Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol.3 No. 4 (2015) 56 - 94
ISSN 2295-3159

Children's understanding of experimental contrast and experimental control: an inventory for primary school

Christopher Osterhaus[a], Susanne Koerber[a], Beate Sodian[b]

[a]Freiburg University of Education, Department of Psychology, Germany

[b]Ludwig-Maximilians-University Munich, Department of Psychology, Germany

Article received 18 October / revised 29 November / accepted 4 December / available online 19 January

Abstract

Experimentation skills are a central component of scientific thinking, and many studies have investigated whether and when primary-school children develop adequate experimentation strategies. However, the answers to these questions vary substantially depending on the type of task that is used: while discovery tasks, which require children to engage in unguided experimentation, typically do not reveal systematic skills in primary school, choice tasks suggest an early use of adequate experimentation strategies. To acquire a more accurate description of primary-school experimentation, this article proposes a novel multiple-select paper-and-pencil inventory that measures children's understanding of experimental design. The two reported studies investigated the psychometric properties of this instrument and addressed the development of primary-school experimentation. Study 1 assessed the validity of the item format by comparing 2 items and an interview measure in a sample of 71 third- and fourth-graders (9- and 10-year-olds), while Study 2 investigated the reliability and the convergent validity of the inventory by administering it to 411 second-, third- and fourth-graders (8-, 9- and 10-year-olds) and by comparing children's performance in the 11-item scale to 2 conventional experimentation tasks. The obtained results demonstrate the reliability and validity of the inventory and suggest that a solid understanding of experimental design first emerges at the end of primary school.

Keywords: experimentation skills; control-of-variable strategy (CVS); scientific thinking; primary (elementary) school; paper-and-pencil assessment

Corresponding author: Christopher Osterhaus, Department of Psychology, Freiburg University of Education, Kunzenweg 21, 79117 Freiburg, Germany. Phone: +49(0)761 / 682-164, Fax: +49(0)761 / 682 - 98164, Email: osterhaus@ph-freiburg.de. DOI: http://dx.doi.org/10.14786/flr.v3i4.220

1. Introduction

Experimentation skills constitute a fundamental component of scientific thinking, and many developmental research studies have investigated children's acquisition of adequate experimentation strategies (e.g., Case, 1974; Inhelder & Piaget, 1958; Kuhn & Phelps, 1982; Siegler & Liebert, 1975; for a review see Zimmerman, 2007). Two central research questions have been whether and when children begin to master the so-called control-of-variables strategy (CVS). This strategy, which is also referred to as the "vary-one-thing-at-a-time" strategy (Tschirgi, 1980), requires informative experiments to contrast a single (focal) variable while keeping all other (non-focal) variables constant. Most studies of experimentation skills probe children's use of CVS in two different task settings: (1) discovery tasks with unrestricted variable configurations (production of CVS) and (2) choice tasks with restricted response options (choice of CVS).

Discovery tasks typically require children to explore the causal relations between different candidate causes and an outcome over a set of experimentation trials (e.g., Kuhn et al., 1995; Schauble, 1990). In each of these trials, mature reasoners form hypotheses about the system of independent variables (i.e., which variables are causal and which are non-causal), and based on these, they use CVS to isolate a single variable, which they contrast in an experiment that controls all non-focal variables (controlled-contrastive experiment). By updating their hypotheses and repeating this procedure over multiple trials, mature reasoners arrive at a final theory of the causal standing of each variable. Discovery tasks hence involve reasoners in multiple phases of constructing scientific knowledge (i.e., reiterative formation of hypotheses, experimentation and data interpretation), and therefore they offer a high ecological validity and are particularly well suited for microgenetic studies that investigate strategy change in detail by repeatedly using the same task with a high density of observations in the short period of time when change is assumed to occur.

Choice tasks, in contrast, typically present children with a specific and directed hypothesis regarding one of the candidate causes. In addition, they include restricted answer options that represent distinct variable configurations from which children are allowed to choose (multiple choice [MC]; e.g., Croker & Buchanan, 2011; Koerber, Mayer, Osterhaus, Schwippert, & Sodian, 2015; Mayer, Sodian, Koerber, & Schwippert, 2014; Tschirgi, 1980). Mature reasoners understand that they need to compare conditions, and their command of CVS is demonstrated by selecting the answer option in which only the focal variable is varied. This requirement of a single choice makes choice tasks easier to administer and hence well suited for large-scale, paper-and-pencil-based assessments, which are important research tools for investigating the large interindividual differences that already exist in primary-school experimentation (Bullock, Sodian, & Koerber, 2009).

Despite their common measurement focus on CVS and their same basic task requirement (i.e., understanding that conditions need to be compared and an appropriate comparison needs to be made), discovery and choice tasks reveal a substantially different picture of primary-school children's experimentation skills. While discovery tasks tend to reveal that only a small number of students *produces* experiments that are designed in accordance with CVS (e.g., 17% initially used CVS in a sample of fifth- and sixth-graders; Schauble, 1990), choice tasks reveal that most primary-school children prefer controlled-contrastive experiments over confounded experiments when they are allowed to *choose* between different experimental designs (e.g., 54% correct CVS choices by fourth-graders in Koerber, Mayer, et al., 2015; around 60% in Bullock & Ziegler, 1999). Although differences between production and choice tasks (i.e., between open-answer and closed-response items) are a common empirical finding for knowledge scales (e.g., Nehm & Schonfeld, 2008), the large discrepancies between the two types of CVS tasks suggest that differences in performance are not solely attributable to the increased probability of correct guessing in choice tasks.

Performance differences between discovery and choice tasks might be attributable to their specific and discrepant task demands and solution strategies (see Table 1). While discovery tasks require the repeated generation of hypotheses, which is an ability that may develop later than do experimentation skills (Piekny & Maehler, 2013) and which requires extensive memory and processing capacities, CVS choice tasks might not assess children's *understanding* of experimentation because they can be solved by lower-level heuristics, such as varying a single variable while keeping the others constant, which does not need to be motivated by children's full understanding of experimentation. Other evidence, such as children's justifications, is therefore required before concluding whether children fully understand the rationale for CVS.

Table 1

Student performance in and characteristics of discovery tasks, choice tasks and understanding of experimental design (UNEX).

	Discovery tasks	Choice tasks	UNEX
Description	iterative experimentation and use of CVS to uncover causal effects	single choice of a (controlled-contrastive) experimental comparison	identification of design errors in (controlled-) contrastive experiments
Task format	open; often computerized	MC; sometimes other formats that do not favour guessing (e.g., Bullock & Ziegler, 1999)	MS
Use of CVS	17% initial use in at ages 11 and 12 (Schauble, 1990)	8–68% (Croker & Buchanan, 2011), 54% (Koerber, Mayer, et al., 2015), and 60% (Bullock & Ziegler, 1999) all at age 10	29% at age 10 (cf. Study 2)
Validity	moderate	low	high
	guessing is not a problem	guessing is a problem (especially for MC items)	guessing is not a problem*
	task requires non-essential skills: tracking of variables (memory and processing skills), hypothesis generation	may be solved by lower-level heuristics	understanding features of experimental contrast and experimental control required
Reliability	high	low	high
	performance estimates stable across problems	performance estimates differ across tasks and contexts	performance estimates stable across problems*
Assessment	time-consuming	rapid; allows for large-scale assessment	rapid; allows for large-scale assessment

Note. *Characteristics investigated in the two studies presented in this article; CVS = control-of-variables strategy; MC = multiple choice; MS = multiple select.

A further criticism of both discovery and choice tasks is that studies of scientific thinking—with their narrow focus on CVS—have not explored the broader context of children's understanding of the experimental method. In addition to CVS, a basic feature of experimentation is that planned comparisons are necessary to test for differences between conditions, as are randomization and an understanding of local control, which is necessary to reduce variation due to extraneous factors and which goes beyond the control of non-focal variables.

