US-China differences in cognition and perception across 12 tasks: Replicability, robustness, and within-culture variation

Anjie Cao*[1], Alexandra Carstensen*[2], Shan Gao[3], & Michael C. Frank[1]

[1] Department of Psychology, Stanford University

[2] Department of Psychology, University of California, San Diego

[3] Department of Psychology, University of Chicago

Author Note

We gratefully acknowledge Alvin Tan, Joseph Outa, and the members of the Language and Cognition Lab at Stanford for comments and assistance. We are grateful for feedback and suggestions from Jenny Yang, Catherine Thomas, Ellen Reinhart, Erik Santoro, Kengthsagn Louis, Leslie Remache, Hazel Markus, and Shinobu Kitayama. We also want to thank the authors who provided the original experiment materials. Experiment 1 was previously reported in abbreviated form in the Proceedings of the Cognitive Science Society as Carstensen et al. (2020).

Correspondence concerning this article should be addressed to Anjie Cao*, 450 Jane Stanford Way, Stanford, 94305. E-mail: anjiecao@stanford.edu

Abstract

Cultural differences between the US and China have been investigated using a broad array of psychological tasks measuring differences between cognition, language, perception, and reasoning. Using online convenience samples of adults, we conducted two large-scale replications of 12 tasks previously reported to show differences between Western and East Asian cultures. Our results showed a heterogeneous pattern of successes and failures: five tasks yielded robust cultural differences, while six showed no difference between cultures, and one showed a small difference in the opposite direction. We observed moderate reliability for all multi-trial tasks, but there was little relation between task scores. As in prior work, cross-cultural differences in cognition (in those tasks showing differences) were not strongly related to explicit measures of cultural identity and behavior. All of our tasks, data, and analyses are openly available for reuse, providing a foundation for future studies that seek to establish a robust and replicable science of cross-cultural difference.

Keywords: replication; cross-cultural differences; cognition; perception; US-China comparison

Word count: 10486

US-China differences in cognition and perception across 12 tasks: Replicability, robustness, and within-culture variation

Introduction

Cross-cultural differences are a striking part of the broader landscape of human variation. Differences in values and behavior across cultures are obvious to even a casual observer, and researchers have attempted to quantify these differences via a wide range of measures. Comparisons between Western and East Asian cultures have been especially well-researched, with differences attested in a wide range of cognitive domains, including visual attention (Chua, Boland, & Nisbett, 2005; Ji, Peng, & Nisbett, 2000; Waxman et al., 2016), executive function (Sabbagh, Xu, Carlson, Moses, & Lee, 2006; B. Tan, 2020), language learning (Chan et al., 2011, 2011; Tardif, 1996; Waxman et al., 2016), relational reasoning (Carstensen et al., 2019; Cheng, 2020; Richland, Chan, Morrison, & Au, 2010; Su, 2020), similarity judgments (Ji, Zhang, & Nisbett, 2004), values (Ji, Nisbett, & Su, 2001; Kwan, Bond, & Singelis, 1997; Spencer-Rodgers, Williams, Hamilton, Peng, & Wang, 2007), preferences (Corriveau et al., 2017; DiYanni, Corriveau, Kurkul, Nasrini, & Nini, 2015; Liang & He, 2012) and self-concepts (Spencer-Rodgers, Boucher, Mori, Wang, & Peng, 2009; Spencer-Rodgers, Boucher, Peng, & Wang, 2009). As a result, Western and East Asian cultures are increasingly treated as cultural poles in efforts to measure cultural differences (Muthukrishna et al., 2020) and to correct for the pervasive bias in psychology research toward US and European samples (Arnett, 2016; Henrich, Heine, & Norenzayan, 2010; Nielsen, Haun, Kärtner, & Legare, 2017).

Despite a long empirical tradition of comparisons between these cultures and an abundance of psychological accounts for observed differences, estimates of differences are difficult to compare quantitatively because of the varying samples, measures, and methods used in different reports. Further, many of the most prominent reports of cross-cultural differences predate the field-wide discussion of methodological issues in psychology research

during the past 10 years (Open Science Collaboration, 2015). For example, much research in this tradition has been exploratory and hence has not followed current guidance regarding limiting analytic flexibility in order to decrease false positives (Simmons, Nelson, & Simonsohn, 2011). Given the importance of evidence about specific cross-cultural differences for constructing theories of culture more broadly (e.g., Markus & Kitayama, 1992, 2010), further investigation of many empirical findings is likely warranted.

Some empirical evidence points to issues in the robustness of cross-cultural measurements. Typically, measures used in this literature are not standardized and do not have published evidence about reliability and validity (Flake & Fried, 2020). The few extant direct comparisons between measures of cultural difference suggest that theoretically related tasks, such as implicit and explicit measures of the same construct, might not cohere (e.g., Kitayama, Park, Sevincer, Karasawa, & Uskul, 2009). Further, in a study with twenty cross-cultural measures used within a single US sample, Na et al. (2010) found a lack of coherence between tasks measuring social orientation and cognitive style, observing only 8 significant correlations between tasks across 90 statistical tests.[1] Finally, more recent work failed to replicate cultural differences on several related measures (Mercier, Yama, Kawasaki, Adachi, & Van der Henst, 2012; Mercier, Zhang, Qu, Lu, & Van der Henst, 2015; Zhou, Gotch, Zhou, & Liu, 2008). Thus, there is a need to explore the reliability of individual tasks as well as the intercorrelations between them.

Our goal in the current study was to collect a large dataset on a range of cross-cultural measures that had previously been used in comparisons of East Asian and Western cultures, enabling investigations of the robustness of these differences in new samples of Chinese and US participants. We decided to gather relatively large and heterogeneous convenience samples using online recruitment, rather than recruiting

[1] These authors interpreted their findings as implying that the measures are orthogonal – indexing different constructs – and concluded that group-level differences between cultures are unlikely to relate to within-group individual differences. However, an alternative possibility is that the reliabilities of many individual tasks are low, a feature which would ensure low correlations between them.

smaller, more matched samples using in-lab recruitment. Our reasoning was that the larger samples that we could access using online recruitment would allow us to conduct highly-powered statistical tests, allowing us to make well-powered tests for cultural differences. Further, larger samples afford the analysis of individual and demographic differences within culture, a topic of considerable interest in this literature (e.g., Na et al., 2010, 2020). Finally, the development of browser-based online versions of prominent cross-cultural tasks would allow their inspection and reuse by other researchers, thus promoting a more cumulative approach to the measurement of cultural differences.

Our experiments were intended to be close replications of the original studies, but differences in format of administration introduced inevitable variation, in some cases more substantial than others. The interpretation of discrepant outcomes between an original study and a replication is complex, given that disparate outcomes can occur for many reasons (Machery, 2020; Nosek & Errington, 2020; Zwaan, Etz, Lucas, & Donnellan, 2018). In our case, interpretation is especially difficult and we explicitly avoid interpreting our results as bearing on the status of the original findings we investigate.

In particular, there are a number of important differences between our experiments and the original studies. First, we recruited online convenience samples from the U.S. and China. Previous work varied in the participants' country of origin (in several cases, Japan for East Asian participants; Canada for Western participants), largely recruited either college students or community members, and was administered more than a decade ago. Within-culture variation and generational differences between our samples and previous samples make results difficult to compare directly. Furthermore, our strategy of constructing a battery of replication studies and administering them uniformly online altered the contexts in which participants engaged with the tasks and in some cases required alterations to the tasks themselves as well.

Accordingly, our replication studies should be viewed as an assessment of robustness:

specifically, we assess whether a set of previously-reported East-West cross-cultural differences can be recovered in online convenience populations. Thus, we reiterate that our work – much like other replication work, but even more so – cannot be taken as an assessment of the veracity of the original findings. Nevertheless, we believe that cross-cultural psychology can be advanced via the identification of tasks that yield cross-cultural differences robustly across a variety of samples and administration formats – we hope our work contributes to this aim. We return to these interpretive issues in the General Discussion.

Our task selection process was initially shaped by an interest in relational reasoning and accounts explaining it with reference to cross-cultural differences in visual attention and social cognition (Duffy, Toriyama, Itakura, & Kitayama, 2009; Kuwabara & Smith, 2012; Moriguchi, Evans, Hiraki, Itakura, & Lee, 2012). Additionally, in Experiment 1, we selected tasks that could potentially be administered to young children as well as adults, for use in future work addressing developmental questions about the relative time course of cross-cultural differences across the visual, social, and cognitive domains. We balanced three desiderata in our task selection, preferentially choosing tasks that (1) had been theoretically or empirically implicated in relational reasoning, (2) were associated with differential performance in US-China comparisons or related cultural contrasts (i.e., East Asian vs. Western cultures), and (3) were relatively short, accessible tasks appropriate for web administration. We also conducted an extensive set of pilot tests to ensure that participants understood instructions and that the tasks yielded interpretable data.

In Experiment 2, we selected a second set of tasks to investigate based in part on the results of Experiment 1. In particular, we repeated a handful of tasks from Experiment 1, in some cases varying task parameters to rule out potential explanations for failures. We then selected a further set of tasks that probed both cross-cultural differences in higher-level cognition (e.g., language and reasoning) and perception, again respecting the desideratum that the tasks should be relatively short and amenable to administration in a

web browser. The final set of tasks included in each experiment is listed in Table 1.

In addition to the goal of replicating individual tasks, our hope was that the dataset we collected could be used to explore the structure of within- and between-culture variation in cognition and perception more broadly. Towards this goal, we included a relatively extensive demographic questionnaire in both of our experiments, with the aim of using these measures to explore variation within our samples. In the final section of the paper, we report a series of exploratory analyses. The first of these assesses the reliability of individual tasks to gauge whether these tasks are reliable enough from a psychometric point of view to support further individual differences analyses. We then report correlations across tasks, aiming to discover covariation between tasks that might indicate that they load on the same construct. Finally, we turn to analyses of whether within-culture demographic variables predict variation in task performance. Overall, a number of tasks revealed acceptable levels of reliability, but tasks did not cluster together and we found relatively limited demographic predictors of within-culture variation.

