

Holstein, Elfriede R.; Lörch, Lucas; Brod, Garvin

Children's cognitive reflection skills predict their conceptual learning

Collabra: Psychology 11 (2025) 143423



Quellenangabe/ Reference:

Holstein, Elfriede R.; Lörch, Lucas; Brod, Garvin: Children's cognitive reflection skills predict their conceptual learning - In: Collabra: Psychology 11 (2025) 143423 - URN: urn:nbn:de:0111-pedocs-347843 - DOI: 10.25656/01:34784; 10.1525/collabra.143423

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-347843>

<https://doi.org/10.25656/01:34784>

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by/4.0/deed.de> - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen. Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License:

<http://creativecommons.org/licenses/by/4.0/deed.en> - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor. By using this particular document, you accept the above-stated conditions of use.



Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

Developmental Psychology

Children's Cognitive Reflection Skills Predict Their Conceptual Learning

Elfriede R. Holstein^{1,2}, Lucas Lörch^{1,2}, Garvin Brod^{1,2,3}

¹ Education and Human Development, DIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany, ² Individual development, IDeA-Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt, Germany, ³ Department of Psychology, Goethe University, Frankfurt, Germany

Keywords: cognitive development, conceptual learning, inquiry-based learning, science learning, cognitive reflection

<https://doi.org/10.1525/collabra.143423>

Collabra: Psychology

Vol. 11, Issue 1, 2025

The developmental Cognitive Reflection Test (CRT-D) measures children's ability to override immediate reactions for more considered responses, a skill crucial for learning non-intuitive concepts, particularly in science. However, the predictive validity of the CRT-D for inquiry-based learning has not been established. This study validated a German version of the CRT-D and investigated its role in predicting children's conceptual learning about balance. Data were collected across three studies with children aged 5 to 10 ($N = 255$), and predictive validity was tested in one study ($n = 116$), in which children participated in a balance scale task assessing conceptual learning before and after an inquiry-based learning session. The CRT-D demonstrated acceptable internal consistency (McDonald's $\omega = .69$), and its factor structure and developmental trends were consistent with other populations. Children's CRT-D scores significantly predicted their conceptual learning, suggesting that cognitive reflection is essential for successful science learning.

1. Introduction

Cognitive reflection, the ability to engage in analytic reasoning and reflect on one's own thinking, has been found to play a significant role in science learning. Various studies have shed light on the association between cognitive reflection and the level of conceptual knowledge about scientific phenomena in adults (Gette & Kryjevskaja, 2019; Gómez-Chacón et al., 2014; Shtulman & McCallum, 2014) and children (Young & Shtulman, 2020a). In addition to the level of knowledge, cognitive reflection may influence the learning processes that lead to such knowledge as it supports the assessment of evidence (Schulz et al., 2007), the identification of knowledge gaps, and the updating of beliefs (Shtulman & Young, 2022). However, this link between cognitive reflection and conceptual learning remains largely unclear. To address this gap, the present study investigated the relationship of children's cognitive reflection skills with their inquiry-based conceptual learning.

1.1. Cognitive Reflection and Its Measurement

Cognitive reflection is the ability to suppress an immediate gut reaction in favor of a deliberate, analytic response (Frederick, 2005; Toplak et al., 2011). It also fosters deeper thinking, supporting the ability to make well-reasoned judgements in complex situations (Young & Shtulman, 2020b, 2020a).

soned judgements in complex situations (Young & Shtulman, 2020b, 2020a).

The cognitive reflection test (CRT; Frederick, 2005) was designed to measure cognitive reflection by presenting questions with intuitively appealing but incorrect answers, requiring deeper analytical thinking and careful reasoning to arrive at the correct answer. The questions are thought to create conflict between "System 1" (instant and effortless) and "System 2" (slow, deliberate, and effortful) thinking, whereby individuals must utilize "System 2" to answer correctly (Frederick, 2005; Meyer et al., 2024). The CRT has been shown to provide incremental predictive validity over knowledge tests for various decision-making tasks, including numeracy, beliefs, and preferences (Meyer et al., 2024). In sum, cognitive reflection is a crucial skill for rational decision making, and the CRT is a promising instrument for assessing it.

However, due to the inclusion of mathematical operations, the CRT is unsuitable for children. Therefore, a child-friendly version of the CRT was developed (Young & Shtulman, 2020b). In US and Chinese populations, this developmental Cognitive Reflection Test (CRT-D) has been found to be a reliable measure and to improve with age, albeit with significant individual differences (Gong et al., 2021; Young & Shtulman, 2020a). In the present study, we used this developmental trend as a benchmark for a Ger-

a Correspondence concerning this article should be addressed to Elfriede R. Holstein, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt, Germany. Email: e.holstein@dipf.de

man version of the CRT-D. We employed the German CRT-D in three studies and used the resulting data to check its psychometric properties and to compare its factor structure and developmental trends with those found in previous studies.

1.2. The Role of Cognitive Reflection in Children's Science Learning

Previous research has shown that CRT scores were significant predictors of science understanding (Shtulman & McCallum, 2014) and were linked to students' comprehension of various secondary school mathematics concepts (Gómez-Chacón et al., 2014). For elementary school children, CRT-D performance has been shown to be a strong predictor of their conceptual understanding of vitalism and mathematical equivalence, even when controlling for age, executive functions, and rational thinking (Young & Shtulman, 2020a).

In addition, there have been initial findings that cognitive reflection is not only related to the level of conceptual knowledge but also to the learning of such knowledge. In a study by Young and Shtulman (2020b), children aged 5 to 12 were asked to judge a series of statements as either true or false. The statements referred to concepts about life and matter and each statement fell into one of four categories: intuitively and scientifically correct, intuitively correct but scientifically incorrect, intuitively incorrect but scientifically correct, intuitively and scientifically incorrect. Between the pre- and posttest involving this statement verification task, children completed a tutorial on either life or matter. The tutorial contained videos on definitions of key characteristics, and on common misconceptions, and why they were false. The authors found that children with higher CRT-D scores showed larger gains in conceptual understanding as a result of the tutorial than children with lower CRT-D scores. In sum, this finding provides initial evidence that cognitive reflection is correlated with learning from direct instruction. However, this study also leaves several important questions about the relation between cognitive reflection and conceptual learning unanswered, which is the gap that the current study attempted to close.

First, there are many instances in which science learning, both in and out of school, involves inquiry learning rather than learning from direct instruction such as video tutorials. Inquiry learning involves activities such as making predictions about outcomes of experiments, observing such outcomes, and drawing inferences about the underlying causal mechanisms. How might these processes be affected by cognitive reflection? When predictions are made, cognitive reflection may enable children to ponder competing hypotheses and the outcomes these hypotheses would suggest. This logic is in line with the notion that cognitive reflection helps learners to pause and reconsider intuitive but incorrect ideas (Young & Shtulman, 2020a), which is essential for updating intuitive theories in line with scientific concepts. When inferences are drawn from presented evidence, cognitive reflection may enable children to suppress the intuitive reaction to discard evidence that conflicts with their prior beliefs (Young & Shtulman, 2024). Cognitive re-

flexion may help children to engage in analytical reasoning about the underlying causal mechanisms and to update their beliefs in line with the evidence. This line of argumentation seems reasonable, as prior research has shown that children engage in reflective processes when learning about causality (Schulz et al., 2007), and that deliberate processing is essential for grasping complex relationships and processes (Vosniadou et al., 2018). Based on these arguments, we expected cognitive reflection to support children's inquiry learning of counterintuitive scientific concepts.

