Three Men Make a Tiger: The Effect of Consensus Testimony on Chinese and U.S. Children’s Judgments about Possibility

Jenny Nissel | ORCID: 0000-0001-7854-2545
PhD Candidate, Department of Psychology, University of Texas at Austin, Austin, TX, USA
Corresponding Author
jenny.nissel@utexas.edu

Hui Li | ORCID: 0000-0002-1595-9113
Associate Professor, School of Education, Central China Normal University, Wuhan, Hubei, China
huilipsy@ccnu.edu.cn

Amanda Cramer
Clinical Research Coordinator, Department of Psychiatry, Washington University in St. Louis, St. Louis, MO, USA
acramer@wustl.edu

Jacqueline D. Woolley | ORCID: 0000-0002-4384-7688
Professor, Department of Psychology, University of Texas at Austin, Austin, TX, USA
woolley@austin.utexas.edu

Abstract

In this study, we ask whether consensus testimony affects children’s judgments of the possibility of improbable and impossible events. Fifty-six U.S. and Chinese 8-year-olds made possibility judgments before and after hearing three speakers affirm or deny the possibility of improbable and impossible events. Results indicated that whereas both U.S. and Chinese children altered their judgments in the direction of the consensus testimony, this effect was stronger for Chinese children. U.S. children were particularly receptive to consensus for improbable events and when the consensus provided correct information, whereas Chinese children were similarly willing to change their
judgment regardless of event type and the validity of the testimony. We propose that the extent of the influence of testimony on possibility judgments varies based on cultural setting. Our findings have potential applications for domains that require evaluation of counterintuitive claims, like religious and scientific education.

Keywords

possibility judgments – testimony – consensus – social cognitive development – counterfactual thinking

1 Introduction

Consider this hypothetical:

If one person were to tell you that there was a tiger running down the street, would you believe him?
If two people were to tell you so, would you believe them?
What if it were three?

Chinese legend has it that a King's advisor posed such an analogy to warn him of the potential pitfalls of falling for an improbability just because enough people say it's true (Zhai, 2008). The resulting idiom “three men make a tiger” is not merely a dusty remnant from a Confucian-era chronicle; “三人成虎” persists in Mandarin today, indicating the sentiment has not lost its relevance. Moreover, similar warnings are peppered throughout Western cultural lore, cautioning readers that the word of many may lead to acceptance of not only improbable but also impossible claims. As a case in point, consider the Emperor's New Clothes, where a growing consensus leads a stark-naked king and his subjects to affirm the existence of clothes visible only to a worthy few (Anderson, 2003). Members of both Eastern and Western cultures throughout history have repeatedly warned against putting too much weight on consensus judgments. That they have deemed it necessary to repeat this warning suggests that despite their efforts, the tendency to believe strange things because those around us do persists.

The present research explores the ontogenetic origins of this propensity. We ask whether Chinese and U.S. children, like their adult analogs in the legends above, change initial judgments of the possibility of improbable and impossible events after hearing consensus testimony. In the remainder of this section,
we explore the development of receptivity to consensus cross-culturally and then review children's understanding of improbable and impossible events. Finally, we provide a rationale for our sample and present our specific research question and hypotheses.

2 Children’s Receptivity to Consensus Testimony

Like the storytellers from whom we borrowed our lead-in, psychologists across history have found human reactions to consensus testimony worthy of attention. Classically, Solomon Asch (1951) demonstrated that Swarthmore students were willing to say that the shortest of three lines was actually the longest when in the company of confederates giving this wrong answer.

Of course, consensus is not always a red herring – in fact, it can be a powerful, evolutionarily-rooted learning tool as adults and children navigate the world, encounter new activities, and learn new skills within their respective cultures. Consensus can be used to learn how to do something: for example, chimpanzees in neighboring communities use consistently distinct nut cracking techniques, specific by community – not for ecological expediency but presumably because of cultural influence (Luncz, Mundry, & Boesch, 2012). Similarly, homo sapiens children pick up on cultural norms when learning new tasks. For example, preschoolers are oftentimes willing to choose less functional tools to perform particular tasks or perform inefficient actions when adults model such behavior (DiYanni, Corriveau, Kurkul, Nasrini, & Nini, 2015; DiYanni & Kelemen, 2008).

This natural sensitivity to cultural norms might extend not only to actions but also to beliefs. Modern polarized views on the source of climate change in the United States may be an illustration of this tendency to accept the narrative of one’s own ingroup (e.g., Goldberg, van der Linden, Leiserowitz, & Maibach, 2020).

Children too seem to use consensus to form new beliefs. When learning object names, 3-to 5-year-old children associate a novel object with a referent indicated by a majority of informants rather than one indicated by a minority (Corriveau, Fusaro, & Harris, 2009). Furthermore, children find members of the majority more trustworthy as individuals – they continue to trust a single informant from the original majority when pitted against one from the original minority when learning subsequent novel object names.

There is reason to believe that children in different cultures might be differentially sensitive to consensus cues. Consensus carries greater weight in a relatively collectivist culture like China, for example, than in a relatively
individualist culture like the United States. Illustratively, although Caucasian American and Chinese American 3- to 5-year-olds were equally unlikely to copy an inefficient action or use a functionally non-affordant tool when modeled by one actor, Chinese American children were far more likely than Caucasian American children to choose the inefficient or nonfunctional route when it was performed by three models (Corriveau, DiYanni, Clegg, Min, Chin, & Nasrini, 2017). This suggests that Chinese American children were more sensitive to consensus cues than were Caucasian American children, who continued to use task specific cues irrespective of the number of models leading them astray.

Caucasian American children are not insensitive to consensus cues, though: 4- to-7-year-old European American and Taiwanese children learned novel object names endorsed by a consensus whether the consensus was made up of European American or Taiwanese informants (Chen, Corriveau, & Harris, 2013). However, when asked to learn additional object names from a single informant from the consensus group, children only continued to trust the informant when she was from their own culture. This suggests that children use social membership cues when evaluating consensus-based information, reinforcing the cultural nature of the phenomenon. Moreover, developmental trends differed between the two cultures: only European American children showed an increasing trust in consensus testimony with age; Taiwanese children showed a high level regardless of age, suggesting perhaps a stronger weight placed on consensus within the latter culture. Based on this previous research, we expect both U.S. and Chinese children in the present study to shift judgments to some degree based on consensus testimony; however, we expect this shift to be more dramatic among Chinese participants.