We make all code and data from our experiments available for further data collection and analysis in hopes of promoting further cumulative work on measures and theories of cross-cultural variation. All of the tasks used in this study can be previewed at https://anjiecao.github.io/uschinareplication.github.io/

Table 1
Tasks included in each experiment and the final sample size after exclusions.

Experiment	Task	Citation	Task Description	CN	US
1	Ambiguous Relational Match-To-Sample (cRMTS)	Carstensen et al. (2019)	Infer whether an object or relation is causally relevant	186	178
	Picture Free Description	Imada, Carlson, & Itakura (2013)	Describe pictures from memory after a brief study period	169	172

	Ebbinghaus Illusion	Imada, Carlson, & Itakura (2013)	Judge the size of circles in a context designed to bias size judgments	190	180
	Horizon Collage	Senzaki, Masuda, & Nand (2014)	Make an image by dragging and dropping stickers onto a display	187	175
	Symbolic Self-Inflation (Family)	Kitayama et al. (2009)	Draw self and family members as circles	150	114
	Uniqueness Preference	Kim & Markus (1999)	Choose a sticker from five stickers, four of which are the same color	191	180
	Child Causal Attribution	Seiver, Gopnik, & Goodman (2013)	Watch short vignettes and explain the decisions of the characters	177	170
	Raven's Progressive Matrices	Su (2020)	Use analogical reasoning to complete visually-presented patterns	191	180
2	Ambiguous Relational Match-To-Sample (cRMTS)	Carstensen et al. (2019)	Infer whether an object or relation is causally relevant	174	293
	Picture Free Description	Imada, Carlson, & Itakura (2013)	Describe pictures from memory after a brief study period	132	284
	Change Detection	Masuda & Nisbett (2006)	Find differences in the foreground or background of two images	160	253
	Symbolic Self-Inflation (Friends)	Kitayama et al. (2009)	Draw a sociogram with self and friends as nodes, relationships as edges	158	252
	Adult Causal Attribution	Morris & Peng (1994)	Read a crime story and explain the criminal's motivations	114	293

Taxonomic-Thematic Similarity	Ji, Zhang, & Nisbett (2004)	Match items based on taxonomic or thematic similarity (e.g., cow: chicken / grass)	178	295
Semantic Intuition	Li, Liu, Chalmers, & Snedeker (2018)	Decide whether a story refers to a named character (whose actions are mischaracterized) or the person who performed the actions (but had a different name)	181	298
Raven's Progressive Matrices	Su (2020)	Use analogical reasoning to complete visually-presented patterns	181	298

Experiment 1

In Experiment 1, our goal was to evaluate cross-cultural differences in a variety of constructs. We assembled a web-based battery of tasks and tested these on a snowball sample of US and Chinese participants.

Methods

Participants. We recruited participants through snowball sampling seeded at large universities in the US and China, in which participants directly recruited by the researchers were encouraged to recruit their friends and family members through email forwarding and social media sharing. Participants in the US were compensated with $5 gift certificates (USD) and participants in China received ¥35 (CNY).

We recruited 203 and 201 participants each from the US and China, respectively. Since we did not have strong a priori expectations about specific effect sizes, our overall preregistered sample size was chosen to meet or exceed the sample sizes used in prior

reports in the literature from which our tasks were drawn. Our sample size, methods, and main analyses were pre-registered and are available at https:// aspredicted.org/37y6a.pdf.

Our preregistered exclusion plan was to exclude people from the full dataset if they failed quality checks[2] on any one task, unless this excluded 20% or more of our sample. Due to a task demand associated with the Symbolic Self-Inflation task, this criterion would have led to the exclusion of 107 people (US: 66, CN: 41) due to this task alone. This triggered the less restrictive exclusion approach in our preregistration, using task-specific quality checks to exclude participants only from the relevant individual task.

After exclusions, the US sample included 180 participants (49 Male, 120 Female, 9 Non-binary, 2 Declined to answer), with a mean age of 22.02 years old, all of whom were native English speakers. The China sample included 191 participants (60 Male, 127 Female, 1 Non-binary, 3 Declined to answer), with a mean age of 22.42 years old, who were all native speakers of Mandarin Chinese. This sample size is shared among all tasks except for the Symbolic Self-Inflation task, which included 114 US participants and 150 CN participants.

In addition to age, gender, and linguistic background, we collected a range of demographic information including subjective socioeconomic status measured using the MacArthur Ladder (Adler, Epel, Castellazzo, & Ickovics, 2000), level of maternal education, the state or province the participant grew up in, residential mobility, and

[2] We performed exclusion based on three aspects: side bias, missing data, and demographic exclusions.
(1) Side bias: If more than 90% of selections by a participant in a 2AFC task were a single response button (left or right), data from this participant in this task would be excluded.
(2) Missing data: If no data/more than 25% missing/data not codeable/side bias/participant did not follow instructions for any one task, then we would exclude all data from that participant, across tasks. If this led to 20% or more of participants being excluded, we would not apply this exclusion criteria to all data from the participant – only at the level of individual tasks.
(3) Demographic exclusions: We would exclude data from [CN/US] participants who reported living abroad for more than 2 years in regions with predominantly [European/Asian] populations (respectively); and [CN/US] participants who reported speaking or understanding [English/any Chinese language or dialect] with proficiency at or above 3 out of 10. If this led to 20% or more of participants being excluded from either test population (US/CN), we would drop this exclusion criterion for the relevant population and use exploratory regressions to examine how the factor relates to responding in our tasks.

number of international experiences.

We conducted power analysis simulations under a range of assumptions and found that our sample sizes were well-powered to detect medium and large effect sizes, with more than 80% power for effect sizes greater than 0.5 (Cohen's d) for nearly all sample sizes for our studies, even after exclusions. See Appendix A for more details.

Procedure. Participants completed an online, browser-based sequence of eight tasks (see Table 1) and a brief demographic questionnaire. All tasks were implemented in a combination of jsPsych (De Leeuw, 2015) and custom HTML/JavaScript code. Tasks were administered in English for the US sample and in Mandarin Chinese for the China sample. To control for the impact of order-related inattention, task order was randomized across participants with two exceptions: (1) the two drawing tasks (Symbolic Self-Inflation and Horizon Collage) were always back-to-back in random order, and (2) Uniqueness Preference was always the penultimate task (in keeping with the task cover story, which congratulated participants on being nearly done with the experiment). In total, the experiment took about 30 minutes to complete.

Measures. Below, we give a short description of prior findings and methods for each task.

Ambiguous cRMTS. Carstensen et al. (2019) observed cross-culturally distinct developmental trajectories in a causal relational match-to-sample (cRMTS) task, and different preferences in an ambiguous formulation of this task. Specifically, when 3-year-olds saw evidence consistent with both object-based (e.g., blue cubes make a machine play music) and relational (pairs of different objects, AB, make a machine play music) solutions, children in the US sample preferentially chose the object-based solution, while those in China chose the relational solution.

We used this ambiguous version of the task (Carstensen et al., 2019, Experiment 3) to explore whether adults in the US and China also show differing preferences for

object-based or relational solutions. Our participants saw two pairs of objects, AB and AC, activate a machine, and were given a forced choice between an object-based solution (a *same* pair of A objects, AA) and a relational solution (*different* pair BC).

Picture Free Description. Imada, Carlson, and Itakura (2013) found that children around the age of 6 showed cultural differences in describing pictures to others. Relative to US children, Japanese children were more likely to mention the objects in the background first, as opposed to the focal objects in the picture. They also tended to provide more descriptive accounts of the background objects than their US counterparts. In our version of the task, we used a subset of seven images from the original study and adapted the task for adult participants, who studied each image for 5 seconds and then typed a description. We coded the first mentioned item (focal or background) and counted descriptors for focal and background elements.

Ebbinghaus Illusion. Both Japanese adults and children have been found to be more susceptible to the Ebbinghaus Illusion – in which context alters the perceived size of a circle – than Western participants in the US and UK (Doherty, Tsuji, & Phillips, 2008; Imada et al., 2013). We followed the Imada et al. (2013) implementation of the task, with two testing blocks: the No Context block (10 trials) and Illusion block (24 trials). The No Context block establishes baseline accuracy for discriminating which of two orange circles is larger. In the Illusion trials, the two orange circles are flanked by a grid of 8 gray circles, which are all smaller or larger than the center orange circle. The illusion occurs because the orange circles appear larger when flanked by smaller gray circles, leading to distortions in comparing the sizes of the two orange circles with differing contexts (i.e., small or large flankers). Across the 24 Illusion trials, we measured accuracy of circle size judgments as a function of the actual size difference and flanker context (helpful or misleading).

Horizon Collage. Senzaki, Masuda, and Nand (2014) found that school-age children in Japan and Canada showed culture-specific patterns when creating a collage of an outdoor scene. Japanese children drew the horizon higher, used more collage items, and

filled more space with collage items relative to Canadian children. We adapted the task from Senzaki et al. (2014) Study 2, in which participants were prompted to make a collage with stickers. Our participants could drag any of thirty images (line drawings of people, animals, houses, etc.) onto a rectangular "canvas" in the middle of the screen. There was also a sticker "horizon," a horizontal line that spanned the length of the canvas. All stickers, including the horizon, could be clicked and dragged to the canvas to produce "a picture of the outside." Participants were asked to include a horizon and any number of other stickers to create their image. We measured the height of the horizon, the number of stickers used, and the total area occupied by stickers as in Senzaki et al. (2014).