Second, the study by Young and Shtulman (2020b) did not control for executive functions when computing the correlation between cognitive reflection and learning gains. Thus, the correlation could have been caused by executive functions being associated with both cognitive reflection and learning gains. Previous studies have revealed that executive functions are related to both cognitive reflection and science learning, but most of them were conducted with adults. It has been found that executive functions are connected to overriding the intuitive response in the CRT (Carriedo et al., 2020), that cognitive reflection draws on executive resources (Johnson & Tubau, 2014), and that working memory is a reliable predictor of CRT performance (Stuppel et al., 2013). Furthermore, cognitive reflection may rely on executive functions, as children need to suppress an immediate reaction (inhibition), switch to an alternative response (set shifting), and maintain the question in their mind (working memory; Young & Shtulman, 2020b). In addition, executive functions have previously been proposed as a key prerequisite for successful science learning (Vosniadou et al., 2018). Taken together, these findings suggest that both cognitive reflection and science learning rely on executive functions and, thus, may appear to be correlated if executive functions are not accounted for.

To preclude the possibility of such a spurious correlation between cognitive reflection and inquiry-based conceptual learning, we measured and statistically controlled for executive functions in the present study. There are several arguments that support the notion that cognitive reflection influences inquiry learning over and above the effect of executive functions. Some studies did not find an association of components of cognitive reflection with executive functions (Johnson & Tubau, 2014), or high correlations between executive functions and CRT performance (Toplak et al., 2011). This suggests that the relationship between the two concepts may not be as strong or consistent as some studies have indicated. Moreover, research with adults suggests that the CRT's predictive strength for rational thinking—the ability to think logically and reflectively—is largely independent of executive functions (Toplak et al., 2011). Research with children found that CRT-D performance was a strong predictor of children's conceptual understanding even when controlling for executive functions (Young & Shtulman, 2020a). These findings suggest that cognitive reflection may contribute uniquely to conceptual learning. This unique contribution of cognitive reflection may lie in fostering metaconceptual awareness, necessary for detect-

ing conflicts between naïve and scientific concepts and resolving these conflicts (Shtulman & Young, 2022).

Third, in the statement verification task used by Young and Shtulman (2020b), children were assessed with materials that closely resembled the instructional content. In both the instructional content and the assessment, children were directly asked if they believed certain concepts to be true. As a result, they did not need to apply their conceptual knowledge to a new task. Thus, it is unclear if this task reflects true conceptual understanding, or rather the recall of learned facts. To address this shortcoming, we investigated the correlation of cognitive reflection with a transfer task that required children to apply their conceptual understanding to new cases. It has been argued previously that cognitive reflection is related to flexibility (Shtulman & Young, 2022), which enables children to apply newly learned concepts in learning situations. Thus, in our transfer task, we expected that cognitive reflection would enable children to suppress automatic responses based on their long-established, intuitive concepts and to apply newly learned ones instead.

In summary, the goal of this study was two-fold. First, we aimed to validate a German version of the CRT-D. For this purpose, we used data from three different studies in which the CRT-D was employed and checked the internal consistency, the factor structure, and the association with age. We expected that German children would fall for intuitive incorrect answers similar to what has been demonstrated in US and Chinese children, and that they would show similar developmental trends.

Second, we examined cognitive reflection as a predictor of inquiry-based conceptual learning. Since a crucial part of conceptual learning is to suppress intuitive responses, we expected that better cognitive reflection would predict conceptual learning, particularly deeper understanding and transferability. We tested this hypothesis with data from a transfer test in the domain of balance that we employed in one of the studies. We expected to find a significant correlation between the CRT-D scores and the performance in this transfer test while controlling for executive functions.

2. Methods

The data was collected in three different studies with children aged 5 to 10 ($N = 255$, $M_{\text{Age}} = 98.18$ months, $SD = 16.97$, 47.06 % girls) in a large natural history museum in Germany. Task data from Study 1 was part of a pilot study and has not been further analyzed to date. Task data from Study 2 was included in a different manuscript (Lörch & Brod, in preparation), which was not mainly focused on analyzing the CRT data. A more in-depth analysis of Study 3 task data is provided in Holstein et al. (in preparation). Data collection took place in 2023 and 2024. Parents gave written informed consent prior to testing, and ethical approval for all studies was obtained from the local ethics committee of the DIPF, Leibniz Institute for Research and Information in Education. The data and analysis code are openly available on Open Science Framework (<https://osf.io/du9e4/>).

2.1. Material and Procedure

Prior to the start of the three studies, the CRT-D was translated into German by the second author, a German native proficient in both German and English. To ensure translation accuracy, the first author, a bilingual speaker proficient in both German and English, reviewed the translation. Critical cases were discussed and resolved. The CRT-D consisted of nine questions (see Table 1), translated from the Young & Shtulman (2020b) version.

In Study 1, 47 children aged 6–9 learned about water displacement in a virtual experiment. The study was conducted as a pilot for another experiment and investigated the interaction of different degrees of generative learning activity and skills in the control-of-variables strategy on conceptual learning. In Study 2, 78 children aged 5–9 indicated how informative they found different types of comparisons for learning about water displacement before and after seeing their outcome (Lörch & Brod, in preparation). Both experiments lasted about 15 minutes, and the CRT-D was administered at the end of the procedure. In Study 3, 130 children aged 7–10 learned about balance in a prediction-feedback-learning task (see below). Data of 14 children had to be discarded because they did not meet the study's inclusion criteria, mostly due to incomplete data. The experiment lasted approximately 30 minutes, and the CRT-D was administered at the end. In the first half of study 1 and in study 3, the Hearts & Flowers Task (Wright & Diamond, 2014) was used to assess inhibition and set shifting, and the Wechsler digit span task (Petermann & Petermann, 2010) was used to assess working memory. Note that only data from Study 3 was used to investigate the relation between cognitive reflection and conceptual learning because only this study provided data that allowed us to assess deeper conceptual learning.

