From outcome to process: A developmental shift in judgments of good reasoning

Hanna Schleihauf a, b, c, *, 1, Zhen Zhang d, e, *, 1, Alissa Gomez f, Jan M. Engelmann g

a Department of Developmental Psychology, Utrecht University, the Netherlands
b Cognitive Ethology Laboratory, German Primate Center – Leibniz Institute for Primate Research, Göttingen, Germany
c Leibniz ScienceCampus Primate Cognition, 37077 Göttingen, Germany
d CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
e Department of Psychology, University of Chinese Academy of Sciences, Beijing 100190, China
f Department of Psychology, University of California, Berkeley, CA, USA
g Social Origins Laboratory, Department of Psychology, University of California, Berkeley, CA, USA

Keywords:
Outcome
Process
Judgment of good reasoning
Cross-cultural developmental differences
China
United States

ABSTRACT

What does it mean to reason well? One might argue that good reasoning means that the outcome of the reasoning process is correct: reaching the right belief. Alternatively, good reasoning might refer to the reasoning process itself: following the right epistemic procedures. In a preregistered study, we investigated children’s (4–9-year-olds) and adults’ judgments of reasoning in China and the US (N = 256). Participants of all age groups evaluated the outcome when the process was kept constant - they favored agents who reached correct over incorrect beliefs, and they evaluated the process when the outcome was kept constant – they preferred agents who formed their beliefs using valid over invalid procedures. Developmental changes emerged when we pitted outcome against process: young children weighed outcome more heavily than process; older children and adults showed the reverse preference. This pattern was constant across the two cultural contexts, with the switch from outcome to process happening earlier in development in China. These results suggest that children initially value what someone believes, but, with development, come to increasingly value how beliefs are formed.

Judgments about good and bad reasoning are key components of everyday discourse. Whether on Twitter, in town hall meetings, or over dinner with friends, our reasoning can be the target of both approval – “Great point! I see where you are coming from.” – and criticism – “It is so irrational of you to believe that...!” Humans routinely tell each other what they should believe and why. This behavior emerges early – preschool-aged children already evaluate others’ conclusions, accepting correct knowledge claims and criticizing incorrect ones (Fedra & Schmidt, 2018; Mioga, Schultz, Schulz-Hardt, & Rakoczy, 2020). But what, exactly, does it mean to reason well?

At a basic level, one can distinguish between outcome-focused and process-focused evaluations of reasoning (Kolodny, 2005). We can, and frequently do, evaluate both the accuracy of an outcome (whether someone’s belief is correct) and the validity of a process (whether someone formed their belief using an appropriate procedure). Ideally, of course, the accuracy of an epistemic outcome follows from the validity of process. When we inspect the entire body of evidence, weigh arguments on both sides appropriately (process), we should arrive at accurate beliefs (outcome). The problem is that we rarely, if ever, find ourselves in ideal epistemic circumstances: relevant data might be lacking, evidence might be difficult to interpret or deluding, or the optimal solution might simply be beyond our cognitive abilities (the so-called problem of intractability; Arkes, Gigerenzer, & Hertwig, 2016). So accurate outcome and valid process do not entail one another. A valid process can result in inaccurate outcomes, the available evidence may mislead and a usually reliable informant might misinform. Likewise, an invalid process might generate accurate outcome, wishful thinking might match reality and flipping a coin to decide whether A or B are true might, purely by chance, get it right.

When we evaluate others’ reasoning, how do we weigh outcome and process? While adults can clearly assess the validity of deliberative process in itself (König-Kersting, Pollmann, Potters, & Trautmann, 2021;
Levy, 2022; Mercier, 2020), process often takes a backseat to outcome when they are both available for scrutiny. Adults from Western societies exhibit a well-documented – and robust – outcome bias: focusing on outcome information, even when good process data is available. In these studies, participants are usually asked to evaluate two agents who used the same procedure to make a decision, but only one agent, through sheer luck, made a good decision (Brownback & Kuhn, 2019). The outcome bias even emerges in experimental contexts in which participants are explicitly told to evaluate the quality of someone’s thinking (Baron & Hershey, 1988). One open research question is who do adults prefer when outcome and process conflict: an agent who uses a valid procedure, but reaches a wrong outcome (i.e., forms an incorrect belief) or an agent who uses an invalid procedure, but reaches a right outcome (i.e., forms a correct belief).

