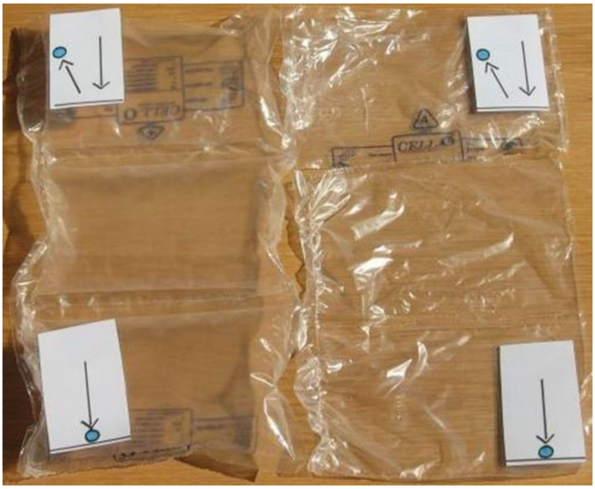

Fig. 2 Representation of p (with/without air) and q (bounces/does not bounce)

the quality of an object p (e.g., with air/without air) and the bouncing behavior q (bounces/does not bounce) were represented (Figs 1 and 2).

6.2.1 Representation of events and event evaluations

(1) Sorting by property After having been presented with the first hypothesis "hard objects bounce," the child received one of the four objects (marble stone, ball of clay, softball, juggling ball) and was told to put it onto one of the two trays according to their property p, in this case hard/not hard (i.e. soft). One tray was covered with an iron plate, representing hard objects (p), and one tray was covered with cottonballs, representing soft objects (–p). The child proceeded and placed the four objects according to their properties. The order of presentation of the four objects was varied randomly. A similar visual representation was used for the second hypothesis "things with air bounce" in training 2, see Fig. 2.

(2) Sorting by bouncing behaviour After all of the objects for one hypothesis had been placed onto the trays, the child was asked to re-arrange the objects according to their bouncing behavior. To this end, each two cards with arrows were additionally placed onto the trays, with an arrow pointing up for q (bounces) or pointing down for–q (does not bounce).

(3a) Presentation of hypothesis[7] The experimenter introduced a boy's hypothesis on objects' bouncing behavior by showing a small figure. "This is Tim. Tim believes: Hard objects bounce." Based on Tim's hypothesis, the child put the figure onto the respective cell of p/q of the visual representation where the four objects had been placed. If the child did not succeed, the experimenter repeated the hypothesis and pointed to the respective tray and arrow, and allowed the child to replace the figure.

(3b) Event evaluation The child evaluated the four objects placed onto the cells in random order by using three smileys (confirming/disconfirming/irrelevant) already employed in the pretest: First, the child picked up the respective object, tested its bouncing behavior, and replaced it onto its cell. Then, the child was asked: "Now, we would like to know whether this [object] shows that Tim's assumption[8] is correct, is incorrect, or whether it has nothing to do with Tim's assumption." The smiley-scale was placed next to the visual representation. On the smiley-scale, tokens could be placed to indicate whether the object was considered to be confirming (happy-looking smiley), disconfirming (sad-looking smiley), or irrelevant (thoughtful-looking smiley) to the given hypothesis. However, unlike in the procedure of the pre- and posttest, the type of scaffolding in the experimental groups PR and PR+M was varied both in training 1 and training 2 (see Sect. 6.2.2). The children of CG did not participate in training sessions 1 and 2 and attended their regular preschool groups (Fig. 3).

6.2.2 *Variation of scaffolding for event evaluations in the two experimental groups*

PR After having completed steps 1–3a, in step 3b the child was encouraged to give an answer to each of the four event evaluations based on step 3a by asking "What do you think?" putting a token on the smiley-scale. In the case of an incorrect answer, the child was given prompts. The child was asked: "What do you think? Is it hard or is

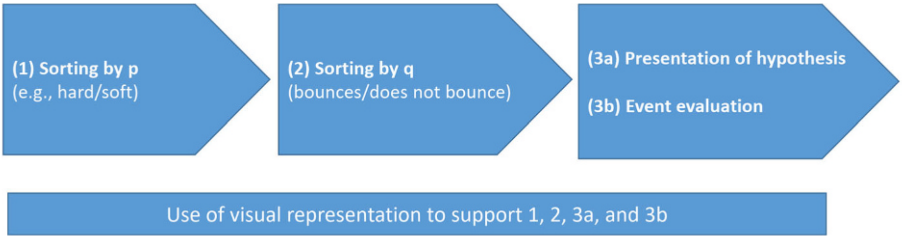
(1) Sorting by p
(e.g., hard/soft)
(2) Sorting by q
(bounces/does not bounce)
(3a) Presentation of hypothesis
(3b) Event evaluation
Use of visual representation to support 1, 2, 3a, and 3b

Fig. 3 Overview of Training Procedure

[7] Prior to (3a) and only in training 1, the experimenter introduced the child to the idea of a hypothesis. By touching an object hidden in a small bag, the child should guess what was in the bag. The child was then allowed to take the object out of the bag and told the experimenter whether his or her guess was correct.

