Replicability of US-China differences in cognition and perception across 12 tasks

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] 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. 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 a selection of 12 tasks previously reported to show cross-cultural differences. Five of these tasks showed robust cross-cultural differences, while six showed no difference and one showed a small difference in the opposite direction. Tasks showing cross-cultural differences tended to have multiple trials measuring high-level reasoning and language; those that did not show cross-cultural difference included measures of attention/perception or implicit social processes or were initially designed to show differences in children. 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.

Keywords: replication; cross-cultural differences

Word count: FIXME

Replicability of US-China differences in cognition and perception across 12 tasks

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 the United States and China – often as exemplars of 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; 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, the US and China 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 two 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 claims about specific cross-cultural differences for constructing theories of culture more broadly (e.g., Markus & Kitayama, 1992, 2010), replication 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 for exploration of the reliability of individual tasks as well as the intercorrelations between them.

Our goal in the current study was to replicate a set of cross-cultural measures that had previously been used in comparisons of East Asian and Western cultures (most often comparisons between US and either Chinese or Japanese participants). We made the decision to pursue the strategy of gathering relatively large and heterogeneous convenience

[1] These authors interpreted their findings as imply 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.

samples using online recruitment, rather than recruiting 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 either reject or accept the null hypothesis of no cultural difference between measures. Further, larger samples would afford the analysis of individual and demographic differences within culture, a topic of considerable interest in this literature (e.g., Na et al., 2010). 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.

The interpretation of any replication result is complex, given that disparate outcomes between an initial study and a replication can occur for many reasons – including but not limited to differences in experimental methods, sample or population differences, and simple sampling variation in the outcomes Machery (2020). Our strategy of pursuing online convenience samples limits the interpretation of our replication results: nearly all of the tasks we selected were previously administered in person, and the populations sampled in previous reports varied but were largely convenience samples of either college students or community members. More generally, our strategy of constructing a battery of replication studies and administering them uniformly means that specific decisions about sampling and administration are not matched with the original studies. Thus, our replication studies should be taken as an assessment of whether a set of previously-reported cross-cultural differences can be recovered in convenience populations recruited online, rather than as assessments of the veracity of the original findings. Nevertheless, we believe that the field of 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; e.g., 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 four 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 (e.g., East Asian vs. Western cultures), (3) were relatively short, accessible tasks appropriate for web administration, and (4) were vision or social cognition accounts for relational reasoning. We further 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. 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 relatively large dataset that we collected could be used to explore the structure of within- and across-cultural 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 assess the reliability of individual tasks, aiming to gauge whether individual tasks are reliable enough from a psychometric point of view to support further individual differences analyses. We then report across-task correlations, 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 few demographic predictors of within-culture variation.

Table 1

Tasks included in each experiment and the final sample size after exclusion.

Experiment	Task	Relevant Citation	Task Description	CN	US
1	Ambiguous Relational Match-To-Sample (RMTS)	Carstensen et al. (2019)	Infer whether an object or relation is causally relevant	N = 167	N = 169
	Picture Free Description	Imada, Carlson, & Ktakura (2013)	Describe pictures from memory after a brief study period	N = 167	N = 169
	Ebbinghaus Illusion	Imada, Carlson, & Itakura (2013)	Judge the size of circles in a context designed to bias size judgments	N = 167	N = 169
	Horizon Collage	Senzaki, Masuda, & Nand (2014)	Make an image by dragging and dropping stickers onto a display	N = 167	N = 169
	Symbolic Self-Inflation (Family)	Kitayama et al. (2009)	Draw self and family members as circles	N = 141	N = 110
	Uniqueness Preference	Kim & Markus (1999)	Choose a sticker from five stickers, four of which are the same color	N = 167	N = 169
	Child Causal Attribution	Seiver, Gopnik, & Goodman (2013)	Watch short vignettes and explain the decisions of the characters	N = 167	N = 169
	Raven's Progressive Matrices	Su (2020)	Use analogic reasoning to complete visually-presented patterns	N = 167	N = 169
2	Ambiguous Relational Match-To-Sample (RMTS)	Carstensen et al. (2019)	Infer whether an object or relation is causally relevant	N = 174	N = 293
	Picture Free Description	Imada, Carlson, & Itakura (2013)	Describe pictures from memory after a brief study period	N = 132	N = 284
	Change Detection	Mausda & Nisbett (2007)	Find differences in the foreground or background of two images	N = 160	N = 253
	Symbolic Self-Inflation (Friends)	Kitayama et al. (2009)	Draw a sociogram with self and friends as nodes, relationships as edges	N = 158	N = 252
	Adult Causal Attribution	Morris & Peng (1994)	Read a crime story and explain the criminal's motivations	N = 114	N = 293
	Taxonomic-Thematic Similariy	Ji, Zhang, & Nisbett (2004)	Match items based on taxonomic or thematic similarity (e.g., cow: chicken / grass)	N =178	N = 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)	N = 181	N = 298
	Raven's Progressive Matrices	Su (2020)	Use analogical reasoning to complete visually-presented patterns	N = 181	N = 298