Less is known about how children consider outcome and process when forming judgments of good reasoning. Do children, like adults, show an outcome bias and evaluate someone’s reasoning more positively if they arrive at a correct rather than an incorrect belief (outcome)? While we know of no study that directly tested this question, relevant indirect evidence comes from investigations of children’s selective social learning. In these studies, children are typically presented with two informants who vary along a relevant dimension (e.g., prior reliability) and are asked the endorse the testimony of one of them (for a review, see Harris, Koenig, Corriveau, & Jaswal, 2018). From a young age, children preferentially attend to and learn from more accurate and knowledgeable informants, which can be interpreted as a preference for individuals who hold correct beliefs, i.e., reached a correct outcome (Butler, Gibbs, & Tavassolie, 2020; Harris et al., 2018; Hermes, Behne, Bich, Thiebert, & Rakoczy, 2017; Koenig, Tiberius, & Hamlin, 2019; Oved, Heyman, & Barner, 2014).

In terms of process-focused evaluations, previous literature provides evidence that, starting from around age 4, children judge beliefs differently as a function of how they are formed. At this age, children selectively trust beliefs that are supported by strong reasons over beliefs that are supported by weak or circular reasons (Corriveau & Kurkul, 2014; Langenhoff, Engelmann, & Srivivasan, 2022; Mercier, Bernard, & Clément, 2014; Schleinhauf, Herrsch, Fischer, & Engelmann, 2022). Relatedly, 4- and 5-year-old children are more likely to forgive incorrect beliefs (outcomes) if the informants had access only to ambiguous evidence (in which case even a valid process would lead to wrong outcomes, Kondraad & Jaswal, 2012). And 6- and 7-year-olds (4- and 5-year-olds to a lesser extent) selectively endorse the testimony of an agent who used a valid epistemic procedure (looking inside a box to determine its content) compared to an agent who applied an invalid epistemic procedure (looking only at the box) (Butler, Schmidt, Tavassolie, & Gibbs, 2018, 2020). In addition, by age 5, children provide valid reasons for their own claims, thereby communicating their own thinking process to others (Köyken & Engelmann, 2022; Köyken, Mammen, & Tomasello, 2016; Köyken, Rosenbaum, & Tomasello, 2014). These results suggest that starting from age 4, but at least by age 6, children are sensitive to the validity of an epistemic process.

A small number of previous studies have investigated how children evaluate reasoning when both outcome- and process-related information are available. Butler et al. (2020, see also Brosseau-Liard and Birch, 2010) gave 4- to 7-year-old children the option of learning either from an agent who had proven reliable in the past (in an object labeling task) but failed to acquire relevant information in a novel task (determining the location of a hidden object) or from an agent who had proved unreliable in the past but acquired relevant information in the novel task. When faced with this conflict, 6- to 7-year-old (but not 4- to 5-year-old) children were slightly above chance in choosing the previously unreliable informant who gathered sufficient information. These results raise the possibility that children younger than 6 years weigh outcome and process similarly, while older children consider process as more important than outcome. However, the goal of this study was not to compare the relative influence of process and outcome on children’s evaluation of reasoning, but rather to determine how past reliability (on an unrelated task) and present epistemic practices interactively influence children’s trust judgments. Finally, and most relevant to the current investigation, Dombreg, Köyken, and Tomasello (2019) studied whether 5- and 7-year-old children evaluate an agent who uses a valid epistemic procedure (selectively endorsing the testimony of an informant who provided a good reason) more positively than an agent who uses an invalid epistemic procedure (endorsing the testimony of an informant who provided a bad reason), even if both agents arrive at the wrong conclusion (outcome). The results show that 7-year-old children, and 5-year-olds to a lesser extent, selectively choose the agent using a valid epistemic procedure as a cooperative partner.