In Study 3 (henceforth called the conceptual learning task), children ($n = 116$, $M_{\text{Age}} = 105.28$ months, $SD = 14.66$, 45.70 % girls) learned about torque via a balance scale (Holstein et al., in preparation). Balance is a suitable domain to investigate conceptual learning because children have strong intuitive theories about balance (Siegler & Chen, 1998; van der Maas & Jansen, 2003), which often do not align with the scientific understanding of balance. That is, children typically believe that weight determines the outcome of the balance scale, i.e. the side with more weight goes down (weight theory). Other intuitive theories include: distance determines the outcome (distance theory); weight primarily determines the outcome but ties are decided by distance (weight-bias theory); distance primarily determines the outcome but ties are decided by weight (distance-bias theory); the sum of weight and distance determines the outcome (sum theory); and balancing weights by shifting them away from the fulcrum until they are equal (buggy theory). The correct concept is torque, where weight is multiplied by its location (Siegler & Chen, 1998; van der Maas & Jansen, 2003). An overview of the task is presented in Figure 1. The stimuli consisted of digital balance scale items that were based on Siegler (1976) balance scale task. The balance scale consisted of a fulcrum with an arm

that can rotate around it (held motionless by two wooden blocks). There were four pegs to the left and the right of the fulcrum and a maximum of four weights were placed on each side. Five out of the original six Siegler (1976) trial types were included in this study. In weight trials, only weight varies and the side with more weight goes down; in distance trials, only distance varies and side with the larger distance goes down; in the balance trials, weight and distance are identical and the scale remains balance; in conflict-weight trials, both weight and distance vary and the side with more weight goes down; in conflict-distance trials, both weight and distance vary and the side with a larger distance goes down.

The conceptual learning task consisted of four phases: Pretest, learning phase, posttest, and transfer test. The pretest consisted of 10 trials (without feedback, McDonald's $\omega = .80$) and was used to assess children's intuitive theories about balance. For each trial, children stated whether they thought that the left or right side of the scale would go down or whether the scale would remain balanced when the blocks were removed. In the 30 trials of the learning phase, children generated predictions about which side of the scale goes down or whether it balances and received feedback (prediction condition) or saw the outcome and the stated their expectations retrospectively (postdiction condition). The posttest (10 trials, McDonald's $\omega = .88$) was identical to the pretest; however, the position and number of weights were mirrored.

The transfer test was administered after the posttest and consisted of 13 items (McDonald's $\omega = .73$). The transfer test was designed to test deeper conceptual understanding via receptive and productive transfer of the learned concepts. The first three tasks of the transfer test assessed receptive transfer, and items were similar to those in the pretest, learning phase, and posttest. However, different items were placed on the scale as weights. Tasks four, five, and six each consisted of three items and assessed productive transfer. For task four (3 items), children were shown an adult sitting on one side of a seesaw and were asked to determine how many children needed to be placed on the other side to balance it. They were informed that three children weighed as much as one adult and that there were three possible positions for the children. For task five (3 items), children were tasked with placing blocks on the empty side of the beam to achieve the outcome that the left or right side of the scale goes down. Weights were positioned at specific locations on the other side of the beam, and children were told that they could place a maximum of three weights on the empty side. For task six (3 items), children had to determine where the fulcrum must be moved to so that the balance beam tilts to the left or right. Weights were given on both sides, and children were required to adjust the fulcrum to achieve the desired outcome. The last task assessed explicit knowledge, and children were tasked with explicitly stating the relevant dimensions for a balance scale. Due to the qualitative nature of the final item, it was not included in the analysis evaluating children's transfer test performance.

The posttest was used to assess children's learning, whereas the transfer test was used to evaluate how well children could apply what they learned to new, unpracticed scenarios. In the pretest, learning phase, and posttest, trials could be either incongruent or congruent: Incongruent trials conflicted with children's intuitive theories, while congruent trials aligned with them. Trials were selected to ensure that, given the range of intuitive theories identified from prior research, there was at least one set of incongruent trials for each child. Of primary interest for the current analyses was the transfer test performance, as this test measured deeper conceptual understanding, i.e., children's ability to use acquired knowledge in contexts beyond those presented in the learning phase.

2.2. Analyses

The analysis of the validity of the German version of the CRT-D consisted of two parts: the analysis of psychometric properties, and the analysis of developmental trends. The first part comprised the analysis of the internal consistency and the factor structure in our German sample, and the comparison of the factor structure across countries. The internal consistency indicates how well the different questions measure the same construct, and it was evaluated using McDonald's ω . The analysis of the factor structure indicates if the items indeed measure one latent factor (cognitive reflection) as expected. To assess the factor structure, we performed a confirmatory factor analysis with one latent factor and checked the overall fit of the data to that factor structure using the root mean square error of approximation (RMSEA) and chi-square goodness-of-fit tests.

Then, we compared the factor structure in our German sample with the one in samples from the US (Young & Shtulman, 2018) and China (Gong et al., 2021). For this purpose, we re-analyzed data available at <https://osf.io/yhn7b/> (Gong et al., 2021). To ensure comparability between these studies and ours, we excluded children outside the age range of our studies ($n = 6$ in both samples). The US sample ($n = 87$, $M_{\text{Age}} = 92.75$ months, $SD = 20.40$, 47.13 % girls) was recruited from public parks and administered the English CRT-D. The Chinese sample ($n = 87$, $M_{\text{Age}} = 104.04$ months, $SD = 19.49$, 50.57 % girls) was recruited from public playgrounds and administered the Chinese CRT-D. It is important to note that the version of the CRT-D used in the US and China includes a slightly different item set, with one question differing from the version we used ("What do you put in a toaster?" (correct: bread; intuitive: toast); not included). Using this data in addition to ours, we performed a multi-group confirmatory factor analysis to test for measurement invariance of the CRT-D items across countries. In other words, we investigated whether the items of the CRT-D loaded similarly onto the latent factor of cognitive reflection across the three countries. The analysis tested for configural, metric, and scalar invariance. That is, the analysis tested the similarity across countries of the factor structure, the factor loadings, and the factor loadings + intercepts, respectively. Model fit was assessed via comparative fit index (CFI), RMSEA, and standardized root mean square

