Educational Research for Policy and Practice
https://doi.org/10.1007/s10671-023-09332-9

ORIGINAL ARTICLE

Measuring scientific reasoning in kindergarten and elementary school: validating the Chinese version of the Science-K Inventory

Christopher Osterhaus[1] · Xiya Lin[2] · Susanne Koerber[3]

Received: 14 December 2021 / Accepted: 17 January 2023

Abstract
Scientific reasoning is a twenty-first century skill that is important for economic growth and social prosperity. A growing body of research documents that basic scientific reasoning skills develop much earlier than initially assumed, with many young elementary school and even kindergarten-aged children showing emergent scientific reasoning skills. Many studies on early scientific reasoning have been conducted in Western countries, and there is a lack of validated instruments that can be used in cross-cultural work. The present paper reports on the findings of a study assessing the psychometric properties of the shortened Chinese version of the Science-K(indergarten) Inventory (SC-SKI). The SC-SKI consists of 10 items that assess children's understanding of the nature of science, as well as their experimentation and data interpretation skills. Sixty-nine 6- to 7-year-olds from urban and rural schools in the Hunan province (China) participated in the study. The results showed an acceptable reliability of the SC-SKI (McDonald's $\omega_t = 0.60$). The ability estimates obtained for children's scientific reasoning (average performance was 47.5% correct) were comparable to those measured in German 6-year-olds (45.1% correct), and the urban sample outperformed the rural sample, supporting the ability of the SC-SKI to detect expected performance differences in young children's scientific reasoning. A significant correlation between scientific reasoning and language skills ($r = 0.54$, $p < 0.05$) confirms earlier findings and indicates construct validity. Taken together, the present study shows that the SC-SKI is a reliable and valid instrument that can be used to measure scientific reasoning in Chinese-speaking 6- to 7-year-olds.

Keywords Scientific reasoning · Measurement · Science-K Inventory (SKI) · Chinese

✉ Christopher Osterhaus
christopher.osterhaus@uni-vechta.de

[1] Developmental Psychology in Education, Faculty of Education and Social Sciences, University of Vechta, Driverstr. 22, 49377 Vechta, Germany

[2] Ludwig-Maximilians-Universität München, Munich, Germany

[3] Freiburg University of Education, Freiburg, Germany

Published online: 11 February 2023

1 Introduction

Scientific reasoning is defined as intentional knowledge seeking (Kuhn, 2002), and it comprises various components, such as experimentation skills (how to design an informative experiment?), data interpretation skills (how to make sense of patterns of covariation data or confounded data?), or nature of science understanding (what is it that scientists do and what kind of questions do they ask?) (Koerber et al., 2005). Scientific reasoning is an important twenty-first century skill (Trilling & Fadel, 2009). In modern knowledge societies, mature scientific reasoning skills are necessary in many professional occupations, and also, they allow citizens to make well-informed decisions with respect to socio-scientific issues, such as climate change or health crises (Ratcliffe & Grace, 2003; Sadler, 2004). While early developmental work on children's and adolescents' scientific reasoning focused on students' shortcomings (Inhelder & Piaget, 1958), there is a growing body of research that shows that children as young as 6-year-olds perform better than chance on many scientific reasoning tasks (Koerber & Osterhaus, 2019, 2021).

In early elementary school, for instance, children reliably differentiate a conclusive from an inconclusive test of a hypothesis (Sodian et al., 1991). Also, they can reason about the informativeness of different kinds of evidence (Köksal et al., 2021). In kindergarten (children aged 4 to 6 years), children reveal an emergent understanding of the nature of science (Samarapungavan et al., 2008) and they successfully select informative interventions that allow to draw appropriate causal inferences (Lapidow & Walker, 2020). Kindergarteners also select unconfounded experiments when they are presented with conflicting evidence (van Schijndel et al., 2015), and 4-year-olds show a rudimentary command of the control-of-variables strategy (i.e., vary one variable at a time while keeping all others constant) (van der Graaf et al., 2015). The ability to successfully interpret data also emerges in kindergarten when children begin to draw correct inferences from simple covariation data, such as perfect and unconfounded patterns of covariation data (Koerber et al., 2005; Piekny & Maehler, 2013).

Most studies of early scientific reasoning are conducted with Western samples, and studies in Asian countries are rare. The rationale of the present study is therefore to investigate the reliability and convergent validity of a scientific reasoning inventory that was originally developed in Germany, the Science-K Inventory (SKI) (Koerber & Osterhaus, 2019), when applied to a Chinese sample of 6- and 7-year-olds. In addition, we compare the performance of an urban Chinese sample with the performance of a rural Chinese sample, we ask whether there are gender differences in the scientific reasoning of Chinese 6- and 7-year-olds, and we investigate whether subcomponents of scientific reasoning are related.

The Science-K Inventory (SKI) is a closed-response instrument that comprises 30 items on children's experimentation skills, data interpretation, and nature of science (NoS) understanding. Items on experimentation skills tap children's ability to differentiate a conclusive from an inconclusive test of a hypothesis (Sodian et al., 1991), as well as their mastery of the control-of-variables strategy (i.e., vary one thing at a time, keep nonfocal variables constant). Children who can differentiate conclusive from inconclusive evidence, for instance, understand that—when trying to find out if someone is good at doing puzzles—this person should piece together a puzzle with many (conclusive evidence) and not just few (inconclusive evidence) pieces. Similarly, when trying to find out if cacao powder dissolves better in warm or cold milk, children with a command of the control-of-variables strategy understand that one should compare how well it dissolves in equal serves of warm and cold milk, rather than comparing how well it dissolves in a large glass of warm milk and a small glass of cold milk. Items on data interpretation measure children's ability to make sense of simple patterns of covariation data and to understand that one cannot draw valid inferences with respect to a single variable from confounded data. For instance, a runner could observe that she runs faster when wearing her new running shoes and her new running suit. Children who understand that confounded data do not allow to draw valid inferences understand that this pattern of data will not allow to decide whether it is the shoes or the suit that influences how fast the runner runs. Finally, NoS items tap children's understanding of what scientists do (they try to find out something about the world) and which types of questions they ask (questions whose answers provide explanations).

The SKI was validated in a German study. Koerber and Osterhaus (2019) used the SKI in a study in kindergarten and applied it to 227 six-year-olds. The administration of the SKI was completed during three individual interview sessions, lasting each approx. 20 min. During these interviews, trained researchers guided the children through all multiple-choice questions and recorded their answers. A scale analysis of the data obtained in this study showed that the SKI is a reliable instrument (Cronbach's $\alpha = 0.78$), and 6-year-olds performed significantly better than chance, with a mean correct performance of 42.5% correct (Koerber & Osterhaus, 2019). In line with the many studies that have associated scientific reasoning with children's language skills (Koerber et al., 2017; Osterhaus et al., 2017; van de Sande et al., 2019; van der Graaf et al., 2018), also performance on the SKI was related to language skills, with correlation coefficients across studies ranging between 0.36 (Koerber & Osterhaus, 2021) and 0.41 (Koerber & Osterhaus, 2019).

1.1 This study

In the present study, we investigate the reliability and convergent validity of a shortened 10-item scale of the SKI that was translated into Mandarin. The 10 items of this shortened Chinese version of the Science-K Inventory (SC-SKI) were selected based on data from the Koerber and Osterhaus (2019) study that was conducted in Germany. In particular, we selected an item pool that would cover the broad aspects of scientific reasoning while simultaneously resulting in a reliable and balanced scale. To address the convergent validity of this shortened Chinese scale, we measured children's language skills (i.e., their vocabulary understanding). Previous work (e.g., Koerber et al., 2017; Mayer et al., 2014; Osterhaus et al., 2017; van de Sande et al., 2019; van der Graaf et al., 2018) has revealed substantial associations between scientific reasoning and young children's language skills, which are well expected and indicative of the broad association between scientific reasoning and (verbal) reasoning in general. Among children's language skills, vocabulary understanding may be a particularly relevant aspect, especially for NoS. NoS requires that children understand science-specific terminology, including an understanding of what it means to 'investigate' something or to 'make an assumption' (Osterhaus et al., 2017).

To investigate the ability of the SC-SKI to detect expected performance differences, we compared the performance of an urban sample to a rural sample from Hunan Province, China. Because of the rural–urban disparity in schooling that prevails in China (e.g., Zhang, 2017), we reasoned that the children from urban areas should outperform children from more rural areas, which is a finding that, if it holds, will lend support to the usefulness of the SC-SKI and its ability to detect meaningful individual differences.

In addition, we assessed whether there are gender differences in this sample of young Chinese elementary school students. Some previous work (e.g., Lazonder et al., 2020) has observed such differences in older elementary school children, with males outperforming females, whereas other studies did not find such differences in elementary school (e.g., Osterhaus et al., 2017). Previous work has also identified significant associations between scientific reasoning subcomponents. For instance, Osterhaus et al. (2017) report a significant factor correlation of 0.54 for elementary school children's experimentation skills and their NoS. In the present study, we therefore assess the correlation between subcomponents, asking whether significant associations emerge in this sample of Chinese elementary school children.