Symbolic Self-Inflation. Kitayama et al. (2009) found a difference between Western and East Asian cultures in the size of circles participants drew to represent themselves relative to other people in their social networks. Japanese participants drew circles of similar sizes to represent themselves and others, while those from Western countries (US, UK, Germany) tended to draw their "self" circles larger than those representing others, indicating a symbolic self-inflation in the three western cultures compared to Japan. We adapted this task, asking participants to draw themselves and the family members they grew up with as circles by clicking and dragging the mouse on a rectangular "canvas" to draw circles of varying sizes. They then labeled each circle for the person it represented. To maximize comparability with the original pen and paper task, the circles had no starting size – they could only be produced by clicking and dragging the mouse. Participants labeled each circle by clicking on it. When they clicked, a text box would appear in the middle of the circle. The size of the "canvas" was defined in pixels so that it varied across displays using different resolutions; this provides a screen equivalent to the in-person requirement that participants with corrected-to-normal vision participate using their normal method of visual correction (e.g., glasses). Participants completed the task at their own pace. We measured the diameter of each circle and calculated a percent inflation score for each participant by dividing the diameter of the self circle by the average

diameter of circles for all others.

Uniqueness Preference. Kim and Markus (1999) tested East Asians' and Americans' preferences for harmony or uniqueness by asking them to pick one gift pen from five options. In the condition that we replicated, the options differed only in the barrel colors, with four that were the same and one that was unique. They found that European Americans were more likely to choose the unique colored pen than East Asian participants. We adapted our task to better fit the format of our online experiment by showing a virtual "sticker book" to measure progress through all tasks in our study. At the end of each task, participants received a virtual sticker. For the uniqueness preference task, we let them select one of five dinosaur stickers that were identical except for color: four blue and one yellow (with repeated and unique colors randomized between participants). Choice of the unique vs. repeated color was recorded.

Child Causal Attribution. Previous work has shown that participants from South Korea and the U.S. attribute behaviors differently in situations where there is evidence in favor of situational explanations (Choi, Nisbett, & Norenzayan, 1999). Similarly, Chinese participants and media are more likely than their US counterparts to attribute a person's behaviors to situational context as opposed to individual traits (Morris, Nisbett, & Peng, 1995; Morris & Peng, 1994). We adapted the deterministic situation condition in Seiver, Gopnik, and Goodman (2013), a task originally designed for children. In this task, two children both engage in one activity and avoid another, suggesting that situational constraints (e.g., the latter activity being dangerous) may be guiding their decisions. Participants watched a series of four short, animated vignettes in which two children both played in a pool and neither child played on a bicycle. We then asked participants to explain in text why each child did not play on the bicycle, making for two test trials per participant. We used the prompt question from Seiver et al. (2013), which explicitly pits person attributions against situational ones: "Why didn't Sally play on the bicycle? Is it because she's the kind of person who gets scared, or because the

bicycle is dangerous to play on?" We coded each response for per-trial count of (a) personal and (b) situational attributions.

Raven's Standard Progressive Matrices. As an additional attention check as well as an exploratory measure of relational reasoning assessing performance rather than preference, we included the 12 questions from Set E of Raven's Standard Progressive Matrices. Su (2020) found cross-cultural differences between adults in the US and China in performance on this set. This set of questions was selected because it is the most difficult subset and also the one most dependent on true analogical reasoning, without alternative heuristic approaches like visual pattern completion.

Analytic approach. Data and analysis scripts are available at https://github.com/anjiecao/CCRR_writeups

The papers that we drew on for our tasks used a heterogeneous set of analytic methods. Rather than planning to replicate these specific analyses, we instead attempted to follow current best practices by using linear mixed effects models with maximal random effect structure as a unified analytic framework (Barr, Levy, Scheepers, & Tily, 2013). We fit a separate model to each task. In case of convergence failure, we followed lab standard operating procedures: pruning random slopes first and then random intercepts, always maintaining random intercepts by participant. For linear models, we report p-values derived from t-scores. For linear mixed models, we report p-values derived from z-scores, which is appropriate for relatively large samples (Blouin & Riopelle, 2004). Our key tests of interest were typically either the coefficient for a main effect of country (US/China) or an interaction of country and condition.

Results

Ambiguous cRMTS. To examine whether adults in the US and China show differing preferences for object-based or relational solutions, we ran a mixed-effects logistic

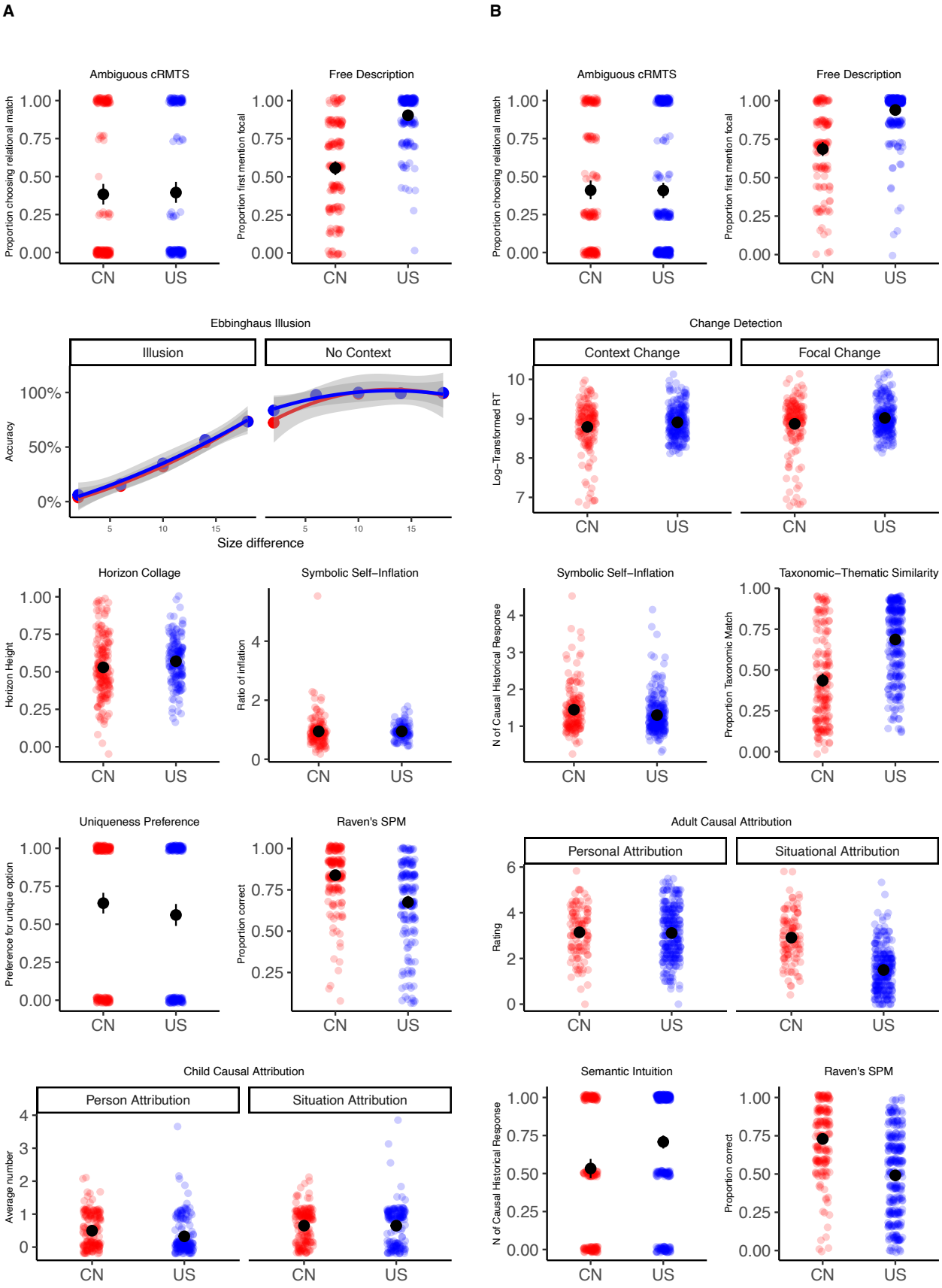

Figure 1. Results from each task. Results from the CN sample are plotted in red, and results from the US sample in blue. Panels A and B include results from Experiments 1 and 2, respectively. Results for other measures used in the Free Description and Horizon Collage are plotted in Appendix B.

regression predicting response choice (object or relation) with country (US or China) as a fixed effect. There was no main effect of country on response choice (object or relation; US: $M = 0.39$, $SD = 0.48$; CN: $M = 0.38$, $SD = 0.47$; $\beta = 0.04$, $SE = 0.84$, $z = 0.05$, $p = 0.96$). The preference for object-based solutions seen in US preschoolers in an ambiguous context and the corresponding preference for relational solutions observed in China did not extend to adults in our sample.

Our US results replicate findings by M. Goddu and Walker (2018), who reported that US adults are at chance in this paradigm. It seems likely that adults in both groups of our study are aware of the ambiguous evidence and their near-chance selections reflect (reasonable) uncertainty.

Picture Free Description. Based on Imada et al (2013), we expected Chinese participants would be more likely to mention background objects first and provide more descriptive accounts for background objects relative to focal objects, in comparison with US participants. Our results extend previous findings with the former metric (first mention; US: $M = 0.90$, $SD = 0.17$; CN: $M = 0.56$, $SD = 0.30$) but not the latter (number of descriptive accounts; for focal objects: US: $M = 1.06$, $SD = 0.51$; CN: $M = 0.87$, $SD = 0.44$; for background objects: US: $M = 1.30$, $SD = 0.93$; CN: $M = 0.94$, $SD = 0.72$).

For first mention, we ran a mixed-effects logistic regression predicting the type of first mention (object or relation) with country (US or China) as a fixed effect. We found a main effect of country ($\beta = 3.39$, $SE = 0.34$, $z = 9.98$, $p < 0.01$). For descriptive accounts, we ran a mixed-effect Poisson regression model predicting the number of descriptive accounts, with description type (focal or background), country (US or China), and their interaction as fixed effects. There was a significant main effect of culture (with US participants providing more descriptions overall: $\beta = 0.35$, $SE = 0.13$, $t = 2.67$, $p < 0.01$). The culture effect interacted with the description types, but the effect was in the opposite direction, with U.S participants providing more background descriptions than focal descriptions, relative to Chinese participants ($\beta = -0.15$, $SE = 0.07$, $t = -1.97$, $p < 0.05$).