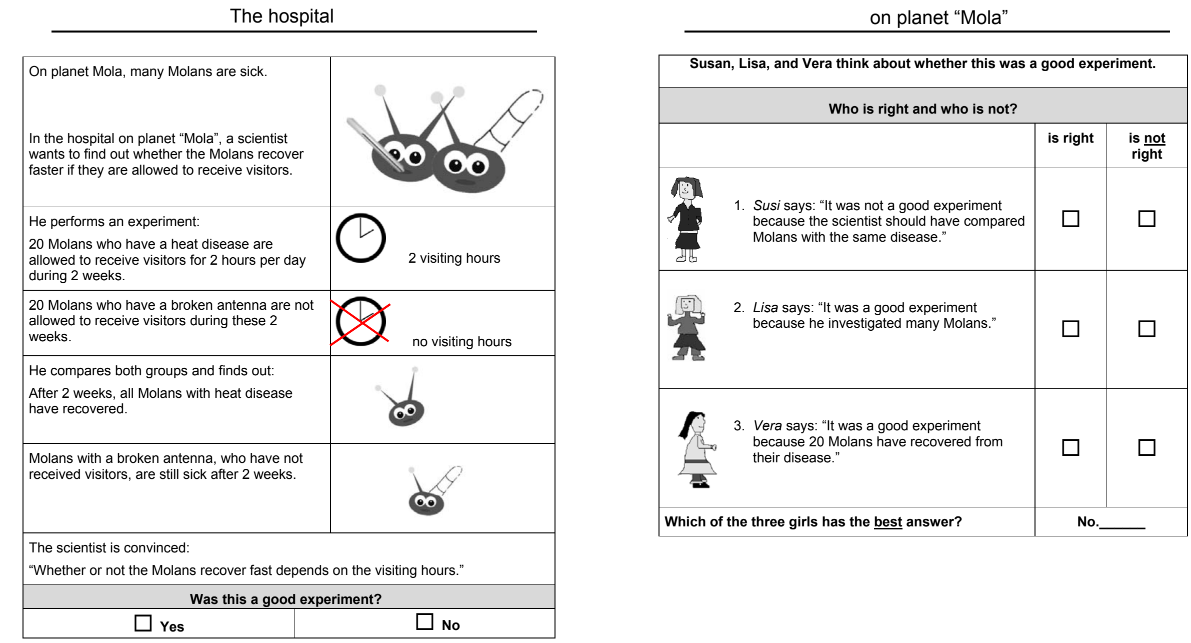
Bullock et al. (2009) developed an interview that investigates children's metaconceptual understanding of experimental design (UNEX) in participants aged 12–22 years. This instrument asks children to review a set of fictitious experiments that contain diverse design errors that violate the principles of local control or experimental contrast. Children need to recognize that ill-designed controlled-contrastive experiments do not provide an adequate test of hypothesis because variation due to extraneous factors is not reduced or non-focal variables are not controlled for (violation of CVS); they also must understand that hypothesis testing is impossible in ill-designed contrastive experiments because a single observation is made (either due to the missing variation of the focal variable, or the lack of an initial measurement in a pre-post design) and the principle of experimental contrast therefore is ignored. This latter experiment type provides no information about the control of non-focal variables, and the 'violated' design feature is simply the principle of experimental contrast, which should be easier to understand than 'experimental control'.

Primary-school children's UNEX has not been investigated previously. However, Bullock and her colleagues found that sixth-graders (the youngest age group interviewed with such an instrument) solved around 50% of tasks correctly, by identifying the experimental design error and justifying their opinion about whether the experiment was appropriately designed. It therefore appears that even very young children can show UNEX if they are questioned in an age-appropriate format that offers contextual support, such as providing a graphic representation of fictitious experiments and offering answer alternatives in a closed-response format.

The present studies investigated whether a paper-and-pencil version of UNEX can provide valid and reliable measurements in primary-school children, in order to determine whether early abilities can be identified in this age group. Our inventory of primary-school UNEX uses 11 closed-response, multiple-select (MS) items (examples are provided in Figure 1, and Appendix 1 provides the full item set). For each item the children have to answer whether or not they consider a fictitious experiment to be well designed, and whether or not they agree with each of three separate justifications concerning the good or bad quality of the experiment (which includes one that identifies the design error in question).

More conventional MC items provide a single choice, whereas the three separate decisions required by our MS procedure have two important advantages: (1) the probability of correct guessing is substantially reduced (i.e., 12.5% instead of 33% in the case of three statements), and (2) the MS format investigates potential inconsistencies in children's understanding of experimentation. MC items only make it possible to conclude that children consider their chosen answer to be superior to the non-selected alternatives, whereas MS items give additional information about their view on all response options (i.e., although children may recognize the design error, they may still hold naïve beliefs regarding the production of effects).

The correct and incorrect answer options used in the present studies are based on children's answers to open items in prestudies and they draw on a conceptual-development model of scientific thinking (Koerber, Mayer, et al., 2015). Specifically, each item includes naïve, intermediate and advanced-level answers. Naïve-level answers reveal no understanding of hypothesis testing, instead referring to the production of an effect (cf. answer option 3 in Figure 1), intermediate-level answers demonstrate a first understanding of the necessity of hypothesis testing and the existence of relevant design features (e.g., testing and sample size are important; cf. answer option 2), and advanced-level answers identify the specific design error and recognize that it restricts the information that can be drawn from the experiment (cf. answer option 1).



The hospital

On planet Mola, many Molans are sick. In the hospital on planet “Mola”, a scientist wants to find out whether the Molans recover faster if they are allowed to receive visitors.	
He performs an experiment: 20 Molans who have a heat disease are allowed to receive visitors for 2 hours per day during 2 weeks.	2 visiting hours
20 Molans who have a broken antenna are not allowed to receive visitors during these 2 weeks.	no visiting hours
He compares both groups and finds out: After 2 weeks, all Molans with heat disease have recovered.	
Molans with a broken antenna, who have not received visitors, are still sick after 2 weeks.	
The scientist is convinced: “Whether or not the Molans recover fast depends on the visiting hours.”	
Was this a good experiment?	
□ **Yes**	□ **No**

on planet “Mola”

Susan, Lisa, and Vera think about whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Susi* says: “It was not a good experiment because the scientist should have compared Molans with the same disease.”	□	□
2. *Lisa* says: “It was a good experiment because he investigated many Molans.”	□	□
3. *Vera* says: “It was a good experiment because 20 Molans have recovered from their disease.”	□	□
Which of the three girls has the best answer?	**No.______**	

Figure 1. Sample item.

While research has shown that children’s (dis-)agreement with conceptually different levels on MS items matches the levels found in an interview conducted *after* exposure to the MS answer options (MS item then open interview [I-after]; cf. Koerber, Osterhaus, & Sodian, 2015), little is known about the relation between children’s MS choices and the beliefs they hold *before* being presented with the MS item (open interview then MS item [I-before]). An instrument’s reliability and validity depends on a close relation between initial beliefs and the levels identified in the MS item.

The present work therefore first investigated (in Study 1) whether the MS item format is valid, resulting in ascriptions of levels that are significantly related with those levels found in I-before, and then (in Study 2) addressed whether the inventory results in a reliable scale with satisfactory content and convergent validities. Because primary-school children’s UNEX has not been investigated previously, Study 2 also addressed the abilities and development of primary-school children.

2. Study 1

Study 1 investigated the validity of the MS item format by comparing children’s performance in two interview measures, which were conducted before (I-before) and after (I-after) the presentation of the closed-response answer options. Study 1 also investigated the relation between the novel MS and conventional MC formats, as well as the relation between MC and the interview, both before and after the presentation of the MS task (i.e., I-before and I-after), to obtain a better understanding of the relative performances of diverse testing formats.

2.1 Methods

2.1.1 Participants

The 71 included primary-school children comprised 56 third-graders (mean age 8 years, 8 months, *SD*=5 months; 31 girls) and 15 fourth-graders (mean age 10 years, 1 month; *SD*=4 months;

7 girls). Children were recruited from three predominantly middle-class schools in Germany. Parental informed consent was obtained for all children.

2.1.2. Materials

The children were presented with and interviewed about (see Procedure) two items from our inventory: one contrastive and one controlled-contrastive experiment. Both experiments were set up in an artificial context in order to reduce interferences of children's content knowledge on design evaluations.