1.2 Aims and objectives

The present study had five main aims: (1) to address the reliability of the SC-SKI, (2) to show its convergent validity by investigating the association between children's performance on the SC-SKI and their language skills, (3) to investigate differences in performance between an urban and a rural Chinese sample, (4) to investigate whether there are gender differences in the scientific reasoning performance of males and females, and (5) to investigate whether there are significant associations between scientific reasoning subcomponents.

2 Methods

2.1 Participants

The participants were 69 first-year elementary school students (31 females, 38 males) from two schools in the Hunan Province, China: one located in an urban area ($n = 53$) and one located in a rural area ($n = 16$). All children were aged 6 to 7 years; there were 43 six-year-olds and 26 seven-year-olds ($M = 6.38$, $SD = 0.49$). All children were at the end of their first semester. First-year curricula were similar across schools, and the children received instruction in Chinese, mathematics, and English education. Science-related courses were not offered at neither school. Parental and teachers' informed consent and child assent were obtained for all participants.

2.2 Materials

Scientific reasoning Scientific reasoning was assessed using ten closed-response items from the shortened Chinese version of the Science–K(indergarten) Inventory (SC-SKI) (Koerber & Osterhaus, 2019). The items were chosen based on pilot studies from Germany (Koerber & Osterhaus, 2019, 2021), and they were translated into Mandarin (see Appendix). The SKI was translated by one of the researchers; a back-translation was done to confirm the accuracy of the translation.

The full version of the SKI (Koerber & Osterhaus, 2019) is a 30-item instrument developed to assess emerging scientific reasoning abilities in kindergarten and early elementary school. The items are administered in individual interviews, and the children are assessed for their abilities in experimentation and data interpretation, and their understanding of the nature of science. The shortened Chinese SC-SKI comprises 4 items on experimentation, 3 items on data interpretation, and 3 items on nature of science understanding. Items on experimentation assessed children's ability to differentiate a conclusive from an inconclusive test (items Exp-1 and Exp-2), as well as children's understanding of the control-of-variables strategy (items Exp-3 and Exp-4). Items on data interpretation tapped children's ability to understand that confounded data patterns do not allow to draw conclusions with respect to a hypothesis (items Dat-1 to Dat-3). And items on nature of science understanding tested children's understanding of what scientists do (item NoS-1) and what kinds of questions they ask (items NoS-2 and NoS-3). All items were presented with three answer options, and no corrective feedback was given.

Full credit (1 point) was given when the children selected the correct answer. All other (wrong) answers were awarded with 0 points.

Language (vocabulary) Children's language skills were assessed using 10 items from the vocabulary test of the Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 2003) that were translated by the researchers: shoe, bike, hat, nail, gasoline, donkey, seesaw, to participate, diamond, to hate. The researchers read out the word to the children and asked them to explain what they meant (e.g., 'Could you explain the word shoe to me?' or 'What is a shoe?'). When the children were hesitant to respond or when they gave an incomplete answer, the experimenters would repeat the question or elaborate and ask children to tell them more about the word (e.g., 'Please tell me more about a shoe.'). All sessions were recorded for subsequent coding.

For nouns, full credit (2 points) was given if the children either provided a correct synonym, a main function of the object, the main characteristic of the object (or several characteristics of the object), or a correct classification that is stated in the Xinhua Dictionary (11th edition). For verbs, full credit (2 points) was given if the children provided a precise description of the activity. Partial credit (1 point) was given if the children provided a correct but simple explanation or a synonym that was not the same as the word given (e.g., poultry-pigeon), if they explained it using an unconventional function (e.g., knife-to kill a person), provided a secondary character, mentioned the word when trying to explain the word, or used an action to represent the word. No credit (0 points) was given for answers that simply restated the question or that were wrong. Children would also receive 0 points when they spoke dialect. The test was discontinued after the children had answered 5 consecutive items wrong. The composite scores reported are the average of the total possible.

2.3 Procedure

A trained researcher conducted the individual interviews. Interviews were conducted online; data collection took place between December 1, 2020, and January 1, 2021. Scientific reasoning and language skills were assessed during two separate sessions. In the urban school, the two sessions took place on two separate days; in the rural school, both sessions were conducted on the same day. Researchers did not provide corrective feedback during or after a session. The order in which the three components of scientific reasoning (experimentation, data interpretation, and nature of science understanding) were assessed was counterbalanced across participants.

2.4 Analysis plan

The data were analyzed using IBM® SPSS® Statistics version 27 and R 4.0.4. Average scientific reasoning performance was computed as percent correct (on the SC-SKI), and we used a *t*-test to test whether children's average performance significantly differed from chance (here 33.3%). Item difficulty and discrimination were calculated using the 'sjPlot' package for R (Lüdecke, 2021), and reliability (McDonald's ω_t) was computed using the 'psych' package for R (Revelle, 2022). In particular, a factor analysis was performed and omega total was computed for the general factor, as well as for three subfactors. Performance differences between groups (males vs females, urban vs rural sample) were assessed using *t*-tests (SPSS); correlations were calculated based on composited scores (SPSS). Correlations between subcomponents of scientific reasoning (i.e., experimentation and data interpretation skills, NoS) were computed based on composite scores.

3 Results

3.1 Core ability

The core performance data are given in Table 1. Dat-1 and Dat-2 (interpreting confounded data), as well as NoS-3 (what questions do scientists ask?) were difficult, and none of the children in the urban or rural sample achieved an average score of > 25%

Table 1 Percent Correct per Item in the Urban and Rural Samples

Item	Aspect	Urban		Rural		Diff.	Discr.
		M	*SD*	*M*	*SD*		
Exp-1	Conclusive experiment	52.8	50.4	37.5	50.0	0.49	0.36
Exp-2	Conclusive experiment	43.4	50.0	12.5	34.2	0.36	0.38
Exp-3	Control of variables	88.7	32.0	12.5	34.2	0.71	0.42
Exp-4	Control of variables	84.9	36.1	37.5	50.0	0.74	0.12
Dat-1	Confounded data	20.8	40.9	6.3	25.0	0.17	0.09
Dat-2	Confounded data	15.1	36.1	6.3	25.0	0.13	0.19
Dat-3	Confounded data	49.1	50.5	25.0	44.7	0.43	0.06
NoS-1	What do scientists do?	94.3	23.3	81.3	40.3	0.91	0.21
NoS-2	Which questions do they ask?	69.8	46.3	43.8	51.2	0.64	– .02
NoS-3	Which questions do they ask?	15.1	36.1	18.8	40.3	0.16	0.08

correct. The mean performance (in percent correct) across all items was 53.4% ($SD = 30.3$) in the urban school; it was 28.0% ($SD = 23.1$) in the rural school. In the urban school, more than 75% of the children gave a correct answer to items Exp-7 and Exp-8 (both assessing children's understanding of the control-of-variables strategy) and item NoS-1 (what do scientists do?). In the rural school, more than 75% of the children gave a correct answer to only item NoS-1. Average performance across groups differed from chance guessing [$t(68) = 6.653$, $p < 0.05$]. However, this was not true for the performance of the children from the rural area, who did not perform significantly better than expected based on guessing [$t(15) = -1.559$, $p > 0.05$]. Descriptively, there was a gender difference, with boys ($M = 49.0\%$, $SD = 31.3$) achieving a higher average performance than girls ($M = 46.5\%$, $SD = 25.1$). However, this descriptive difference was nonsignificant [$t(9) = 0.663$, $p > 0.05$], which is a finding that is in line with prior work showing no gender differences in early scientific reasoning (Koerber & Osterhaus, 2019; Koerber et al., 2015). The reliability of the SC-SKI was good, with McDonald's $\omega_t = 0.60$ for the entire test, and ω_t being 0.43, 0.56, and 0.62 for the three factors. Item discrimination and difficulty indices are given in Table 1.

The average mean score (on a scale from 0 to 2 points) for the vocabulary test was 1.439 ($SD = 0.488$). There was no difference in performance between boys ($M = 1.444$, $SD = 0.19$) and girls ($M = 1.44$, $SD = 0.13$) [$t(9) = -0.045$, $p > 0.05$). Children from the urban sample ($M = 1.64$, $SD = 0.10$) outperformed children from the rural area ($M = 0.77$, $SD = 0.34$), with children from the rural area providing less accurate explanations [$t(9) = 10.214$, $p < 0.05$].

3.2 Correlational analysis

The correlation between the composite scientific reasoning score and vocabulary score was significant and of substantial magnitude, with $r = 0.54$, $p < 0.05$. However, not all three components of scientific reasoning were significantly correlated with language skills (see Table 2). The correlation between nature of science understanding and language skills was insignificant ($r = 0.11$, $p = 0.4$), as was the correlation between data interpretation and experimentation skills, and data interpretation and nature of science understanding ($r = 0.15$, $p = 0.20$, and $r = 0.02$, $p = 0.80$, respectively).