Though the evidence cited suggests that children will adjust outward behavioral displays and belief statements to accord more closely with consensus, such adjustments do not necessarily require actual changes in underlying beliefs themselves. As a case in point, a child-friendly version of Asch’s (1951) paradigm yielded expected results: 3- and 4-year-old Caucasian and Asian American children both deferred to an inaccurate perceptual judgment endorsed by a majority of informants some of the time; Asian American children were significantly more likely to defer to this inaccurate majority than were their Caucasian American peers (Corriveau & Harris, 2010). However, the authors questioned the extent to which children’s beliefs were actually affected in this scenario and suggested that children’s decisions to go with the majority might have merely been gestures of “respectful deference” (Corriveau & Harris, 2010, p. 443). That is, children may have changed their behavior, but because of the clear conflict between perceptual and cultural information, they may have not in turn changed their beliefs. As a test of this, children were
later given the opportunity to construct a bridge with the longest line used in the earlier task. Caucasian and Chinese American children never selected the short, consensus-endorsed line, suggesting that despite their earlier assertions, their underlying beliefs remained unchanged. However, follow-up work indicated that at least in a public setting (when compared to a private one), Asian children were less likely to criticize obviously incorrect perceptual judgments made by a consensus, and more likely to misremember these judgments as being correct (Corriveau, Kim, Song, & Harris, 2013). In this case, the authors suggest that respectful deference – above and beyond simply reflecting an effort to appease social companions – actually resulted in a recalibration of underlying beliefs. Thus, the extent to which behavior change actually represents belief change in experimental paradigms seems to vary based on a number of factors including culture and social pressures, and merits further investigation.

In the current study, we continue the exploration of potential cultural differences in children’s use of consensus information. We focus our investigation on the effect of consensus testimony on children’s concepts of possibility: improbable and impossible events. As these events are, by definition, rare or unrealizable, children are unlikely to be able to draw from their own experience when evaluating the possibility of these events. In these circumstances, children might rely on the testimony of others to guide them, and, as in the research reviewed above, consensus might be expected to have a powerful effect on their judgments.

3 Understanding of Possibility in the U.S.

Many U.S. children believe in the existence of cultural fantastical beings who possess impossible attributes and perform impossible actions (e.g., Santa Claus; Goldstein & Woolley, 2016). This might suggest that children would believe in other impossible things, like selectively invisible clothes, or improbable things, like tigers running through city streets. However, research suggests that children as young as 3 understand impossibility: they consistently deny that impossible events – like eating lightning for dinner – could happen in the real world (e.g., Johnson & Harris, 1994; Rosengren, Kalish, Hickling, & Gelman, 1994; Shtulman & Carey, 2007). Oddly, children exhibit similar skepticism toward events that are unlikely nevertheless possible: when faced with improbable hypotheticals like finding an alligator under a bed, 4-year-olds deny that such events are actually possible (e.g., Shtulman & Carey, 2007; Shtulman, 2009).
Though understanding of improbability develops extensively between ages 4 and 8, even 8-year-olds’ understanding is far from adult level. In the seminal work on the subject, Shtulman and Carey (2007; Study 1) found that 8-year-olds correctly categorized improbable events only 65% of the time. Since then, researchers have replicated and extended this finding, suggesting that children’s ideas about these events are relatively intractable even in face of testimony: neither prompts to imagine improbable events nor explanations of how these events can happen moved children to change their point of view (e.g., Lane, Ronfard, Francioli, & Harris; Woolley, Ghossainy, & Cornelius, 2015).

Further work, however, shows that the picture is more nuanced and children can in fact reason accurately about possibility in certain circumstances. Various cognitive mechanisms have been offered to explain children’s conceptions of improbable events, springing from Shtulman and Carey (2007)’s original proposal that if children can’t imagine how an event could happen, they will be reluctant to say it could occur. Woolley and Cornelius (2017) posed an alternative theory, suggesting that children use context to guide their probability judgments. This is consistent with Weisberg and Sobel’s (2012) observation that even 4-year-olds implicitly recognize improbable events as members of a distinct class. Similarly, Goulding and Friedman (2021) offered the compatible explanation that children use a similarity heuristic to judge such events – when provided with examples of improbable events that actually occurred, children will infer that similar (but not dissimilar) events are also possible.

4 Understanding of Possibility in China

In contrast to the relative wealth of research on children’s ability to separate the impossible from the real in the U.S. (for review see Woolley & Nissel, 2020), there are few similar studies in China. One study suggests that Chinese children might differentiate these events to a lesser degree. When shown televised events, Chinese 4-year-olds overestimated the possibility of impossible events yet underestimated the possibility of ordinary events (Li, Boguszewski, & Lillard, 2015). However, television itself provides a potentially confusing context for making possibility judgments – and fantasy-reality confusion when watching television is also well-documented among children in the U.S. (e.g., Flavell, Flavell, Green, & Korfmacher, 1990; Richert & Lillard, 2004). Furthermore, any confusions exhibited by Chinese 4-year-olds decreased by age 6 (Li et al., 2015), and further work by Li, Liu, Woolley, and Zhang (2019) indicates that Chinese 8-year-olds are as good as Chinese adults at distinguishing televised
impossible events from real events. The limited evidence thus suggests that Chinese and U.S. children should be similarly capable of separating impossible and real events, though this needs to be tested outside of a televised context.

Regarding improbable events however, limited evidence suggests that Chinese children’s conceptions of these events may follow a distinct developmental trajectory from that in the U.S. The single study to our knowledge addressing Chinese children’s assessments of improbability detected no developmental change between 5 and 10 – both the youngest and oldest children asserted that improbable events could happen around half the time (Cui, Payir, Davoodi, Harris, & Corriveau, 2021). This is particularly interesting in light of the dramatic developmental changes that occur in U.S. children across this age span; however, we are unable to draw any direct cultural comparisons. First, improbable events consisted of events like finding loaves of bread in the forest, events which are different from and arguably more mundane than the canonical Shtulman and Carey (2007) set (e.g., making a mug shaped building). Second, the study was conducted in China only, so we do not know how children in the U.S. would have responded to a similar paradigm.

Thus, how children in the U.S. and China might differentially evaluate impossible and improbable events remains open to question. Moreover, the extent to which testimony may influence possibility judgments in both countries in unknown.