[8] The term assumption is used in adults' statements to refer to the more technical term of hypothesis.

it not hard?” and “Does it bounce or not?” (redirecting attention to p and q). The child then restated the object’s properties accordingly and pointed to the respective cells of the visualization. The experimenter then said: “Remember what Tim believes.” Thereafter, the experimenter repeated the hypothesis and pointed to the small figure placed onto the respective cell of p/q (redirecting attention to hypothesis). The child then again answered using the smiley scale. In case of another incorrect answer, the experimenter asked the child to touch the object and asked: “Please touch the object again. Is it hard or soft?” The interviewer repeated the prompts until the child’s answer was correct. The same procedure was used in training 1 and training 2.

PR + M After having completed steps 1–3a, the experimenter modelled the correct reasoning by thinking aloud in each of the four event evaluations in step 3b. Before the adult explained his or her thinking, the child touched the respective object and tested its bouncing behavior, replacing it onto its cell. For the hypothesis “hard objects bounce” and an object of p/–q, the experimenter said: “I take this ball of dry clay. Now I look whether Tim’s assumption, hard objects bounce, is true for this ball. First, I try to decide whether it is hard or soft. This is the first rule: To think about what this object is like.” The experimenter then touched the object, gave it to the child, and said, “I notice that this is hard”. In the next step, the adult said: “Then I try out whether this ball of clay bounces or not. This is the second rule: To think about the question whether an object bounces or not. Now, I let it fall on the floor. And I can see: This ball does not bounce. Now, I have a look at Tim’s assumption. He assumes that hard things bounce. Therefore, Tim talks about hard objects. My object is hard. Tim says that those things bounce. My ball of clay is hard, but it does not bounce. Therefore, Tim is not right. Thus, I select the sad smiley.” All events of training 1 were given model explanations following the same rationale. On training day 2, the modeling was faded in the last three event evaluations to only employing the procedure of PR. Thus, the group of PR + M received additional modeling support in training 1 and in one event of training 2 which faded into adaptive prompting.

7 Results

In preliminary analyses, we checked whether there were significant differences between the experimental groups on the control measures. There were no significant differences on the variables of domain-general scientific reasoning, F (2, 60) = 0.94, $p = 0.39$, $\eta^2 = 0.03$, on inhibition (time), F (2, 60) = 1.02, $p = 0.37$, $\eta^2 = 0.03$, and inhibition (sum score), F (2, 60) = 0.15, $p = 0.84$, $\eta^2 = 0.01$, see Table 3.

Event evaluations (hypothesis 1a) Table 4 displays the frequencies of answers for the respective event evaluations of p/q, p/–q, –p/q, and –p/–q in the three groups in the pretest and the posttest. In the pretest, events of p/q were predominantly judged correctly as confirming evidence. Events of p/–q were predominantly judged correctly as disconfirming, whereas the events of –p/q and –p/–q were judged inconsistently as confirming, disconfirming, or irrelevant.

Table 3 Means and standard deviations of control variables

	PR Mean (SD)	PR + M Mean (SD)	CG Mean (SD)
Inhibition (time)	74.3 (26.6)	78.6 (22.4)	75.9 (28.3)
Inhibition (sum score)	1.71 (2.67)	1.33 (2.37)	2.62 (3.76)
Domain-general scientific reasoning	8.38 (2.29)	7.42 (2.50)	8.14 (2.24)

Range of scores for inhibition (time) = 0–24; inhibition (sum) = 0–158; domain-general reasoning = 0–12