Table 2 Correlation between Scientific Reasoning Components and Language Skills

	Language skills	Data interpretation	Experimentation	NoS
Language skills	–			
Data interpretation	0.35*	–		
Experimentation	0.51*	0.15	–	
NoS	0.11	0.02	0.25*	–

*$p < .05$

4 Discussion

What are the psychometric properties of the SC-SKI, a 10-item scientific reasoning test for Mandarin-speaking children in early elementary school? That was the main question of the present study that found that the SC-SKI reveals a good reliability, as well as convergent validity and the ability to detect expected performance differences between young children in an urban and a rural sample.

To investigate convergent validity, we studied the association between children's performance on the SC-SKI and language skills, which are two constructs that have been firmly associated across many studies (Koerber et al., 2017; Mayer et al., 2014; Osterhaus et al., 2017; van de Sande et al., 2019; van der Graaf et al., 2018). In line with earlier findings with 6-year-olds that documented correlation coefficients between 0.36 (Koerber & Osterhaus, 2021) and 0.41 (Koerber & Osterhaus, 2019), we found a correlation of 0.54 between the SC-SKI and a Mandarin vocabulary test. Finding this well expected association, which points to the close association between scientific reasoning and general (verbal) reasoning, as well as to the need for children to master science-specific vocabulary, is evidence of the convergent validity of the SC-SKI.

To investigate whether the SC-SKI is able to detect meaningful and expected performance differences between a sample of young children from an urban sample and those from a rural sample, we compared the performances of these two samples. Because of the disparities in school performance between urban and rural areas in China (e.g., Zhang, 2017), we reasoned that the children from the urban school should reveal a better performance than children from the rural area if the instrument was valid. And this was indeed the case: Children from the urban area performed significantly better than children from the rural area, whose performance did not exceed chance level. This finding supports the usefulness of the SC-SKI and its ability to detect meaningful individual differences, and at the same time, it shows the importance of fostering scientific reasoning skills from early on, which is likely to happen more frequently in the urban than rural school.

The present study also identified some areas of improvement for the SC-SKI. In particular, item discrimination indices were low for two data interpretation and two NoS items, suggesting necessary improvements to some of the items included in the SC-SKI. It is worth noting, however, that item discrimination was overall good, and in particular for items assessing children's experimentation skills. A potential explanation for the poor discrimination of some of the data interpretation items, which assessed children's ability to recognize confounded data and to understand that no conclusions can be drawn from this type of data, lies in their relative difficulty: Especially in the rural sample, only few children (< 10%) solved these items correctly. This finding is in line with prior research, showing that this particular aspect of data interpretation is rather challenging for young children (Osterhaus et al., 2020) and substantially more difficult than the interpretation of simple and conclusive patterns of covariation data (see Koerber et al., 2005; Piekny & Maehler, 2013).

Although the reliability of the SC-SKI was good, significant correlations did not emerge between all scientific reasoning subcomponents (i.e., experimentation and data interpretation skills, NoS). While we found a significant correlation between experimentation skills and NoS (0.25), children's data interpretation skills were uncorrelated with any other scientific reasoning subcomponent. The significant association between experimentation and NoS confirms earlier findings from studies with German elementary school children (Osterhaus et al., 2017), which showed a substantial association between these constructs. The finding of the present study that children's data interpretation skills were uncorrelated with any other scientific reasoning subcomponent can best be explained by the floor effect in data interpretation skills: Many children struggled with these items and especially the rural sample performed poorly. Future studies that address the relation between subcomponents of scientific reasoning should therefore include children of a more-diverse ability spectrum, assessing the scientific reasoning skills of older children who have already developed more-profound data interpretation skills.

For educators, these findings have several implications: First, when educators want to foster children's scientific reasoning skills or make use of them in the classroom (e.g., in inquiry-based learning), they need to make sure that they select the subcomponents that are appropriate given the children's ability level. In particular, data interpretation skills (understanding that confounded data patterns make it impossible to draw valid inferences) seem hard to acquire for young children, and hence elementary school teachers should focus on experimentation when they want to engage young children (i.e., first or second grade students) in scientific reasoning activities. Second, finding that subcomponents do not fully cohere suggests that skills should not be fostered in isolation, but educators should highlight what these different aspects of scientific reasoning have in common so that it will be easier for children to transfer their skills from one subcomponent to the other.

In line with previous work (Koerber & Osterhaus, 2019; Koerber et al., 2015), we did not observe any gender differences in scientific reasoning performance. Some researchers (e.g., Lazonder et al., 2020) have observed such differences in older elementary school children. It may well be that gender differences—if they exist—develop late in development, once children are confronted more with stereotypical images of scientists (who may often be depicted as males), which may result in a stronger identification with science topic in boys than girls. Future research should address this question, and cross-cultural studies may be particularly helpful in this respect, as cultures differ in how they portray science and scientists. The availability of a validated measure of scientific reasoning, such as the SC-SKI, is a necessary prerequisite to make possible and foster this type of research.

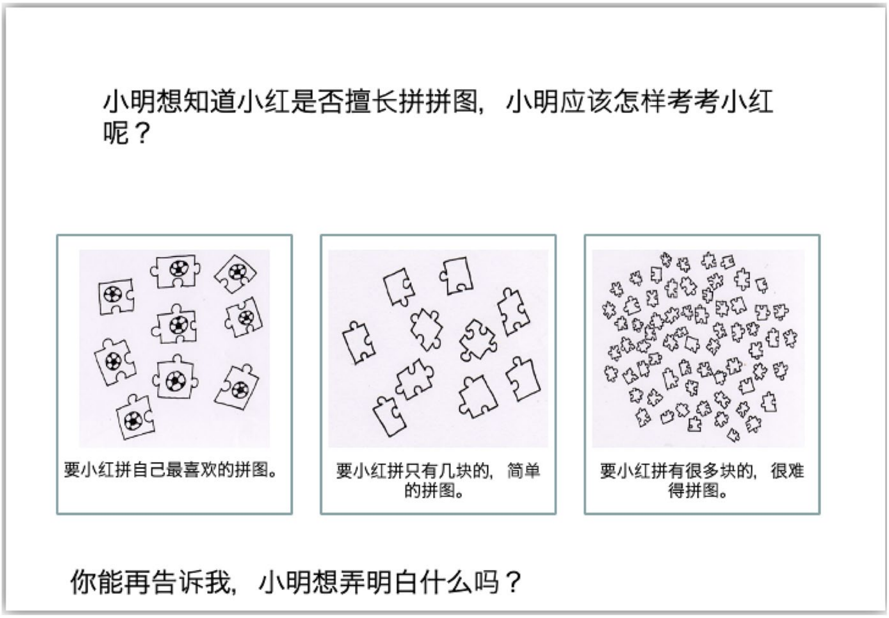
There are two shortcomings of the present study: First, we studied a relatively small sample of children. Needless to say, future work must study larger and more representative groups of children, including those from different provinces and age groups. Second, our urban and rural samples were not equal in size. Future work should draw larger samples from rural schools and address the specific learning histories of the children— in both urban and rural schools to account for variation between classrooms. The groundwork for such research is laid here, and the availability of the SC-SKI can be expected to promote such research.

5 Conclusion

The shortened Chinese version of the Science-K(indergarten) Inventory (SC-SKI) is overall a reliable and valid instrument to measure the scientific reasoning skills of Chinese 6- and 7-year-olds. Although further item improvements are necessary for some of the items, the SC-SKI is a valuable instrument and point of departure to foster cross-cultural research on young children's scientific reasoning, which is an important twenty-first century skill.

Appendix

See Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20.



小明想知道小红是否擅长拼拼图，小明应该怎样考考小红呢？

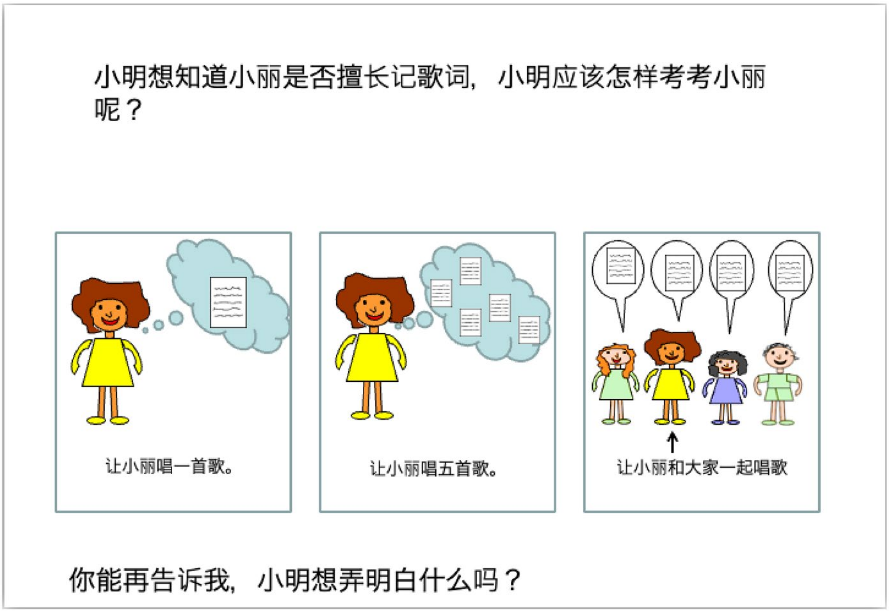
要小红拼自己最喜欢的拼图。

要小红拼只有几块的，简单的拼图。

要小红拼有很多块的，很难得拼图。

你能再告诉我，小明想弄明白什么吗？

Fig. 1 Item Exp-1, assessing children's ability to distinguish a conclusive from an inconclusive test. Tom wants to find out if Mia is good at doing puzzles. What should he ask Mia to do? Piece together her favorite puzzle (1; incorrect), piece together a puzzle with few pieces (2; incorrect), or piece together a puzzle with many pieces (3; correct)?