Item 1 (contrastive experiment) presented the children with a story about a scientist who wants to test the hypothesis that yellow fertilizer increases plant growth significantly more than blue fertilizer. The scientist administers yellow fertilizer to 100 plants. He observes that all plants got bigger blossoms and therefore concludes that the yellow fertilizer works better than the blue one. Children were asked whether or not this was a good experiment, and evaluated three explanations for their opinion (MS) on three hypothesized levels: (1) an explanation based on the production of effects (naïve level; "It was a good experiment because all plants that received the yellow fertilizer became bigger"), (2) an explanation that contained a design feature that was not the crucial to the experiment's validity (intermediate level; "It was a good experiment because he tested the yellow fertilizer on many plants") and (c) the correct explanation that identified the design error in question (advanced level; "It was not a good experiment because he only tested the yellow fertilizer"). After evaluating each of the three levels (MS), the children also indicated which of the three explanations they considered the best (MC).

Item 2 (controlled-contrastive experiment) presented the children with a story about two grandmothers who use lake or river water to water their plants. While those plants watered with lake water grow well, plants that are watered with river water are withering. Children had to evaluate whether the experimental data was sufficient to support the hypothesis that lake water causes plants to grow well, or whether more information would be needed (i.e., control of non-focal variables). Analogously to Item 1, the three explanations reflected the distinct levels; however, while the naïve level as was the case for Item 1 referred to the production of effects and the advanced level identified the design error in question, the intermediate level for this item included a reference to a potential causal mechanism rather than to a non-crucial design feature (i.e., "Lake water contains more minerals than lake water. Therefore, lake water is better for plants").

2.1.3 Procedure

Both items were read out loud to the children in a one-on-one interview. Children marked their answers in their own booklets. Interviews were conducted at two points during the presentation of the item: (1) after the children's initial design evaluation and before presenting the MS answer options (I-before; "Why was the experiment good/bad?"), and (2) after presenting the answer options and the children choosing the best one (I-after; "Why did you consider this answer to be the best one?"). Both instances included follow-up questions such as "Why did this [the reason children named] make the experiment a good/bad one?" or "Would you have done anything differently?"

Table 2

Coding of MS and interviews

	Levels children agreed to		
	Naïve	Intermediate	Advanced
Coding of MS *(final level)*			
Naïve	x	---	---
Naïve	x	x	---
Naïve	x	x	x
Naïve	x	---	x
Naïve	---	---	---
Intermediate	---	x	---
Intermediate	---	x	x
Advanced	---	---	x
Coding of interviews	*Sample answers*		
Naïve	“because the fertilizer made the plants look more beautiful” (production of effect)		
Intermediate	“because he [the scientist] tried it on so many plants” (reference to non-relevant design feature) “because lake water is often dirty, whereas not much rubbish is thrown into river water” (reference to mechanism)		
Advanced	“because he [the scientist] did not try both fertilizers” (correct identification of design error)		

2.1.4 Transcription and coding of children’s answers

Interviews were audiotaped, transcribed verbatim and coded by two independent raters (see Table 2 for coding examples and the MS coding). Using a strict criterion, the lowest level which children agreed to was taken as the final level in the coding of the MS item (e.g., if a child accepted the naïve and intermediate levels simultaneously, the MS item was coded as naïve). The interrater kappa reliability values for I-before and I-after were .94 and .88, respectively, for Item 1, and .73 and .90 for Item 2. Since some children gave invalid answers (especially on I-before; e.g., they refused to answer or only gave answers that were irrelevant to the question), some of the subsequent analyses involved a smaller sample.

2.2. Results and discussion

2.2.1 Core performance

Core performance data for Items 1 and 2 (Table 3) and a Wilcoxon signed-rank test of I-before revealed that, as expected, more children recognized the design error in the contrastive (*Mdn*=1) than the controlled-contrastive (*Mdn*=0) experiment, Z=2.92, p<.01, r=.30.

2.2.2 Comparison of the different formats

To investigate whether meaningful relations between the different formats in the two items existed (i.e., whether there was a high agreement in the levels assigned), we assessed Spearman correlations (*rho*) and, in addition, computed Wilcoxon signed-ranks statistics (*Z*) when non-significant correlations suggested low convergence, which may have been a result of over- or underestimation in one of the formats.

For Item 1 (Table 4), correlations between formats were significant for all comparisons, except for I-before and MC, where the correlation did not reach significance. However, as a Wilcoxon signed-ranks test revealed, there was no significant difference in difficulty of MC or I-before, and none of the formats led to a systematic over- or underestimation. Rather, 24% of all children performed better on MC than on I-before while 27% were assigned a higher level in I-before than in MC (cf. Table 4).

For Item 2 (Table 5), only the correlation between I-before and the MS task was significant. A Wilcoxon signed-ranks test revealed that the MC question (*Mdn*=1) significantly overestimated performance with respect to I-before. Similarly, MS was significantly more difficult than MC, just as I-before was more difficult than I-after, suggesting that presenting answer options significantly decreased the difficulty of identifying the design error in this second item with a more complex experimental design.

Table 3

Core performance data for Items 1 and 2 (contrastive and controlled-contrastive experiment)

Item / format	Naïve level	Intermediate level	Advanced level	Total
Item 1				
I-before	15 (33)	11 (24)	19 (42)	45 (100)
MS	62 (87)	1 (1)	8 (11)	71 (100)
MC	23 (32)	14 (20)	34 (48)	71 (100)
I-after	18 (26)	29 (41)	23 (33)	70 (100)
Item 2				
I-before	25 (51)	16 (33)	8 (16)	49 (100)
MS	53 (75)	13 (18)	5 (7)	71 (100)
MC	10 (14)	26 (37)	35 (49)	71 (100)
I-after	21 (30)	26 (37)	23 (33)	70 (100)

Notes. Data are *n* (%) values. I-before = open interview then MS item; I-after = MS item then open interview.

Table 4

Convergence between different item formats for Item 1 (contrastive experiment)

	Same level (a=b)				Overestimation (b>a)				Underestimation (b<a)						
Test pair (a/b)	Naïve-	Int.-	Adv.-	Total	Naïve-	Naïve-	Int.-	Total	Adv.-	Adv.-	Int.-	Total	*n*	*rho*	*Z*
	naïve	int.	adv.		int.	adv.	adv.		int.	naïve	naïve				
I-before & I-after	4 (9)	5 (11)	14 (31)	23 (51)	7 (16)	4 (9)	6 (13)	17 (38)	1 (2)	4 (9)	0 (0)	5 (11)	45 (100)	$.43^{**}$	---
I-before & MC	5 (11)	3 (7)	14 (31)	22 (49)	4 (9)	6 (13)	1 (2)	11 (24)	0 (0)	5 (11)	7 (16)	12 (27)	45 (100)	.24	-.08
I-before & MS	15 (33)	0 (0)	5 (11)	20 (44)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (31)	11 (24)	25 (55)	45 (100)	$.38^{*}$	---
MC & MS	23 (32)	1 (1)	8 (11)	32 (44)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	26 (36)	13 (18)	39 (54)	71 (100)	$.31^{**}$	---

Notes. Data are *n* (%) values. Int. = intermediate; adv. = advanced; *Z* = Wilcoxon signed-rank test. Over- and underestimation of the second comparison (b) relative to the first (a). *Rho* and *Z* are computed for children with complete observations on both measures for each test pair (i.e., children who did not give a valid answer on a measure that was part of the respective test pair were excluded from the respective analysis).

* *p*<.05. ** *p*<.01.

Table 5

Convergence between different item formats for Item 2 (controlled-contrastive experiment)

	Same level (a=b)				Overestimation (b>a)				Underestimation (b<a)						
Test pair (a/b)	Naïve-	Int.-	Adv.-	Total	Naïve-	Naïve-	Int.-	Total	Adv.-	Adv.-	Int.-	Total	*n*	*rho*	*Z*
	naïve	int.	adv.		int.	adv.	adv.		int.	naïve	naïve				
I-before & I-after	7 (14)	6 (12)	6 (12)	19 (38)	10 (20)	8 (16)	5 (10)	23 (46)	2 (4)	0 (0)	5 (10)	7 (14)	49 (100)	.23	3.29^{**1}
I-before & MC	4 (8)	7 (14)	6 (12)	17 (34)	9 (18)	12 (25)	7 (14)	28 (57)	2 (4)	0 (0)	2 (4)	4 (8)	49 (100)	.15	4.31^{***2}
I-before & MS	22 (45)	1 (2)	2 (4)	25 (51)	2 (4)	1 (2)	2 (4)	5 (10)	4 (8)	2 (4)	13 (27)	19 (39)	49 (100)	$.64^{***}$	---
MC & MS	9 (13)	5 (7)	5 (7)	19 (27)	1 (1)	0 (0)	0 (0)	1 (1)	7 (10)	23 (32)	21 (30)	51 (72)	71 (100)	.24	6.34^{***3}

Notes. Data are *n* (%) values.