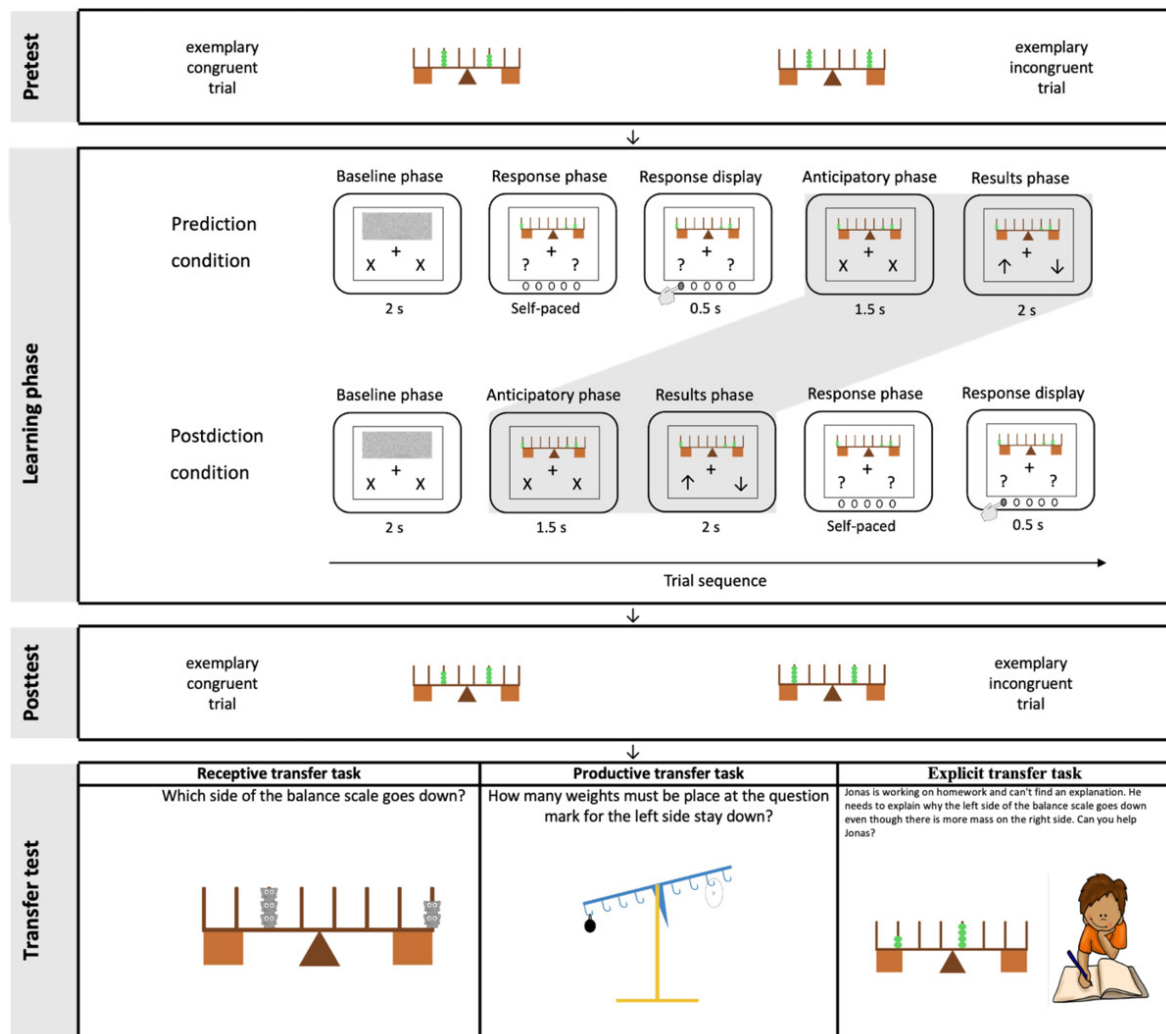


Figure 1. Schematic overview of the experimental procedures of the conceptual learning task.

The figure shows an exemplary congruent and incongruent trial in the prediction and the postdiction conditions for a child with an intuitive weight theory as well as exemplary transfer test items.

residual (SRMR). For this analysis, we included only the questions that were available in all three countries (seven questions).

To compare the developmental trend of the German CRT-D with those in studies from the US and China, we performed two linear regression models in which we predicted CRT-D scores with an interaction of age and country. The first model contained data from Germany and the US (eight questions), the second model contained data from all three countries (seven questions). We chose this approach as some questions were not used in the Chinese CRT-D. The regression models allowed us to check if the CRT-D scores increased with age in our German sample and if this trend was similar to the ones in the US and Chinese samples.

To investigate the relation between cognitive reflection and inquiry-based conceptual learning, we computed a regression model in which we predicted transfer test performance with CRT-D scores while controlling for age and executive functions.

3. Results

3.1. Psychometric Properties of the German CRT-D

Our analysis of the psychometric properties showed that German CRT-D showed acceptable internal consistency (McDonald's $\omega = .69$) and a good fit to a measurement model with one latent factor ($\chi^2(12) = 16.48$, $p < .170$, RMSEA = .038, 90% CI [0.00, 0.08]). Moreover, we found that overall, most children provided either the correct (31.72 %) or the intuitive incorrect (60.57 %) response, suggesting that the German CRT-D maintains the fundamental structure of the original CRT-D. A more detailed look at the questions revealed that two questions did not elicit the desired intuitive incorrect response in many children (see Table 1), namely questions five and seven. Excluding these questions improved internal consistency (McDonald's $\omega = .74$) and fit to the one-factor model ($\chi^2(3) = 1$, $p < .800$, RMSEA = 0, 90% CI [0.00, 0.07]). As the present study includes cross-cultural comparisons, questions five and seven were not excluded in the following analyses. However, it

Table 1. CRT-D questions with correct and intuitive incorrect answers, as well as the distribution (correct, intuitive incorrect, other) of children's responses.

Question	Correct answer	Intuitive incorrect answer	Correct	Intuitive incorrect	Other
1. If you're running a race and you pass the person in second place, what place are you in? (Wenn du an einem Wettlauf teilnimmst und du überholst die Person auf dem zweiten Platz, auf welchem Platz bist du dann?)	Second (Zweiter)	First (Erster)	21.96	74.51	3.53
2. Emily's father has three daughters. The first two are named Monday and Tuesday. What is the third daughter's name? (Emilias Vater hat drei Töchter. Die ersten beiden heißen Montag und Dienstag. Wie ist der Name der dritten Tochter?)	Emily (Emilia)	Wednesday (Mittwoch/ Wochentage)	12.55	80.78	6.67
3. A farmer has 5 sheep, all but 3 run away. How many are left? (Ein Bauer hat 5 Schafe. Alle außer 3 laufen weg. Wie viele bleiben übrig?)	3	2	38.43	56.08	5.49
4. If there are 3 apples and you take away 2, how many do you have? (Wenn es 3 Äpfel gibt und du zwei davon wegnimmst, wie viele hast du dann?)	2	1	16.86	77.65	5.49
5. What do cows drink? (Was trinken Kühe?)	Water (Wasser)	Milk (Milch)	71.37	27.84	0.78
6. What weighs more, a pound of rocks or a pound of feathers? (Was wiegt mehr, ein Kilo Steine oder ein Kilo Federn?)	Same weight (Gleich viel)	Rocks (Steine)	17.65	80.39	1.96
7. What hatches from a butterfly egg? (Was schlüpft aus dem Ei eines Schmetterlings?)	Catepillar/ Larva (Raupe/ Larve)	Baby Butterfly (Baby Schmetterling)	82.35	11.37	6.28
8. Who makes Christmas presents at the North Pole? (Wer macht die Weihnachtsgeschenke am Nordpol?)	Elves, Parents, no one (Elfen, Eltern, niemand)	Santa (Weihnachtsmann)	10.20	76.47	13.34
9. Anna is playing foursquare with her three friends: Eeny, Meeny, and Miny. Who is the fourth player? (Anna spielt Uno mit ihren zwei Freunden. Die drei Spieler heißen Ene, Mene, und?)	Anna (Anna)	Mo (Mu/Miste)	14.17	60.24	25.59

may be advisable to omit questions five and seven when conducting the German CRT-D.