小明想知道小丽是否擅长记歌词，小明应该怎样考考小丽呢？

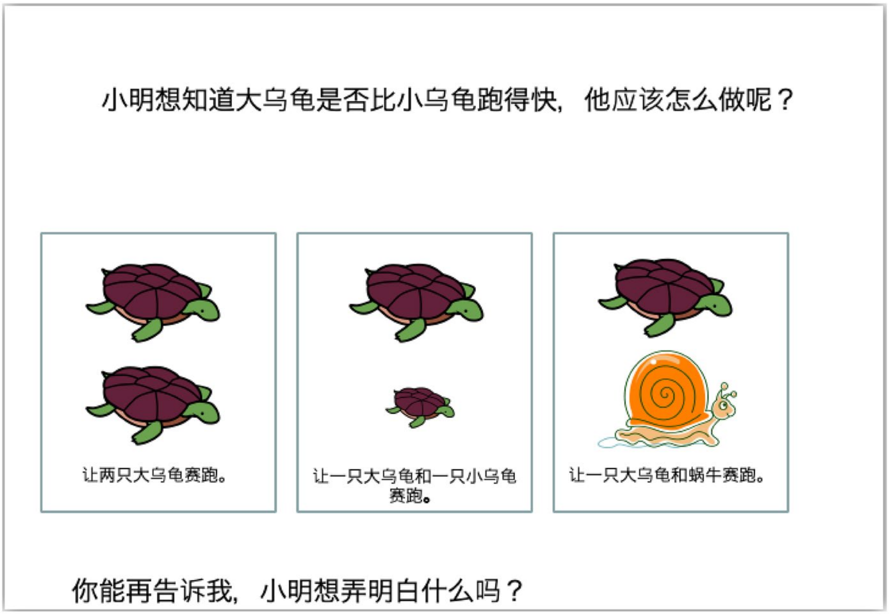
让小丽唱一首歌。

让小丽唱五首歌。

让小丽和大家一起唱歌

你能再告诉我，小明想弄明白什么吗？

Fig. 2 Item Exp-2, assessing children's ability to distinguish a conclusive from an inconclusive test. Tom wants to find out if Mia is good at memorizing songs. What should he ask Mia to do? Sing one song by herself (1; incorrect), sing 5 songs by herself (2; correct), or sing 1 song with everyone else (3; incorrect)?



小明想知道大乌龟是否比小乌龟跑得快，他应该怎么做呢？

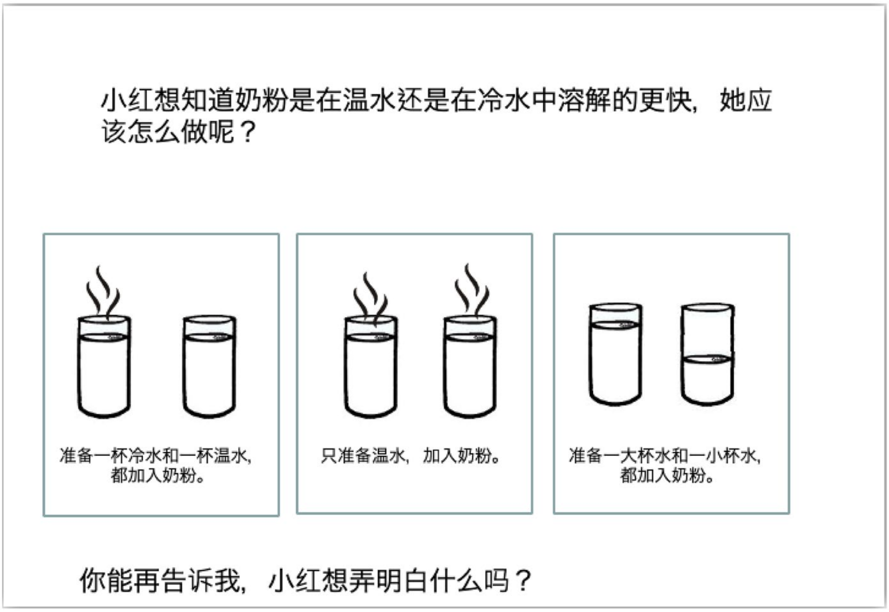
让两只大乌龟赛跑。

让一只大乌龟和一只小乌龟赛跑。

让一只大乌龟和蜗牛赛跑。

你能再告诉我，小明想弄明白什么吗？

Fig. 3 Item Exp-3, assessing children's understanding of the control-of-variables strategy. Tom wants to find out if big turtles run faster than small turtles? What does he have to do? Have two big turtles race each other (1; incorrect), have a big turtle race against a small turtle (2; correct), or have a big turtle race a slug (3; incorrect)?



小红想知道奶粉是在温水还是在冷水中溶解的更快，她应该怎么做呢？

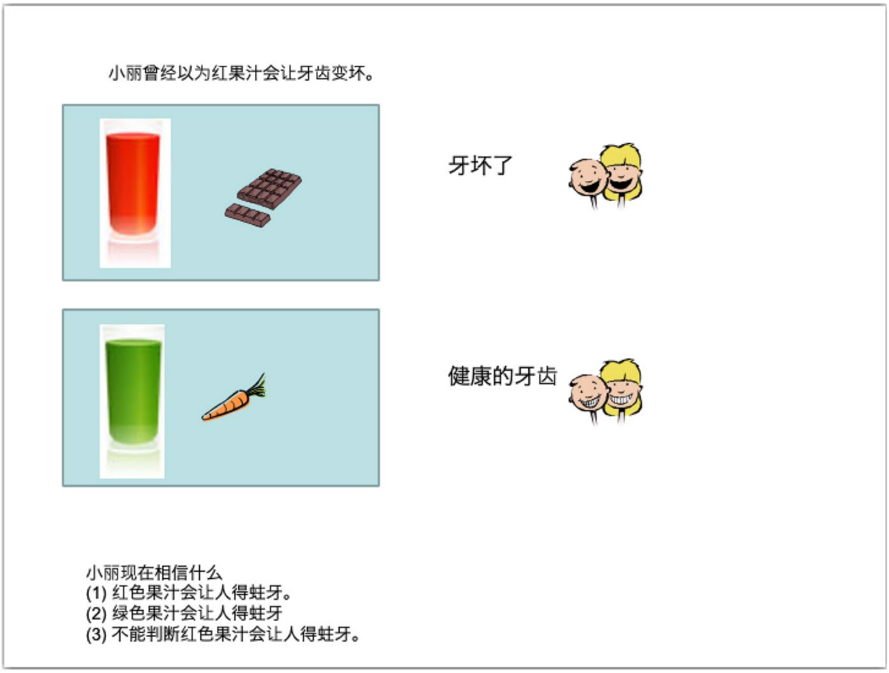
准备一杯冷水和一杯温水，都加入奶粉。

只准备温水，加入奶粉。

准备一大杯水和一小杯水，都加入奶粉。

你能再告诉我，小红想弄明白什么吗？

Fig. 4 Item Exp-4, assessing children's understanding of the control-of-variables strategy. Mia wants to find out if cocoa powder dissolves better in warm or in cold milk. What does she have to do? Put cacao powder in a glass of warm and cold milk (1; correct), put cacao powder in two glasses of warm milk (2; incorrect), or put cacao powder in a glass with a lot of milk and a glass with little milk (3; incorrect)?