We performed repeated measures MANOVAs with the between-subjects factor of Group (PR, PR + M, CG) and the within-subjects factors of Event (p/q, p/–q, –p/q, –p/–q) and Time (pre, post). We found a significant multivariate effect of Time, $F\ (4, 57) = 3.37$, wilks $\lambda = 0.81$, $p = 0.015$, $\eta^2 = 0.19$; Time × Group, $F\ (8, 114) = 2.63$, wilks $\lambda = 0.71$, $p = 0.011$, $\eta^2 = 0.156$, but not Group, $F\ (8, 114) = 1.05$, wilks $\lambda = 0.87$, $p = 0.40$, $\eta^2 = 0.07$. In univariate follow-up analyses it was shown that there were significant effects of Time for –p/q, $F\ (1, 60) = 8.80$, $p = 0.004$, $\eta^2 = 0.128$ and for –p/–q, $F\ (1, 60) = 8.80$, $p = 0.038$, $\eta^2 = 0.070$, but not for p/q, $F\ (1, 60) = 2.73$, $p = 0.10$, $\eta^2 = 0.04$ and for p/–q, $F\ (1, 60) = 0.81$, $p = 0.37$, $\eta^2 = 0.01$. There were significant effects for the events Time × Group for p/q, $F\ (2, 60) = 4.78$, $p = 0.12$, $\eta^2 = 0.14$. Children in PR + M and CG did not differ in their gains, but the PR + M outperformed the PR, $p < 0.05$, and CG outperformed the PR, $p < 0.01$. There were also significant effects for the events of –p/–q with Time × Group, $F\ (2, 60) = 4.55$, $p = 0.014$, $\eta^2 = 0.13$; Children in PR + M outperformed the PR, $p < 0.01$, and the CG, $p < 0.05$, there were no significant differences between PR and CG. In events of –p/q, Time × Group, $F\ (2, 60) = 2.51$, $p = 0.09$, $\eta^2 = 0.077$, PR + M outperformed PR, $p < 0.05$, but there were no significant differences between PR + M and CG, nor between PR and CG. In sum, as expected, the significant overall effect of Time by Group shows that children differed in their evaluations of the four events from pretest to posttest according to group. Children in PR + M showed higher performance in the difficult event evaluations of –p/q and –p/–q compared to PR; however, not in all cases did the CG differ significantly from the two intervention groups.

Reasoning patterns (hypothesis 1b) Table 5 displays the frequency of reasoning patterns in the pre- and posttest across the three groups on the scores of consistency of reasoning pattern and maximum reasoning pattern across the three tasks of the pretest and the posttest. The majority of reasoning patterns in the pretest were classified as inconsistent (53% across groups). As expected, advanced reasoning patterns of biconditional and conditional reasoning were found rarely in the pretest (3% across groups). Table 5 also shows that advanced reasoning patterns occur with higher frequency in the score of maximum reasoning pattern. Specifically, conditional reasoning is coded with 9% (PR + M), 0% (PR) and 9% (CR) in the pretest and 43% (PR + M), 9% (PR), and 9% (CG) in the posttest.

To investigate hypothesis 1b, we performed a repeated measures ANOVA with the within-subjects variable of Time (pre, post) and the between-subjects variable of Group (PR, PR + M, CG) and the score of consistent reasoning pattern as dependent variable. We found a significant effect of Time, $F\ (1, 60) = 3.85$, $p = 0.05$,

Table 4 Means (standard deviations) on the four events in the pre- and posttest

	PR			PR + M			CG		
Event	Confirming	Irrelevant	Disconfirm	Confirming	Irrelevant	Disconfirm	Confirming	Irrelevant	Disconfirm
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Prestest									
[p/q]	2.95 (0.22)	0.00 (0.00)	0.05 (0.22)	2.67 (0.73)	0.19 (0.51)	0.14 (0.48)	2.48(0.87)	0.19 (0.51)	0.33 (0.58)
[p/–q]	0.00 (0.00)	0.38 (0.74)	2.62 (0.74)	0.29 (0.78)	0.29 (0.46)	2.43 (0.87)	0.48 (0.93)	0.24 (0.54)	2.29 (1.00)
[–p/q]	1.52 (1.08)	0.38 (0.67)	1.10 (1.09)	1.23 (1.15)	0.57 (0.98)	1.14 (1.15)	1.14 (1.06)	0.43 (0.75)	1.43 (1.08)
[–p/–q]	0.10 (0.30)	0.48 (0.87)	2.43 (0.93)	0.57 (0.81)	0.43 (0.68)	2.00 (1.10)	0.67 (0.80)	0.43 (0.75)	1.90 (0.83)
Posttest									
[p/q]	2.71 (0.78)	0.19 (0.09)	0.09 (0.30)	2.90 (0.44)	0.09 (0.43)	0.00 (0.00)	2.95 (0.22)	0.00 (0.00)	0.05 (0.22)
[p/–q]	0.10 (0.30)	0.43 (0.87)	2.48 (0.87)	0.33 (0.58)	0.43 (0.81)	2.24 (1.00)	0.14 (0.36)	0.62 (0.86)	2.24 (1.04)
[–p/q]	0.95 (1.28)	0.43 (0.81)	1.62 (1.32)	1.00 (1.14)	1.38 (1.28)	0.62 (0.74)	0.81 (1.12)	0.81 (1.12)	1.38 (1.20)
[–p/–q]	0.33 (0.73)	0.43 (0.93)	2.24 (1.14)	0.43 (0.93)	1.29 (1.31)	1.29 (1.27)	1.05 (1.20)	0.48 (0.81)	1.48 (1.12)