5 Effects of Consensus Testimony on Children’s Possibility Judgments

Research in the U.S. suggests that although children's possibility judgments of improbable events may be largely resistant to testimony, a particular subtype may defy this rule. Lane, Ronford, and El-Sherif (2018) presented 4- to 8-year-olds with first-hand or second-hand testimony about improbable events (e.g., “I saw someone drink onion juice” vs. “someone told me they drank onion juice”). As in previous work, first-hand testimony did nothing to change children's initial judgments. Second-hand testimony, however, had a polarizing impact on these judgments – heightening skepticism amongst the oldest children but heightening credulity among the youngest. Lane et al. (2018) suggest that the younger children assumed that the speaker conveying second-hand information wouldn't be doing so unless both he and the person he was talking about believed it. Essentially, they interpreted these statements as evidence of consensus and were thereby willing to change their original judgment. If Lane et al. (2018)'s interpretation is correct, consensus testimony may influence children’s possibility judgments. Thus, even if children's possibility judgments
are resistant to testimony from one individual, they may be open to testimony when it comes from a consensus.

6 Research Question and Hypotheses

In the present study, we directly compare the effect of consensus testimony on U.S. and Chinese children's possibility judgments. First, we compare children's judgments of the possibility of improbable and impossible events at baseline. Second, we ask whether children change these initial possibility judgments after hearing either positive or negative consensus testimony.

Regarding baseline judgments, we expect U.S. children to perform similarly to previous studies (i.e., they will differentiate between impossible and improbable events, though performance on improbable items will fall below adult level). It remains open to question how Chinese children will judge these events, as there is less existing research with this population. However, we have little reason to expect baseline differences; we anticipate Chinese children will perform similarly to U.S. children, given the robustness of this finding across multiple studies of children in the U.S.

In response to testimony, we anticipate that children from both countries will show willingness to shift judgments to some degree. However, we expect Chinese children to show greater sensitivity to testimony overall, for both event types. We expect children from both countries to show more willingness to shift judgments when evaluating improbable versus impossible events, as they should be less certain of the reality status of these events to begin with. However, we expect this differentiation between event type to be more pronounced in U.S. participants. In sum, we expect both U.S. and Chinese participants to utilize both event type and testimony when making their judgments; however, we expect the relative weight placed on these factors to vary by nationality.

7 Method

7.1 Participants
The planned sample size was 30 participants per country; the final sample consisted of 56 children ranging in age from 7;5 to 9;3 (M = 8;2) from the two countries. To our knowledge, prior empirical work has not yet compared the influence of consensus testimony on children's possibility judgments in the U.S. and China, so we began by establishing a baseline comparison. We focused our investigation specifically on 8-year-olds, as research suggests that
they possess nuanced yet still developing conceptions of possibility and have also had more exposure to respective cultural influences than younger children. The breakdown of sample demographics from each country is detailed below.

7.2 U.S.
Twenty-five U.S. children (44% female), ranging in age from 7;11 to 9;3 ($M = 8;4, SD = .38$), participated. Due to unintentional over-recruitment, 45 children were initially tested. However, experimenter error at the Chinese site resulted in all Chinese participants receiving only one of our two orders. Thus, to ensure methodological consistency across cultures, we retained only the 25 participants who received the stimuli in the order that the Chinese participants received. The racial ethnic breakdown of the sample was 76% white, 12% Hispanic, and 12% Asian. Children were residents of a midsize Southwestern city with a large university, and were recruited from the participant database of a research laboratory at the university. Maternal education levels indicated that 4% held associates degrees, 4% attended some college, 44% held bachelor’s degrees, and 36% held graduate or professional degree. The study was approved by the Institutional Review Board of the university (2012-01-0094: Effects of Conflicting Testimony on Decision about Reality Status). Data were collected from May to October of 2019. Parents signed a consent form and children signed an assent form before participation.

7.3 China
Thirty-one Chinese children (58% female), ranging in age from 7;5 to 9;3 ($M = 8;0$), participated. Demographic information was not collected; however, the sample was mainly Han Chinese – who currently account for 92% of the population in mainland China (Gao et al., 2020). Children were residents of a large central Chinese university city, students at an elementary school associated with the university, and were recruited from an after-school program on the university campus. At least one parent of each participant was affiliated with the university. Data were collected from May to October of 2019. The study was approved by the Institutional Review Board of the university (2017-025-112: The effect of consensus on children’s reality judgement). Parents signed a consent form before their children participated.

8 Materials

Surveys consisted of two parts, a pretest and a posttest. The pretest contained 10 realistic photographic images (located and downloaded from google
images) illustrating 10 different events (two ordinary, e.g., a person drinking milk; four improbable, e.g., a person with a peacock as a pet, and four impossible, e.g., a tree with money growing on it; see Table 1 for a complete listing of items). Each illustration was followed by a possibility question containing a verbal description of each event and a certainty question (specific wording included in Procedure section). Ordinary events were included not for analysis purposes (we did not expect participants to differ on these items), but to break response sets that might have been created by judging multiple extraordinary events in a row (refer to Appendix A for ordinary event means and pairwise comparisons).

The posttest contained the same 10 event images, immediately followed by three recorded videos of undergraduate-aged women providing consistent (positive or negative) testimony as to whether each event could occur. Thirty videos were created per country, so each speaker was seen only once per study to avoid confounding inferences about speaker reliability. These videos were followed by the accompanying questions, repeated from the pretest. The U.S. version was in English featuring American students in the videos; the Chinese version was in Mandarin featuring Chinese students in the videos. All materials were directly translated from English to Mandarin by a researcher fluent in both languages. The images were consistent across cultures, though the tree with money growing on it was edited to sprout the appropriate national currency.

### Table 1: Ordinary, improbable, and impossible events used

<table>
<thead>
<tr>
<th>Ordinary</th>
<th>Improbable</th>
<th>Impossible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wear a yellow hat</td>
<td>Eat pickle flavored ice cream</td>
<td>Open a window with one’s mind</td>
</tr>
<tr>
<td>Drink a glass of milk</td>
<td>Own a peacock for a pet</td>
<td>Travel back in time</td>
</tr>
<tr>
<td></td>
<td>Paint polka dots on an airplane</td>
<td>Find a tree with money growing on it</td>
</tr>
<tr>
<td></td>
<td>Make a building shaped like a teapot</td>
<td>Walk through a brick wall</td>
</tr>
</tbody>
</table>

9 Design

The design was a mixed model with event type (2: improbable, impossible) and testimony type (3: baseline, positive, negative) as within subject factors.
and culture (2: U.S., China) as the between subjects factor. Each participant saw each event type and heard each testimony type, in counterbalanced combinations. Two counterbalanced surveys were created such that the items paired with positive consensus in one survey were paired with negative consensus in the other. Due to experimenter error, only one item/consensus pairing combination was used for the Chinese participants. To ensure operational equivalence, the final comparison sample in the U.S. consisted of the subset of participants run in the same version of the survey that the Chinese participants saw.