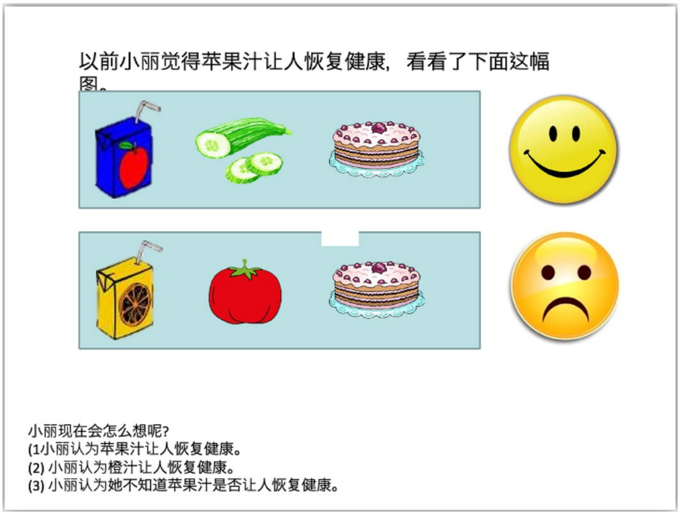
小丽曾经以为红果汁会让牙齿变坏。

牙坏了

健康的牙齿

小丽现在相信什么
(1) 红色果汁会让人得蛀牙。
(2) 绿色果汁会让人得蛀牙
(3) 不能判断红色果汁会让人得蛀牙。

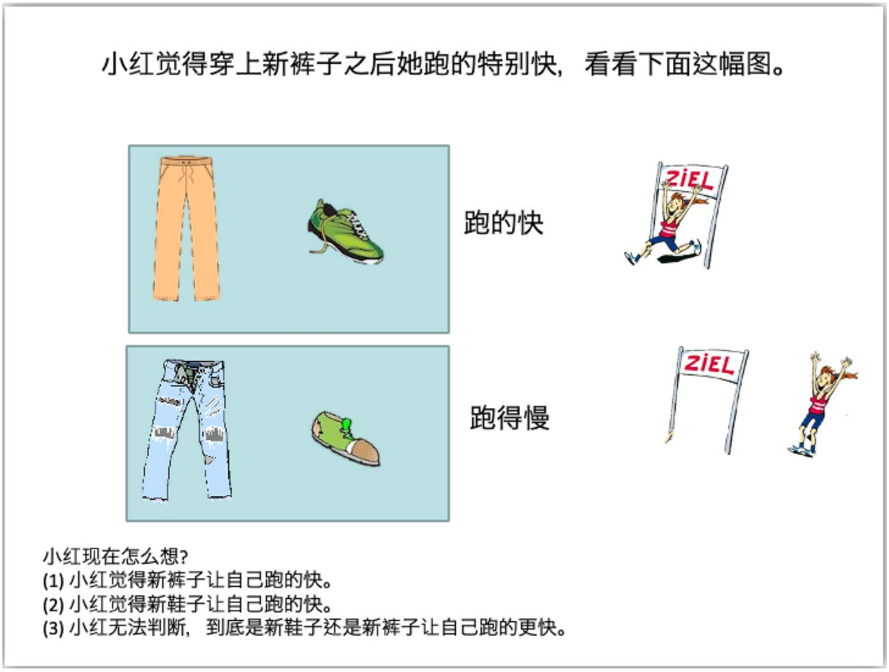
Fig. 5 Item Dat-1, assessing children's understanding of confounded data. Mia believes that red juice makes your teeth fall out. What does she believe after seeing the data? Red juice makes your teeth fall out (1; incorrect), green juice makes your teeth fall out (2; incorrect), or you cannot tell if red juice makes your teeth fall out (3; correct)?



以前小丽觉得苹果汁让人恢复健康，看看了下面这幅图。

小丽现在会怎么想呢?
(1小丽认为苹果汁让人恢复健康。
(2) 小丽认为橙汁让人恢复健康。
(3) 小丽认为她不知道苹果汁是否让人恢复健康。

Fig. 6 Item Dat-2, assessing children's understanding of confounded data. Mia believes that apple juice makes sick people feel better. What does she believe after seeing the data? Apple juice makes sick people feel better (1; incorrect), orange juice makes sick people feel better (2; incorrect), or you cannot tell if apple juice makes sick people feel better (3; correct)?



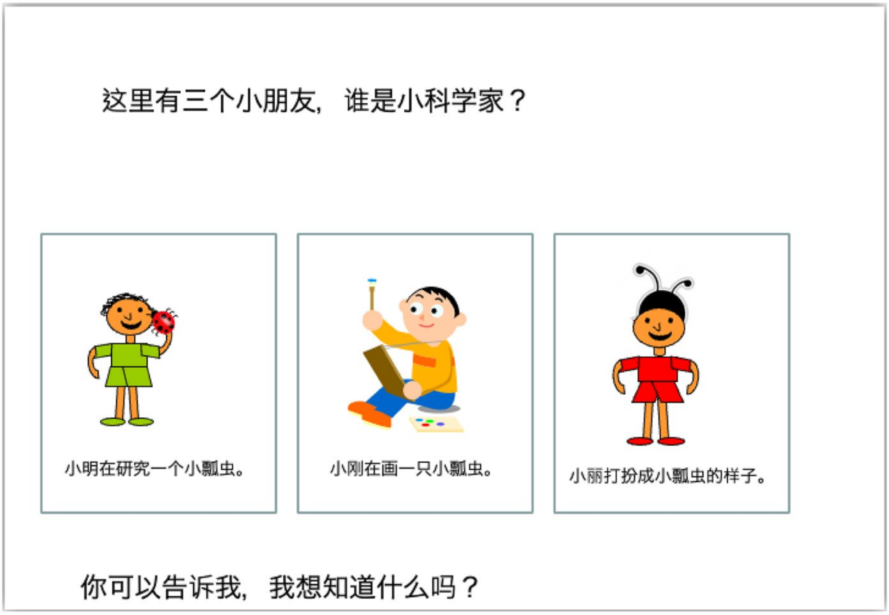
小红觉得穿上新裤子之后她跑的特别快，看看下面这幅图。

跑的快

跑得慢

小红现在怎么想?
(1) 小红觉得新裤子让自己跑的快。
(2) 小红觉得新鞋子让自己跑的快。
(3) 小红无法判断，到底是新鞋子还是新裤子让自己跑的更快。

Fig. 7 Item Dat-3, assessing children's understanding of confounded data. Mia believes that she can run fast because she has a new pair of trousers. What does she believe after seeing the data? That she runs fast because of her new trousers (1; incorrect), that she runs fast because of her new shoes (2; incorrect), or that you cannot tell if she runs fast because of her new trousers (3; correct)?



这里有三个小朋友，谁是小科学家？

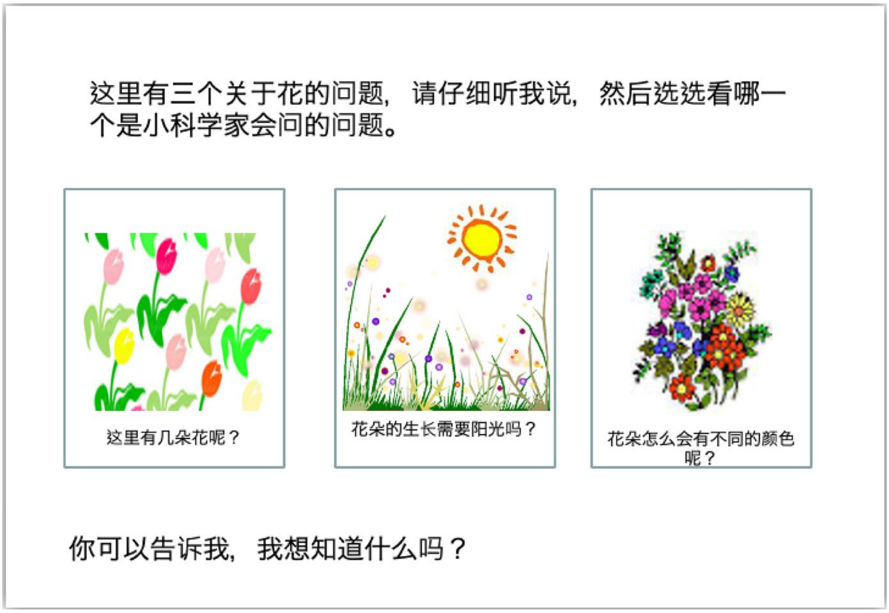
小明在研究一个小瓢虫。

小刚在画一只小瓢虫。

小丽打扮成小瓢虫的样子。

你可以告诉我，我想知道什么吗？

Fig. 8 Item NoS-1, assessing children's understanding of what scientists do. Who of the three children is like a scientist? Tom investigates a lady bug (1; correct), Jan paints a lady bug (2; incorrect), or Nick dresses up like a lady bug (3; incorrect)?