The mixed results between the first mention and descriptive accounts measures suggest that there is some complexity in linking broader theoretical accounts to specific measures; we interpret this result with caution and include the task in Experiment 2 to follow up further.

Ebbinghaus Illusion. To test whether perception of the Ebbinghaus Illusion varied across populations in our sample, we ran a mixed-effects logistic regression predicting accuracy on each trial, with country (US or China), context (No Context or Illusion context), and circle size difference (the percent difference in diameters) as fixed effects, along with their interactions. We found main effects of context (with worse performance in the Illusion context; $\beta = 5$, $SE = 0.27$, $z = 18.64$, $p < 0.01$) and circle size difference (worse performance for smaller differences; $\beta = 0.34$, $SE = 0.01$, $z = 29.13$, $p < 0.01$). There was a marginally significant main effect of country in the opposite of the predicted direction (US participants performed worse: $\beta = 0.52$, $SE = 0.25$, $z = 2.13$, $p < 0.05$) but no interactions with country (All $\beta < 0.01$; All $p > 0.05$).

In sum, we failed to replicate cultural differences found between Western and East Asian participants in susceptibility to the Ebbinghaus Illusion.

Horizon Collage. In the Horizon Collage task, three key measurements were calculated from the "collage" participants created: the height of the horizon (in proportion to the height of the frame), the number of stickers, and the total area that the stickers covered (following the original analysis, we added up the area occupied by each individual sticker). Senzaki et al. (2014) found that Japanese children tended to put the horizon higher, include more stickers, and cover more area in their collage, compared with Canadian children. We ran a fixed effect linear model with culture as the main predictor for each of the measurements. Culture significantly predict the horizon height, but the effect was in the opposite direction direction, with U.S. participants putting the horizon sticker at higher than the Chinese participants (Sticker height: US: $M = 0.57$, $SD = 0.15$; CN: $M = 0.53$, $SD = 0.20$; $\beta = 0.04$, $SE = 0.02$, $t = 2.16$, $p < 0.05$). Culture did not

significantly predict any of the other two measurements (Sticker number: US: $M = 11.41$, $SD = 5.78$; CN: $M = 11.53$, $SD = 5.73$; Sticker area: US: $M = 16.83$, $SD = 8.28$; CN: $M = 17.18$, $SD = 8.47$; Both $\beta < 0.03$; Both $p > 0.1$).

Our experiment contrasted Chinese and US adults, rather than Japanese and Canadian children. Although Senzaki et al. (2014) found that the cultural differences were more salient in older children than younger children, suggesting that cultural differences might increase with development, interpretation of our failure to replicate is still qualified by differences in culture and medium of administration.

Symbolic Self-Inflation. To test whether US adults show greater symbolic self-inflation than Chinese adults, we ran a linear regression predicting percent inflation (the diameter of the self circle divided by the average diameter of circles for others) with country (US or China) as a fixed effect. No difference was found in the degree of symbolic self-inflation between US and Chinese adults based on percent inflation (US: $M = 0.95$, $SD = 0.26$; CN: $M = 0.94$, $SD = 0.53$; $\beta = < 0.01$, $SE = 0.05$, $t = 0.03$, $p = 0.98$).

As an exploratory analysis, we also considered whether the number of circles drawn for others could be a potential confounding factor for the effect. Due to the limited space on the canvas, people who drew more circles might draw each circle smaller, resulting in less symbolic self-inflation. We found culture to be a significant predictor of the number of others circles drawn, with US participants drawing more others circles than Chinese participants (US: $M = 3.49$, $SD = 1.26$; CN: $M = 3.12$, $SD = 1.29$). However, when controlling for the number of circles drawn, there was still no difference between US and Chinese adults on the percent inflation ($\beta = -0.01$, $SE = 0.05$, $t = -0.22$, $p = 0.82$).

One possible explanation for our null result is that there are cultural differences between Japan and China in self-concept; Japanese samples typically demonstrate characteristics previously associated with East Asian cultures in general, with Chinese samples deviating from these characteristics at times (Bailey, Chen, & Dou, 1997; Church

et al., 2012, 2014). Talhelm et al. (2014) found demographic variation within China on this task, with participants from wheat-growing regions showing greater self-inflation compared to those from rice-growing regions, who drew circles for themselves that were comparably sized to other circles (with an average inflation near zero). The performance of both Chinese and US participants in our study resembles that of the wheat-region participants in Talhelm et al. (2014), which is consistent with the possibility that our sample was biased toward wheat-region participants. We address this interpretation in later demographic analyses, but do not find support for it. Alternatively, our null results could be attributed to differences between our task design and that of Kitayama et al. (2009). Instead of asking participants to draw their social network, our design asked participants to draw themselves and the family members they grew up with. During the coding process, we noticed that people from both cultures tended to draw older people, e.g., their parents, as larger circles, which might have resulted in larger circles for others than for the self in both cultures, masking any US-China difference in the degree of self-inflation. We follow up on this possibility by changing the task prompt in Experiment 2.

Uniqueness Preference. We examined cross-cultural preferences for uniqueness by running a simple logistic regression predicting each participant's single choice (minority or majority color) with country (US or China) as a fixed effect; we used logistic regression rather than mixed effects logistic regression due to the absence of repeated observations. There was no cross-cultural difference in the probability of choosing the unique sticker (US: $M = 0.56$, $SD = 0.50$; CN: $M = 0.64$, $SD = 0.48$; $\beta = -0.32$, $SE = 0.21$, $z = -1.52$, $p = 0.13$).

The difference between our result and that of the original study by Kim and Markus (1999) may be related to our use of an online format in our study. In the original study, participants were asked to pick a gift pen from five physical pens with different barrel colors. It could be that Asian American participants in the previous study chose the more common color because they wanted to leave a choice for the next participant in the face of

resource scarcity, rather than because they were expressing values or identities influenced by East Asian cultural mandates favoring interpersonal harmony and similarity. Our finding is also consistent with previous work demonstrating that tendencies toward conformity in East Asian samples are linked to reputation management (Yamagishi, Hashimoto, & Schug, 2008); it may be that our online experiment did not establish a sufficient social context to motivate participants' concern about reputation, and accordingly failed to motivate reputation management in the form of a conformity preference.

Child Causal Attribution. To test whether Chinese participants tended to make more situational attributions than US adults, we ran a mixed-effects Poisson regression predicting the number of attributions included in each explanation, with attribution type (situational or personal), country (US or CN), and their interaction as fixed effects. We found a main effect of attribution type (situational attribution: US: $M = 0.65$, $SD = 0.61$; CN: $M = 0.65$, $SD = 0.52$; personal attribution: US: $M = 0.33$, $SD = 0.54$; CN: $M = 0.50$, $SD = 0.52$; $\beta = 0.27$, $SE = 0.10$, $z = 2.65$, $p < 0.01$). Neither the interaction nor the main effect of culture was significant (both $\beta < 0.3$; $p > 0.05$).

The failure to find cross-cultural differences in attribution could be related to the style of the tasks, which were relatively repetitive and originally designed for children; in Experiment 2, we follow up with a causal attribution task designed for adults.

Raven's Standard Progressive Matrices. As an exploratory measure of relational reasoning, we ran a mixed-effects logistic regression predicting per-trial accuracy, with country as a fixed effect, random intercepts for each subject and question, and by-question random slopes for country. We found a main effect of country, with Chinese participants outperforming those from the US (US: $M = 0.68$, $SD = 0.25$; CN: $M = 0.84$, $SD = 0.17$; $\beta = -1.36$, $SE = 0.24$, $z = -5.77$, $p < 0.01$).

Our findings replicate Su (2020) in finding an advantage for Chinese participants on Raven's Matrices. In our context, we also interpret the relatively high scores we observed

as evidence that participants were engaging fully with our tasks.

Discussion

We did not observe cross-cultural differences in the majority of tasks in Experiment 1. The only exceptions were in Picture Free Description and our exploratory measure of performance in relational reasoning (Raven's SPM). We also found cultural difference in the opposite direction in one of the measures in the Horizon Collage task. Many of our tasks did not have a manipulation check and could yield null results simply by virtue of inattention. However, the results of Raven's SPM (and the Ebbinghaus Illusion) suggest that participants were engaged in our tasks and performed at a high objective level. In addition to minor methodological changes that we made, interpretation of our failure to replicate individual tasks in many cases could be due to (1) differences in administration (online vs. in-person), (2) differences in participant recruitment (e.g., university pool vs. snowball sampling), (3) differences in target age (adults vs. children), and (4) differences in sample (e.g. Japanese vs. Chinese adults in the East Asian group).

Our failure to find robust differences between Western and East Asian cultures in this initial selection of tasks was dispiriting. We designed Experiment 2 to extend Experiment 1 by recruiting a different sample and identifying followup or replacement tasks that we hoped would yield a broader set of cross-cultural differences.

Experiment 2

Experiment 2 was designed to follow up on Experiment 1 and further evaluate cross-cultural differences across a battery of tasks. Because several of our tasks in Experiment 1 yielded no evidence for cross-cultural differences, we replaced these with alternative tasks selected to address similar or related constructs. We replaced the Ebbinghaus Illusion with a measure of Change Detection that has been argued to index

context sensitivity (Masuda & Nisbett, 2006). We replaced the child-appropriate causal attribution task with a version designed for adults (Morris & Peng, 1994). We also included two tasks measuring linguistic or semantic intuitions more broadly (Taxonomic/Thematic Similarity and Semantic Intuition), following up on the detection of cross-cultural differences in the Picture Free Description task. Although our goal in Experiment 2 was to evaluate a further set of tasks, we also included the Ambiguous cRMTS, Picture Free Description, and Raven's Progressive Matrices tasks to replicate our results from Experiment 1, and we included a modified version of Symbolic Self-Inflation to address several issues with the earlier version of the task.