Prior studies thus present suggestive evidence that, from age 4, children evaluate more positively agents who hold true rather than false beliefs (outcome), and who use valid rather than invalid epistemic procedures (even if both procedures result in wrong outcomes). However, to date no study has systematically investigated how epistemic outcome and process influence judgments of good reasoning across development. In addition, prior research on the development of judgments of reasoning has focused nearly exclusively on participants from WEIRD (Western, Educated, Industrialized, Rich, Democratic) backgrounds (Henrich, Heine, & Norenzayan, 2010; Nielsen, Haun, Kättner, & Legare, 2017). Whether the relative importance of outcome and process in determining sound reasoning varies across different cultural backgrounds is not known. Prior research on cross-cultural epistemology has argued that individuals from China and the US – besides many similarities – also show subtle differences in the value placed on certain epistemic practices in their everyday life (Wellman, Fang, Liu, Zhu, & Liu, 2006). Generally speaking, while Western epistemology seems to focus on truth and belief (outcome), Chinese epistemology focuses on “pragmatic knowledge acquisition” (process) (Li, 2002; Nisbett, 2003; Wellman et al., 2006). This is reflected, for example, in how the environment is explored and perceived. As Markus and Kitayama (1991, p. 246) note: “If one perceives oneself as embedded within a larger context of which one is an interdependent part, it is likely that other objects or events will be perceived in a similar way”. While people from North America seem to show stronger attention, relatively speaking, to objects and object features, people from East Asia seem to attend instead to relations and relational structures (Carstensen et al., 2019; Chiu, 1972; Nisbett, Peng, Choi, & Norenzayan, 2001). This difference has bearing on the current investigation. Appreciating the special relevance of process in judgments of reasoning requires understanding the relationship between a given process and a given outcome. Thus, it is possible that Chinese (compared to US American) participants put greater emphasis on the evaluation of process than the evaluation of outcome when judging others’ reasoning.

In sum, prior research leaves open (1) how the validity of outcome and of process, as well as conflicts between outcome and process, influence judgments of good reasoning; (2) whether the weighing of outcome and process changes across development; and, finally, (3) whether certain patterns in evaluating reasoning are robust across cultural contexts.

These questions are the focus of the present studies. Participants were presented with a series of stories in which two agents attempted to determine the location of a missing animal (see Fig. 1). In the outcome condition, one agent correctly inferred the animal’s location whereas the second agent made an incorrect claim. In the process condition, one individual used a valid procedure (e.g., searching for evidence) and one agent used an invalid procedure (e.g., flipping a coin) to determine the animal’s whereabouts. Finally, in the process-vs-outcome condition, one agent used an invalid procedure and arrived at the right conclusion, while the second agent used a valid procedure and arrived at the wrong conclusion (because the available evidence was misleading). Following these stories, participants were asked two questions: “Who do you think is doing a better job?” (performance evaluation) and “Who would you ask for help if you lost something of value?” (partner choice). We investigated
these questions in a Chinese and a US developmental sample (4–5-year-olds, 6–7-year-olds, 8–9-year-olds, adults).

We hypothesized a developmental trend from outcome- to process-based evaluations of good reasoning. Based on the above-mentioned findings, we also predicted an earlier switch from outcome to process in the Chinese compared to the US children, potentially even leading to a reduced outcome-bias in Chinese compared to US-American adults.

1. Methods

1.1. Participants

We tested four age groups (4–5-year-olds, 6–7-year-olds, 8–9-year-olds, and adults, see Table S1) in China and the United States. The required sample size ($N = 256$, 32 individuals per age group) was determined with a power simulation, expecting a significant two-way interaction effect of the factors condition and age group. This sample size led to an average power of $1 - \beta = 0.75$ (for details see preregistration). To reach this predetermined sample size, we tested 264 individuals. Data of one 5-year-old (US), one 7-year-old (US), and one 8-year-old (China) were excluded due to experimenter errors or children’s unwillingness to participate. Data of five adults (US) were excluded because participants started, but did not finish, the online survey.

We tested six trials per individual (1536 trials in total). For the statistical analysis of the first dependent variable (performance evaluation), we had to exclude eleven trials of seven children (five 8-year-olds and two 9-year-olds from China), because they did not want to choose any of the characters. Thus, 1525 trials were included in the final sample. For the statistical analysis of the second dependent variable (partner choice), we had to exclude twelve trials of ten children (four 8-year-olds and one 9-year-old from China, three 5-year-olds, one 7-year-old and one 9-year-old from the US). Thus, in total, 1524 trials were included in the final sample.

Most Chinese children and adults came from Beijing or surrounding areas and were recruited over the communication software WeChat. Children from the United States came from the San Francisco Bay Area and were recruited through a database to which they were assigned following their parents’ written consent. Adult participants came from all over the United States and were recruited over the crowdsourcing marketplace MTurk.