10 Procedure

In both countries, participants were tested individually in research laboratories. Participants viewed the survey on an iPad, and the researcher entered answers into the survey directly. Children were informed that we needed their help figuring out whether some different things could happen in real life or not.

10.1 Pretest

First, participants were presented with 10 events (2 ordinary, 4 improbable, and 4 impossible), illustrated with photographic images. After viewing each picture, they were asked a possibility question (e.g., “Could a person have a peacock for a pet in real life?”) followed by a certainty question (“How sure are you?”). Children were asked to select one of two multiple choice answers for the possibility question (“yes” or “no”) and one of two answers for the certainty question (“a little sure” or “really sure”).

10.2 Posttest

Subsequently, children were told that the experimenter asked some other people whether these things could happen. They were instructed to listen to what the informants had to say before telling the experimenter what they thought. They were then presented with each event again and a description of the question that was asked (e.g., “I asked them whether someone could eat pickle flavored ice cream in real life”), followed by three videos of female undergraduates giving consistent positive testimony (“Yes, it could happen”) or consistent negative testimony (“No, it could not happen”). Ordinary events were always paired with positive testimony to establish informant reliability. Of the four improbable events, two appeared with positive consensus and two appeared with negative consensus. Likewise, of the four impossible events,
two appeared with positive and two appeared with negative consensus. For each event, after listening to the testimony, children were again asked the possibility and certainty questions. At the end of the study, children and parents were debriefed. As thanks for their participation, U.S. children were given the opportunity to select a small toy and Chinese children were given stickers.

11 Results

11.1 Coding

11.1.1 Possibility Judgment Scores
We created scaled possibility judgment scores for each item by combining each participant’s response to each forced-response possibility question and associated certainty question (1 = no, really sure; 2 = no, a little sure; 3 = yes, a little sure; 4 = yes, very sure). Analysis of scaled belief scores as a continuous variable are precedented in previous research (e.g., Woolley & Van Reet, 2006) and have been argued to increase variance and enable greater detection of relations between variables (Lane et al., 2018).1

11.1.2 Change Scores
To capture the degree to which each participant shifted their judgment after hearing testimony, we created change scores, derived by subtracting a participant’s baseline possibility judgment score on each item from the post-testimony possibility judgment score on the same item.

11.1.3 Correction/Regression Scores
We created correction and regression scores to capture whether individuals changed their baseline possibility judgment in the direction of contradictory testimony. We included only cases where individuals heard testimony that contradicted their initial judgment. Each response was coded as 1 if the individual changed their judgment, and 0 if the individual retained their baseline judgment.

11.2 Bonferroni Corrections
Prior to beginning our main analyses, we planned to conduct 3 sets of multiple pairwise comparisons based on hypotheses. As such, to safeguard against Type 1

---

1 Refer to Harpe (2015), Sullivan & Artino (2013), and Carifio & Perla (2008) for justification of a parametric approach to analyzing Likert scaled data.
error, we used stringent Bonferroni corrected \( p \)-values to determine significance for all pairwise comparisons. See Table 2 for detailing of corrections used by set.

11.3 Possibility Judgments

11.3.1 Full Model

As illustrated in Fig. 1, U.S. and Chinese children seemed to adjust their baseline judgments in response to consensus, albeit to different degrees. To explore these patterns statistically, we ran a repeated measures ANOVA on possibility judgment with timing (2: before testimony, after testimony), event type (2: improbable, impossible) and testimony type (2: positive, negative) as the within-subjects variables and country (2: U.S., China) as the between-subjects variable. The analysis yielded main effects of timing, \( F(1,54) = 12.55, p < .001, \eta_p^2 = .19 \), event type, \( F(1,54) = 231.53, p < .001, \eta_p^2 = .81 \), testimony type, \( F(1,54) = 27.149, p < .001, \eta_p^2 = .34 \), and country, \( F(1,54) = 13.75, p < .001, \eta_p^2 = .20 \). These effects were subsumed by significant higher-order interactions, including a significant four-way interaction between country, timing, event type, and testimony type, \( F(1,54) = 5.49, p = .027, \eta_p^2 = .09 \). All model effects are documented in Table 3. To explore these effects, we first conducted a series of pairwise comparisons examining possibility judgments in the absence of testimony (baseline judgments). Following, we explored response to testimony by event type in each country before comparing responses across cultures.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Pairwise comparisons and associated adjusted significance level by set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set (by dv)</td>
<td>Number of Comparisons</td>
</tr>
<tr>
<td>Possibility Judgment</td>
<td>14</td>
</tr>
<tr>
<td>Change Score</td>
<td>8</td>
</tr>
<tr>
<td>Correction Score</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Model effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>( F(174) )</td>
</tr>
<tr>
<td>Country</td>
<td>13.75</td>
</tr>
<tr>
<td>Timing</td>
<td>12.55</td>
</tr>
<tr>
<td>Event</td>
<td>176.51</td>
</tr>
<tr>
<td>Consensus</td>
<td>27.15</td>
</tr>
<tr>
<td>Country*Timing</td>
<td>13.05</td>
</tr>
</tbody>
</table>
Table 3: Model effects (cont.)

<table>
<thead>
<tr>
<th>Source</th>
<th>$F(174)$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country*Event</td>
<td>75.77</td>
<td>$p &lt; .001**$</td>
</tr>
<tr>
<td>Country*Consensus</td>
<td>15.30</td>
<td>$p &lt; .001**$</td>
</tr>
<tr>
<td>Timing*Event</td>
<td>.280</td>
<td>$p = .599$</td>
</tr>
<tr>
<td>Timing*Consensus</td>
<td>130.92</td>
<td>$p &lt; .001**$</td>
</tr>
<tr>
<td>Event*Consensus</td>
<td>.43</td>
<td>$p = .516$</td>
</tr>
<tr>
<td>Country<em>Timing</em>Event</td>
<td>.004</td>
<td>$p = .947$</td>
</tr>
<tr>
<td>Country<em>Timing</em>Consensus</td>
<td>11.15</td>
<td>$p = .002*$</td>
</tr>
<tr>
<td>Country<em>Event</em>Consensus</td>
<td>.002</td>
<td>$p = .963$</td>
</tr>
<tr>
<td>Timing<em>Event</em>Consensus</td>
<td>13.85</td>
<td>$p &lt; .001**$</td>
</tr>
<tr>
<td>Country<em>Timing</em>Event*Consensus</td>
<td>5.19</td>
<td>$p = .027*$</td>
</tr>
</tbody>
</table>

Note. * Indicates significance $p < .05$. ** Indicates significance $p < .001$. 