To test whether the German CRT-D had a similar factor structure than the US and the Chinese CRT-D, we conducted a multi-group confirmatory factor analysis. The analysis tested for measurement invariance at the metric, configural, and scalar levels. The results showed that the configural model, which tests whether the overall factor structure is the same across countries without imposing any constraints on the factor loadings or intercepts, provided an adequate fit to the data (CFI = .909, RMSEA = .052, SRMR = .053). When comparing this configural invariance model to a metric invariance model, which constrains the factor loadings to be equal across countries, there was no significant difference in model fit ($\Delta\chi^2(12) = 13.40$, $p = .341$). This suggests that the factor loadings of the CRT-D items 1-7 were equivalent across the three countries. The metric invariance model provided an adequate fit to the data (CFI = .901, RMSEA = .047, SRMR = .064). This model is depicted in [Figure 2](#). Additionally, when comparing the

metric invariance model to the scalar invariance model, which constrains factor loading and intercepts to be equal across countries, there was a significant decline in model fit for the scalar invariance model ($\Delta\chi^2(12) = 105.5$, $p < .001$) and a high RMSEA (.233), indicating poor model fit. This suggests that the intercepts of some items differ between countries, which may reflect cultural, educational, or linguistic differences, and could indicate that the average level of cognitive reflection differs between countries. In summary, these analyses showed that the German CRT-D is reliable and measures cognitive reflection in a similar manner than the US and the Chinese CRT-D.

3.2. Developmental Trends

As a next step in our analyses, we investigated whether the German CRT-D was correlated with age and how this correlation compared to the ones found in other studies. As can be seen in [Table 2](#), CRT-D score in our German sample was significantly correlated with age ($r = .29$, $p < .001$). This association is depicted in [Figure 3](#).

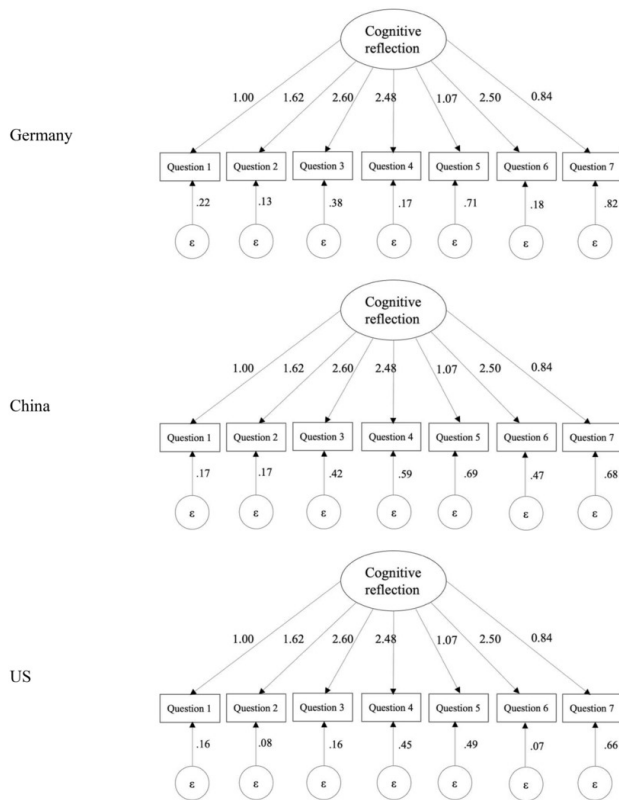


Figure 2. Multi-group confirmatory factor analysis showing the metric invariance model of the CRT-D across Germany (top), China (middle), and the US (bottom).

Factor loadings and residual variances are depicted next to the arrows. Here, the factor loadings are constrained to be equal across the countries, demonstrating the consistency of the CRT-D items 1-7 and the latent construct of cognitive reflection. This suggests that the measurement of cognitive reflection is comparable across the three countries.

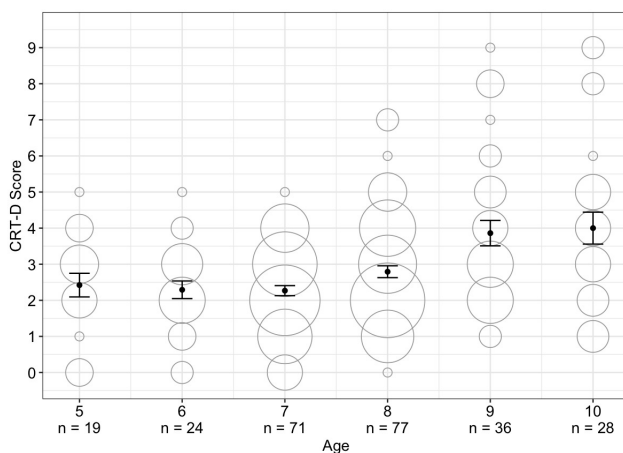


Figure 3. CRT-D score distribution by age (in years).

Circle size is proportionate to the number of participants for a given CRT-D score at each age. Black dots represent means. Error bars indicate standard deviations.

Moreover, in line with our expectations, developmental trends were comparable to those observed in other countries. CRT-D scores were compared for available questions in each country, and for the age range 50 to 132 months.

When comparing the German CRT-D with the US CRT-D (questions 1–8), we examined how CRT-D score changed with age in both countries. The results showed that CRT-D scores significantly improved with age ($b = 0.17$, standardized $\beta = .17$, $SE = 0.06$, $p = .008$; see Figure 4). This suggests that, in both Germany and the US, the number of correct responses increased with age. Further, when comparing CRT-D scores across Germany, China and the US (questions 1–7), CRT-D scores significantly improved with age ($b = 0.17$, standardized $\beta = .17$, $SE = 0.06$, $p = .006$; see Figure 4). This indicates that age was associated with higher CRT-D scores across all three countries. Additionally, there was a main effect for being from the US ($b = -0.47$, standardized $\beta = -.47$, $SE = 0.12$, $p < .001$), indicating that children from the US performed worse than their German counterparts. Furthermore, the interaction term between age and being from China was significant ($b = 0.36$, standardized $\beta = .36$, $SE = 0.11$, $p = .002$), suggesting that the positive effect of age on CRT-D score was stronger for children from China than for children from Germany. A table with all parameter estimates can be found in Appendix A. The similarity of the developmental trend in our German sample with the ones in samples from the US and China provides further evidence for the validity of the German CRT-D.

3.3. The Relationship between CRT-D and Conceptual Learning

The main question of the present paper was whether CRT-D predicts inquiry-based conceptual learning. Inquiry-based learning was operationalized via an inquiry-based learning session, and a transfer task within a balance scale paradigm. The transfer task, in particular, was designed to assess deeper conceptual understanding and the ability to apply learned knowledge to novel contexts.

Children's knowledge of balance improved substantially across the task, as indicated by a strong increase in the percentage of responses consistent with the torque theory (pretest: 3.98 %, $SD = 8.81$; posttest: 24.44 %, $SD = 32.72$). On the transfer test, children scored an average of 6.48 ($SD = 2.27$) out of 12, indicating that a substantial portion of the children gained a deeper understanding of balance. Performance improved with age on all tests (pretest: $r(114) = .21$, $p = .021$, posttest: $r(114) = .28$, $p = .003$, transfer test: $r(114) = .40$, $p < .001$), indicating developmental progress in conceptual understanding.