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.

Experiment 1

Methods

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.

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 original preregistered exclusion plan was to exclude people from the full dataset if they failed quality checks on any one task. However, due to a task demand associated with the Symbolic Self-Inflation task, this criterion would have led to the exclusion of 85 people (US: 59, CN: 26) due to this task alone. As a result, we deviate from our preregistration and include participants in the broader dataset even if they failed the quality check for the Symbolic Self-Inflation task.

After exclusions, the US sample included 169 participants (44 Male, 114 Female, 9 Non-binary, 2 Declined to answer), with a mean age of 21.79 years old, all of whom were native English speakers. The China sample included 167 participants (51 Male, 112 Female, 1 Non-binary, 3 Declined to answer), with a mean age of 22.49 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 110 US participants and 141 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 number of overseas experiences.

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 the methods for each task; further details are available in Supplemental Materials and code for tasks is available at FIXME.

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 an 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 descriptions. 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 tended 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). In this task, 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 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 would draw the horizon higher and put more collage items in the picture, 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 (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. 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 – four were the same and one was unique. They found that European Americans were more likely to choose the unique colored one 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: four blue dinosaurs and one yellow. Choice of the unique vs. repeated color was recorded.

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 media is more likely than U.S. media 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) person and (b) situation 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 was the most difficult subset and also the one most dependent on true analogical reasoning (without alternative heuristic approaches like visual pattern completion).

Analyitic approach. Our sample size, methods, and main analyses were pre-registered and are available at https:// aspredicted.org/37y6a.pdf. Data and analysis scripts are available at FIXME

The specific 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 standard operating procedure of pruning random slopes first and then random intercepts, always maintaining random intercepts by participant. We report p-values derived from approximating t-scores 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 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: $M = 0.37$, $SD = 0.47$; $\beta = 0.14$, $SE = 0.89$, $z = 0.16$, $p =$