Figure 1: Effect of positive (+) and negative (−) consensus on possibility judgments of impossible and improbable events by culture

Note. * Base (+) indicates the baseline ratings for those items that were later accompanied by positive (+) testimony; Base (−) indicates baseline ratings for items that were later accompanied by negative (−) testimony

**1 = Could not happen, really sure; 2 = Could not happen, a little sure; 3 = Could happen, a little sure; 4 = Could happen, really sure

Error bars represent standard error of the mean.
11.3.2 Baseline Possibility Judgments

As can be seen in the left half of Figure 1, U.S. children appeared to be sensitive to whether the event was improbable or impossible at baseline. As participants were unaware that each item would later be accompanied by testimony of a particular valence, and our hypothesis predicted differentiation by item type regardless of valence of testimony, we compared judgments of all improbable events to judgments of all impossible events. A paired samples t-test on possibility judgment by event type showed that U.S. children differentiated between event types: they judged improbable ($M = 3.23$ (out of 4), $SD = .62$) events as more possible than impossible events ($M = 1.22$, $SD = .40$; $t(24) = 14.03$, $p < .001$). As can be seen in the right half of Figure 1, Chinese participants similarly differentiated between event types at baseline, judging improbable events ($M = 1.98$, $SD = .62$) as more possible than impossible events ($M = 1.38$, $SD = .33$; $t(30) = 5.38$, $p < .001$). Comparing responses across countries, U.S. and Chinese possibility judgments were at similar levels for impossible events ($p = .001$). However, U.S. children were significantly more willing than Chinese children to say that improbable events were possible, $t(54) = 7.57$, $p < .001$, suggesting that U.S. participants differentiated between event types at baseline to a greater extent.

11.3.3 U.S. Children’s Response to Consensus

11.3.3.1 Improbable Events

As can be seen in Figure 1, U.S. children appeared to be moderately sensitive to testimony – whether positively or negatively valanced – when evaluating the possibility of improbable events. Paired samples t-tests showed that possibility judgments after positive consensus ($M = 3.66$, $SD = .64$) were significantly higher than possibility judgments on those same items at baseline, $M = 2.82$, $SD = .85$; $t(24) = 5.98$, $p < .001$. Inversely, negative consensus resulted in judgments ($M = 2.74$, $SD = 1.04$) that were significantly lower than those at baseline, $M = 3.64$, $SD = .57$; $t(24) = -4.99$, $p < .001$.

11.3.3.2 Impossible Events

In contrast to their relative willingness to recalibrate responses when the events were improbable, U.S. children appeared to be unmoved by consensus testimony when the events were impossible. Both positive and negative consensus resulted in judgments that were no different from those at baseline, ($p = .210$; $p = .083$, respectively). It should be noted that performance on these items was near floor to begin with ($M_{baseline}=1.22$, $SD = .40$), leaving little room for negative testimony to have an effect. However, given U.S. children’s
receptivity to testimony on improbable items, this relatively diminished receptivity to testimony on impossible items suggests that they seemed use the nature of the event to determine how to evaluate subsequent testimony.

11.3.3.3 Change on Impossible vs. Improbable Events
Indeed, as seen in Figure 2, change scores revealed that U.S. children responded to testimony differently based on event type: paired samples \( t \)-tests showed that U.S. children changed significantly more with both positive and negative consensus on improbable (\( M_{\text{change}+} = .84, \text{SD} = .70, M_{\text{change} -} = -.90, \text{SD} = .91 \)) as compared to impossible events (\( M_{\text{change}+} = .22, \text{SD} = .85, M_{\text{change} -} = -.10, \text{SD} = .46 \)); positive: \( t(24) = 3.03, p = .006 \); negative: \( t(24) = 4.62, p < .001 \).

![Figure 2](image_url)

FIGURE 2 Change in possibility judgment scores in response to positive (+) and negative (−) consensus
Note. Error bars represent standard error of the mean.

11.3.3.4 Baseline Certainty
This increased malleability on improbable events may be partially due to lower certainty about their reality status to begin with. Thus, we examined children’s baseline certainty (before testimony) on both event types. We coded baseline certainty as 1 (a little sure) or 2 (very sure) regardless of the baseline answer provided. Indeed, using a paired samples \( t \)-test, we found that U.S. children were significantly more certain of their initial answers for impossible events
Children's relatively decreased baseline certainty on improbable items may thus have contributed to their willingness to shift judgments on these events.

11.3.4 Chinese Children’s Response to Consensus

11.3.4.1 Improbable Events

As shown in Figure 1, as with U.S. children, Chinese children’s possibility judgments for improbable events were swayed by testimony to some degree. Paired samples t-tests revealed that positive testimony significantly heightened judgments ($M_{positive} = 3.23, SD = .58; M_{baseline} = 1.73, SD = .81; t(30) = 6.29, p < .001$). Negative testimony lowered possibility judgments only at trend level when corrected for multiple comparisons ($p = .006$).

11.3.4.2 Impossible Events

Chinese children exhibited a similar pattern of receptivity to consensus on impossible items, as shown in Figure 1. Again, positive consensus significantly increased judgments: $t(30) = 2.71, SD = 1.12, M_{positive} = 2.71, SD = .50, and M_{baseline} = 1.27, SD = .50$, and negative consensus appeared to lessen baseline judgments though was not significant when corrected for multiple comparisons ($p = .01$). Of course, as with U.S. participants, this was not surprising given floor level baseline judgments of these events.

11.3.4.3 Change on Improbable vs. Impossible Events

Unlike U.S. participants, Chinese children seemed to respond similarly to consensus for both improbable and impossible events (see Figure 2). Paired samples t-tests on change scores across item type confirm this statistically: shifts in judgment for improbable events were of the same magnitude as those for impossible events with both positive ($p = .703$) and negative testimony ($p = .024$). Chinese participants, then, seemed to prioritize testimony over event type, responding similarly to consensus regardless of whether the target event was improbable or impossible.