In Experiment 2, we made use of crowd-sourcing services – rather than snowball sampling – as our participant recruitment channel. Each of these sampling strategies has strengths and weaknesses. The strength of snowball sampling (used in Experiment 1) is the ability to sample from comparable university/student populations, but crowd-sourcing services allow us to access convenience populations more easily and scale more flexibility. There are also two additional rationales. First, in Experiment 1 our samples were quite young (due to seeding our sampling with university students through email and social media). A younger sample may be less enculturated because they are less experienced or more exposed to international media and influences, and thus less likely to show distinct cross-cultural differences. Second, we were concerned that being recruited by friends and family (as in a snowball sample) might prime interdependent thinking among our participants, leading to decreased cross-cultural differences.

Methods

Participants. We recruited participants through online crowdsourcing websites. For the US, we used Prolific and applied the following screening criteria: a) US nationality, b) born in the US, and c) currently residing in the US. For China, we used Naodao (www.naodao.com), a platform designed for conducting online experiments in mainland

China. Participants in US received $12.25 in compensation and in China ¥35.

We recruited 304 participants from the U.S. and 185 participants from China. 10 participants were excluded because they did not meet our demographic inclusion criteria[3]. Following our preregistration (available at https://osf.io/u7mzg), we applied a task-based exclusion procedure in which we excluded a participant's responses in a particular task if they a) showed a response bias for a single response button or value[4], b) had missing data on more than 25% of trials[5], or c) failed to meet the inclusion criteria for that task as specified in the preregistration[6].

Similar to Experiment 1, we collected demographic information from participants, including subjective socioeconomic status, the state or province the participant grew up in and the one they currently reside in, residential mobility, number of international

[3] Demographic exclusions: we would exclude data from [CN/US] participants who reported living abroad for more than 2 years in regions with predominantly [European/Asian] populations (respectively); and [CN/US] participants who reported speaking or understanding [English/any Chinese language or dialect] with proficiency at or above 3 out of 10. If this led to 20% or more of participants being excluded from either test population (US/CN), we would drop this exclusion criterion for the relevant population and use exploratory regressions to examine how the factor relates to responding in our tasks.

[4] If more than 90% of selections by a participant in a 2AFC task were a single response button (left/right in RMTS or top/bottom in taxonomic/thematic similarity), data from this participant in this task would be excluded. If 100% of selections by a participant in a scalar (e.g., Likert scale used in causal attribution or in the demographics section) or multiple choice task (e.g. Raven's SPM) used a single response button/value, data from this participant in this task would be excluded.

[5] If no data/more than 25% missing/data not codeable/side bias/participant did not follow instructions for any one task, then we would exclude all data from that participant, across tasks. If this led to 20% or more of participants being excluded, we would not apply this exclusion criteria to all data from the participant – only at the level of individual tasks.

[6] Additional task and trial-level exclusions:
(1) Change detection: trials where the participant incorrectly identified the change would not be analyzed (but will not be counted as missing data unless a response was not attempted).
(2) Free description: trials where the response was not code-able would be discarded and considered missing data.
(3) Causal attribution: none.
(4) Symbolic self-inflation: data from participants who failed to draw exactly one "self" circle would be excluded, as well as data from participants who drew only a "self" circle.
(5) Taxonomic/thematic similarity task: There were two unambiguous catch trials intermixed with regular trials in this task (e.g., Choose cat: cat, dog). If a participant missed either unambiguous catch trial, all triad data from that participant would be excluded.
(6) Semantic intuition: Participants would be excluded for missing any of the 5 control questions.
(7) Ambiguous RMTS: none.
(8) Raven's SPM: none.

experiences, education, and undergraduate area of study (STEM or non-STEM). We also administered scales to collect explicit measures of participants' cultural identities and behaviors (Cleveland & Laroche, 2007; Cleveland, Laroche, & Takahashi, 2015; Strizhakova & Coulter, 2013).

The sample size for each task after exclusions and the descriptive statistics for each demographic question are reported in Table 1.

Procedure. Similar to Experiment 1, participants completed eight tasks and a brief demographics questionnaire online. The experiment was administered in English for the US sample and in Mandarin Chinese for the Chinese sample, with the exception of the Adult Causal Attribution task. As in previous work, this task was administered in English, and only Chinese participants who self-identified as being able to read English participated in it. To control for the impact of order-related inattention, task order was randomized across participants with two exceptions: (1) the Free Description task always occurred before (not necessarily immediately) Change Detection (because Change Detection included a manipulation check that explicitly asked about focal objects, which could bias responding in Free Description), and (2) the two story-based tasks (Semantic Intuition and Adult Causal Attribution) always occurred together in a fixed order at the end of the study, with Semantic Intuition first and Adult Causal Attribution last. Adult Causal Attribution was always the last task (if run) because it was administered in English and we did not wish to prime CN participants with English stimuli before any of the other tasks, all of which were run in Mandarin.

Measures.

Tasks from Experiment 1. We replicated three tasks from Experiment 1 using identical procedures: Ambiguous cRMTS, Picture Free Description, and Raven's Standard Progressive Matrices.

Symbolic Self-Inflation. Participants were asked to draw themselves and their friends as circles, as opposed to drawing themselves and their family members as circles in Experiment 1. They were also asked to draw lines between any two people who are friends, as in the original study by Kitayama et al. (2009). They then labeled each circle to indicate the person it represents. We calculated a percent inflation score for each participant by dividing the diameter of the self circle by the average diameter of circles for others.

Adult Causal Attribution. We speculated that the lack of cross-cultural differences in Causal Attribution in Experiment 1 might be due to the simplistic nature of our task, which was designed for use with young children. Therefore, in Experiment 2 we used a paradigm designed for adults, in which participants were asked to read a crime narrative from a news report that included substantial information on a criminal's background and the events leading up to their crime, and then rate the relevance of various situational and personal factors (Morris & Peng, 1994). In the original study, both Chinese participants and US participants read stories in English. We followed this procedure by selecting the subset of our Chinese participants who self-identified as comfortable reading short stories in English to participate. In the task, participants were told that they would read news stories and answer questions to help social scientists understand the factors that contribute to murders. Participants were randomly assigned to read one of two stories (Iowa shooting or Royal Oak shooting). After the stories, they were asked to write a short explanation for the murderer's behaviors. Then, they rated a list of statements about causes of the murder on a 7-point Likert scale. The statements included items that describe personal and situational factors, and we measured endorsement of these two factor types.

Change Detection. Masuda and Nisbett (2006) found differences in attention allocation between Japanese and US participants in a change detection paradigm. They found that Japanese participants were significantly faster than US participants in identifying changes in the background of images. We followed their original procedure and used the same stimuli. In this task, participants were presented with 30 pairs of images.

On each trial, two pictures would alternate on the screen, each presented for 560ms with a blank screen in between images for 80ms. The two pictures were almost identical with subtle differences, either in the focal object (e.g., a tractor in daylight with its lights on or off) or the background (e.g., a cloud with slightly different locations in the sky). Participants were instructed to press a key when they spotted the difference, and then describe the difference in a text box. If they did not detect a difference within 60 seconds, the trial timed out. Only trials in which participants correctly identified the changes were included in the analysis. After 30 trials, participants saw each pair of images again, this time side-by-side on the screen. They were asked to identify the focal object(s) in the pictures by typing into a text box. These responses were used as a manipulation check to ensure that participants in both cultures construed focal objects similarly.

We coded change descriptions to exclude trials in which participants did not identify the change, and checked agreement on focal objects across cultures. We measured how quickly participants identified the difference on trials in which they reported the difference correctly.

Taxonomic-Thematic Similarity. Ji et al. (2004) showed that Chinese participants are more likely to match items based on thematic similarity, whereas US participants are more likely to match items based on taxonomic similarity. In this task, participants were presented with triads containing a cue word and two match options. In each test set, one option was a taxonomic match (e.g., monkey - elephant) and the other a thematic match (e.g., monkey - banana). In each filler set, the cue item and the options were broadly similar, thematically and taxonomically, making for a more ambiguous decision (e.g., monkey: elephant, tiger). Participants completed a two-alternative forced choice task in which they chose one match for each cue item.

The findings of Ji et al. (2004) were replicated in more recent work (Le, Frank, & Carstensen, 2021); we used a subset of testing materials from Le et al. (2021), with 15 test triads, 15 filler triads, and 2 attention check questions. The order of the triads was

randomized between subjects. We measured taxonomic vs. thematic match selections on each of the test trials.

Semantic Intuition. Li, Liu, Chalmers, and Snedeker (2018) found cultural differences in semantic intuitions about ambiguous referents in Chinese and US participants. Specifically, Chinese participants were more likely to determine the referent of a name based on the description of the speaker (the descriptivist view) whereas US participants were more likely to determine the referent based on the original usage (the causal-historical view). In the study, participants read five separate stories and judged the correctness of statements referring to a character after each story. Two comprehension check questions were included for each story. We followed the original procedure closely and used the same materials. We measured participants' semantic intuition as their judgment on the correctness of statements referring to the critical characters.

Results

Ambiguous cRMTS. Our analysis was identical to that in Experiment 1. We did not observe a main effect of country on participants' preference for object vs relational matches (proportion relational match: US: $M = 0.41$, $SD = 0.44$; CN: $M = 0.41$, $SD = 0.42$; β = -0.01, $SE = 0.48$, z = -0.03, $p = 0.98$). As in Experiment 1, we did not find evidence that the differential preferences observed in preschoolers extend to adults. It seems likely that adults in both populations are aware of the mixed evidence for the relational and object solution and that their responses reflect sensitivity to this ambiguous design.[7]

Picture Free Description. US participants were more likely to initially mention the focal objects than the background objects (first mention: US: $M = 0.94$, $SD = 0.14$; CN: $M = 0.69$, $SD = 0.26$). We used the same regression analysis as in Experiment 1 and

[7] Our reliability analysis shows that adults expressed this uncertainty only at the population level: individuals tended to be consistent in choosing the same solution type across all four test trials, with ambiguity expressed as disagreement between participants.

found a main effect of country ($\beta = 3.09$, $SE = 0.32$, $z = 9.61$, $p < 0.01$). Our results replicate the first mention finding in Experiment 1 with a comparable effect size (standardized mean difference; Experiment 1: 1.50[1.26, 1.74]; Experiment 2: 1.57[1.34, 1.80]).