In the Chinese sample, participants spoke Mandarin as their native language. Parents of participating children had on average high levels of education (18% had a doctorate, 42% a master’s degree, 39% a bachelor’s degree, 1% a high school degree). Information on annual household income was not collected; however, the high education level of the parents indicates at least medium socioeconomic backgrounds. All participating children were Asian (95% Han, 3% Mongolian, 2% Manchu). The majority (84%) of participating adults were undergraduate and graduate students without income. The other participants came from medium to high educational backgrounds (7% had a master’s degree, 9% a bachelor’s degree) and had low to medium socioeconomic status (annual household income: 4% between $47,000 and $55,000, 6% between $14,000 and $18,000, 3% between $4500 and $8000, 3% less than $1500). All participating adults were Asian (94% Han ethnicity, 3% Manchu, 3% Uygur).

Fig. 1. Example stimuli for each condition with shortened instructions. Participants saw two different stories per condition.
In the US sample, all participants spoke English. The education level of the parents of the participating children was on average high (19% had a doctorate, 33% a master’s degree, 34% a bachelor’s degree, 9% a high school degree, and 5% did not finish high school). Children came from medium to high socioeconomic backgrounds (annual household income: 43% above $200,000, 34% between $100,000 and $200,000, 9% between $80,000 and $100,000, 9% between $60,000 and $80,000, 5% less than $60,000) and were of mixed descent (40% White, 27% Asian, 17% Asian/White, 10% Hispanic, and 6% African American). Participating adults came from mixed education backgrounds (2% had a doctorate, 6% a master’s degree, 64% a bachelor’s degree, 28% a high school degree) and various socioeconomic backgrounds (annual household income: 5% above $200,000, 5% between $100,000 and $200,000, 8% between $80,000 and $100,000, 17% between $60,000 and $80,000, 50% less than $60,000). Most of them were White (81% White, 5% Asian, 8% Hispanic, and 6% African American).

1.2. Materials and procedure

Child participants in both cultures were tested online via the video-communication software Zoom by a native speaking experimenter. All sessions were recorded. The procedure for adult participants was identical, with minor modifications. To participate, adults followed a link to a Qualtrics survey. They were then instructed that they would see a series of videos and answer questions about what they had seen. In these videos, the voice of a pre-recorded experimenter led them through the slides as it was done for the children. To respond, participants could click on one of two potential answers or type in their response. The stimuli were first created in English, then translated to Mandarin, and back translated to English to check accuracy.

Participants were presented with picture-book like stories on PowerPoint slides. Each participant saw 6 stories in total: 2 stories in the outcome condition, 2 stories in the process condition, and 2 stories in the process-vs-outcome condition. The order of the stories was counterbalanced. All stories started with the introduction of two agents (always of the same gender within one story), whose pet had run away. The content of the stories differed according to condition. See Fig. 1 for an overview.

In the outcome condition, participants were presented with two possible hiding locations (e.g., a pile of grey stones and a pile of red stones) and a pet hiding behind one of them (e.g., a bunny sitting behind the red stones). Then, both agents tried to find their pet using the same invalid procedure (using a spinning wheel or a counting-out rhyme [dian bing dian jiang for Chinese participants; eeyy, meeny, miny, moe for US participants]). Following this procedure, one agent chose the correct hiding location, while the other agent chose the incorrect location.

In the process condition, participants saw two hiding locations and a clue pointing to one of them (e.g., footprints). However, the animal had gone somewhere else and was not hiding behind one of the two possible options anymore. In this condition, one agent used an invalid procedure (using a spinning wheel or a counting-out rhyme), while the other agent used a valid procedure (looking for the pet’s footprints or an object the pet could have lost). Since the animal had left, both agents drew wrong conclusions. We counterbalanced which of the two different valid and invalid procedures participants saw.

In the process-vs-outcome condition, participants saw two possible hiding locations, the pet hiding behind one of them, and a clue leading to the pet’s location. On the next slide, the pet moved to the other hiding location, which transformed the visible clue into misleading evidence. When the two agents looked for their pet, one used an invalid procedure (using a spinning wheel or a counting-out rhyme), leading them to the correct hiding location (where the pet was). The other agent used a valid procedure (looking for the pet’s footprints or an object the pet could have lost), leading them to the wrong hiding location (where the pet was not).