To get an overview of correlations between the main variables of the present study, Table 2 provides correlations between CRT-D scores, age, executive functions, and conceptual learning. We found that CRT-D scores correlated with pretest performance on the conceptual learning task ($r(114) = .19$, $p = .040$). This finding is consistent with previous studies that found a correlation between cognitive reflection and the level of conceptual knowledge.

To investigate whether CRT-D scores correlated with posttest performance while controlling for pretest scores, residuals from a linear regression model predicting posttest performance from pretest scores were calculated, as the residuals reflect posttest performance independent of

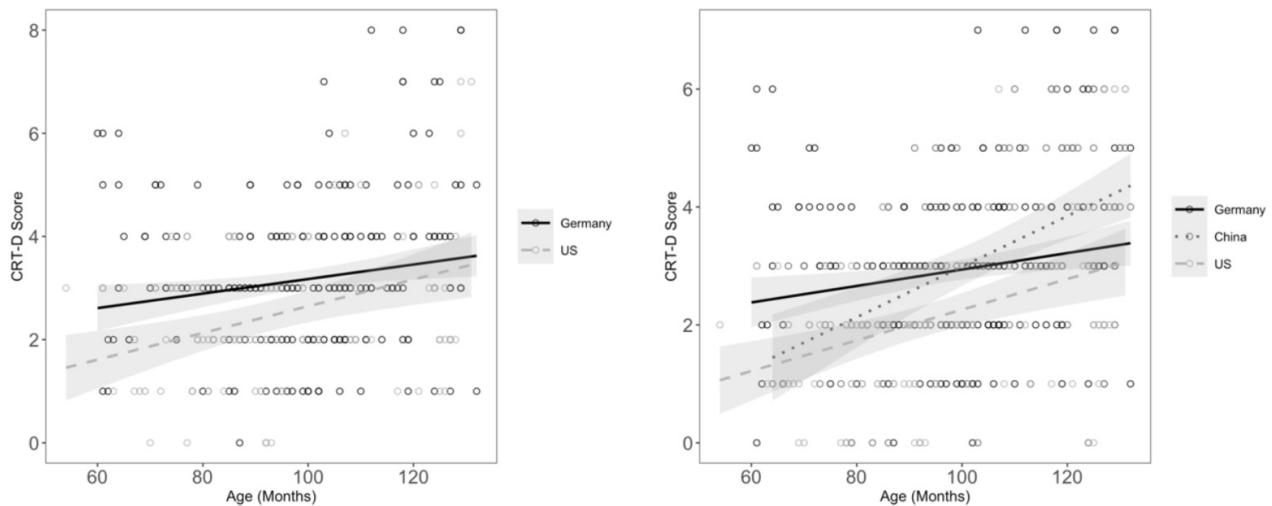


Figure 4. Developmental trends between Germany and the US (left) and Germany, China, and the US (right).

Eight out of nine questions were included for the comparison between Germany and China, whereas seven questions were included for Germany, China, and the US.

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7
1. CRT-D (<i>N</i> = 255)	1.32	1.50							
2. Age (months) (<i>N</i> = 255)	98.18	16.97	.29*** [0.17, 0.41]						
3. Inhibition (<i>n</i> = 149)	-0.41	0.60	.21* [0.05, 0.37]	.20* [0.03, 0.34]					
4. Set shifting (<i>n</i> = 149)	-1.16	0.58	.12 [-0.04, 0.29]	.01 [-0.16, 0.18]	.48*** [0.34, 0.61]				
5. Backward digit span (<i>n</i> = 152)	6.00	1.75	.32*** [0.17, 0.45]	.42*** [0.28, 0.56]	.22** [0.07, 0.37]	-.01 [-0.18, 0.16]			
6. Pretest: Conceptual understanding (<i>n</i> = 116)	3.98	8.81	.19* [0.00, 0.35]	.21* [0.03, 0.40]	.12 [-0.06, 0.31]	.17 [0.00, 0.35]	.13 [-0.05, 0.31]		
7. Posttest: Conceptual understanding (<i>n</i> = 116)	24.44	32.72	.11 [-0.06, 0.29]	.28** [0.10, 0.45]	.10 [-0.11, 0.28]	.06 [-0.13, 0.25]	.07 [-0.11, 0.24]	.37*** [0.19, 0.53]	
8. Transfer test (<i>n</i> = 116)	6.48	2.27	.47*** [0.32, 0.61]	.40*** [0.23, 0.55]	.13 [-0.04, 0.30]	-.02 [-0.22, 0.18]	.37*** [0.20, 0.54]	.29** [0.12, 0.45]	.32*** [0.15, 0.49]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Table 2. Means, standard deviations, and correlations with confidence intervals.

pretest performance. The residuals were then correlated with CRT scores, but this correlation was not significant ($r(114) = .01, p = .931$).

Most importantly, to analyze if CRT-D scores predicted the transfer of conceptual knowledge to new tasks—our main indicator of deeper understanding following inquiry-based learning—we computed a regression model in which transfer test performance was predicted by CRT-D score, pretest score, age, and executive functions. This analysis showed that, as expected, cognitive reflection was related to conceptual learning even after controlling for pretest performance, age, and executive functions ($b = 0.35$, stan-

dardized $\beta = .31, SE = 0.10, p = .001$), suggesting a robust link between cognitive reflection and successful conceptual learning through inquiry. Figure 5 depicts this association. A table with the full information on the regression model can be found in Appendix A.

4. Discussion

Cognitive reflection is considered an important skill for learning counter-intuitive science concepts. A previous study by Young and Shtulman (2020b) found a correlation between cognitive reflection and conceptual learning from

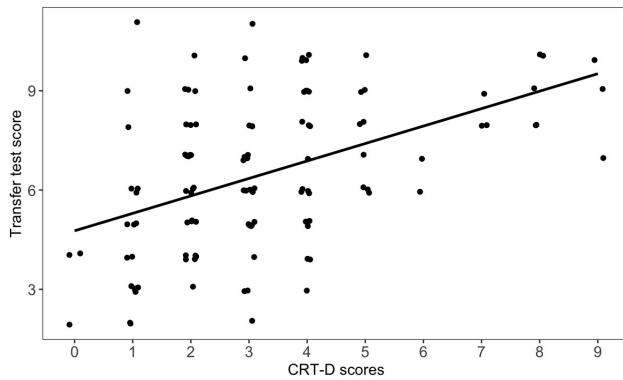


Figure 5. Raw correlation between CRT-D scores and transfer test performance.

The regression line indicates the linear trend, with a correlation of $r(114) = .47, p < .001$. This plot reflects the statistical association between cognitive reflection and transfer test performance without adjusting for pretest, age, or executive functions.

direct instruction in children. We extend this finding by demonstrating the importance of cognitive reflection skills for children's successful inquiry-based learning. In particular, our study demonstrates that cognitive reflection skills promote a deeper understanding of a scientific concept and that their importance is maintained even when executive functions are controlled.