11.3.4.4 Baseline Certainty

We asked whether this willingness to shift for both types of events could be attributed to being equally uncertain about the reality status of both types of events to begin with. Using a paired samples t-test on baseline certainty for both events, we found that Chinese children, like U.S. children, were significantly more certain of their initial answers for impossible events ($M = 1.79$, $SD = .31$) than they were for improbable events ($M = 1.57, SD = .36; t(24) = 3.41, p = .002$).
effect of consensus testimony on possibility judgments

That this difference in certainty exists despite the similar change patterns for both events, suggests that among Chinese participants, judgment shifts for impossible items cannot be attributed solely to baseline uncertainty.

11.3.5 Response to Consensus between Countries

As can be seen in Figure 1, U.S. and Chinese children display contrasting response patterns to testimony for each event type. For impossible events, independent samples $t$-tests revealed that Chinese participants showed significantly higher possibility judgment scores after positive consensus than did U.S. children ($t(54) = 4.95, p < .001$). Possibility judgments after negative consensus, at floor to begin with, were equivalent between countries. Of course, direct comparisons of judgments after testimony between countries are insensitive to the baseline differences that emerged, and fail to capture whether participants changed their initial responses after hearing testimony. An independent samples $t$-test on change scores across countries, however, confirmed that Chinese participants shifted with positive consensus to a greater degree than did U.S. participants, $t(54) = 4.09, p < .001$, as can be seen in Figure 2.

For improbable events, Chinese participants showed significantly lower possibility judgment scores than did their U.S. counterparts after negative consensus, ($t(54) = 4.61, p < .001$), though judgment scores after positive consensus were relatively equivalent between countries when corrected for multiple comparisons, $p = .017$. This is notable given that U.S. baseline possibility judgments of these events were significantly higher than those of Chinese participants, suggesting greater shifts among the Chinese participants (Figure 2). Independent samples $t$-tests examining mean judgment change on improbable items by culture confirm this, showing that positive consensus heightened Chinese participants’ possibility judgments ($M_{change} = 1.28, SD = .68$) significantly more than U.S. participants’ judgments, $t(54) = 2.93, p = .005$), though judgment shifts with negative consensus were equal between countries ($p = .78$).

11.4 Receptivity to Contradictory Testimony

Recall that because participants were randomly assigned to hear either negative or positive consensus for each item irrespective of their baseline judgments, not all participants heard testimony that conflicted with their original response. Thus, for the next set of analyses, we excluded all cases where particular individuals did not have the opportunity to change their judgments, including only cases where individuals heard testimony that contradicted their initial judgment. The final subsample for this analysis consisted of 15
U.S. children and 30 Chinese children who had at least one opportunity to correct an initial judgment, and the entire U.S. ($n = 25$) and Chinese ($n = 31$) sample who had at least one opportunity to regress from an initially correct judgment (see Table 2 for the composition of all cases used in the analysis). Each response was coded as 1 if the individual changed their judgment, and 0 if the individual retained their baseline judgment. A response was termed a regression if the participant changed an initially correct judgment to an incorrect one after hearing incorrect testimony or a correction if the participant changed an initially incorrect judgment to a correct one after hearing correct testimony.

### 11.4.1 Full Model

As suggested by Figure 3, U.S. and Chinese participants appeared to display unique patterns of corrections and regressions following conflicting testimony. A repeated measures ANOVA with change direction (2: correction, regression) as the within subject variable and culture as the between subject variable (2: U.S., China) revealed a significant main effect of change direction, $F(1,43) = 43.912, p < .001, \eta^2_p = .51$ along with a significant interaction between change direction and culture, $F(1,43) = 14.20, p < .001, \eta^2_p = .25$. An independent samples $t$-test on corrections across cultures showed that U.S. and Chinese participants exhibited an equivalent tendency to correct their answers when
given the opportunity, $M_{US} = .39$, $SD = .24$, $M_{China} = .85$, $SD = .33$; $t(43) = .86$, $p = .393$. (It should be noted that corrections were driven primarily by responses to improbable items, as shown in Table 2.) An independent samples $t$-test on regressions across cultures, however, revealed that Chinese participants ($M = .66$, $SD = .40$) were significantly more likely than U.S. participants ($M = .27$, $SD = .37$; $t(54) = 4.61$, $p < .001$) to move from a correct to an incorrect answer after hearing testimony to that effect. Furthermore, whereas U.S. participants made significantly more corrections than regressions, $t(14) = 7.39$, $p < .001$; Chinese participants appeared to correct and regress at a relatively equivalent rate, $t(29) =2.33$, $p = .027$ (not significant when corrected for multiple comparisons).

### Table 4 Frequency of corrections and regressions for each item type by culture

<table>
<thead>
<tr>
<th></th>
<th>Agree with consensus</th>
<th>Disagree with consensus</th>
<th>Total cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U.S.</td>
<td>China</td>
<td>U.S.</td>
</tr>
<tr>
<td><strong>Corrections</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improbable</td>
<td>18</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>Impossible</td>
<td>3</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td><strong>Regressions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improbable</td>
<td>16</td>
<td>18</td>
<td>32</td>
</tr>
<tr>
<td>Impossible</td>
<td>7</td>
<td>36</td>
<td>44</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>54</td>
<td>76</td>
</tr>
</tbody>
</table>

11.4.1.1 **Baseline Certainty**

To attempt to explain these patterns, we examined whether baseline certainty predicted participants’ willingness to regress or correct in the face of contradictory testimony. Using this subset of participants who had the opportunity to change their answers, we conducted a $t$-test comparing children’s baseline certainty when they agreed with the contradictory consensus after hearing it to their baseline certainty when they disagreed with it. We found that when U.S. children retained their initial answer in the face of contradictory testimony, they were more certain of this answer to begin with ($M = 1.77$, $SD = .34$) than when they changed their answer to accord with the testimony ($M = 1.37$, $SD = .38$; $t(13) = 2.85$, $p = .014$). Similarly, Chinese children who retained their initial answers were more certain of this answer to begin with ($M = 1.81$, $SD = .22$) than when they changed their answer ($M = 1.50$, $SD = .42$; $t(15) = -2.70$, $p = .017$).
Discussion

Though in multiple studies children have clung to their conceptions of event possibility even in the face of testimony, our data suggests that both Chinese and U.S. children’s judgments about possibility can indeed be influenced by consensus endorsements. U.S. and Chinese children were particularly responsive to consensus testimony about improbable events, and both U.S. children and Chinese children corrected initially incorrect answers at equivalent rates after hearing correct testimony. Thus, as in object naming studies or as implied in Lane et al. (2018), consensus testimony seems to carry a weight above and beyond testimony of a single individual in the domain of possibility.