** *p*<.01. *** *p*<.001 . ^{1} *r*=.29. ^{2} *r*=.39. ^{3} *r*=.53

Overall, MC items reliably identified advanced conceptions from I-before (14 out of 19 for Item 1, 6 out of 8 for Item 2) while MS identified naïve conceptions particularly well (15 out of 15 for Item 1, 22 out of 25 for Item 2). However, with respect to I-before, the MC item produced several false-positive advanced conceptions (i.e., many children were assigned the advanced level even if they still held naïve beliefs), while the MS pattern revealed a low sensitivity for intermediate- and advanced-level answers from I-before. Although this pattern may indicate a validity problem of MS, our interview data suggest that the low performance for MS is attributable to the difficulty children experience in overcoming naïve beliefs. For example, several children gave a correct answer for I-before, correctly identifying the design error in question (e.g., "It was not a good experiment because he didn't try the blue fertilizer"), but they still agreed with the naïve answer when MS answer options were presented (e.g., "It was good experiment because the scientist found that all plants got bigger blossoms"). Although this indicates that these children understand that the missing experimental contrast means that the hypothesis cannot be tested, they do not yet understand that hypothesis testing is the only purpose of experimentation and still uphold naïve beliefs regarding the production of effects.

Together these results indicate that the presentation of answer options (in MC) may lead to an overestimation of performance when the item difficulty is high (controlled-contrastive experiment) and children are allowed to choose the best answer. However, making children indicate whether or not they agree with each answer option (in MS) reveals potentially conflicting views in their thinking, and means that their competence is only confirmed once they have formed coherent advanced conceptions. Therefore, MS is superior to MC when investigating profound—as opposed to beginning—competencies.

3. Study 2

Study 2 investigated the reliability and the content and convergent validities of our 11-item inventory. The content validity was investigated by testing whether contrastive experiments (contrastive UNEX) are, as hypothesized, easier than controlled-contrastive experiments (controlled-contrastive UNEX), while convergent validity was assessed by comparing UNEX to the performance in two conventional CVS choice tasks. In addition, Study 2 assessed interindividual differences in the development of UNEX in primary school.

3.1 Methods

3.1.1 Participants

The 411 included primary-school children comprised 128 second-graders (mean age 8 years, 3 months; *SD*=7 months; 66 girls), 137 third-graders (mean age 9 years, 4 months; *SD*=7 months; 62 girls) and 146 fourth-graders (mean age 10 years, 2 months; *SD*=9 months; 78 girls). Children were recruited from 25 classrooms in predominantly middle-class schools in Germany. Parental informed consent was obtained for all children. Study 2 was part of a larger study of the cognitive development of primary-school children.

3.1.2 Materials

UNEX. The 11 UNEX items were designed analogously to the sample items presented in Study 1, and comprised 6 contrastive and 5 controlled-contrastive experiments (Appendix 1 provides the full item set). The contrastive experiments each included two experiments involving the manipulation of a single value of the focal variable, a missing control group and a missing reference to baseline. The controlled-contrastive experiments did not control non-focal variables.

CVS choice tasks. We adopted two tasks from the literature (Bullock & Ziegler, 1994; Chen & Klahr, 1999) that both presented children with three variables (each with two possible values) and a hypothesis regarding one of them. Our adaptation of Bullock and Ziegler's cars task involved the children testing the hypothesis that the speed of a car depends on the size of its tires (wide vs. narrow; non-focal variables: car and spoiler types). Our adaptation of Chen and Klahr's slopes task involved the children testing the hypothesis that the steepness of a slope influences the distance a marble travels after rolling down it (steep vs. flat; non-focal variables: size and position of the marble). All 8 (=2^3) variable combinations in each item were presented graphically, and the children were asked to select the 2 cars or slopes that they wanted to compare.

3.1.3 Procedure

All items were illustrated and presented in booklets that the children worked on individually in a whole-class testing procedure (taking approximately 30 minutes). Tasks and (verbal) answer options were read out by an experimenter and displayed in a PowerPoint presentation. Test assistants ensured that the children worked on their own booklets only, and assisted children who had any questions regarding the procedure.

3.2 Results and discussion

3.2.1 Core performance

Core performance data for all items (see Table 6) indicate a high frequency of naïve MS answers (all >60%), and lower frequencies of intermediate and advanced conceptions (all <20% and <25%, respectively).

3.2.2 Scale analysis

The reliability of the scale was determined by fitting a partial credit model (Masters, 1982) to the children's responses. This model assumes that developmental progression is unidirectional and that categories (i.e., naïve, intermediate and advanced) are hierarchical, reflecting the assumption of our conceptual-development model. All but one item (Item U5, a contrastive experiment) had a good fit to the model (i.e., 0.85> *infit* mean-square statistic [MNSQ] <1.15; see Table 6). Removing this poorly fitting item yielded a scale with an expected a posteriori estimate based on a plausible values (EAP/PV) reliability of .82 (weighted-likelihood estimator person separation reliability=.55, Cronbach's α=.85; scale mean=5.35, *SD*=5.85, minimum=0, maximum=22). All items satisfied the partial credit model's requirements of increasing point-biserial correlations and ability estimates per category. However, none of our items fulfilled a third criterion (ordered delta parameters). It is debated in the literature as to whether this is a mandatory requirement for model fit to hold (Adams, Wu, & Wilson, 2012), but the unordered delta parameters are consistent with children's low frequency of choosing the intermediate level. We therefore dichotomized the data by collapsing the naïve and intermediate levels. A Rasch model fitted to the resulting binary data revealed a good fit for all items (0.85< *infit* MNSQ <1.15) and an EAP/PV reliability of .71 (weighted-likelihood estimator person separation reliability=.62, Cronbach's α=.72). All subsequent analyses are based on these binary data.

Table 6

Percentage of answers on the naïve, intermediate and advanced level for the 11 UNEX items of Study 2 overall, and the item difficulty, discrimination and item fit.

Item	Experiment type	Level			Difficulty	Discri-	*Infit*
		Naïve	Inter.	Adv.			
		(0)	(1)	(2)		mination	MNSQ
U1	Contrastive	62.4	19.9	17.7	–.14	.57	1.10
U2	Contrastive	67.8	9.3	22.9	–.32	.66	1.03
U3	Contrastive	74.2	7.6	18.2	.20	.69	0.92
U4	Contrastive	73.2	4.4	22.4	–.14	.66	1.03
U5	Contrastive	65.2	10.4	24.3	–.48	.80	0.74[#]
U6	Contrastive	73.2	5.6	21.2	.01	.45	0.96
U7	Controlled-contrastive	73.7	15.1	11.2	.56	.43	1.04
U8	Controlled-contrastive	72.7	4.6	22.7	–.15	.63	1.05
U9	Controlled-contrastive	72.3	11.5	16.2	.69	.49	0.90
U10	Controlled-contrastive	68.1	12.9	19.9	–.04	.69	0.91
U11	Controlled-contrastive	74.0	5.6	20.44	–.06	.72	0.89

Notes. Difficulty, discrimination and *infit* mean-square statistic (MNSQ) are based on an analysis of the partial credit model; large negative item difficulties indicate easy items, while large positive values indicate difficult ones; *infit* MNSQ should be between .85 and 1.15 for an item to show a good fit to the partial credit model; [#] indicates an item with a poor fit according to *infit* MNSQ. Item numbering applies to Study 2 and the full item set given in the Appendix.

3.2.3 Developmental patterns and interindividual differences

A univariate analysis of variance revealed a significant main effect in UNEX for grade, $F(2,402)=21.75$, $p<.001$, partial $\eta^2=.10$. Interestingly, while fourth-graders recognized significantly more design errors than did third-graders, $F(1,402)=24.14$, $p<.001$, partial $\eta^2=.06$ (see Figure 2), there was no difference between grades 2 and 3, $F(1,402)=1.89$, $p>.05$. There were also no differences between boys and girls, $F(1,402)=3.18$, $p>.05$.