The first major finding of our study is that our German CRT-D is a valid measure of cognitive reflection. Across three studies, German children similarly fell for intuitive incorrect answers as has been shown for Chinese and US children. The factor loadings of the CRT-D items were equivalent across the three countries, which suggests that the test measures similar properties across the countries. Moreover, German children also displayed a similar age-related increase in CRT-D scores. All countries showed a positive correlation between age and CRT-D scores; differences in the strength of this association could be due to differences in sampling procedure across the three studies. Cognitive reflection in our study followed a general developmental trajectory that has also been observed for related cognitive and metacognitive abilities (Diamond, 2013; Zelazo, 2015). Moreover, the CRT-D appears to be robust in measuring cognitive reflection across different cultural groups.

The second major finding of our study is that cognitive reflection was related to inquiry-based conceptual learning. We tested children's learning in the domain of balance with a balance scale task that included a pretest, posttest, and transfer test. During the learning phase, children predicted which side of the balance scale would go down when the blocks were removed and received feedback which allowed them to draw inferences about balance. In the transfer test, children had to apply their conceptual understanding about balance to a number of new tasks. In short, we found that the CRT-D score predicted transfer test performance even when controlling for pretest score, age, and executive functions. These findings suggest that cognitive reflection is important for inquiry-based conceptual learning.

Of note, we found the CRT-D score to be correlated with pretest performance, which is in line with previous studies that found cognitive reflection to be related to the level of conceptual knowledge (Gómez-Chacón et al., 2014; Shtulman & McCallum, 2014; Young & Shtulman, 2020a). However, it did not correlate with posttest performance (when controlling for pretest performance), which contrasts with the correlation found for the transfer test. While clearly speculative, this finding suggests that cognitive reflection is particularly important for deeper conceptual understanding and conceptual revision, which was required in the transfer test but not in the posttest. Specifically, cognitive reflection appears to support children most when they need to apply their understanding in a novel context (transfer), whereas performance on the posttest—which resembled items in the learning phase—can be achieved with less reflective effort. Alternatively, the learning phase might have attenuated differences between children with high and low cognitive reflection skills, but those differences might have reemerged when given the more demanding transfer task.

One rather surprising finding was that cognitive reflection was significantly correlated only with working memory and inhibition, but not with set shifting. This is in contrast with findings from Young and Shtulman (2020a), who reported correlations between CRT-D scores and inhibition, set shifting, and working memory. A possible reason for this difference could lie in tasks differences. Young & Shtulman (2020a) used the Toolbox Flanker Inhibitory Control and Attention Test to assess inhibition, and a verbal fluency task to assess set shifting. In contrast, we used the Hearts & Flowers Task (Wright & Diamond, 2014) to assess both inhibition and set shifting. While these tasks assess the same executive functions, they differ slightly in their focus. The Flanker Test assesses inhibitory control and attentional focus in response to neutral distractors; it measures a child's ability to inhibit automatic responses to irrelevant information and focus attention on the relevant target (Zelazo et al., 2013). The Hearts & Flowers Task, on the other hand, assesses response inhibition and impulse control; it evaluates a child's ability to suppress automatic responses to conflicting stimuli (Wright & Diamond, 2014). Additionally, verbal fluency is often regarded as a reflection of self-directed cognitive flexibility, as it involves both internal set shifting and proactive executive planning (Munakata et al., 2012), whereas the Hearts & Flowers Task involves shifting between response sets, which requires mental flexibility to switch between rules (Wright & Diamond, 2014). These differences could explain the lack of correlation between CRT-D scores and set shifting in our data. Additionally, the differences in the age range in this study compared to Young & Shtulman (2020a; 5 to 10 vs. 5 to 12 years) could explain this lack of correlation. Older children tend to perform better on both the CRT-D (Young & Shtulman, 2020a) and executive function tasks (Wright & Diamond, 2014), which likely increases variability in their data. The narrower age range in our sample may have reduced this variability, making correlations less detectable. Taken together, our results suggest that while cognitive reflection is linked to working memory, the relationship with other executive functions

may be more nuanced and influenced by age and task-specific factors.

4.1. Limitations

This study has several limitations that should be considered when interpreting the findings. First, the sample was not a representative one, as data was collected in a science museum, which may attract participants with particular interests or demographics that differ from the general population. Additionally, the age range of participants in the German sample (5–10 years) was narrower than that in other countries such as China and the United States, where the age range extended from 5 to 12 years. This discrepancy may impact the generalizability of findings, particularly regarding developmental trends across a broader age span.

Additionally, due to cultural and linguistic differences across countries some items did not elicit the desired intuitive response consistently in our German sample. Consequently, the number of usable questions varied by country, complicating the assessment of average cognitive reflection levels across different populations. This limitation is important to consider when interpreting findings from cross-cultural studies. In this study, we limited the cross-cultural analysis to the questions available in all countries. However, future studies could consider fine-tuning omitted questions to better account for cultural and linguistic nuances in each country, ensuring greater consistency across diverse populations.

4.2. Conclusions and Future Directions

Taken together, this study indicates that the CRT-D is a valuable tool for predicting children's real-world conceptual learning. Future research is needed to better understand the mechanisms by which cognitive reflection influences conceptual learning. Incorporating process data such as response times or neural activity could allow for a deeper understanding of where children with higher CRT-D skills differ in learning from children with lower CRT-D skills. With the help of this data, researchers could further develop more comprehensive assessments of children's cognitive abilities and tailor interventions to support their cognitive and academic development. Future research could also benefit from exploring how the CRT-D relates to emerging areas like critical online reasoning. Given the increasing importance of digital literacy and critical thinking in online environments, examining the connection between cognitive reflection and online reasoning skills could offer valuable insights into how these cognitive abilities intersect in the context of information-rich, digital settings.

The CRT-D could also be valuable for educators. Its high face validity makes it a suitable instrument for teaching educators about the importance of cognitive development for children's learning success. Moreover, educators could use problems similar to those in the CRT-D to teach children to question their intuitive responses. Afterwards, discussing the results could emphasize how the children might have been misled by intuition and how more careful reflection could have led to the correct answer. Overall, the CRT-D has great potential not only as a research tool for understanding cognitive reflection and conceptual learning, but also as a practical tool for developing children's critical thinking skills.

CRediT Author Statement

EH: Conceptualization, Resources, Investigation, Formal Analysis, Data Curation, Writing – Original Draft.

LL: Conceptualization, Resources, Investigation, Writing – Review & Editing.

GB: Conceptualization, Funding acquisition, Supervision, Writing – Review & Editing.

Funding Information

This research was funded by a grant of the German Research Foundation to GB.

Ethics Statement

All research involving human participants was approved by the authors' institutional ethics committee and conducted in accordance with the principles outlined in the Declaration of Helsinki.

Competing Interests

The authors declare no competing interests.

Data Accessibility Statement

The data and the analytic code necessary to reproduce the analyses as well as the materials are publicly available at the following URL: <https://osf.io/du9e4/>.