However, significant cultural differences also emerged. First, at baseline, U.S. children asserted that improbable events could occur significantly more frequently than did Chinese children, thus appearing to differentiate the two types of events to a greater degree. In response to testimony, Chinese participants adjusted possibility judgments regardless of whether the event in question was improbable or impossible. By contrast, U.S. participants adjusted judgments for improbable events – in both directions – but were generally resistant toward re-evaluating their initial judgments of impossible events. For both types of events, Chinese children changed with testimony significantly more than U.S. children. Essentially, though both Chinese and U.S. participants used testimony and event type to some degree when making possibility judgments, Chinese participants seemed to prioritize testimony while U.S. participants seemed to prioritize event type.

12.1 Baseline Possibility Judgments

Because no previous studies had compared U.S. and Chinese children’s possibility judgments, we solicited such responses before presenting children with testimony. That participants from each country differentiated between improbable and impossible events was expected; however, the direct contrast between U.S. and Chinese baseline possibility judgments of improbable events was unexpected.

One possible explanation for this cultural divergence might derive from differences in how such events are portrayed and categorized in Chinese versus Western media. While Western literature falls into distinct genres, and magical events often cluster into a fantasy oeuvre; Chinese literature tends to blur Western boundaries, and magical events are not quarantined to a specific genre (Gu, 2006). Thus, Western children may develop expectations about the co-occurrence and operation of particular events based on exposure to these events in literature (Kibbe, Kreisky, & Weisberg, 2018), while Chinese children may form alternate expectations based on exposure to different combinations
of stimuli. This might explain why U.S. children distinguished improbable and impossible events more than did Chinese children both at baseline and in the presence of testimony.

Furthermore, exploratory individual item analysis of yes – no possibility judgments suggests that while U.S. participants’ possibility judgments cluster clearly by item type (improbable items range from 52–96% while impossible items range from 0–8%), Chinese evaluation does not (See Appendix B for breakdown of binary responses by item). Perhaps Chinese children considered the possibility of certain items, such as opening a window with one’s mind, figuratively. This is consistent with results from adult work that suggests that divergence on Chinese and U.S. evaluations of specific items can be attributed to differences in nonliteral interpretation among Chinese participants (e.g., Gong & Shtulman, 2021). In current work in our labs, we are asking children to explain why these events are possible (or not); we expect these findings to be highly informative in determining the criteria that children use to arrive at their baseline judgments.

In the current study, items used were extracted from Shtulman and Carey (2007) and were selected by U.S. researchers to measure the developmental trajectory of U.S. children’s evaluation of improbable and impossible events. A future cross-cultural study examining the developmental trajectory of conceptions of possibility might use pilot ratings from adults from each nation for item selection to establish an equivalent baseline.

12.2 Effect of Consensus Testimony on Possibility Judgments

Both U.S. and Chinese children shifted these baseline possibility judgments in response to testimony, albeit to different degrees. First, however, with regard to improbable events, Chinese and U.S. response patterns were more similar than different. Children from both countries displayed significantly higher possibility judgments after positive consensus, suggesting that they regarded positive consensus as valuable information. Similarly, negative consensus resulted in possibility judgments that were significantly lowered from baseline for U.S. children and marginally lowered for Chinese children.

Differences emerged with respect to impossible events. Chinese children appeared to be significantly more responsive to consensus about impossible events than did U.S. children. For U.S. children, neither type of consensus resulted in possibility judgments that were significantly different than baseline. For Chinese children, on the other hand, positive consensus resulted in judgments that were significantly higher than those at baseline, while negative consensus resulted in judgments that were marginally lower (though did not reach significance when corrected for multiple comparisons). Furthermore, although both U.S. and Chinese children corrected initially incorrect answers
at equivalent rates after hearing correct testimony, Chinese children changed initially correct answers to incorrect ones after hearing incorrect testimony significantly more frequently than did U.S. children.

Taken together, though both U.S. and Chinese children incorporated testimony into their final possibility judgments, Chinese children exhibited greater levels of receptivity to consensus testimony overall. U.S. children modulated their receptivity to consensus based on the nature of the underlying event, incorporating the testimony into their possibility judgments only for improbable events. Chinese children, by contrast, seemed to focus primarily on the nature of the consensus, adjusting possibility judgments in the direction of the consensus similarly for both event types.

Chinese children’s sensitivity to testimony and U.S. children’s sensitivity to the nature of the event could each be explained by cultural variations in openness to experience. Chinese children may be exhibiting cultural openness (often characterized by interest in cultural stimulation), while U.S. children may be exhibiting intellectual openness (characterized by interest in ideas, theories; Woo, Chernyshenko, Longley, Zhang, Chiu, & Stark, 2004). In a lower-level structural analysis of openness to experience, Woo et al. (2004) suggest that Chinese adults fall higher on the first dimension, whereas U.S. adults fall higher on the second. Both may affect learning – for example, openness to cultural experience may result in greater attunement to cultural input, while openness to intellectual experience may result in greater reliance on imagination and analytical reasoning. If Chinese and U.S. children share these characteristics with adults, Chinese children may have prioritized testimony over preexisting conceptions, while U.S. children may have prioritized preexisting conceptions over testimony when the two were in conflict.

12.3 Limitations and Directions

Although children often changed their possibility judgments in response to testimony, we cannot know from the present data whether these changes represent actual belief change or just the desire to conform – the robust impulse to behave like those around us regardless of a possibly contradictory underlying belief. Children in our study may be simply respectfully deferring, responding out of a sense of social obligation as opposed to genuine belief, and our phenomenon may simply be another illustration of the tendency to conform. However, the story may be more nuanced.

As Jaswal and Kondrad (2016) have proposed, children’s decisions to use testimony when making judgments may be epistemically motivated (i.e., reflective of actual belief change), or socially motivated, reflecting the pursuit of nonbelief-related but perhaps more salient social goals. Our data suggest that U.S. and Chinese children may have prioritized these sometimes competing
epistemic and social goals differentially, leading to possibility judgments reflective of true belief to different degrees.

Among U.S. children, our analyses of certainty judgments indicated that uncertainty may have motivated judgment change, with children changing their responses more on improbable than on impossible events. Furthermore, in response to contradictory testimony, U.S. changed initially incorrect answers at greater rates than initially correct answers. These together suggest that epistemic belief change may have been involved. By contrast, Chinese children changed their answers at similar rates for both improbable and impossible events and regardless of whether the testimony was correct or incorrect, suggesting that their decision to incorporate testimony may have been socially motivated, that is, they may have said what they thought they were expected to say rather than what they believed.