3.2.4 Content validity

An explanatory item response model (De Boeck & Wilson, 2004) with a fixed person effect (age) revealed that controlled-contrastive experiments were, as hypothesized, more difficult than contrastive experiments (see Table 7 for model fit, Table 8 for parameter estimates).

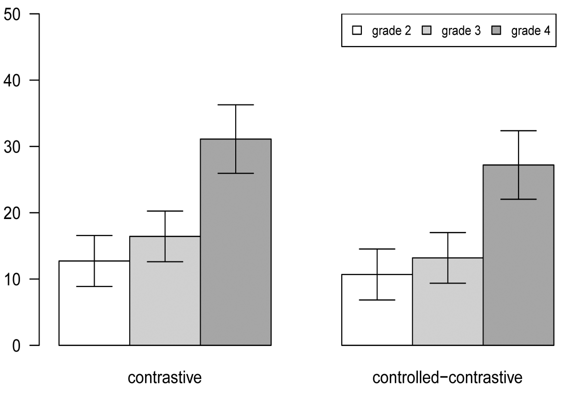


Figure 2. Percentage of correct answers for contrastive and controlled-contrastive experimental designs per grade. Error bars indicate 95% confidence intervals.

Table 7

Model comparisons for the explanatory item response model

Model	Effects (fixed)	Effects (random)	AIC	BIC	-2LL	df	LR test
M0 (1PL)	---	Intercept	3493.3	3506.1	3489.3		
M1	Experiment type	Intercept	3483.3	3502.6	3477.4	2	11.87***
M2	Experiment type + age	Intercept	3465.4	3491.0	3457.4	1	20.01***

Notes. M0 = reference model (1-parameter logistic model); AIC = Akaike information criterion; BIC = Bayesian information criterion; -2LL = deviance; df = degrees of freedom; LR test = Likelihood ratio test.
*** *p*<.001.

Table 8

Coefficients for M2

Model	B	SE	*p*
Intercept	–7.20	1.14	< .001
Age	.54	.12	< .001
Controlled-contrastive (reference: contrastive)	–.33	.01	< .001

3.2.5 Convergent validity

One-third (130) of the children chose CVS for both the cars and slopes tasks; however, only 55 children (14%) chose CVS consistently across the 2 items. The cars task was performed correctly by 20%, 33% and 44% of second-, third- and fourth-graders, respectively; the corresponding rates for the slopes task were 18%, 35% and 46%. The overall performance was thus slightly worse than that reported by Bullock and Ziegler (1999), where approx. 40% and 60% of third- and fourth-graders, respectively, correctly solved a CVS choice task. These differences are probably due to differences in sociodemographic characteristics between the previous urban sample and our more rural sample (cf. Koerber, Mayer, et al., 2015).

Interestingly, whereas the change-all strategy (vary all—including non-focal—variables) was the third most frequently chosen strategy for the cars task (20%), it was the least popular strategy for the slopes task (4%). In contrast to the cars task, where low performance resulted mostly from children disregarding the necessity of experimental control, errors in the slopes task were primarily due to children choosing an incorrect focal variable (i.e., size or position of the marble instead of slope steepness). While both these errors reflect unsuccessful coordination of hypothesis and evidence (cf. Kuhn, 2011), they suggest that the content domain of the specific task and children's knowledge thereof influence the expression of this interference between children's hypotheses and their construction of evidence.

A binomial regression revealed that performance in the cars task was significantly predicted by controlled-contrastive UNEX, $\chi^2(1)=4.67$, *p*=.03. Specifically, children with a more profound controlled-contrastive UNEX were more likely to use CVS than any other strategy, β=.13, *t*(1)=4.62, *p*=.03, odds ratio=1.13. However, contrastive-UNEX experiments did not predict strategy choice in the cars task, β=.02, *t*(1)=.11, *p*=.74. In contrast, the use of CVS in the slopes task was predicted by contrastive UNEX, $\chi^2(1)=8.16$, *p*=.004. Specifically, children with a high contrastive UNEX showed an increased use of CVS, β=.13, *t*(1)=8.02, *p*=.005, odds ratio=1.14, while there was no positive effect of controlled-contrastive

UNEX, β=–.08, $t(1)=2.03$, $p=.15$. This finding is consistent with the descriptive data for the slopes task, which suggest that few children choose an experiment in which all non-focal variables are varied. These results might be due to children only contrasting non-focal variables when their individual influences are unknown to them and they want to find out about them in a single test (cf. Schauble, 1990).

While 68% and 56% of the children with mastery in the controlled-contrastive UNEX (≥four items correctly solved) solved the cars and slopes tasks correctly, 70% and 68% of the incompetent children (<four items solved correctly) did not apply CVS to these tasks, $\chi^2(1)=16.90$ and $\chi^2(1)=6.43$, respectively; both $p<.05$. Conversely, around 30% of the children with no mastery in controlled-contrastive UNEX solved the cars or slopes task, suggesting that choice tasks lead to an overestimation of performance, even when the probability of correct guessing is small due to there being a large number of answer alternatives (28 in Study 2).

To summarize, these results support the reliability and the content and convergent validities of our inventory. In addition, the data show that children's UNEX is rudimentary at the beginning of primary school, but that it increases to a percentage of correct answers of around 30% by the end of primary school. This performance level is lower than that in choice tasks, which focus on children's beginning understanding of experimentation (e.g., 54% used CVS in an MC task by Koerber, Mayer, et al., 2015), but higher than that in discovery tasks (e.g., 17% initial use in Schauble, 1990).

4. General discussion

The findings of the two studies presented in this article suggest that our inventory of primary-school UNEX is a reliable measure of both content and convergent validities, and that MS is a well-suited and valid item format that yields a reliable performance estimate comparable to that obtained when using an interview measure. Our results further suggest that developmental progressions in UNEX take place in the late primary-school years, where a solid UNEX first emerges.

4.1 Validity of the item format

In contrast to conventional choice tasks in CVS, our inventory does not employ the more common MC item format, instead relying on MS, which requires children to independently agree or disagree with a set of statements (qualitatively different levels in the present studies). Study 1 revealed that this item format yields a reliable estimate of children's abilities that is comparable to an open-interview measure that is presented *before* exposure to the answer options (I-before). While Koerber, Osterhaus, and Sodian (2015) demonstrated a significant relation between children's MS choices in scientific thinking tasks and their performance in an interview measure that is conducted *after* exposure to the answer alternatives (I-after), the present studies are the first to validate the MS format against I-before. The significant relation between MS and I-before is an important result because our data indicate that revealing the answer alternatives of an item to children may lead to an overestimation of the performance for I-after, especially when the item content is difficult. As with the results of Koerber, Osterhaus, and Sodian (2015), Study 1 additionally corroborates the finding that the MS format is especially well suited for obtaining a reliable estimate of consolidated rather than beginning abilities.

4.2 Reliability and validity of the inventory

While CVS choice tasks are often strongly influenced by non-essential task characteristics, resulting in large performance discrepancies between items (Croker & Buchanan, 2011; see also the finding for the cars and slopes tasks in Study 2), the solving rates did not vary substantially between different items in our inventory, and a Rasch analysis indicated that the scale had a good reliability. This is an important finding since a solid and thoroughly tested instrument is needed for large-scale studies of the interindividual differences in scientific thinking and experimentation skills that already exist in primary school (Bullock et al., 2009). While inventories of primary-school children's general scientific thinking have recently been proposed and tested (Koerber, Mayer, et al., 2015), our inventory of primary-school UNEX is the first instrument that exclusively focuses on experimentation skills in early primary school; for example, Hammann, Phan, Ehmer, and Grimm (2008) provided an instrument that was only suited to grade 5 and above.