Across all conditions, at the end of each story, the participants were asked two questions: (1) “Who do you think is doing a better job looking for the [name of pet]?” (performance evaluation), and (2) “Imagine that you lost one of your favorite toys (for children) / something that’s important to you (adults). Who would you want to ask for help looking for this toy (for children) / this item (for adults)?” (partner choice). If participants did not want to decide on one of the characters, they were asked who they would choose if they had to make a choice. If they still did not want to make a choice, we coded a missing response. Following each question, participants were also asked to justify their responses (“Why do you think so?”). The responses to the why questions are analyzed descriptively in the supplementary material.

After the sixth story, the experimenter thanked the participant (and their parent) and gave them a short summary of the purpose of the study. Children in the US received a certificate for their participation. The parents of Chinese children as well as adult participants of both cultures were financially compensated for their participation.

1.3. Coding and reliability

We used a binary coding system for both dependent variables, performance evaluation and partner choice. A second coder, who was blind to the hypotheses of the study, coded 25% of all trials. Inter-rater reliability was perfect (Cohen’s kappa: $K = 1$) for the performance evaluation and almost perfect for the partner choice variable (Cohen’s kappa: $K = 0.99$).

Additionally, we coded whether participants refused choosing one of the agents after the first prompt or mentioned that they would rather not choose at all because they think that both agents did an equally good or bad job.

1.4. Statistical analysis

To statistically investigate the effects of condition, age group, and culture, we attempted to fit two Generalized Linear Mixed Models (GLMMs, one for each dependent variable) that contained all the predictors of interest (as preregistered). However, these models did not converge. This was due to complete separation issues (Field, Miles, & Field, 2012). The participants’ responses were too extreme in the outcome condition and the process condition (almost all participants in the outcome condition favored the agent who came to the correct conclusion and almost all participants in the process condition favored the agent who used a valid procedure, see Figs. 2 and 3). Thus, we focused on the condition with enough variance: the process-vs-outcome condition. Additionally, we ran simulation analyses with all conditions in which we addressed the complete separation problem (these analyses are reported in the supplementary material). The two analytical approaches led to the same conclusions. All analyses were conducted in R (R-Core-Team, 2020).

To analyze the process-vs-outcome condition individually, we fitted a GLMM with only the data of the process-vs-outcome condition. The full model contained the predictors age group, culture, and the two-way interaction between them. To account for within-subject variation and repeated measurements we further included the random intercept of individual identity. The model converged without any problems. To avoid an increased type 1 error risk due to multiple testing, we first tested the overall effect of all test predictors. Thus, we compared the full model with a null model comprising only the random effects to examine whether the inclusion of the test predictors provided a better fit to the data than participant identity alone. To determine the effects of each predictor alone, we further compared the full model with the corresponding reduced models that lacked the predictor of interest using likelihood ratio tests.

We also visually inspected the estimates and confidence intervals (calculated with 1000 parametric bootstraps) of the fitted models. If the confidence intervals do not include 0.5, the estimates can be considered
to be significantly different from chance level.

To investigate developmental patterns across and within cultures, we performed post-hoc pairwise comparisons using the package \textit{emmeans} (Lenth, 2018). We focused on pairwise comparisons of the same age groups across cultures, and on the comparison of adjacent age groups within each culture.

2. Results

2.1. DV 1: performance evaluation

Inspection of Fig. 2 reveals that, when judging who of the two agents did a better job, children and adults of both cultures had a strong
preference for the agent who came to a correct conclusion in the outcome condition. Similarly, children and adults of both cultures had a strong preference for the agent who used a valid procedure in the process condition. Age differences emerged in the process-vs-outcome condition: In both cultures younger children showed a tendency to choose the agent who used an invalid procedure but arrived at the right conclusion, whereas older children and adults preferred the agent who used a valid procedure but arrived at the wrong conclusion. In China, the shift from outcome to process evaluation seems to happen at an earlier age.