Editors: Jonathan Kominsky (Associate Editor)

Submitted: April 07, 2025 PDT. Accepted: July 31, 2025 PDT.

Published: September 17, 2025 PDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

References

- Carriedo, N., Corral, A., Montoro, P. R., & Herrero, L. (2020). A developmental study of the bat/ball problem of CRT: How to override the bias and its relation to executive functioning. *British Journal of Psychology*, 111(2), 335–356. <https://doi.org/10.1111/bjop.12400>
- Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, 64(1), 135–168. <https://doi.org/10.1146/annurev-psych-113011-143750>
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Gette, C. R., & Kryjevskaja, M. (2019). Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses. *Physical Review Physics Education Research*, 15(1), 010118. <https://doi.org/10.1103/PhysRevPhysEducRes.15.010118>
- Gómez-Chacón, I. M., García-Madruga, J. A., Vila, J. Ó., Elosúa, M. R., & Rodríguez, R. (2014). The dual processes hypothesis in mathematics performance: Beliefs, cognitive reflection, working memory and reasoning. *Learning and Individual Differences*, 29, 67–73. <https://doi.org/10.1016/j.lindif.2013.10.001>
- Gong, T., Young, A. G., & Shtulman, A. (2021). The Development of Cognitive Reflection in China. *Cognitive Science*, 45(4), e12966. <https://doi.org/10.1111/cogs.12966>
- Holstein, E., Theobald, M., Colantonio, J., Bonawitz, E., & Brod, G. (2025). *Generating predictions promotes complex, conceptual learning*.
- Johnson, E. D., & Tubau, E. (2014). The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36.
- Lörch, L., & Brod, G. (2025). *Does positive testing depend on the strength of children's prior beliefs?*
- Meyer, A., Attali, Y., Bar-Hillel, M., Frederick, S., & Kahneman, D. (2024). Cognitive reflection is a distinct and measurable trait. *Proceedings of the National Academy of Sciences*, 121(49), e2409191121. <https://doi.org/10.1073/pnas.2409191121>
- Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing Cognitive Control: Three Key Transitions. *Current Directions in Psychological Science*, 21(2), 71–77. <https://doi.org/10.1177/0963721412436807>
- Petermann, F., & Petermann, U. (Eds.). (2010). *Hamburg-Wechsler-Intelligenztest für Kinder - IV: HAWIK-IV; Manual* (3., erg. Aufl. Huber [u.a.]).
- Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, 43(5), 1124–1139. <https://doi.org/10.1037/0012-1649.43.5.1124>
- Shtulman, A., & McCallum, K. (2014). Cognitive Reflection Predicts Science Understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36. <https://escholarship.org/uc/item/4t79p8pj>
- Shtulman, A., & Young, A. G. (2022). The development of cognitive reflection. *Child Development Perspectives*, cdep.12476. <https://doi.org/10.1111/cdep.12476>
- Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, 8(4), 481–520. [https://doi.org/10.1016/0010-0285\(76\)90016-5](https://doi.org/10.1016/0010-0285(76)90016-5)
- Siegler, R. S., & Chen, Z. (1998). Developmental Differences in Rule Learning: A Microgenetic Analysis. *Cognitive Psychology*, 36(3), 273–310. <https://doi.org/10.1006/cogp.1998.0686>
- Stupple, E. J. N., Gale, M., & Richmond, C. R. (2013). Working Memory, Cognitive Miserliness and Logic as Predictors of Performance on the Cognitive Reflection Test. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7), 1275–1289. <https://doi.org/10.3758/s13421-011-0104-1>
- van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, 85(2), 141–177. [https://doi.org/10.1016/S0022-0965\(03\)00058-4](https://doi.org/10.1016/S0022-0965(03)00058-4)
- Vosniadou, S., Pnevmatikos, D., Makris, N., Lepenioti, D., Eikospentaki, K., Chountala, A., & Kyrianakis, G. (2018). The Recruitment of Shifting and Inhibition in On-line Science and Mathematics Tasks. *Cognitive Science*, 42(6), 1860–1886. <https://doi.org/10.1111/cogs.12624>
- Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00213>
- Young, A. G., & Shtulman, A. (2018). Developing A Cognitive Reflection Test for School-Age Children. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 1238–1243.
- Young, A. G., & Shtulman, A. (2020a). Children's Cognitive Reflection Predicts Conceptual Understanding in Science and Mathematics. *Psychological Science*, 31(11), 1396–1408. <https://doi.org/10.1177/0956797620954449>
- Young, A. G., & Shtulman, A. (2020b). How Children's Cognitive Reflection Shapes Their Science Understanding. *Frontiers in Psychology*, 11, 1247. <https://doi.org/10.3389/fpsyg.2020.01247>
- Young, A. G., & Shtulman, A. (2024). Children's cognitive reflection predicts successful interpretations of covariation data. *Frontiers in Developmental Psychology*, 2, 1441395. <https://doi.org/10.3389/fdpys.2024.1441395>

Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. *Developmental Review*, 38, 55–68. <https://doi.org/10.1016/j.dr.2015.07.001>

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH TOOLBOX COGNITION BATTERY (CB): MEASURING EXECUTIVE FUNCTION AND ATTENTION. *Monographs of the Society for Research in Child Development*, 78(4), 16–33. <https://doi.org/10.1111/mono.12032>

Appendix A

Table A1. Parameter estimates of the regression model that predicted CRT-D score with an interaction of age and country (Germany/US). Age is in month and scaled. The reference level for country was Germany.

	Estimate	SE	t
Intercept	0.10	0.06	1.73
Age	0.17	0.07	2.65**
Country (US)	-0.38	0.12	-3.12**
Age*Country (US)	0.14	0.11	1.28

Note. * $p < .05$; ** $p < .01$

Table A2. Parameter estimates of the regression model that predicted CRT-D score with an interaction of age and country (Germany/US/China). Age is in month and scaled. The reference level for country was Germany.

	Estimate	SE	t
Intercept	0.08	0.06	1.37
Age	0.17	0.06	2.76**
Country (China)	0.002	0.12	0.02
Country (US)	-0.47	0.12	-3.99***
Age*Country (China)	0.36	0.11	3.16**
Age * Country (US)	0.15	0.11	1.36

Note. * $p < .05$; ** $p < .01$, *** $p < .001$

Table A3. Parameter estimates of the regression model that predicted transfer test performance with CRT-D score and age while controlling for pretest, age, and executive functions.

	Estimate	SE	t
Intercept	0.41	1.50	0.27
CRT-D score	0.35	0.10	3.37**
Pretest residuals	0.01	0.02	0.57
Age	0.03	0.02	2.24*
Inhibition	0.35	0.33	1.06
Switching	-0.31	0.37	-0.85
Working memory	0.19	0.12	1.49

Note. * $p < .05$; ** $p < .01$

Supplementary Materials

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/143423-children-s-cognitive-reflection-skills-predict-their-conceptual-learning/attachment/298699.docx?auth_token=fuutBU0SOAVd1g5jr3j-