这里有三个关于花的问题，请仔细听我说，然后选选看哪一个是小科学家会问的问题。

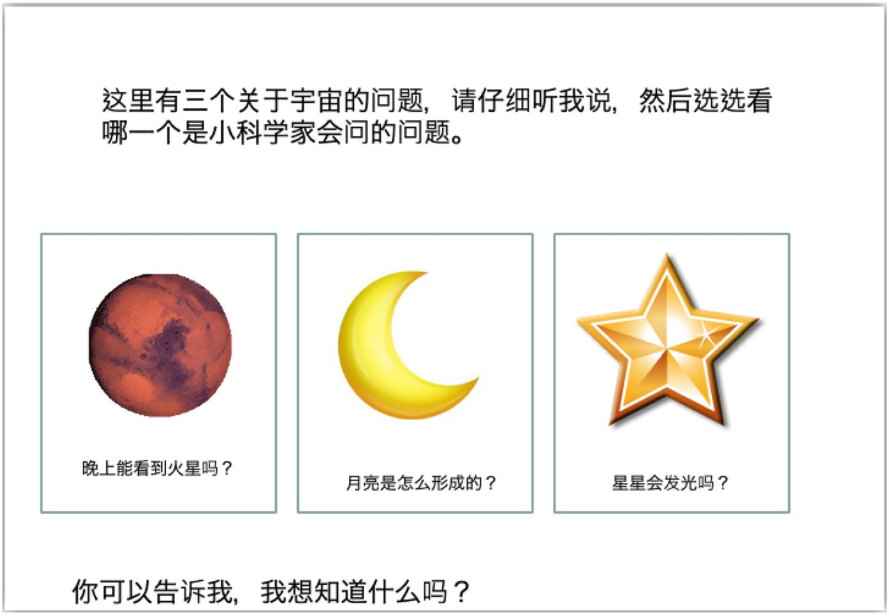
这里有几朵花呢？

花朵的生长需要阳光吗？

花朵怎么会有不同的颜色呢？

你可以告诉我，我想知道什么吗？

Fig. 9 Item NoS-2, assessing children's understanding of what kind of questions scientists ask. Which of the following three questions is from a scientist? Are there many flowers (1; incorrect), do flowers need sun to grow (2; correct), or how do flowers get their color (3; incorrect)?



这里有三个关于宇宙的问题，请仔细听我说，然后选选看哪一个是小科学家会问的问题。

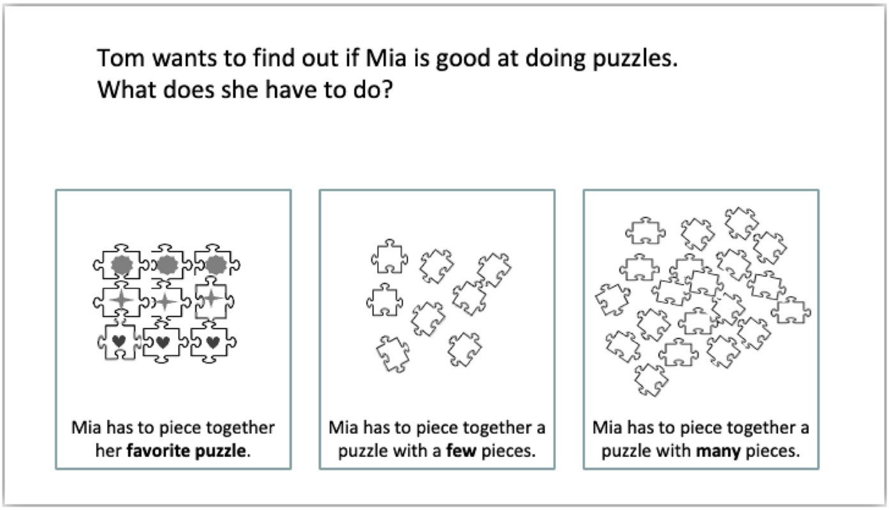
晚上能看到火星吗？

月亮是怎么形成的？

星星会发光吗？

你可以告诉我，我想知道什么吗？

Fig. 10 Item NoS-3, assessing children's understanding of what kind of questions scientists ask. Which of the following three questions is from a scientist? Can you see mars at night (1; incorrect), how was the moon formed (2; correct), or do stars shine bright (3; incorrect)?



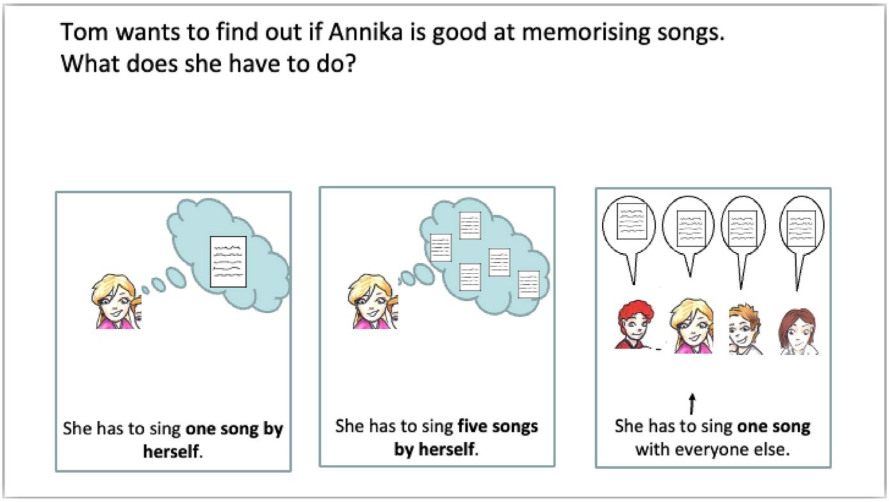
Tom wants to find out if Mia is good at doing puzzles.
What does she have to do?

Mia has to piece together her **favorite puzzle**.

Mia has to piece together a puzzle with a **few** pieces.

Mia has to piece together a puzzle with **many** pieces.

Fig. 11 Item Exp-1, assessing children's ability to distinguish a conclusive from an inconclusive test. Tom wants to find out if Mia is good at doing puzzles. What should he ask Mia to do? Piece together her favorite puzzle (1; incorrect), piece together a puzzle with few pieces (2; incorrect), or piece together a puzzle with many pieces (3; correct)?



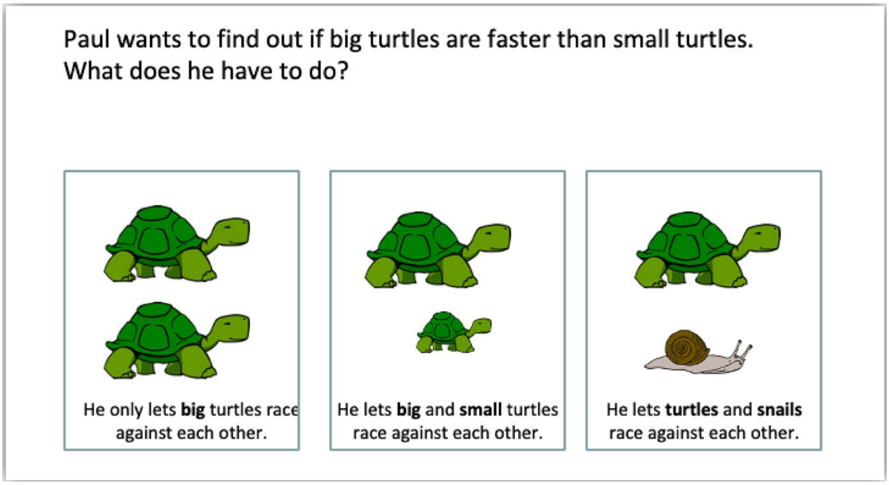
Tom wants to find out if Annika is good at memorising songs.
What does she have to do?

She has to sing **one song by herself**.

She has to sing **five songs by herself**.

She has to sing **one song** with everyone else.

Fig. 12 Item Exp-2, assessing children's ability to distinguish a conclusive from an inconclusive test. Tom wants to find out if Mia is good at memorizing songs. What should he ask Mia to do? Sing one song by herself (1; incorrect), sing 5 songs by herself (2; correct), or sing 1 song with everyone else (3; incorrect)?



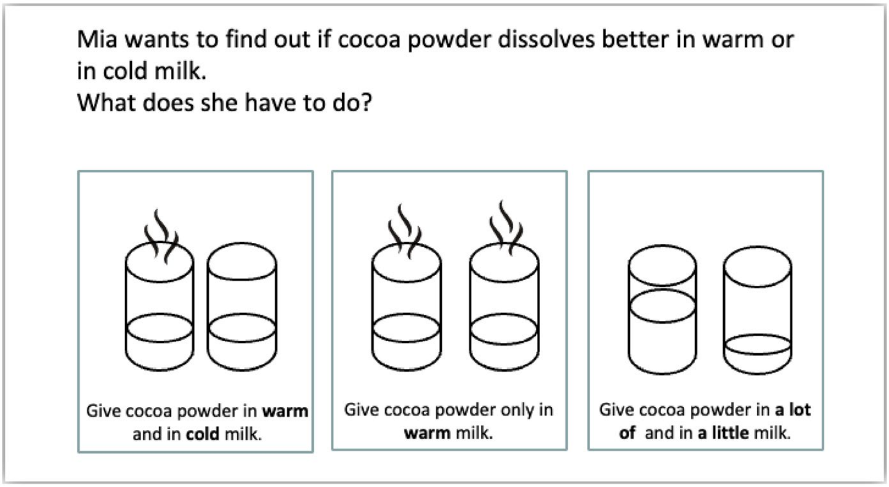
Paul wants to find out if big turtles are faster than small turtles.
What does he have to do?