Coding of events: correct answer = 1, incorrect answer = 0 across three pretest/posttest tasks (range 0–3)
confirming positive smiley, *irrelevant* neutral smiley, *disconfirm.* = *disconfirming* negative smiley

Table 5 Frequencies (percent) of consistent reasoning of patterns and maximum reasoning patterns by group and time

	PR		PR + M		CG	
	Consistent	Maximum	Consistent	Maximum	Consistent	Maximum
Prestest						
Conditional	0 (0%)	0 (0%)	1 (5%)	2 (9%)	0 (0%)	2 (9%)
Biconditional	0 (0%)	2 (9%)	0 (0%)	2 (9%)	1 (5%)	2 (9%)
Conjunctive	7 (33%)	11 (52%)	2 (9%)	6 (28%)	3 (14%)	7 (33%)
Equivalence	0 (0%)	0 (0%)	2 (9%)	2 (9%)	2 (9%)	0 (0%)
Consequence-focus	4 (19%)	7 (33%)	6 (28%)	5 (24%)	2 (9%)	5 (24%)
Inconsistent	10 (48%)	1 (5%)	10 (48%)	4 (19%)	13 (62%)	5 (24%)
Posttest						
Conditional	2 (9%)	2 (9%)	6 (29%)	9 (43%)	1 (5%)	2 (9%)
Biconditional	0 (0%)	1 (5%)	0 (0%)	1 (5%)	0 (0%)	1 (5%)
Conjunctive	7 (33%)	9 (43%)	0 (0%)	2 (9%)	2 (9%)	7 (33%)
Equivalence	0 (0%)	1 (5%)	0 (0%)	1 (5%)	3 (14%)	3 (14%)
Consequence-focus	5 (24%)	5 (24%)	5 (24%)	4 (19%)	4 (19%)	2 (9%)
Inconsistent	7 (33%)	3 (14%)	7 (33%)	4 (19%)	11 (52%)	6 (29%)

$\eta^2 = 0.60$, but no significant interaction of Time × Group, F (2, 60) = 0.249, $p = 0.78$, $\eta^2 = 0.008$, with means of PR + M ($M = 0.61$, $SD = 0.74$), PR ($M = 0.52$, $SD = 0.51$) and CG ($M = 0.43$, $SD = 0.60$) in the pretest and means of PR + M ($M = 1.0$, $SD = 1.34$), PR ($M = 0.86$, $SD = 0.85$) and CG ($M = 0.57$, $SD = 0.75$) in the posttest. As the majority of children were not assigned to a consistent reasoning pattern across three tasks, we performed the same analysis with the score of maximum reasoning pattern as a dependent variable. We found no significant effects of Time, F (1, 60) = 1.91, $p = 0.17$, $\eta^2 = 0.03$ and Time × Group, F (1, 60) = 2.51, $p = 0.09$, η^2 = 0.078, with means of PR + M ($M = 1.09$, $SD = 0.83$), PR ($M = 1.05$, $SD = 0.38$) and CG ($M = 1.05$, $SD = 0.86$) in the pretest and PR + M ($M = 1.71$, $SD = 0.77$), PR ($M = 1.09$, $SD = 0.77$) and CG ($M = 0.95$, $SD = 0.86$) in the posttest. In planned simple contrasts, the group of PR + M showed higher mean gains in the maximum reasoning pattern from pretest to posttest than the CG ($p = 0.038$) and PR ($p = 0.096$). The groups of PR and CG did not differ from each other.

8 Discussion

In the present study, we investigated to what extent preschoolers will profit from scaffolding when evaluating empirical evidence with regard to a given hypothesis. The coordination of theory and evidence is regarded a central aspect of scientific reasoning, with the evaluation of hypotheses as integral to inquiry science activities intended to promote conceptual learning from early on (Grimm et al. 2018; Leuchter 2017). Following the deductive reasoning paradigm (e.g., Barrouillet et al. 2008), we employed tasks in the domain of elasticity ("why do balls bounce?") and presented preschoolers with hypotheses (e.g., “Tim believes: heavy objects bounce.”) and each four events of the antecedent p (present/absent) and the consequent q (present/absent) that were to be evaluated. In an experimental training study with pre-post design, we evaluated whether support with adaptive prompts (PR) or with adaptive prompts including modeling (PR + M) will advance reasoning in preschoolers (van de Pol et al. 2010; Pea 2004). In prior studies it was found that elementary school children are able to advance their deductive reasoning in the domain of elasticity both in a training setting (Robisch et al. 2014) and in a classroom setting (Grimm et al. 2018) by means of adaptive prompts, whereas modeling did not contribute to this effect.