We also deviated from our pre-registered analysis plan and coded for the number of descriptive accounts directed at the focal objects and background objects using the same coding schemes as Experiment 1. We ran the same mixed-effect Poisson regression model predicting the number of descriptive accounts, with the interactions between description type (focal or background) and country (U.S. or China) as the predictor. Interestingly, we did not replicate the results in Experiment 1, but found patterns similar to the results in Imada et al. (2013). We found an interaction between country and type of description ($\beta = 0.16$, $SE = 0.07$, $z = 2.24$, $p < 0.05$). Chinese participants, in contrast to U.S. participants, provided more descriptive accounts of the background objects relative to the focal objects (for focal objects: US: $M = 0.66$, $SD = 0.82$; CN: $M = 0.50$, $SD = 0.64$; for background objects: US: $M = 0.69$, $SD = 1.07$; CN: $M = 0.61$, $SD = 0.80$).

In summary, these results extend Imada et al.'s (2013) findings to Chinese adults.

Change Detection. We ran a linear mixed-effects model predicting the reaction time to correctly identified changes in the pictures, with country (U.S. or China) and type of change detected (focal or background) as main effects, as well as their interaction. We did not find evidence for an interaction between culture and type of change detected ($\beta = 0.04$, $SE = 0.03$, $z = 1.40$, $p = 0.16$). Participants in both countries identified changes to the context faster than changes to focal objects (context changes: $M = 10{,}101.87$, $SD = 4{,}257.15$; focal object changes: $M = 10{,}646.54$, $SD = 4{,}816.10$; $\beta = 0.07$, $SE = 0.02$, $t = 3.45$, $p < 0.01$). Chinese participants identified both types of change more quickly than US participants (US: $M = 10{,}689.49$, $SD = 4{,}406.73$; CN: $M = 9{,}875.67$, $SD = 4{,}733.57$; $\beta = 0.12$, $SE = 0.05$, $t = 2.27$, $p < 0.05$).

As an exploratory analysis, we also retroactively analyzed the coded accuracy of the participants' responses. Interestingly, we found main effects of culture and the type of change, as well as an interaction between culture and the type of change. Participants across both countries were more accurate in identifying changes of the focal objects than the context objects, and Chinese participants were more accurate in identifying the changes than the U.S. participants on average (US focal: $M = 0.90$, $SD = 0.30$; US context: $M = 0.85$, $SD = 0.36$; CN focal: $M = 0.94$, $SD = 0.24$; CN context: $M = 0.89$, $SD = 0.32$). However, the difference between Chinese participants and U.S. participants was larger in focal changes than in background changes, with Chinese participants being more accurate than U.S. participants in the focal changes than in the context changes (β = -0.22, $SE = 0.09$, z = -2.53, $p < 0.05$). This interaction is different from our predictions: if we extrapolate from the original study, we should expect to see a difference in the background context, with participants performing similarly on background change trials but Chinese participants showing higher accuracy on background changes than U.S. participants.

In sum, we did not replicate the findings of Masuda and Nisbett (2006).

Symbolic Self-Inflation. In Experiment 1, we did not find a significant difference in the degree of symbolic self-inflation between adults in the US and China. Here, we observed a pattern contrary to the prediction: US adults showed less self-inflation than Chinese adults (US: $M = 1.30$, $SD = 0.51$; CN: $M = 1.45$, $SD = 0.65$; β = -0.15, $SE = 0.06$, t = -2.56, $p < 0.05$). We did not replicate the findings of Kitayama et al. (2009) with Japanese participants in either of our experiments.

Adult Causal Attribution. We ran a mixed-effects linear regression predicting endorsement of each potential cause with country (US or China) and attribution type (personal or situational) as fixed effects, as well as their interaction. We found an interaction in the predicted direction: Chinese participants endorsed situational attributions to a greater extent than their counterparts in the US (situational ratings: US: $M = 1.71$, $SD = 0.80$; CN: $M = 3.17$, $SD = 0.89$; personal ratings: US: $M = 3.12$, $SD =$

1.10; CN: M = 3.14, SD = 1.07; β = -1.39, SE = 0.14, t = -9.71, $p < 0.01$). This result replicates the original findings by Morris and Peng (1994), and suggests that the measure of causal attribution in Experiment 1 (which was designed for use with child participants) may not be appropriate for measuring cross-cultural differences in causal attribution among adults.

Taxonomic-Thematic Similarity. We used a mixed-effects logistic regression predicting response (taxonomic or thematic match) with country (US or China) as a fixed effect. There was a significant effect in the predicted direction: participants in the US were more likely to choose taxonomic matches than participants in China (proportion taxonomic matches: US: M = 0.69; SD = 0.46; CN: M = 0.44; SD = 0.50; β = 2.02, SE = 0.89, t = 2.27, $p < 0.05$). This finding replicates the findings of Ji et al. (2004) and Le et al. (2021).

Semantic Intuition. We ran a mixed-effects logistic regression predicting response (descriptive or causal-historical) with country (US or China) as a fixed effect, and found that US participants made significantly more causal-historical choices than Chinese participants (proportion causal historical choice: US: M = 0.71; SD = 0.46; CN: M = 0.53; SD = 0.50; β = 1.59, SE = 0.37, t = 4.37, $p < 0.01$). We also replicated the item effect identified by Li et al. (2018), though this was not among our preregistered analyses. In sum, We replicated Li et al. (2018) with new samples of adults in the US and China.

Raven's Standard Progressive Matrices. We replicated the findings from Experiment 1. Chinese participants scored higher on Raven's Standard Progressive Matrices than US participants (US: M = 0.49, SD = 0.27; CN: M = 0.73, SD = 0.23; β = -1.82, SE = 0.25, z = -7.39, $p < 0.01$).

Discussion

Overall, Experiment 2 was more successful than Experiment 1 in documenting cross-cultural differences between participants in the US and China. This success can be

attributed to the inclusion of the successful tasks from Experiment 1 (e.g., Free Description and Raven's Standard Progressive Matrices), and the exclusion of tasks designed for young children (e.g., Child Causal Attribution, Horizon Collage).

Exploratory analyses

We conducted a set of exploratory analyses to consolidate results from the two experiments. We first performed a miniature meta-analysis with the tasks from both experiments. Then, we assessed the reliability of the tasks that included multiple trials, the relationships between tasks, and finally, how explicit cultural identities and demographic factors relate to task performance.

Mini meta-analysis

As our first exploratory analysis, we identified the key effect of interest from our pre-registration (usually a main effect of culture or an interaction of culture, depending on task) and converted the coefficient into a standardized measure of effect size (standardized mean difference; SMD) via the method described by Westfall, Kenny, and Judd (2014). Because there is no "correct" direction for any task except Raven's SPM, we show the absolute value of the effect sizes (Figure 2).

Across our two experiments, we saw consistent and generally large differences (SMD > 0.6) in Free Description, Raven's SPM, Adult Causal Attribution, Semantic Intuition, and Taxonomic-Thematic Similarity tasks. Aside from Raven's SPM, all of these tasks have in common that they are deliberative linguistic tasks that tapped into relatively high-level cognitive constructs. In contrast, we observed effect sizes close to zero for our more aesthetic and perceptual tasks (Change Detection, Ebbinghaus Illusion, and Horizon Collage). We also observed little consistent difference in four other tasks (Ambiguous cRMTS, Symbolic Self-Inflation, Uniqueness Preference, and Child Causal Attribution),

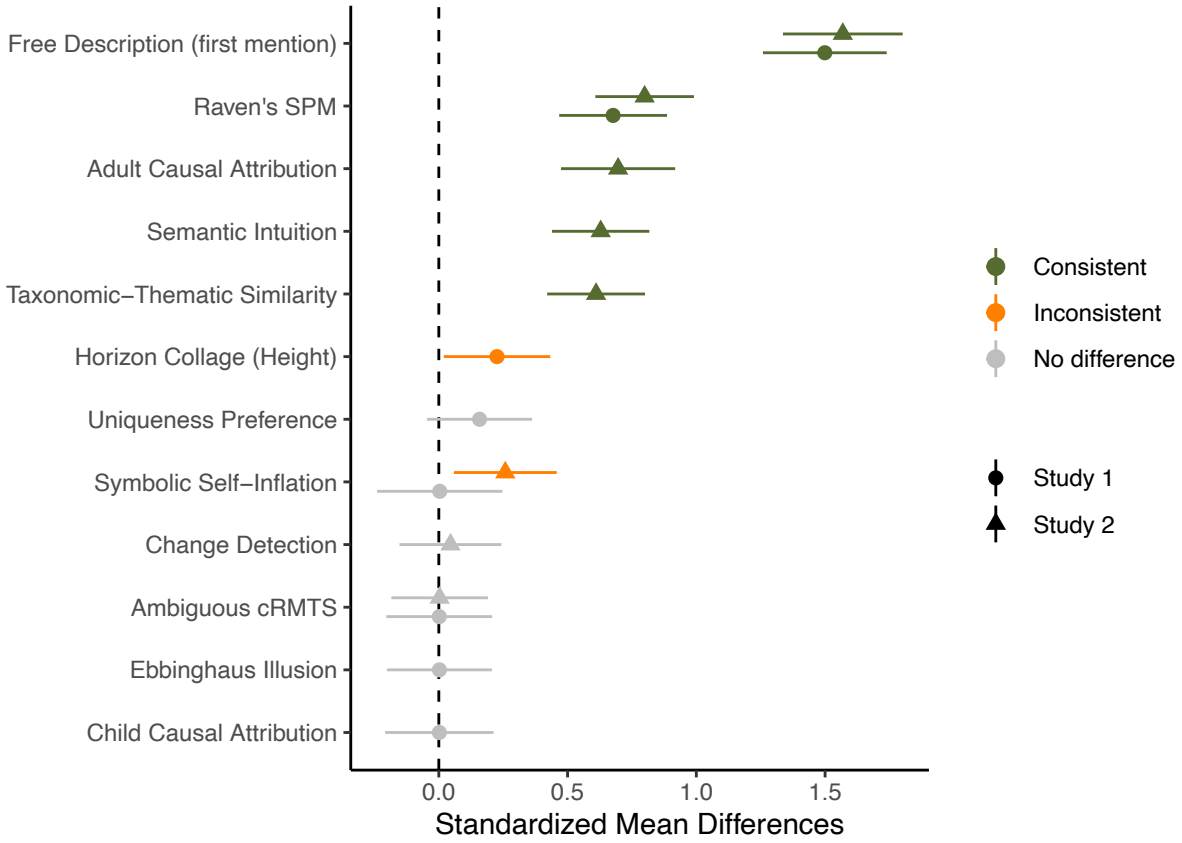

Figure 2. Forest plot of effect sizes (standardized mean difference) for each task across both experiments. Point shape shows experiment number and color indicates whether effects were consistent with prior literature.