He only lets **big** turtles race against each other.

He lets **big** and **small** turtles race against each other.

He lets **turtles** and **snails** race against each other.

Fig. 13 Item Exp-3, assessing children's understanding of the control-of-variables strategy. Tom wants to find out if big turtles run faster than small turtles? What does he have to do? Have two big turtles race each other (1; incorrect), have a big turtle race against a small turtle (2; correct), or have a big turtle race a slug (3; incorrect)?



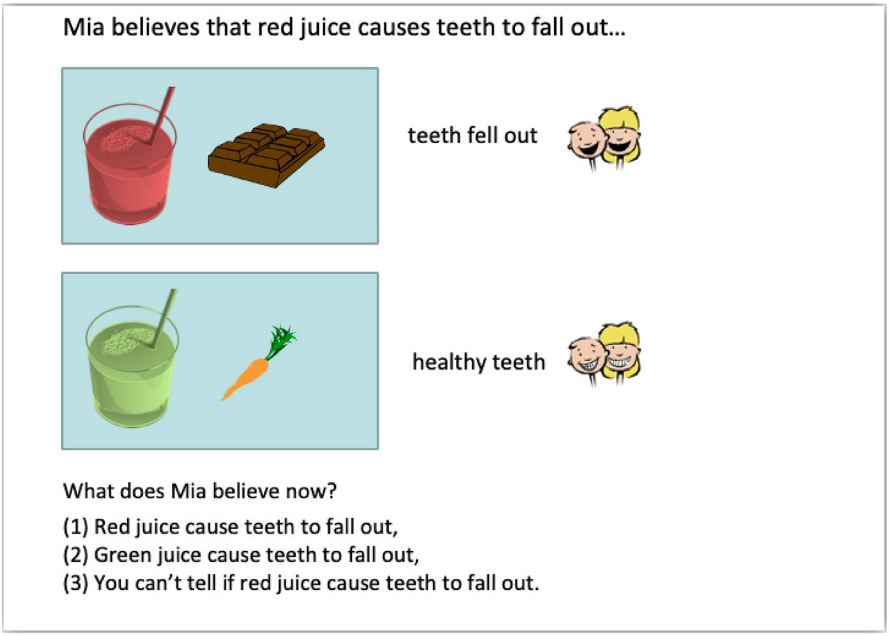
Mia wants to find out if cocoa powder dissolves better in warm or in cold milk.
What does she have to do?

Give cocoa powder in **warm** and in **cold** milk.

Give cocoa powder only in **warm** milk.

Give cocoa powder in **a lot of** and in **a little** milk.

Fig. 14 Item Exp-4, assessing children's understanding of the control-of-variables strategy. Mia wants to find out if cocoa powder dissolves better in warm or in cold milk. What does she have to do? Put cacao powder in a glass of warm and cold milk (1; correct), put cacao powder in two glasses of warm milk (2; incorrect), or put cacao powder in a glass with a lot of milk and a glass with little milk (3; incorrect)?



Mia believes that red juice causes teeth to fall out...

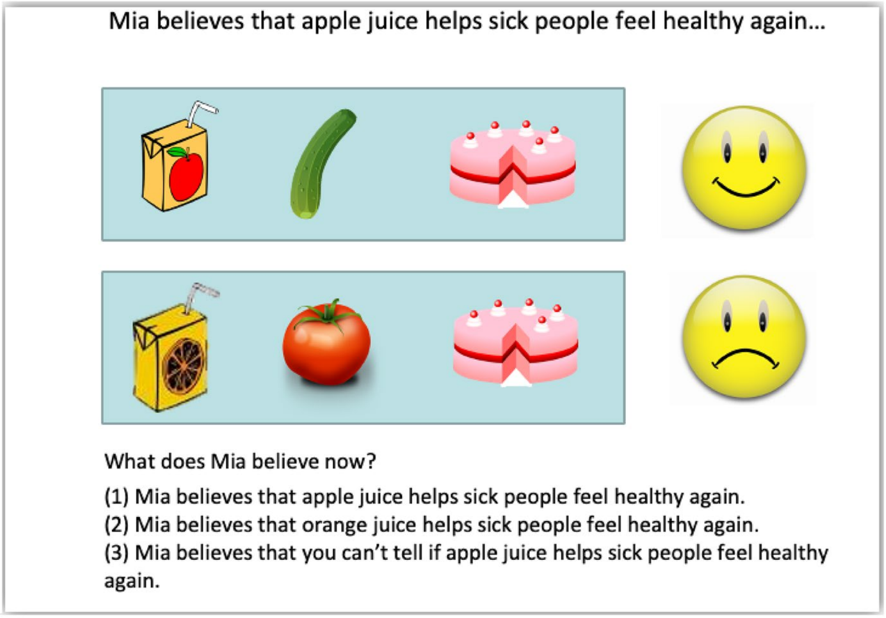
teeth fell out

healthy teeth

What does Mia believe now?

(1) Red juice cause teeth to fall out,
(2) Green juice cause teeth to fall out,
(3) You can’t tell if red juice cause teeth to fall out.

Fig. 15 Item Dat-1, assessing children’s understanding of confounded data. Mia believes that red juice makes your teeth fall out. What does she believe after seeing the data? Red juice makes your teeth fall out (1; incorrect), green juice makes your teeth fall out (2; incorrect), or you cannot tell if red juice makes your teeth fall out (3; correct)?

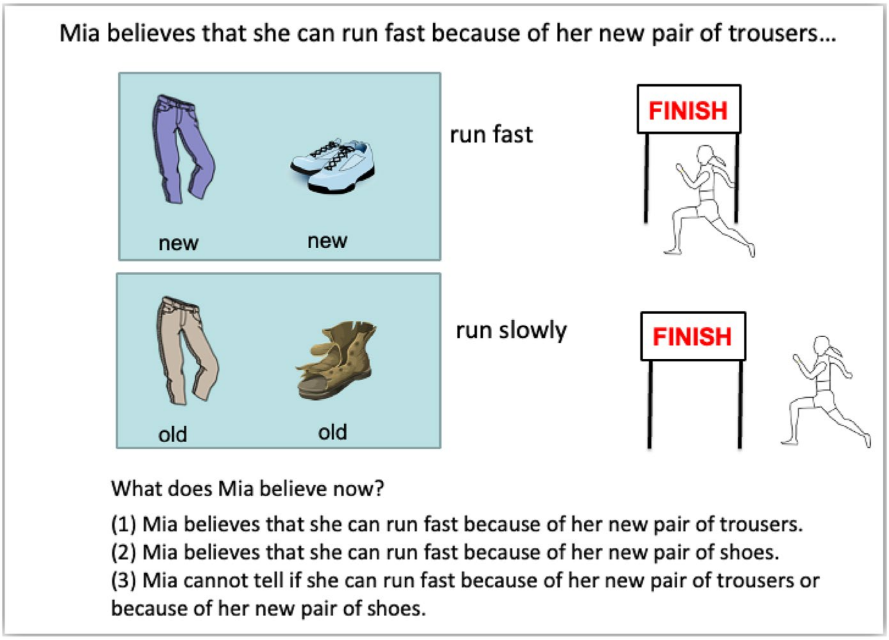


Mia believes that apple juice helps sick people feel healthy again...

What does Mia believe now?

(1) Mia believes that apple juice helps sick people feel healthy again.
(2) Mia believes that orange juice helps sick people feel healthy again.
(3) Mia believes that you can’t tell if apple juice helps sick people feel healthy again.

Fig. 16 Item Dat-2, assessing children’s understanding of confounded data. Mia believes that apple juice makes sick people feel better. What does she believe after seeing the data? Apple juice makes sick people feel better (1; incorrect), orange juice makes sick people feel better (2; incorrect), or you cannot tell if apple juice makes sick people feel better (3; correct)?

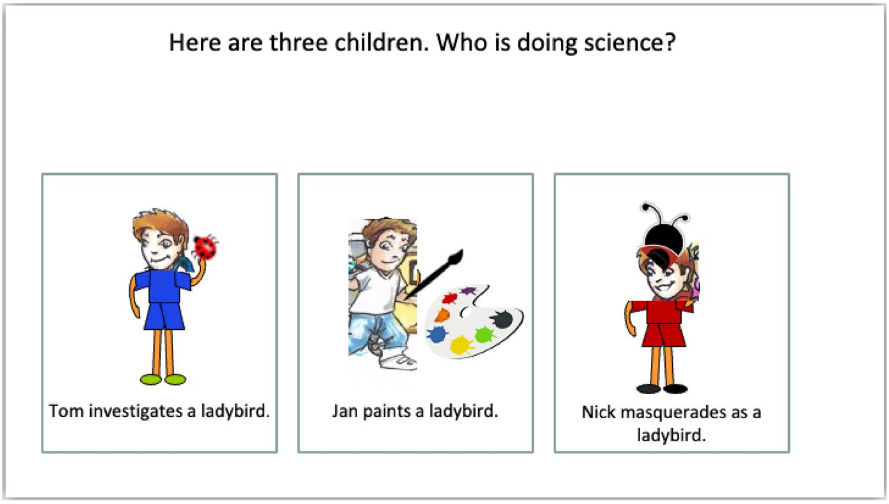


Mia believes that she can run fast because of her new pair of trousers...