The statistical analysis of the process-vs-outcome condition confirmed that age groups varied significantly in whether they focused on the agent’s achieved outcome or the validity of the procedure used when judging who did a better job \( (\chi^2 = 105.39, df = 3, p < .001) \). The difference between age groups did not vary strongly enough between cultures to lead to a significant interaction effect (age group x culture: \( \chi^2 = 3.10, df = 3, p = .747 \)). As preregistered, we ran post-hoc pairwise comparisons indicating that the shift from outcome- to process-focused evaluations emerges earlier in children from China than in children from the United States. In the sample from China, the significant jumps occurred between the three younger age groups: the 4-5-year-old children differed significantly in their responses from 6 to 7-year-old children (\( p = .014^* \)), and 6-7-year-old children differed significantly from 8 to 9-year-old children (\( p = .030^* \)), but the 8-9-year-old children did not differ significantly from adults (\( p = .744^* \)). In the sample from the United States, 4-5-year-old children did not differ significantly in their responses from 6 to 7-year-old children (\( p = .111 \)), but 6-7-year-old children differed significantly from 8 to 9-year-old children (\( p = .016^* \)), and 8-9-year-old children differed significantly from adults (\( p = .054^* \)). This trend was not strong enough to lead to significant pairwise comparisons of the same age groups across cultures with our current sample size (see Table 1). However, an inspection of Fig. 3 shows that while 6-7-year-olds in the US were significantly below chance to choose the agent who used a valid procedure (i.e., they focused on outcome), approximately half of the 6-7-year-olds in China applied a process-focused evaluation.

We also checked on how many trials children refused to answer the question. In the outcome condition, Chinese children did not want to decide which of the two agents did a better job in 10 trials. In the process condition this was the case in 1 trial. None of the US children refused to choose.

### Table 1

<table>
<thead>
<tr>
<th>Performance evaluation</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contrast one age group with the next within a culture</strong></td>
<td></td>
</tr>
<tr>
<td>USA 4-5-year-olds</td>
<td>6-7-year-olds</td>
</tr>
<tr>
<td>6-7-year-olds</td>
<td>8-9-year-olds</td>
</tr>
<tr>
<td>8-9-year-olds</td>
<td>adults</td>
</tr>
<tr>
<td>China 4-5-year-olds</td>
<td>6-7-year-olds</td>
</tr>
<tr>
<td>6-7-year-olds</td>
<td>8-9-year-olds</td>
</tr>
<tr>
<td>8-9-year-olds</td>
<td>adults</td>
</tr>
<tr>
<td><strong>Contrast age groups across cultures</strong></td>
<td></td>
</tr>
<tr>
<td>4-5-year-olds USA</td>
<td>4-5-year-olds China</td>
</tr>
<tr>
<td>6-7-year-olds USA</td>
<td>6-7-year-olds China</td>
</tr>
<tr>
<td>8-9-year-olds USA</td>
<td>8-9-year-olds China</td>
</tr>
<tr>
<td>adults USA</td>
<td>adults China</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Partner choice</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contrast one age group with the next within a culture</strong></td>
<td></td>
</tr>
<tr>
<td>USA 4-5-year-olds</td>
<td>6-7-year-olds</td>
</tr>
<tr>
<td>6-7-year-olds</td>
<td>8-9-year-olds</td>
</tr>
<tr>
<td>8-9-year-olds</td>
<td>adults</td>
</tr>
<tr>
<td>China 4-5-year-olds</td>
<td>6-7-year-olds</td>
</tr>
<tr>
<td>6-7-year-olds</td>
<td>8-9-year-olds</td>
</tr>
<tr>
<td>8-9-year-olds</td>
<td>adults</td>
</tr>
<tr>
<td><strong>Contrast age groups across cultures</strong></td>
<td></td>
</tr>
<tr>
<td>4-5-year-olds USA</td>
<td>4-5-year-olds China</td>
</tr>
<tr>
<td>6-7-year-olds USA</td>
<td>6-7-year-olds China</td>
</tr>
<tr>
<td>8-9-year-olds USA</td>
<td>8-9-year-olds China</td>
</tr>
<tr>
<td>adults USA</td>
<td>adults China</td>
</tr>
</tbody>
</table>

Note. p values are fdr-corrected.

2.2. DV 2: partner choice

Inspection of Fig. 4 reveals that participants’ responses to the partner choice question were highly similar to the performance evaluation question. Here, the earlier occurring developmental shift in Chinese children becomes even more apparent.