perhaps for reasons idiosyncratic to each. We return to the broader question of generalization across task types in the General Discussion. But we note that the majority of tasks showing significant differences between cultures differ in the predicted direction, consistent with prior work. These five tasks are color-coded green in Figure 2, while the two that show differences opposite to predictions are indicated with orange.

We conducted three additional exploratory analyses to consolidate results from the two experiments. First, we assessed the reliability of the tasks that included multiple trials. Second, we examined whether there was shared variance between tasks. Finally, we examined how explicit cultural identities and demographic factors relate to task performance.

Reliability assessment

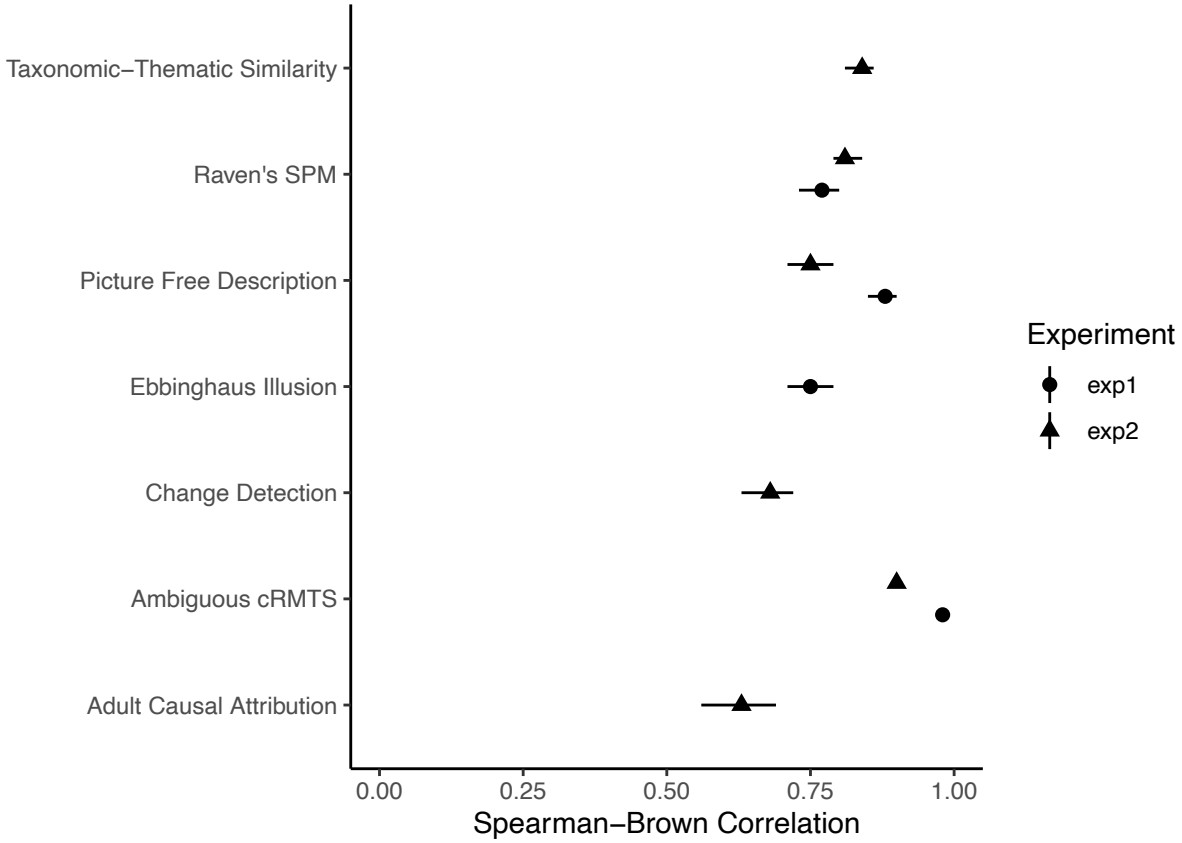

Figure 3. Spearman-Brown adjusted reliabilities for tasks with four or more trials. Point shape shows experiment number. Error bars show 95% confidence intervals.

One question motivating our work was whether the individual tasks we used were reliable enough – had low enough measurement error – to be used for further investigation of individual differences. The gold standard for evaluating whether a task yields stable within-person measurements is test-retest reliability (simply because test-retest gives a direct estimate of stability over time), but this method was outside the scope of our study. Instead, we used a split-half approach, asking whether participants' answers on individual questions relate to one another. Specifically, we used a permutation-based split half approach (Parsons, 2021) in which we made 5000 random splits of items into two simulated "halves" and then computed the within-person correlation between scores on these two halves, averaging across simulated runs. To estimate the reliability of the full-length

instrument, we used the Spearman-Brown "prophecy" formula.

Since the split-half approach is only suitable for tasks with multiple trials, we removed tasks with fewer than four trials from the analysis. For tasks with more than one condition, we focused on the condition that was predicted to show cultural differences (i.e., the Illusion context in the Ebbinghaus Illusion; situational factors in Adult Causal Attribution; background change scenes in Change Detection).

Figure 3 shows the corrected split-half reliabilities for all tasks in both of our experiments. Overall, the reliabilities were acceptable (all Spearman-Brown Correlations > 0.6). We further investigated whether there was cultural variation in the reliability of tasks. For most tasks, the reliabilities were relatively similar (within 0.1 of one another), but there were three tasks where reliability was lower for US participants than Chinese participants: Change Detection (US - CN = -0.19), Adult Causal Attribution (US - CN = -0.33), Free Description in Experiment 1 (US - CN = -0.23).

Relations between individual tasks

One (perhaps simplistic) interpretation of the prior literature on cultural variation is that there is a general tendency toward holistic or analytic reasoning that varies across cultures and explains variation in tasks. This single dimension might correspond to broad (or focused) attention and contextualized, relational reasoning (or an emphasis on focal people or objects). As a first step towards investigating this interpretation, we explored whether there was a single dimension of individual variation in our data that corresponded to this general axis of cross-cultural difference. Because some data was missing, largely due to task-related exclusions, we treated the missing data using two approaches: listwise deletion and imputation with means. These approaches yielded comparable results, so here we report correlations from listwise deletion.

Correlations between task scores were quite low on average, suggesting limited

support for a single factor explanation. Across both experiments, the largest absolute magnitude of correlations observed were -0.29 (Taxonomic-Thematic Similarity and Adult Causal Attribution in Experiment 2), -0.28 (Free Description and Raven's SPM in Experiment 2), and -0.24 (Adult Causal Attribution and Free Description in Experiment 2). All other correlations were between -0.23 and 0.23. Hence, the amount of shared variation between tasks was quite limited and our attempts at exploratory factor analysis discovered structures with many distinct factors and very low loading on the first factor.

We also applied the same correlation analysis within each culture. Again, we found limited correlations between tasks. This pattern replicated the findings from a previous work showing negligible relationships among a battery of tasks that revealed cultural differences (Na et al., 2010) (See Appendix C).

Demographic variation and explicit measures of cultural identity

As a final exploratory analysis, we asked whether demographic variation or variation in cultural identity predicted responding in our tasks. Our approach to these questions was to fit a set of exploratory regression models for each task, predicting task scores as a function of an individual scale and its interaction with culture. This approach allowed us to explore both within- and between-culture effects in a single model. Our predictors were 1) the summed score for our global/local cultural identity and consumption measures (with local items reverse-scored, such that higher scores represent more global identities and consumption patterns); 2) geographic information about where participants grew up Markus & Conner (2014); and 3) a range of demographic factors, including age, gender identity, residential mobility, number of international experiences, maternal education level, and subjective socioeconomic status as measured by the MacArthur Ladder (Adler et al., 2000).

Task performance and global identity. We fit models predicting task performance based on culture and its interaction with global (vs local) identity for tasks in

Experiment 2 (we did not collect these scales in Experiment 1). Two of these relationships were statistically significant at .01 < p < .05 (Adult Causal Attribution: $p = 0.05$; Taxonomic-Thematic Similarity: $p = 0.04$) but neither of these relationships survived Bonferroni correction for multiple comparisons.

Task performance and geographic origin. We next considered whether regions within each country were meaningful predictors of task performance. We fit models predicting task performance based on the regions that participants reported growing up in. For China, provinces were categorized as rice-cultivating regions or wheat-cultivating regions based on Talhelm et al. (2014). For the US, states were categorized based on either the coastal locations (West Coast, East Coast, and Inland) or broad geographic locations (West, South, Northeast, Midwest), following the categorization reported in Carstensen, Saponaro, Frank, and Walker (2022).