With regard to our first hypothesis, we found that preschoolers who were supported by adaptive prompts including the modeling of advanced reasoning gained significantly in their correct evaluations of irrelevant events of –p/–q compared to the control group and a group receiving adaptive prompts only. Similarly, they gained in the evaluation of events of –p/q, outperforming the group receiving prompts only. The mere use of adaptive prompts, thus, led to mixed results, as this group did not differ from the control group in the comparison of irrelevant event evaluations. Apparently, the adaptive support by hints and questions, additionally supported by the use of a visual representation that was used in both experimental groups, did not suffice to promote young children’s deductive reasoning when evaluating irrelevant events. Rather, it was the explication of reasoning strategies by the experimenter’s modeling of thinking processes and event evaluations in the group of PR + M that fostered preschoolers’ irrelevant event evaluation—those evaluations that have proven to be most difficult to elementary school and early secondary school students in prior investigations (e.g, Robisch et al. 2014; Troebst et al. 2011).

While differences between the two experimental groups were consistently found for irrelevant event evaluations, the differences between PR + M and the performance of the control group were not consistently significant. Possibly, despite power analyses based on a previous study in elementary school, the statistical power to detect differences in performance was not sufficient given that there were a priori differences in performance in the pretest between the experimental groups. In addition, the control group’s performance in event evaluations of p/q was raised significantly without intervention. While this result might be explained by the control group’s lower mean performance in the pretest, it also points to gains merely by the repeated exposure to reasoning tasks. Apparently, especially the relatively low cognitive challenge of matching a given hypothesis with the characteristics of p and q may be achieved without training involving scaffolding features. Overall, our results show a differential picture of the three groups’ gains on the four event evaluations from pretest to posttest. As expected, the events of p/q were evaluated correctly to a high degree already in the pretest. Apart from evaluations of p/q, children’s performance with regard to p/–q as disconfirming evidence was rather high at pretest, with no significant differential gains across groups.

How may our results with regard to gains in event evaluations of p/q, p/–q, –p/q, and –p/–q be explained? According to the Dual Process Mental Models Theory (Gauffroy and Barrouillet 2009), the construction of mental models constitutes the cognitive basis for deriving decisions with regard to conditional statements. The construction of a model representing the antecedent (p) and consequent (q) of a given statement in an affirmative mode constitutes a basic first step. Our data shows that preschool children are able to correctly evaluate events of p/q, presumably based on an initial model, even in the pretest. In a second step of "fleshing out," further mental models are derived, concerning the states of –p/q and –p/–q. It is assumed that the states not explicitly represented in the fleshed-out mental models would lead to disconfirming event evaluations in cases of p/–q. We found that preschool children indeed show rather high performance with regard to the evaluation of disconfirming events in the pretest. The fleshing out of additional models as a differential effect of training, therefore, presumably concerned the processes of analytical reasoning in the construction of additional mental models of –p/q and –p/–q. Here, the significant gains of the training group of PR + M may be interpreted in terms of the facilitation of mental models of irrelevant events that are compatible with a given hypothesis, yet do not show its truth and therefore are irrelevant to the assumption (Barrouillet et al. 2008). This process of fleshing out may also have affected performance with regard to the evaluation of disconfirming evidence, focusing children's attention onto irrelevant events and thereby hindering additional gains in event evaluations of p/–q.

As the mental models theory is based on the interpretation of reasoning patterns rather than single event evaluations, a look at the systematicity of children's responses across our three tasks of the pretest and posttest is illuminating, as investigated in our hypothesis 1b. Our analyses of consistent reasoning patterns show that a large proportion of children's responses across the three tasks of the pretest were indeed inconsistent (48% for PR and PR + M, respectively, 62% for CG). That is, they did not follow the patterns of conjunctive, biconditional, or conditional reasoning, nor the patterns of equivalence and consequence-orientation found in prior studies with elementary school children (Robisch et al. 2014). Given that elementary school children's reasoning patterns without training were inconsistent to a high extent, this result is not surprising (Troebst et al. 2011). However, our score of maximum reasoning pattern revealed that the majority of children followed the reasoning patterns of conjunctive reasoning, consequence-focus, or equivalence at least once across the three pretest tasks, thus showing young children's tendency for affirming events of p/q (conjunctive) or one element (equivalence; consequence-focus) when evaluating evidence without training. The code of maximum reasoning also shows that, as expected in hypothesis 1b, significant training gains in advanced reasoning patterns were evident in the group of PR + M compared to CG and in tendency compared to PR. Moreover, 43% of children in the group of PR + M were able to reason according to the pattern of conditional reasoning at least once in the posttest, whereas this was the case for only 9% of the PR and 9% of CG. Altogether, the score of maximum reasoning pattern may be better suited for an analysis of young children's reasoning with fully contextualized assumptions (e.g., "Tim believes: large balls bounce") than the consistency scores typically used in analyses of abstract task content (e.g. "if the circle is white, the triangle is black."). Yet, further research comparing children's deductive reasoning in context-lean and context-rich domains of science is needed to disentangle differences in task content on deductive reasoning patterns.