However, children in both cultures who displayed less certainty during their baseline judgments of event possibility were more likely to change their response after hearing contradictory testimony. If children, in a polite gesture of respectful deference, merely repeated the answer they heard while holding underlying beliefs intact, they should have changed at rates irrespective of their initial certainty level. However, as the subset of participants who heard contradictory testimony was not representative of our entire sample, we caution drawing generalizations from these findings. In the end, to what degree these changes in possibility judgments reflect epistemic belief change and social acquiescence is a question for future research.

A further limitation of our research is that we only included one age group. Because there is significant development with respect to concepts of probability, future studies should probe the developmental trajectory of the influence of consensus on possibility judgments across cultures. Are younger children more willing to accede to consensus or more willing to defy it? How does this differ between cultures? Research in this vein seems particularly important as children grow up in a world where consensus information is readily available not only in person, but also through media sources, most notably, the internet and social media.

Our findings extend the growing body of research demonstrating the powerful role of consensus testimony in children’s learning (e.g., Corriveau et al., 2008; Corriveau et al., 2017; DiYanni et al., 2015). Yet additional work is needed to explore how various forms of testimony affect children’s possibility judgments. Although research indicates that a single informant’s claims that an improbable or impossible event can happen would not be enough to change their judgment (e.g., Lane et al. 2018; Lane et al., 2016; Woolley et al., 2015), it seems important to know whether it might do so in this context and whether the results would differ across cultures. Because our trials all involved three
informants, we also don’t know whether, for example, one or even two infor-
mants would have produced the same results, or what effects multiple infor-
mants relaying conflicting information might have. Moreover, the question
remains whether our effect is unique to the domain of possibility, or whether
this effect extends to similar abstract concepts such as – to name a few –
scientific facts, superstitions, or religious beliefs.

Acknowledgements

We thank Jennifer Clegg-Petz, Christine Coughlin, Kelsey Kelley, and Isabella
Schloss for their comments on a draft of this manuscript. We thank the follow-
ing research assistants for their help in facilitating data collection: Paola Baca,
Zoe Bass, Silviya Bastola, Cameron Bates, Katherine Bos, Ashlea Cooper, Helen
Wang Ding, Grace Holt, Avery Largent, Melanie Lown, Yao Lu, Claire Perkins,
Lexi Reichstein, Hayden Waterstreet, and Madeline Young (U.S.); Jie Huang,
Haotong Jiang, Xiaoqi Qiu, Jiawen Wang, Qinyin Wang, Yijia Wang, Shasha
Yi, and Fangqiong Zhang (China). When the manuscript for this article was
drafted, three authors self-identified as U.S. White American, and one author
self-identified as Chinese. This research did not receive any specific grant from
funding agencies in the public, commercial, or not-for-profit sectors.

References

Abingdon, UK: Routledge.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of
Carnegie Press.

Likert scales. *Medical Education, 42*(12), 1150–1152.

Chen, E. E., Corriveau, K. H., & Harris, P. L. (2013). Children trust a consensus com-
posed of outgroup members – but do not retain that trust. *Child Development, 84*(1),
269–282.


Corriveau, K. H., & Harris, P. L. (2010). Preschoolers (sometimes) defer to the majority


Appendix A

Table A.1 Ordinary event means and pairwise comparisons

<table>
<thead>
<tr>
<th></th>
<th>Means (SD)</th>
<th>Pairwise Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. baseline poss.</td>
<td>3.98 (.10)</td>
<td>–</td>
</tr>
<tr>
<td>judgment (USO)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U.S. poss. judgment</td>
<td>3.98 (.10)</td>
<td>t(30) = 3.86</td>
</tr>
<tr>
<td>post-testimony (US+)</td>
<td></td>
<td>p = .001*</td>
</tr>
<tr>
<td>China baseline poss.</td>
<td>3.73 (.28)</td>
<td>t(54) = 4.26</td>
</tr>
<tr>
<td>judgment (CO)</td>
<td></td>
<td>p &lt; .001**</td>
</tr>
<tr>
<td>China poss. judg.</td>
<td>3.92 (.19)</td>
<td>t(54) = 1.46</td>
</tr>
<tr>
<td>post-testimony (C+)</td>
<td></td>
<td>p = .15</td>
</tr>
</tbody>
</table>

Note. * indicates significance after Bonferroni correction. We expected U.S. and Chinese children to perform at ceiling on these items before and after testimony. We found, surprisingly, that Chinese responses indicated less certainty prior to testimony (when compared to both their own responses after testimony and U.S. responses prior to testimony). The effect was driven by 15 Chinese participants, who said they were only “a little sure” that someone could wear a yellow hat in real life. By contrast, all participants indicated that they were “really sure” that someone could drink milk in real life. Our finding may have been due to the accidental inclusion of the yellow hat item: we learned post hoc that attendees of the University school where we conducted our research wear yellow hats when on class field trips. Thus, they are only able to wear these hats in school-related contexts with the permission of an authority figure. However, the inclusion became an interesting accident, as it seems to suggest that Chinese children may in fact base their possibility judgment in this case largely on context. Thus, wearing a yellow hat is possible, but only in some circumstances. The fact too that testing occurred within a school environment may have made this issue especially salient. Of course, other interpretations are also possible, such that children interpret “in real life” to mean outside of a school-related context, which is also entirely feasible given that these children were all tested during the school day. As we have no records of spontaneous justifications made by participants at the Chinese location, we do not address either speculation further.
Appendix B

<table>
<thead>
<tr>
<th>Item type</th>
<th>Item</th>
<th>U.S.</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinary</td>
<td>Wear a yellow hat</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Drink a glass of milk</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Improbable</td>
<td>Eat pickle flavored ice cream</td>
<td>.72</td>
<td>.13</td>
</tr>
<tr>
<td></td>
<td>Own a peacock for a pet</td>
<td>.52</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>Paint polka dots on an airplane</td>
<td>.92</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td>Make a building shaped like a teapot</td>
<td>.96</td>
<td>.61</td>
</tr>
<tr>
<td>Impossible</td>
<td>Open a window with one's mind</td>
<td>.08</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>Travel back in time</td>
<td>.04</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>Find a tree with money growing on it</td>
<td>0.0</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Walk through a brick wall</td>
<td>.04</td>
<td>.10</td>
</tr>
</tbody>
</table>