5 out of the 48 models we ran showed statistically significant relationships between regions and task performance. In Experiment 1, US coastal location was a significant predictor for the Free Description task. Participants who grew up in Inland regions ($N = 54$) or on the East Coast ($N = 27$) were more likely to mention the focal object first when describing the pictures than participants who grew up on the West Coast ($N = 84$; Inland: $p = 0.02$; East Coast: $p = 0.05$). In Experiment 2, both coastal location and broad geographic location were significant predictors for Raven's SPM, with participants from the East Coast ($N = 89$) and Inland regions ($N = 159$) scoring higher than participants from the West Coast ($N = 46$; Inland: $p < 0.01$; East Coast: $p = 0.05$), and participants from the Midwest ($N = 78$) and South ($N = 94$) scoring higher than participants from the West ($N = 59$; Midwest: $p < 0.01$; South: $p = 0.04$). In addition, both region categories predicted performance in Change Detection. East Coast participants ($N = 75$) took longer to respond than West Coast ($N = 42$) participants ($p = 0.02$), and Northeastern participants ($N = 52$) took longer to respond than participants who grew up in the West ($N = 52$; $p < 0.01$). However, none of these relationships survived Bonferroni correction.

Basic demographic effects. We fit 192 exploratory regression models to see if basic demographic factors could predict task performance. The demographic factors we explored were age, gender identity, residential mobility, number of international experiences, maternal education level, and subjective socioeconomic status as measured by the MacArthur Ladder (Adler et al., 2000). 26 were statistically significant, but only one model survived Bonferroni correction. Change detection was predicted by age in the US sample, with older participants taking longer to respond than younger participants (adjusted $p < 0.01$). Given some of the models could be considered as conceptual replications of previous work, we reported selected models with significant uncorrected results in Appendix D. We did not find any systematic patterns across the uncorrected significant results.

General Discussion

The world's cultures are strikingly different, and psychologists have long sought to measure and characterize this variation, with differences between Western and East Asian cultures as a case study of particular interest. These efforts have given rise to a rich literature documenting cultural differences in a wide range of psychological tasks. Across two experiments, we selected a collection of tasks that had previously been shown to yield differences between Western and East Asian samples and replicated them with two relatively large online samples of participants from the US and China. In this discussion, we first consider the limitations of our study since these contextualize the remainder of our conclusions. Next, we consider the interpretation of our results within individual tasks. We end the discussion with a summary of the key findings of this work.

General Limitations

As discussed above and in the introduction, we did not design our experiments to replicate prior work directly, and hence one important limitation of our work is simply that

it cannot be used as a test of the reliability of prior findings. Instead, our measures provide estimates of US-China differences on a range of constructs, specifically for online convenience samples. These estimates are likely biased downward – towards the null hypothesis of no difference between cultures – by several features of our experimental design.

Online experiments (especially grouped into a long battery as ours were) likely receive slightly less attention than in-person studies, though differences between in-person and online studies have tended to be small in US samples (Buhrmester, Kwang, & Gosling, 2016). Further, participants did perform relatively accurately on those tasks that had correct answers (e.g., Raven's SPM, the Ebbinghaus Illusion), and in our exploratory analysis, we found relatively high reliabilities on all tasks. Finally, our pre-registered exclusion criteria removed participants who performed poorly. Thus, we do not believe that participants were inattentive overall.

A second potential issue is the collection of reaction time measures through online experiments (e.g. Ebbinghaus Illusion and Change Detection). Due to variations in individual participants' hardware and internet conditions, reaction times could potentially be more variable online than in in-person experiments. However, the validity of online reaction time measures has been shown for many different classic cognitive psychology tasks, suggesting browser-based experiments typically yield useful reaction time data (Crump, McDonnell, & Gureckis, 2013). It is worth noting also that the only task for which reaction times were the key dependent measure in our studies was Change Detection, in which average reaction times were upwards of 10 seconds. The accuracy and reliability of reaction times gathered through jsPsych, the specific JavaScript library that we used here, has been found to be high for even 10–100 millisecond level contrasts (Anwyl-Irvine, Dalmaijer, Hodges, & Evershed, 2021; Pinet et al., 2017). Therefore, there is no reason to believe that the variability in online reaction time measures was a key cause of non-replication.

A more substantial limitation of our estimates of US-China differences comes from variation in our sampling strategy between cultures. In Experiment 1, we used the same snowball sampling procedure, but this procedure may have yielded different samples due to differences in social networks or norms about sharing study information across cultures. In Experiment 2, because the platform we used to recruit US participants (Prolific) was not accessible in China, we used a different platform to recruit Chinese participants (Naodao). Prolific and Naodao have different levels of popularity and different participant pools, resulting in some asymmetry between the US and Chinese samples. Despite these differences between samples both across and within experiments, we do not see indications that our estimates were dramatically biased by our sampling decisions. First, our results were largely comparable in the tasks that were included in both experiments (e.g. Picture Free Description; Raven's SPM; and Ambiguous cRMTS). Second, in our exploratory analyses we did not find strong associations between participant demographics and cross-cultural effects (with some small exceptions discussed in that section). Finally, we reran all of our preregistered analyses with an age-matched subset of U.S. participants in Experiment 2 and found our results were qualitatively identical.

, with U.S. samples being older on average and exhibiting a wider range in age than Chinese samples

Thus, while our samples are certainly not representative samples of US or Chinese national populations – indeed to our knowledge, nearly all work to date has used convenience samples of one type or another – they appear to yield stable cross-sample estimates that do not reflect large biases due to sampling strategy or demographics.

One of the main ways in which our samples may not have been representative is that they are likely to be more globalized than the population on average simply by being young (and thus less acculturated) and having access to a computer. Contra this concern, variation in local cultural identity did not strongly relate to variation in any of our tasks, but interestingly, we observed the strongest local identities (within our Chinese sample) among the youngest participants.

Last but not least, another difference between our experiments and previous work was the lack of an experimenter, and some of our tasks may have been particularly sensitive to the presence of an experimenter. In a web experiment, participants are often isolated in front of their own computer. In contrast, in an in-person experiment, participants must interact with and perform the task in front of experimenters who are often from the same social group. Indeed, as we discuss below, in the Uniqueness Preference pen choice task, cross-cultural differences are dependent on the presence of an experimenter (Yamagishi et al., 2008).

Task-specific Limitations

In addition to the general limitations discussed above, there are features of our experimental adaptations that may have affected performance in specific tasks. In this section, we highlight concerns about these issues and discuss their implications for interpreting the results of these tasks. See Table 2 for a summary of this task-specific discussion.

In the case of the Uniqueness Preference task, it is possible that adapting the task to an online format in which resource scarcity was not strictly real and task choices had no lasting effect (in the form of a new pen), may have trivialized the choice and undermined the incentive for prosocial, harmonious behavior or expression. This possibility is consistent with the chance responding we observed in both groups. Alternatively, our results could be seen as a conceptual replication of Yamagishi et al. (2008), who argue that differences in this task are moderated by the likelihood of evaluation, with no differences in pen choice observed in the absence of an experimenter.

The ambiguous developmental tasks, Ambiguous cRMTS and Child Causal Attribution, may have been too heavy-handed in their key manipulations; both were designed to highlight ambiguity for young children, but it may be that their explicit cues

and repetitive instructions impressed this ambiguity too strongly for adult audiences, resulting in the adults' near-chance responding – a reasonable response to such marked ambiguity. Cultural differences in causal reasoning and attribution may only manifest when the task design is age-appropriate. Consistent with this view, we did replicate previously attested differences in the Adult Causal Attribution task in Experiment 2, and other recent work has shown cross-cultural differences in causal attribution among 4- to 9-year-olds in Germany, Japan, and Ecuador using a design similar to the Child Causal Attribution task (Jurkat, Iza Simba, Hernández Chacón, Itakura, & Kärtner, 2022).

Last but not least, variation within the broad cultural constructs of East Asia and the West could explain some of our findings, as a failure to extend previous work. Some of the tasks we included originally compared participants from other parts of East Asia and the West (e.g., Horizon Collage, Symbolic Self-Inflation, Change Detection; but c.f. Masuda, Ishii, and Kimura (2016) for an alternative account of mixed findings in change detection paradigms). For example, the Taxonomic-Thematic Similarity task replicated previously attested cross-cultural differences between the US and China both here and in other work (Le et al., 2021) but these differences failed to generalize to a US-Vietnam comparison, despite the cultural, historical, and geographic similarities between China and Vietnam. This variation suggests that similar psychological tendencies could be expressed differently under distinct sociocultural contexts and traditions, even across regions and countries that share many similarities. As another example, responding in the Horizon Collage could be modulated by variation between countries: Chinese and Japanese aesthetic traditions differ, so while Chinese and Japanese people may share a preference for highly contextualized information, this preference may be typically expressed through distinct visual techniques.

Conclusion

We conducted two experiments to examine the robustness of several classic experimental paradigms in cross-cultural psychology. Our results showed a heterogeneous

pattern of successes and failures: some tasks yielded robust cultural differences across both experiments, while others showed no difference between cultures. We estimated the reliability of the tasks to be moderate, with only minor variation in reliability across cultures. We also explored the effects of a range of demographic variables, including explicit identification with global identity, regional differences within cultures, and several demographic characteristics. All of these had minimal relation to task performance.

Our goal here was not to perform direct replications that would shed light on the replicability of specific findings. Instead, since our methods, administration medium, sample, and analytic approach differed from the prior literature, our hope was to examine the robustness of these paradigms as a method for measuring US-China differences in an online context. Our work has several strengths relative to the prior literature, including larger samples of participants from the US and China, two broad groups of tasks implemented openly online (and reusable by future researchers), and a preregistered analysis plan that allows for the unbiased estimation of cross-cultural effects. In sum, we hope that our work here provides a foundation for future studies that seek to establish a robust and replicable science of cross-cultural difference.