Taken together, our results imply that deductive reasoning in a contextualized evaluation task may be advanced by modeling that is combined with adaptive prompts in individualized settings. This result is in line with English (1997) who found that deductive reasoning improved only in conditions with an explication of p/q and adaptive prompts in context-lean reasoning tasks. Similarly, in inquiry science contexts, the dimensions of instructional guidance based on strategies of modeling and explanations, have been suggested as a fruitful way for enriching more implicit approaches such as prompting (Lazonder and Harmsen 2016; van der Graaf et al. 2019). Thus, the results of our study point to potential benefits of combining the different instructional approaches of direct instruction and adaptive verbal support for challenging scientific reasoning processes. In our study, adaptivity of prompts was defined as the provision of support only in cases of a child's incorrect answer. This type of adaptive prompting thus incorporated essential characteristics of scaffolding based on diagnosed (lack of) student understanding. Thus, prompts were only provided to those children that showed difficulties in answering questions of event evaluations during training 1 and 2. While we consider this adaptivity of prompts as a strength of our study as the personal relevance of prompting for task solution was given, we do not know whether prompting would have led to differing effects had it been provided without contingency on individual understanding.

Several further limitations of this study need to be pointed out. Although we employed an experimental design with high comparability of participants across conditions, our sample was relatively small so that effects with regard to the performance of the control group without training could not be disentangled entirely. Specifically, it may be that the mere confrontation with evidence-based reasoning tasks lead to performance gains in events of combination of p and q. Therefore, conducting replication studies scrutinizing effects with regard to control and experimental groups taking into account differing individual preconditions might be of interest. Furthermore, the investigation of deductive reasoning patterns with scores of consistency and scores of maximum performance needs to be validated with regard to varying age groups. This way, an integration with results of existing developmental approaches such as Barrouillet et al. (2008) may be achieved, delineating age-specific and context-specific differences in reasoning. Finally, this study was conducted in a highly controlled and individualized training context of one-to-one scaffolding of an expert with a child. Whether effects of a combination of modeling and prompting will transfer to inquiry science contexts is a question for future research. Following Grimm et al. (2018), the transfer and integration of scientific reasoning tasks into educational contexts involving argumentation and whole-class scaffolding seems to be a promising approach. Results of our study may therefore also delineate research on properties of learning environments for further aspects of scientific reasoning such as hypothesis generation and argumentation (Lecare 2104; Köksal-Tuncer and Sodian 2018) in settings of early science education.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Baddeley, A.D. (1983). Working Memory. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *302*(1110), 311–324.

Barkl, S., Porter, A., & Ginns, P. (2012). Cognitive training for children: Effects on inductive reasoning, deductive reasoning, and mathematics achievement in an Australian school setting. *Psychology in the Schools*, *49*(9), 828–842.

Barrouillet, P., & Camos, V. (2001). Developmental increase in working memory span: resource sharing or temporal decay? *Journal of Memory and Language*, *45*(1), 1–20. https://doi.org/10.1006/jmla.2001.2767.

Barrouillet, P., & Lecas, J.F. (1999). Mental models in conditional reasoning and working memory. *Thinking and Reasoning*, *5*, 289–302.

Barrouillet, P., Gauffroy, C., & Lecas, J.-F. (2008). Mental models and the suppositional account of conditionals. *Psychological Review*, *115*(3), 760–771.

Brush, T., & Saye, J. (2002). A summary of research exploring hard and soft scaffolding for teachers and students using a multimedia supported learning environment. *The Journal of Interactive Online Learning*, *1*(2), 1–12.

Chantal, P. de, & Markovits, H. (2017). The capacity to generate alternative ideas is more important than inhibition for logical reasoning in preschool-age children. *Memory & Cognition*, *45*, 208–220. https://doi.org/10.3758/s13421-016-0653-4.

Chantal, P.-L. de, Gagnon-St-Pierre, E., & Markovits, H. (2019). Divergent thinking promotes deductive reasoning in preschoolers. *Child Development*, *91*(4), 1081–1097. https://doi.org/10.1111/cdev.13278.