In addition to the reliability of our inventory, Study 2 revealed its content and convergent validities. The content validity was supported by the finding that contrastive experiments were—as predicted by our hypothesis—easier than controlled-contrastive experiments, which in addition to understanding experimental contrast requires an understanding of experimental control. The convergent validity was supported by our finding that UNEX predicted the performance in the cars and slopes tasks. However, while choosing CVS for the cars task was significantly related to controlled-contrastive UNEX, for the slopes task this choice was predicted by contrastive UNEX. This result shows a problem of conventional CVS choice tasks, and suggests that contrastive UNEX suffices for correctly solving tasks that include non-focal variables whose effects children hold strong beliefs about (e.g., how the size of a marble and its position on the slope influences the distance it travels). According to this interpretation, children only vary non-focal variables when their influences are unknown to them (e.g., influences of the car and spoiler types on car speed) and they want to uncover all individual effects in a single experimental test (cf. Schauble, 1990). Therefore, controlled-contrastive UNEX seems to be especially important when children perform experiments that involve non-focal variables whose effects are unknown to them.

4.3 Development of UNEX in primary school

Our theoretical, conceptual-development model, on which item construction was based, suggests that UNEX develops from naïve to intermediate to more advanced levels of understanding. However, while the partial credit model generally fitted the data, the descriptive statistics in Study 2 suggested that a small percentage of primary-school children perform at an intermediate level. This finding of an overall naïve classification is mostly due to many children selecting the naïve level even when they agreed with the intermediate level (cf. also Study 1). Therefore, although an important step towards a more mature UNEX is understanding that hypothesis testing is necessary and that there are features of experimental design that influence the quality of an experiment, this realization does not appear to necessarily overcome naïve beliefs.

A beginning understanding of UNEX first emerges late in primary school, where we found a correct performance rate of about 30%. This estimate, which lies between discovery (e.g., 17% initially using CVS in grades 5 and 6; Schauble, 1990) and choice (e.g., 54% in grade 4; Koerber, Mayer, et al., 2015) tasks, differs significantly from performance in early primary school (grades 2 and 3), where performance is still low, typically at 10–15%. Because experimentation is not explicitly taught in German primary schools, it is important to identify the mechanisms underlying this substantial development in children's UNEX during primary school. Potential mechanisms include children's metacognitive development (Kuhn, 2000; Lockl & Schneider, 2002) and increases in their executive control, which may allow children to inhibit an experiential processing of the experimental designs in favour of analytical processing (cf. Amsel et al., 2008; Klaczynski, 2000).

Future research needs to address these issues and reveal how development in UNEX can be understood and promoted in primary school. Our development of a well-tested instrument that is psychometrically sound and can be used in large-scale studies is an important prerequisite for these studies, and it should promote future work on this important component of scientific thinking.

Keypoints

- A novel 11-item inventory of primary-school children's understanding of experimental design yields a valid and reliable estimate of children's experimentation skills.
- The multiple-select item format serves as a strict criterion, revealing profound—rather than beginning—competencies.
- A coherent advanced understanding of experimentation emerges in primary school.

Acknowledgments

We are grateful to all research assistants for their help in the data collection, as well as to all teachers, children, and parents for their friendly collaboration and support of this research.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*, 547–573. doi:10.1177/00131644114321

Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., & Walker, R. (2008). A dual-process account of the development of scientific reasoning: The nature and development of metacognitive intercession skills. *Cognitive Development*, *23*, 452–471. doi:10.1016/j.cogdev.2008.09.002

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood. Findings from the Munich Longitudinal Study* (pp. 173–197). Mahwah, NJ: Erlbaum.

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

Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology*, *6*, 544–573. doi:10.1016/0010-0285(74)90025-5

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*, 1098–1120. doi:10.1111/1467-8624.00081

Croker, S., & Buchanan, H. (2011). Scientific reasoning in a real-world context: The effect of prior belief and outcome on children's hypothesis-testing strategies. *British Journal of Developmental Psychology*, *29*, 409–424. doi:10.1348/026151010X496906

De Boeck, P., & Wilson, M. (2004). *A framework for item response models*. New York, NY: Springer.

Hammann, M., Phan, T. T. H., Ehmer, M., & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education*, *42*, 66–72. doi:10.1080/00219266.2008.9656113

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York, NY: Basic Books.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development*, *71*, 1347–1366. doi:10.1111/1467-8624.00232

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. doi:10.1111/cdev.12298

Koerber, S., Osterhaus, C., & Sodian, B. (2015). Testing primary-school children's understanding of the nature of science. *British Journal of Developmental Psychology*, *33*, 57–72. doi:10.1111/bjdp.12067

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science*, *9*, 178–181. doi:10.1111/1467-8721.00088

Kuhn, D. (2011). What is scientific thinking and how does it develop? In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 472–523). Oxford, UK: Wiley.

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., & Carver, S. M. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, *60*(4), 1–128.

Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. In H. Reese (Ed.), *Advances in child development and behavior* (Vol. 17, pp. 2–44). New, NY: Academic Press. doi: 10.1016/S0065-2407(08)60356-0

Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgements. *International Journal of Behavioral Development*, *26*, 327–333. doi:10.1080/01650250143000210

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi: 10.1007/BF02296272

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning & Instruction, 29*, 43–55. doi:10.1016/j.learninstruc.2013.07.005

Nehm, R. H., & Schonfeld, I. R. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, *45*, 1131–1160. doi:10.1002/tea.20251

Piekny, J., & Maehler, 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*, 153–179. doi:10.1111/j.2044-835X.2012.02082.x

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, *49*, 31–57. doi:10.1016/0022-0965(90)90048-D

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology*, *11*, 401–402. doi:10.1037/h0076579

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, *51*, 1–10. doi:10.2307/1129583

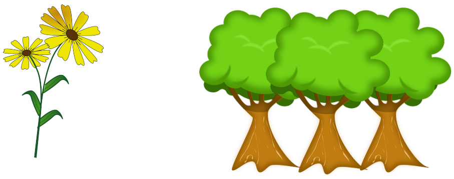
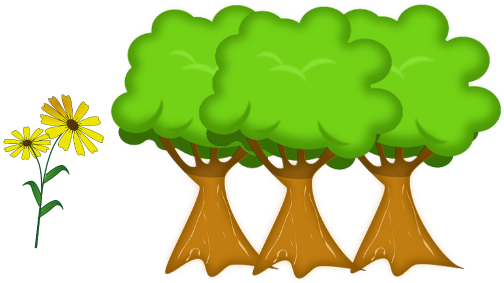
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172–223. doi:10.1016/j.dr.2006.12.001

Appendix 1

Full item set

The scientist …

Task 1: Trees (U1)

A scientist travels to a faraway planet, planet Ogi. There, he observes that all trees are very small.	
The scientist develops a tree medicine that is supposed to help the trees grow. He calls his this medicine Supergrow. The scientist wants to find out whether his Supergrow works and whether it really makes the trees grow. Therefore, he conducts an experiment. He gives his Supergrow to all trees on planet Ogi.	
Six months later, the scientist travels back to planet Ogi. He observes that all trees are huge now. He is convinced: “My Supergrow works!”	

Was this a good experiment?

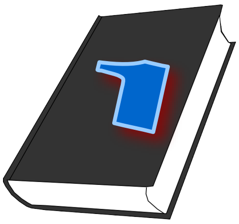
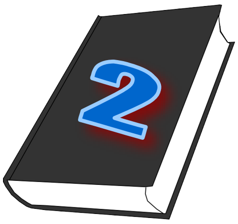
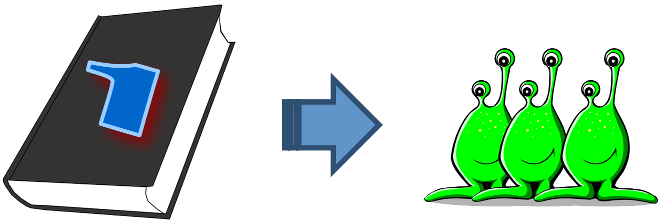
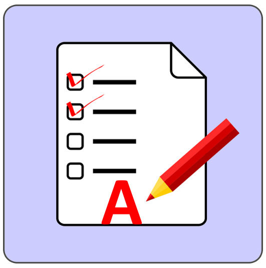
☐ **Yes**	☐ **No**

… and the trees

Susan, Lisa, and Vera wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Susan says*: "It was not a good experiment because he does not know how big the trees would have grown without his Supergrow."	□	□
2. *Lisa* says: "It was a good experiment because he found that all trees have grown huge after receiving his Supergrow."	□	□
3. *Vera* says: "It was a good experiment because you can only see whether things work if you test them."	□	□
Which of the three girls has the best answer?	**No.______**	

The math teacher…

Task 2: math textbook (U2)