Analyzing the process-vs-outcome condition confirmed that age groups varied significantly in whether they focused on the agent’s achieved outcome or on the validity of the procedure used when choosing who they would ask for help in the future (\( \chi^2 = 105.39, df = 3, p < .001 \)). As in the performance evaluation analysis, this difference between age groups did not vary strongly enough between cultures to lead to a significant effect (age group x culture: \( \chi^2 = 3.739, df = 3, p = .291 \)). Also here, post-hoc pairwise comparisons indicated that the shift from outcome- to process-focused evaluations emerges earlier in children from China than in children from the United States. In the sample from the United States, both 4-5-year-old children and 5-6-year-old children focused mostly on the outcome when choosing who to ask for help (\( p = .222 \)). The significant jump from outcome-focused to process-focused responses happened from 6 to 7-year-old children to 8-9-year-old children (\( p = .004^* \)). The responses from 8 to 9-year-old children were not significantly different from adults (\( p = .453 \)). In the sample from China, the significant jumps occurred between 4 and 5-year-old children and 6-7-year-old children (\( p < .001^* \)), and 6-7-year-old children differed marginally from 8 to 9-year-old children (\( p = .070^* \)), but the 8-9-year-old children were not significantly different from adults (\( p = .954 \)). As for the other response variable, with our current sample size, we did not find any significant effects when conducting pairwise comparisons of the same age groups across cultures (see Table 1). However, as can be seen in Fig. 5, while most 6-7-year-olds in the US focused on the outcome of agents’ beliefs, more than half of the 6-7-year-olds in China applied a process-focused evaluation.

Again, we noted the number of trials in which children refused to answer the question. In the outcome condition, children from China refused to pick an agent in 6 trials, in the US in 1 trial. In the process condition, this happened in 1 trial for Chinese children and in 1 trial for US children. In the process-vs-outcome condition, children from the US did not pick an agent in 3 trials.

3. Discussion

Much like walking, deliberation involves taking a route (the process – the way we form our belief) to a certain endpoint (the outcome – the belief we end up holding). When we evaluate others’ reasoning, what do we value more: the route or the endpoint? Our results suggest that a developmental shift takes place in middle childhood. Younger children evaluate individuals who take the wrong route (invalid process) but arrive at the right endpoint (correct outcome) more positively. Older children and adults show a preference for individuals who take the right route (valid process) but arrive at the wrong endpoint (incorrect outcome). This developmental pattern was robust across the two samples studied here – China and US – and the two evaluation questions we asked participants: “Who is doing a better job?”, a question which was intentionally left ambiguous and could be interpreted either in terms of process or outcome, and “Who would you ask for help?”, a future-oriented question that required participants to predict an agent’s usefulness as an
Children and adults showed a very strong and consistent preference for agents who held accurate over inaccurate beliefs (in the *outcome condition*, where we held procedure constant) and for agents who used appropriate over inappropriate epistemic procedures (in the *process condition*, where we held outcome constant). The former result is a clear demonstration that from at least the age of 4, children consider outcome information in their evaluation of reasoning, a possibility that had been raised by previous findings in the selective social learning literature (as reviewed in the introduction). The latter result is noteworthy as it shows that by age 4, children cannot only distinguish non-circular from circular reasons, but can also differentiate other valid from invalid belief-
forming methods and evaluate the former more positively. This finding also demonstrates that the developmental difference observed in the process-vs-outcome condition did not arise because the stories were too complex for the younger children or because they had difficulty assessing the process information, but was due instead to a shift in judgments of good reasoning that occurs in middle childhood. Interestingly, across both cultures, children were slightly more likely to focus on process instead of outcome when asked who they would approach if they needed help in the future compared to when asked who they thought did a better job. The future directedness of the partner choice question might make it easier for children to focus on what matters for long-term success. This should be considered in succeeding studies investigating children’s evaluation of others’ reasoning.