Chinn, C.A., & Brewer, W.F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, *35*(6), 623–654.

Collins, A., Brown, J.S., & Newman, S.E. (1987). Cognitive apprenticeship: teaching the crafts of reading, writing and mathematics. In L. Resnick (Ed.), *Knowing, learning, and instruction* (pp. 453–494). Mahwah: Lawrence Erlbaum.

Duschl, R.A. (2003). Assessment of inquiry. In J.M. Atkin & J. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41–59). Arlington: NSTA.

Edelsbrunner, P.A., Schalk, L., Schumacher, R., & Stern, E. (2018). Variable control and conceptual change: a large-scale quantitative study in primary school. *Learning and Individual Differences*, *66*, 38–53. https://doi.org/10.1016/j.lindif.2018.02.003.

English, L.D. (1997). Interventions in children's deductive reasoning with indeterminate problems. *Contemporary Educational Psychology*, *22*(3), 338–362.

Furtak, E.M., Seidel, T., Iverson, H., & Briggs, D.C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: a meta-analysis. *Review of Educational Research*, *82*(3), 300–329. https://doi.org/10.3102/0034654312457206.

Gauffroy, C., & Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: an integrative developmental theory. *Developmental Review*, *29*(4), 249–282. https://doi.org/10.1016/j.dr.2009.09.002.

Gauffroy, C., & Barrouillet, P. (2011). The primacy of thinking about possibilities in the development of reasoning. *Developmental Psychology*, *47*, 1000–1011.

Gerde, H.K., Schachter, R.E., & Wasik, B.A. (2013). Using the scientific method to guide learning: an integrated approach to early childhood curriculum. *Early Childhood Education Journal*, *41*(5), 315–323. https://doi.org/10.1007/s10643-013-0579-4.

Graaf, J. van der, Segers, E., & Verhoeven, L. (2016). Scientific reasoning in kindergarten: cognitive factors in experimentation and evidence evaluation. *Learning and Individual Differences*, *49*, 190–200. https://doi.org/10.1016/j.lindif.2016.06.006.

Graaf, J. van der, Segers, P.C.J., & Verhoeven, L.T.W. (2018). Experimentation abilities in kindergarten children with learning problems. *European Journal of STEM Education*, *3*(3), 1–10. https://doi.org/10.20897/ejsteme/3873.

Graaf, J. van der, Sande, E. van der, Gijsel, M., & Segers, E. (2019). A combined approach to strengthen children's scientific thinking: direct instruction on scientific reasoning and training of teacher's verbal support. *International Journal of Science Education*, *41*(9), 1119–1138. https://doi.org/10.1080/09500693.2019.1594442.

Grimm, H., Robisch, C., & Möller, K. (2018). Förderung hypothesenbezogener Schlussfolgerungen im naturwissenschaftlichen Sachunterricht durch gezieltes Scaffolding – Gelingt dies unter Feldbedingungen? *Zeitschrift für Grundschulforschung*, *11*(2), 349–363.

Gropen, J., Clark-Ciarelli, N., Hoisington, C., Stacy, B., & Ehrlich, S.B. (2011). The importance of executive function in early science education. *Child Development Perspectives*, *5*(4), 298–304.

Hager, W., & Hasselhorn, M. (1998). The effectiveness of the cognitive training for children from a differential perspective: a meta-evaluaton. *Learning and Instruction*, *8*, 411–438.

Handley, S.J., Capon, A., Beveridge, M., Dennis, I., & Evans, J.S.B. (2004). Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, *10*(2), 175–195.

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal of Educational Research Online*, *11*(2), 169–191.

Hermkes, R., Mach, H., & Minnameier, G. (2018). Interaction-based coding of scaffolding processes. *Learning and Instruction*, *54*, 147–155.

Jansen, H., Mannhaupt, G., Marx, H., & Skowronek, H. (1999). *Bielefelder Screening zur Früherkennung zu Lese-Rechtschreibschwierigkeiten*. Göttingen: Hogrefe.

Johnson-Laird, P.N., & Byrne, R.M.J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review*, *109*, 646–678.

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

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1–48.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667. https://doi.org/10.1111/j.0956-7976.2004.00737.x.

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.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: a comprehensive inventory. *Child Development*, *86*, 327–336. https://doi.org/10.1111/cdev.12298.

Köksal-Tuncer, O., & Sodian, B. (2018). The development of scientific reasoning: hypothesis testing and argumentation from evidence in young children. *Cognitive Development*, *48*, 135–145. https://doi.org/10.1016/j.cogdev.2018.06.011.

Kuhn, D. (2002). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *Handbook of childhood cognitive development* (pp. 371–393). Malden: Blackwell.