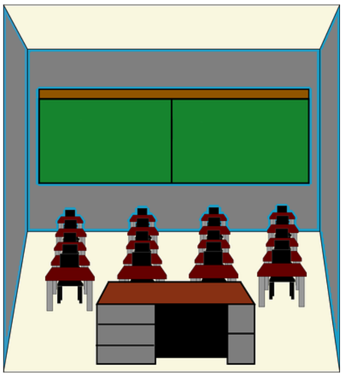
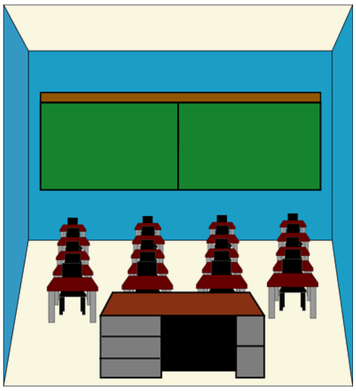
On planet Kroxon, all students learn mathematics at school.	
Their math teacher has two different textbooks. Textbook 1 and Textbook 2.	 Textbook 1  Textbook 2
He wants to find out whether his students learn faster with Textbook 1 or Textbook 2. He performs an experiment and teaches his students with Textbook 1 during 2 weeks.	 Textbook 1 2 weeks
After these 2 weeks, he gives his students a math exam. All students get an A. The math teacher is convinced: “My students learn faster with Textbook 1.”	
Was this a good experiment?	
□ **Yes**	□ **No**

… and his students

Paul, Mark, and Luke wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Paul says*: “It was a good experiment because all students got an A on the exam.”	□	□
2. *Mark says*: “It was not a good experiment because he did not try out Textbook 2.”	□	□
3. *Luke* says: “It was a good experiment because he tested Textbook 1 with many students.”	□	□
Which of the three boys has the best answer?	**No.______**	

The headmaster …

Task 3: Classroom (U3)

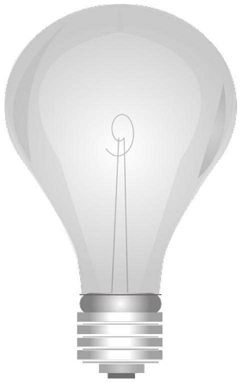
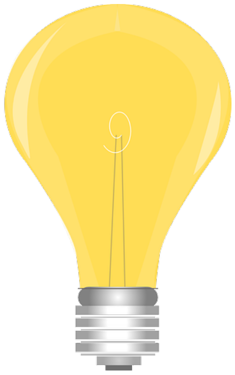
On planet Mawi, the Mawi children learn in grey classrooms. The Mawi children write an exam. All students get very bad grades.	
The principle of the school believes that his students all got very bad grades because they do not learn well in grey classrooms. Therefore, he performs an experiment and paints all classrooms blue.	
After 4 weeks, the Mawi children take the same exam again. Now all students obtain very good grades. The principle is convinced: “My students learn better in blue classrooms.”	
Was this a good experiment?	
□ **Yes**	□ **No**

… and the classroom

Susan, Lisa, and Vera wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Susan* says: “It was a good experiment because the principle painted all classrooms blue and not just one.”	□	□
2. *Lisa* says: “It was a good experiment because all students got better grades on the second exam.”	□	□
3. *Vera* says: “It was not a good experiment because the principle does not know how the students would have performed on the second exam if they had been taught in grey classrooms.”	□	□
Which of the three girls has the best answer?	**No.______**	

The scientist …

Task 4: Light bulb (U4)

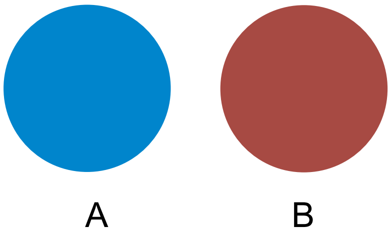
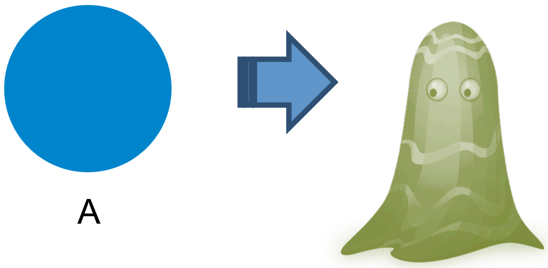
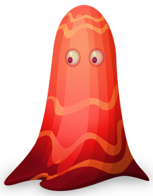
A scientist develops a special machine. He believes that this machine can make broken light bulbs glow again. He wants to find out whether the machine indeed works and therefore, he conducts an experiment.	
He travels to planet Bulby and tells the Bulbians: “Bring me 50 light bulbs. I want to show you something.”	
The Bulbians bring him 50 light bulbs. The scientist puts them in the machine. Then he starts the machine.	
The Bulbians take the light bulbs, bring them to their homes, and try them out. All light bulbs work. Now the scientist is convinced: “My machine works.”	
Was this a good experiment?	
□ **Yes**	□ **No**

… and the machine

Susan, Lisa, and Vera wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Susan* says: "It was a good experiment because all light bulbs work again after they have been in the machine."	□	□
2. *Lisa* says: "It was a good experiment because he put many light bulbs in the machine."	□	□
3. *Vera* says: "It was not a good experiment because he does not know whether or not the light bulbs worked before he put them in the machine."	□	□
Which of the three girls has the best answer?	**No.______**	

The scientist …

Task 5: Krogi (U5)

A scientist travels to planet Krogi. There he meets the Krogians. He observes that the Krogians are all very sick and very pale.	
The scientist wants to help the Krogians. He knows that there are two medicines that might heal the disease of the Krogians: Medicine A and Medicine B.	
The scientist wants to find out which medicine works better: Medicine A or Medicine B? Therefore, he conducts an experiment. He gives Medicine A to 100 Krogians.	
After a week, he observes that all Krogians have recovered from their disease. The scientist is convinced: “Medicine A works better!”	
Was this a good experiment?	
□ **Yes**	□ **No**

… and the Krogians

Chris, Mark, and Luke wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Chris* says: “It was a good experiment because all Krogians who have received Medicine A have recovered.”	□	□
2. *Mark* says: “It was not a good experiment because the scientist does not know how well Medicine B works.”	□	□
3. *Luke* says: “It was a good experiment because you find out about things when you test them.”	□	□
Which of the three boys has the best answer?	**No.______**	

The hair dresser …

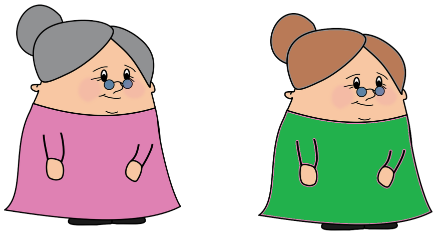
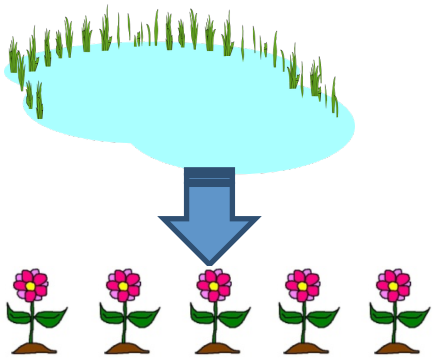
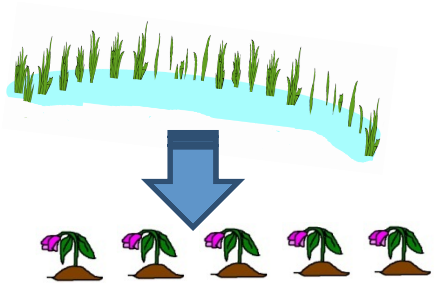
Task 6: Hair dresser (U6)

Mr. Tousle-Head is a hairdresser. He has developed a shampoo that he believes to help against hair loss.	
He performs an experiment to find out whether the shampoo truly works. He gives his shampoo to 50 people. During an entire year, they use his shampoo always when they wash their hair.	
After a year, Mr. Tousle-Head asks all 50 people whether or not they have hair loss. Nobody has hair loss. Mr. Tousle-Head is convinced: “My shampoo works just fine!”	
Was this a good experiment?	
□ **Yes**	□ **No**