The developmental shift replicated across our Chinese and US sample. The tendency to value the processes by which beliefs are formed (process) over the content of beliefs (outcome) might be part of core adult epistemology, a cross-culturally recurrent way of judging good reasoning (for other claims about core epistemology, see Machery et al., 2017; Machery et al., 2017; Nagel, Juan, & Mar, 2013). What explains the observed developmental pattern from outcome to process? There are at least three possible interpretations. First, it is conceivable that young children have a stronger outcome bias than older children and adults. When information about both process and outcome is available, children might selectively pay attention to the latter. The finding that children show an outcome bias in other belief-related contexts, e.g. selectively believing more in a claim when they stand to gain from the claim’s truth, lends support to this interpretation (Oved et al., 2014; Woolley, Boerger, & Markman, 2004). A second possibility is that young children, like older children and adults, in fact weigh process more heavily than outcome, but that, in their eyes, an accurate outcome can rationalize an invalid process. That is, children might take the fact that the invalid belief-forming method (e.g., spinning a wheel) produces an accurate outcome as evidence for the method’s reliability. A third possibility is that children have a stronger tendency to evaluate lucky characters more positively than unlucky ones. The character who gets it right in the process-vs-outcome condition is epistemically lucky: the fact that they arrive at the correct conclusion is determined by factors beyond their control. The character who gets it wrong is epistemically unlucky: they, through no fault of their own, find themselves in circumstances that are epistemically misleading. Prior research has shown that children prefer lucky over unlucky individuals (Olson, Banaji, Dweck, & Spelke, 2006). In the process-vs-outcome condition, some participants justified their choice of the agent who used an invalid procedure but arrived at the correct conclusion by saying things such as: “Because she is lucky and sticks with what she thinks”. Note, however, that such luck-based justifications were generally rare: they occurred only on 21 trials, mostly by 8–9-year-old and adult participants. We know of no studies on children’s concept of epistemic luck. This is an interesting domain for future research.

The age at which children began shifting from outcome to process varied cross-culturally: between 5 and 6 years in the Chinese sample and between 7 and 8 years in the US sample. What might account for the earlier transition in the Chinese sample? As discussed in the introduction, we hypothesized that children’s early learning environments in our Chinese and US sample might vary in ways that are relevant to evaluations of good reasoning. To quote Richard Nisbett: “The Chinese believe in constant change, but with things always moving back to some point of origin. The Westerner, in contrast, to a wide range of events they search for relationships between things; and they think you can’t understand the part without understanding the whole. Westerners live in a simpler, prior state. They pay attention to a wide range of events; they search for weight they place on outcome and process in their performance evaluations. Attributes of good and bad reasoning – or rationality and irrationality – can apply to both what someone believes (outcome) and to how they formed the belief (process). Here we have provided evidence that in a sample from China and a sample from the US, young children evaluate others’ reasoning primarily in terms of outcome (belief accuracy), while older children and adults evaluate reasoning based primarily on the process by which a belief is formed.
Author contribution

H. Schleihauf and J. Engelmann developed the study concept. All authors contributed to the study design. Testing and data collection were performed by Z. Zhang, A. Gomez, and H. Schleihauf. H. Schleihauf performed the data analysis. H. Schleihauf and J. Engelmann drafted the manuscript, and Z. Zhen and A. Gomez provided critical revisions. All authors approved the final version of the manuscript for submission.

Funding

This project has received funding from the European Union’s Framework Program for Research and Innovation Horizon 2020 (2014–2020) under the Marie Skłodowska-Curie Grant Agreement awarded to Hanna Schleihauf. Additionally, Jan Engelmann acknowledges support from the Jacobs Foundation.

Declaration of Competing Interest

H. Schleihauf et al. 

Declaration of Competing Interest 

H. Schleihauf et al. 

Support from the Jacobs Foundation. (2014–2020). Funding authors approved the final version of the manuscript for submission. Testing and data collection were performed by Z. Zhang, A. Gomez, and H. Schleihauf. H. Schleihauf and J. Engelmann drafted the manuscript, and Z. Zhen and A. Gomez provided critical revisions. All authors approved the final version of the manuscript for submission.

Support from the Jacobs Foundation. (2014–2020). Funding authors approved the final version of the manuscript for submission. Testing and data collection were performed by Z. Zhang, A. Gomez, and H. Schleihauf. H. Schleihauf and J. Engelmann drafted the manuscript, and Z. Zhen and A. Gomez provided critical revisions. All authors approved the final version of the manuscript for submission.

Funding

This project has received funding from the European Union’s Framework Program for Research and Innovation Horizon 2020 (2014–2020) under the Marie Skłodowska-Curie Grant Agreement awarded to Hanna Schleihauf. Additionally, Jan Engelmann acknowledges support from the Jacobs Foundation.

Data availability

The preregistration, materials, the data, and the code employed to conduct the data analysis have been made publicly available in an OSF repository [http://doi.org/10.17605/OSF.IO/RJBG6].

Acknowledgements

We thank the kindergarten parents, and children that supported our work. We thank our research assistants Yuhan Wang, Alicina Geng, and Claudine Narayan for their great support during data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105425.

References