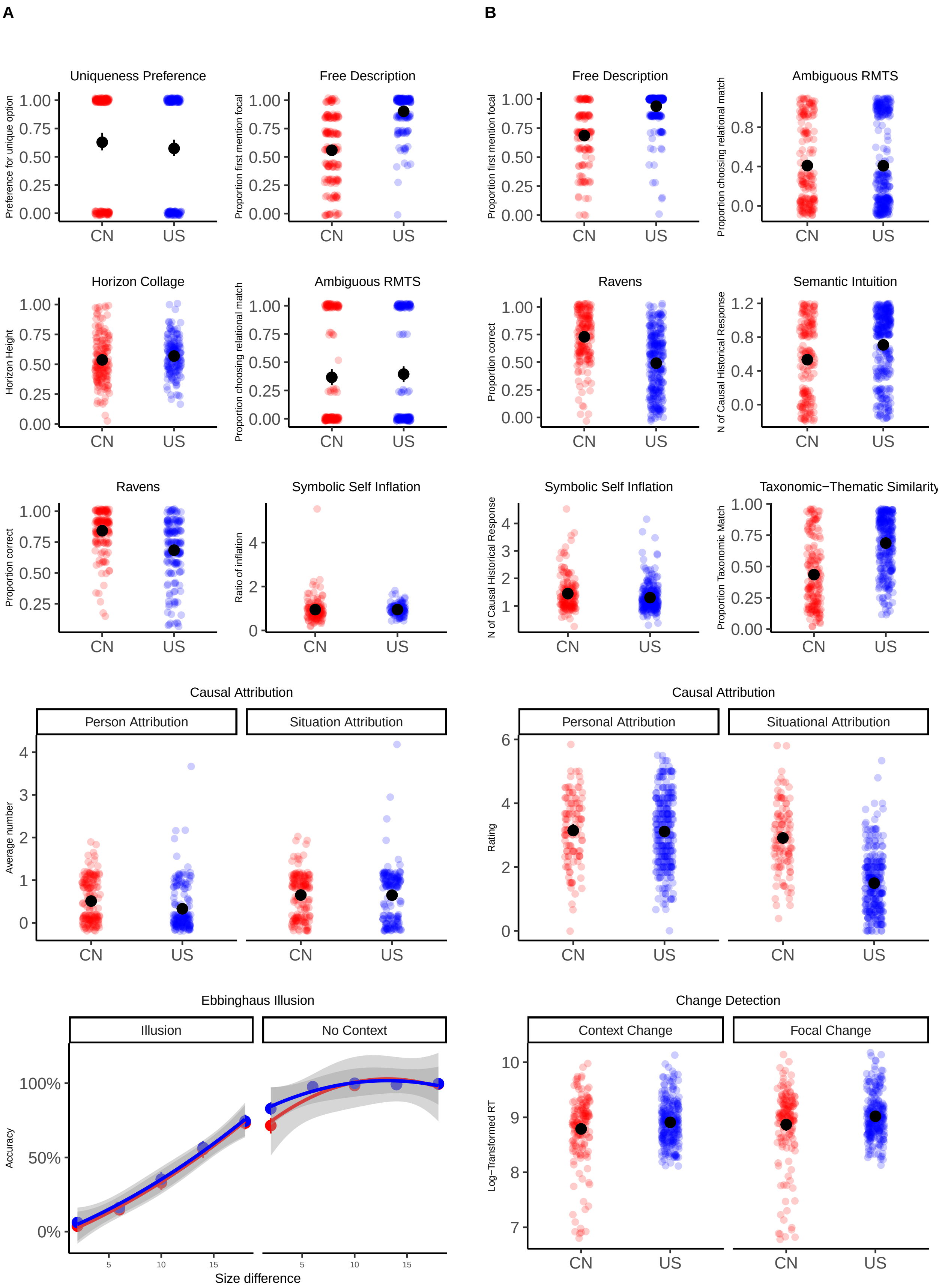

Figure 1. Results from each task. Results from the CN sample are plotted in red, and the US in blue.

0.87). The preference for object-based solutions seen in US preschoolers and the corresponding preference for relational solutions observed in China in an ambiguous context did not extend to adults in our samples.

Our US results replicate findings by 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; Proportion of Relational Match choice: 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: $M = 0.88$, $SD = 0.44$; For background objects: US: $M = 1.31$, $SD = 0.94$; 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 ($\beta = 3.36$, $SE = 0.34$, $z = 9.94$, $p < 0.01$). 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: $\beta = 0.36$, $SE = 0.13$, $t = 2.68$, $p < 0.01$). The culture effect interacted with the description types, but the effect was in the opposite direction, with U.S participants provided more background descriptions than focal descriptions, relative to Chinese participants ($\beta = -0.16$, $SE = 0.07$, $t = -2.16$, $p < 0.05$).

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 of difference in diameters) as fixed effects, along with their interactions. We found main effects of context (with worse performance in the Illusion Context; $\beta = 4.95$, $SE = 0.29$, $z = 17.03$, $p < 0.01$) and circle size difference (worse performance for smaller differences; $\beta = 0.34$, $SE = 0.01$, $z = 27.33$, $p < 0.01$). There was a marginally significant main effect of country at the opposite of the predicted direction (US participants performed worse: $\beta = 0.52$, $SE = 0.26$, $z = 1.95$, $p = 0.05$) but no interactions with country (All $\beta < 0.01$; All $p > 0.05$).

In sum, we failed to replicate cultural differences found between Western (US/UK) and Japanese participants in susceptibility to the Ebbinghaus illusion.