Lawson, A.E. (2010). Basic inferences of scientific reasoning, argumentation, and discovery. *Science Education*, *94*, 336–364.

Lazonder, A.W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Review of Educational Research*, *86*(3), 681–718. https://doi.org/10.3102/0034654315627366.

Legare, C. (2014). The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives*, *8*(2), 101–106. https://doi.org/10.1111/cdep.12070.

Leuchter, M. (2017). *Kinder erkunden die Welt. Frühe naturwissenschaftliche Bildung und Förderung*. Entwicklung und Bildung in der Frühen Kindheit. Stuttgart: Kohlhammer.

McNeill, K.L., & Krajcik, J. (2008). Scientific explanations: characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, *45*(1), 53–78. https://doi.org/10.1002/tea.20201.

Meiser, T., & Klauer, K.C. (2000). Training des deduktiven Denkens [Training of deductive thinking]. In K.J. Klauer (Ed.), *Handbuch Kognitives Training [Handbook cognitive training]* (2nd edn., pp. 211–234). Göttingen: Hogrefe.

Muhonen, H., Rasku-Puttonen, H., Pakarinen, E., Poikkeus, A.M., & Lerkkanen, M.K. (2016). Scaffolding through dialogic teaching in early school classrooms. *Teaching and Teacher Education*, *55*, 143–154. https://doi.org/10.1016/j.tate.2016.01.007.

National Research Council (NRC) (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington: National Academies Press.

Pea, R. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education and human activity. *The Journal of the Learning Sciences*, *13*, 423–453.

Piekny, J., & Mähler, C. (2013). Scientific reasoning in early and middle childhood: the development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *British Journal of Developmental Psychology*, *31*(2), 153–179. https://doi.org/10.1111/j.2044-835X.2012.02082.x.

Piekny, J., Grube, D., & Mähler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *International Journal of Science Education*, *36*(2), 334–354.

Pol, J. van de, Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: a decade of research. *Educational Psychological Review*, *3*, 210–215.

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex learning environment: what have we gained and what have we missed? *Educational Psychologist*, *40*(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1.

Reiser, B.J. (2004). Scaffolding complex learning: the mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences*, *13*(3), 273–304.

Robisch, C., Tröbst, S., & Möller, K. (2014). Hypothesenbezogene Schlussfolgerungen im Grundschulalter fördern. *Zeitschrift für Grundschulforschung*, *7*, 88–101.

Samarapungavan, A., Patrick, H., & Mantzicopoulos, P. (2011). What kindergarten students learn in inquiry-based science classrooms. *Cognition and Instruction*, *29*(4), 416–470. https://doi.org/10.1080/07370008.2011.608027.

Sandoval, W., Sodian, B., Koerber, S., & Wong, J. (2014). Developing children's early competencies to engage with science. *Educational Psychologist*, *49*(2), 139–152. https://doi.org/10.1080/00461520.2014.917589.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*, 753–762.

Steffensky, M. (2017). *Naturwissenschaftliche Bildung in Kindertagesstätten. Weiterbildungsinitiative Frühpädagogische Fachkräfte*. WiFF-Expertisen, Vol. 48. München: Deutsches Jugendinstitut e. V.

Stephan-Gramberg, S. (2015). Förderung der Koordination von Theorie und Evidenz bei Kindern im Elementarbereich. Inauguraldissertation am Fachbereich Erziehungswissenschaften, Goethe-Universität Frankfurt

Tiego, J., Testa, R., Bellgrove, M.A., Pantelis, C., & Whittle, S. (2018). A hierarchical model of inhibitory control. *Frontiers in Psychology*, *9*, 1–25. https://doi.org/10.3389/fpsyg.2018.01339.

Troebst, S., Hardy, I., & Möller, K. (2011). Die Förderung deduktiver Schlussfolgerungen bei Grundschulkindern in naturwissenschaftlichen Kontexten. *Unterrichtswissenschaft*, *39*(1), 7–20.

Vorholzer, A., & von Aufschnaiter, C. (2019). Guidance in inquiry-based instruction—an attempt to disentangle a manifold construct. *International Journal of Science Education*, *41*(11), 1562–1577. https://doi.org/10.1080/09500693.2019.1616124.

Wason, P.C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology*. Harmonswordth: Penguin.

Wason, P.C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23*, 63–71.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, *92*, 941–967.

Wood, D., Bruner, J.S., & Ross, G. (1976). The role of tutoring in problem solving. *Child Psychology*, *17*, 89–100.

Zimmermann, C. (2007). The development of scientific thinking skills in elementary and middle school. *Science Direct*, *2*, 172–223.