What does Mia believe now?

(1) Mia believes that she can run fast because of her new pair of trousers.
(2) Mia believes that she can run fast because of her new pair of shoes.
(3) Mia cannot tell if she can run fast because of her new pair of trousers or because of her new pair of shoes.

Fig. 17 Item Dat-3, assessing children's understanding of confounded data. Mia believes that she can run fast because she has a new pair of trousers. What does she believe after seeing the data? That she runs fast because of her new trousers (1; incorrect), that she runs fast because of her new shoes (2; incorrect), or that you cannot tell if she runs fast because of her new trousers (3; correct)?



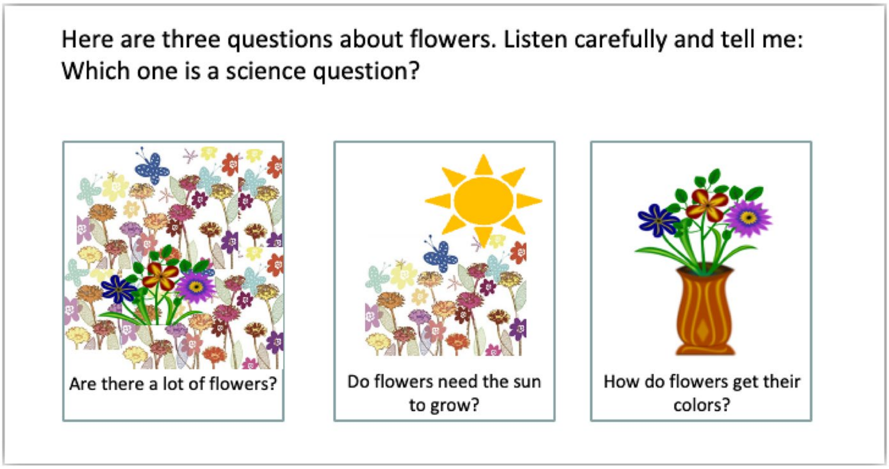
Here are three children. Who is doing science?

Tom investigates a ladybird.

Jan paints a ladybird.

Nick masquerades as a ladybird.

Fig. 18 Item NoS-1, assessing children's understanding of what scientists do. Who of the three children is like a scientist? Tom investigates a lady bug (1; correct), Jan paints a lady bug (2; incorrect), or Nick dresses up like a lady bug (3; incorrect)?



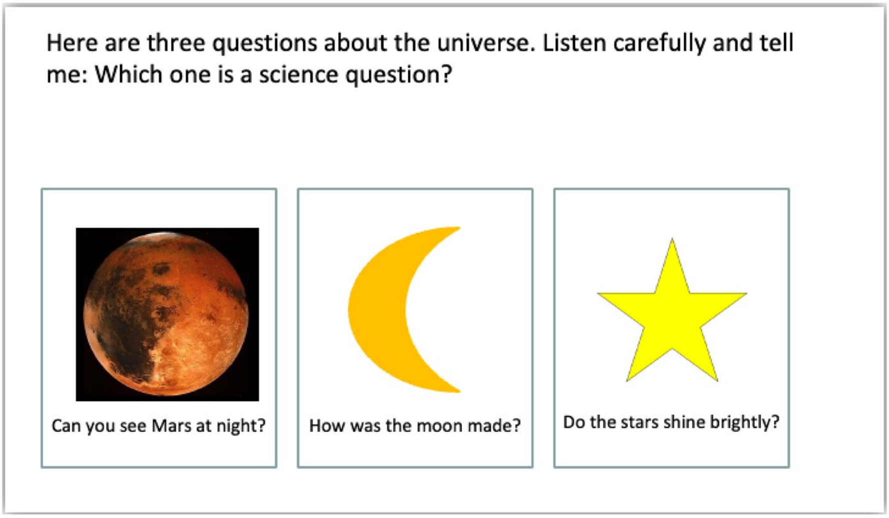
Here are three questions about flowers. Listen carefully and tell me: Which one is a science question?

Are there a lot of flowers?

Do flowers need the sun to grow?

How do flowers get their colors?

Fig. 19 Item NoS-2, assessing children's understanding of what kind of questions scientists ask. Which of the following three questions is from a scientist? Are there many flowers (1; incorrect), do flowers need sun to grow (2; correct), or how do flowers get their color (3; incorrect)?



Here are three questions about the universe. Listen carefully and tell me: Which one is a science question?

Can you see Mars at night?

How was the moon made?

Do the stars shine brightly?

Fig. 20 Item NoS-3, assessing children's understanding of what kind of questions scientists ask. Which of the following three questions is from a scientist? Can you see mars at night (1; incorrect), how was the moon formed (2; correct), or do stars shine bright (3; incorrect)?

Funding Open Access funding enabled and organized by Projekt DEAL.

References

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

Koerber, S., & Osterhaus, C. (2019). Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *Journal of Cognition and Development, 20*(4), 510–533. https://doi.org/10.1080/15248372.2019.1620232

Koerber, S., & Osterhaus, C. (2021). Science competencies in kindergarten: A prospective study in the last year of kindergarten. *Unterrichtswissenschaft, 49*, 117–136. https://doi.org/10.1007/s42010-020-00093-5

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. https://doi.org/10.1024/1421-0185.64.3.141

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*(1), 327–336. https://doi.org/10.1111/cdev.12298

Koerber, S., Sodian, B., Osterhaus, C., Mayer, D., Kropf, N., & Schwippert, K. (2017). Science-P.II. Modeling scientific reasoning in primary school. In D. Leutner, J. Fleischer, J. Grünkörn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments.* London: Springer.

Köksal, Ö., Sodian, B., & Legare, C. H. (2021). Young children's metacognitive awareness of confounded evidence. *Journal of Experimental Child Psychology, 205*, 105080. https://doi.org/10.1016/j.jecp.2020.105080

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

Lapidow, E., & Walker, C. M. (2020). Informative experimentation in intuitive science: Children select and learn from their own causal interventions. *Cognition, 201*, 104315. https://doi.org/10.1016/j.cognition.2020.104315

Lazonder, A. W., Janssen, N., Gijlers, H., & Walraven, A. (2020). Patterns of development in children's scientific reasoning: Results from a three-year longitudinal study. *Journal of Cognition and Development*. https://doi.org/10.1080/15248372.2020.1814293

Lüdecke, D. (2021). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.10, https://CRAN.R-project.org/package=sjPlot.

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

Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children's social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology, 53*(3), 450–462. https://doi.org/10.1037/dev0000260

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*(2), 153–179. https://doi.org/10.1111/j.2044-835X.2012.02082.x

Ratcliffe, M., & Grace, M. (2003). *Science education for citizenship: Teaching socio-scientific issues*. Maidenhead: Open University Press.

Revelle, W. (2022). *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching: THe Official Journal of the National Association for Research in Science Teaching, 41*, 513–536. https://doi.org/10.1002/tea.20009

Samarapungavan, A., Mantzicopoulos, P., & Patrick, H. (2008). Learning science through inquiry in kindergarten. *Science Education, 92*(5), 868–908. https://doi.org/10.1002/sce.20275

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*(4), 753–766. https://doi.org/10.1111/j.1467-8624.1991.tb01567.x

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco: Jossey-Bass.

van de Sande, E., Kleemans, T., Verhoeven, L., & Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learning and Instruction, 62*, 20–26. https://doi.org/10.1016/j.learninstruc.2019.02.002

van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science, 43*(3), 381–400. https://doi.org/10.1007/s11251-015-9344-y

van der Graaf, J., Segers, E., & Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learning and Instruction, 56*, 1–9. https://doi.org/10.1016/j.learninstruc.2018.03.005

van Schijndel, T. J., Visser, I., van Bers, B. M., & Raijmakers, M. E. (2015). Preschoolers perform more informative experiments after observing theory-violating evidence. *Journal of Experimental Child Psychology*, *131*, 104–119. https://doi.org/10.1016/j.jecp.2014.11.008.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children (WISC)–IV*. Uttar pradesh: Psychological Corporation.

Zhang, H. (2017). Opportunity or new poverty trap: Rural-urban education disparity and internal migration in China. *China Economic Review, 44*, 112–124. https://doi.org/10.1016/j.chieco.2017.03.011