… and his shampoo

Susan, Lisa, and Vera wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Susan* says: “It was a good experiment because nobody had hair loss after a year.”	□	□
2. *Lisa* says: “It was a good experiment because he investigated many people.”	□	□
3. *Vera* says: “It was not a good experiment because he does not know whether or not the people had hair loss in the first place.”	□	□
Which of the three girls has the best answer?	**No._______**	

Task 7: Grannies (U7)

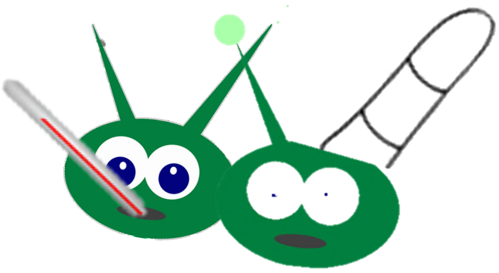
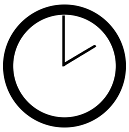
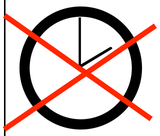
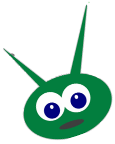
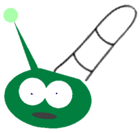
Granny Bubu and Granny Kiki live on planet Iber. Both grannies love flowers. They want to find out what makes their flowers bloom beautifully. Therefore, they conduct an experiment.	 Granny Bubu Granny Kiki
Granny Bubu waters her flowers with lake water. Granny Bubu's flowers bloom beautifully.	
Granny Kiki waters her flowers with river water. Granny Kiki's flowers are withered.	
Granny Bubu is convinced "It's the lake water that makes my flowers bloom so beautifully."	
Was this a good experiment?	
□ **Yes**	□ **No**

… and their flowers

Nick, Thomas, and Simon wonder whether Granny Bubu and Granny Kiki conducted a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Nick* says: “It was a good experiment because all flowers that have been watered with lake water bloom beautifully.”	□	□
2. *Thomas* says: “It was not a good experiment because the grannies do not know whether they do anything else differently, apart from the water they use .”	□	□
3. *Simon* says: “It was a good experiment because you can only find out how good something works if you try it out.”	□	□
Which of the three boys has the best answer?	**No.______**	

The hospital …

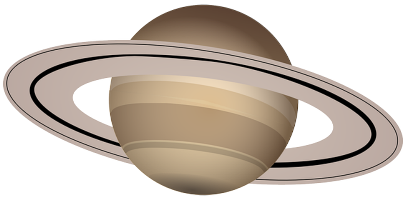
Task 8: The Molans (U8)

On planet Mola, many Molans are sick. In the hospital, a scientist wants to find out whether the Molans recover faster if they are allowed to receive visitors.	
He conducts an experiment: 20 Molans who are suffering from heat disease are allowed to receive visitors for 2 hours per day during 2 weeks.	 2 visiting hours
20 Molans who suffer from a broken antenna are not allowed to receive visitors during these 2 weeks.	 no visiting hours
He compares both groups and finds out: After 2 weeks, all Molans who suffered from heat disease have recovered.	
All Molans who suffer from a broken antenna and who have not received visitors are still sick after these 2 weeks.	
The scientist is convinced: “It depends on the visiting hours whether or not the Molans recover fast.”	
Was this a good experiment?	
□ **Yes**	□ **No**

… on planet "Mola"

Susan, Lisa, and Vera wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Susi* says: "It was not a good experiment because the scientist should have compared 2 groups of Molans with the same disease."	□	□
2. *Lisa* says: "It was a good experiment because he investigated many Molans."	□	□
3. *Vera* says: "It was a good experiment because 20 Molans have recovered from their disease."	□	□
Which of the three girls has the best answer?	**No.______**	

Task 9: Ms Mef (U9)

Mrs. Mef lives on planet Ballinki with her 6 children.	
Mrs. Mef owns an old family recipe for a juice that she believes to be good for children's teeth. She wants to find out whether the juice indeed works, and therefore, she conducts an experiment. Three of her children drink the special juice every night. They have good and healthy teeth. Three of her children never drink the special juice. They have bad and ill teeth.	
Mrs. Mef is convinced: "The special juice makes your teeth healthy!"	

Was this a good experiment?

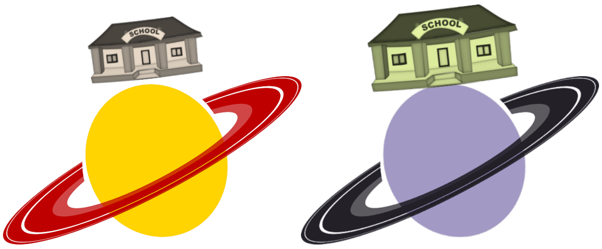
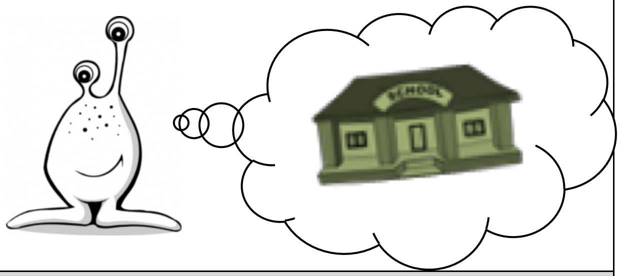
☐ **Yes**	☐ **No**

… and her family recipe

Susan, Lisa, and Vera wonder whether this was a good experiment.		
Who is right and who is not?		
	is right	**is not right**
1. *Susan* says: “It was a good experiment because all children who drank the juice have good teeth.”	□	□
2. *Lisa* says: “It was not a good experiment because Mrs. Mef does not know which other foods or drinks the children consumed.”	□	□
3. *Vera* says: “It was a good experiment because Mrs. Mef can only find out whether her family recipe works if she tries it out.”	□	□
Which of the three girls has the best answer?	**No._______**	

The principle…

Task 10: The principle (U10)

On planet Iber, all Iber children have to go to school. However, they all get very bad grades. The principle of their school has three different ideas about why his students get such bad grades:	
1. Iber teachers are too strict. 2. Iber children do not work hard enough. 3. The Iber school is too small and the children cannot concentrate in the small rooms.	
The principle makes a scientific study. He takes a look on the neighboring planet where children get very good grades to find out how large the schools are over there. Indeed, schools are much larger on the neighboring planet.	
The principle is convinced: “The Iber children get bad grades because the school is too small.”	
Was this a good study?	
□ **Yes**	□ **No**

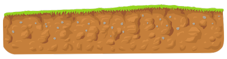
… and his school

Nick, Thomas, and Simon wonder whether this was a good experiment.

Who is right and who is not?		
	is right	**is not right**
1. *Nick* says: “It was a good study because the principle can only find out if he thinks hard about it and checks his ideas.”	□	□
2. *Thomas* says: “It was not a good study because it might also be because of the strict Iber teachers or the lazy Iber children that students on planet Iber get bad grades.”	□	□
3. *Simon* says: “It was a good study because the children on the neighboring planet get good grades.”	□	□
Which of the three boys has the best answer?	**No._______**	

The flowers …

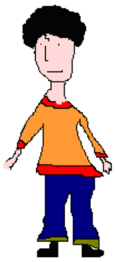
Task 11: Flowers (U11)

All flowers on planet Blossom are sick and withering. A scientist thinks that there may be three reasons for why plants are getting sick:	
1. They do not get sufficient sunlight. 2. There is too much wind. 3. The soil is not fertile.	
The scientist conducts an experiment. He takes 50 flowers and brings them to a different spot where they receive much more sun light. Indeed, here the flowers are not getting sick any more.	
The scientist is convinced: “The reason for why the flowers got sick is that they did not get sufficient sunlight.”	
Was this a good experiment?	
□ **Yes**	□ **No**

… on planet Blossom

Nick, Thomas, and Simon wonder whether this was a good experiment.

Who is right and who is not?

	is right	**is not right**
1. *Nick* says: “It was a good experiment because he brought many plants to a different spot and observed them.”	□	□
 2. *Thomas* says: “It was not a good experiment because he does not know whether wind and soil would have made a difference.”	□	□
3. *Simon* says: “It was a good experiment because the flowers are not sick anymore.”	□	□
Which of the three boys has the best answer?	**No.______**	