Horizon Collage. In the Horizon Collage task, three key measurements are calculated from the "collage" participants created: the height of the horizon (height in proportion to the height of the frame), the number of stickers, and the total area of the stickers covered (following the original analysis, we added up the area occupied by each individual sticker); Japanese children tend to put the horizon higher and include more stickers that 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 did not significantly predict any of the three measurements (Sticker height: US: $M = 0.57$, $SD = 0.15$; CN: $M = 0.54$, $SD = 0.20$; Sticker number: US: $M = 11.51$, $SD = 5.81$; CN: $M = 11.77$, $SD = 5.80$; Sticker area: US: $M = 16.98$, $SD = 8.36$; CN: $M = 17.43$, $SD = 8.60$; All $\beta < 0.03$; All $p > 0.1$).

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 have a larger symbolic self than Chinese adults, we ran a linear regression predicting percent inflation score (calculated by dividing the diameter of the self circle 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 China adults based on percent inflation scores (US: $M =$ 0.95, $SD = 0.26$; CN: $M = 0.95$, $SD = 0.55$; $\beta = 0.36$, $SE = 0.13$, $t = 2.68$, $p < 0.01$).

One possible explanation for our null results is that we adopted a different task design from 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, into larger circles, which might have resulted in overall larger circles for other people than the self-circles in our task for both cultures, masking any US-China difference in the degree of self-inflation. It is possible that there are also 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).

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 not a large cross-cultural difference in the probability of choosing the uniquely colored sticker (US: M = 0.57, SD = 0.50; CN: M = 0.63, SD = 0.48; β = -0.23, SE = 0.22, z = -1.02, p = 0.31).

The difference between our result and that of the original study by Kim and Markus (1999) might be related to the use of 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 the next person to also have room for decision making in the face of resource scarcity, or 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 participant concern about reputation, and accordingly failed to motivate reputation management in the form of a conformity preference.

Causal Attribution. To test whether Chinese participants tended to make more situational attributions, and US adults more personal attributions, we ran a mixed-effects Poisson regression predicting the number of attributions included in each explanation, with attribution type (situation or person), country (US or CN), and their interaction as fixed effects. We found a main effect of attribution type (Situation attribution: US: M = 0.65, SD = 0.61; CN: M = 0.65, SD = 0.52; Person attribution: US: M = 0.33, SD = 0.55; CN: M = 0.51, SD = 0.52; β = 0.24, SE = 0.10, z = 2.37, p < 0.05). Neither the interaction nor the main effect of culture was significant (Both β < 0.3; p > 0.05).

The failure to find cross-cultural differences in attributions could be related to the style of the tasks, which was 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: $M = 0.68$, $SD = 0.24$; CN: $M = 0.84$, $SD = 0.17$; $\beta = -1.31$, $SE = 0.23$, $z = -5.64$, $p < 0.01$).

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.

Experiment 2

Discussion

We did not observe cross-cultural differences in the majority of the tasks in Experiment 1. The only exceptions were in picture description and our exploratory measure of reasoning performance (Raven's Matrices). Many of our tasks did not have a manipulation check and could yield null results simply by virtue of inattention. However, the results of the Raven's task (and the Ebbinghaus Illusion) suggest that participants were engaged in our tasks and performed at a high objective level. Further, in addition to minor methodological changes that we made, interpretation of our failures 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 recruitment), (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 Western vs. East Asian cultural differences 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.

Methods

Experiment 2 was designed to follow up on Experiment 1 and further evaluate cross-cultural differences across a battery of tasks. Since 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 had been argued to index context sensitivity (Masuda & Nisbett, 2006). We replaced the child-appropriate causal attribution task with a task 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 RMTS, 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. We had two rationales. First, in Experiment 1 our samples were quite young (due to the use of email and social media to populations of university students for recruitment). A younger sample might be more exposed to international media and influences and be 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 (Markus & Kitayama, 1992).

Participants. We recruited participants through online crowdsourcing websites. For the US, we used Prolific and applied the following screening criteria: a) U.S. nationality; b) born in the U.S. and c) currently reside in the U.S. For China, we used

Naodao (www.naodao.com), a platform designed for conducting online experiments in mainland China. Participants in U.S. were compensated at the rate of $12.25 per submission and in China ¥35 per submission. 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. 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 in the tasks, b) had missing data on more than 25% of trials or c) failed to meet the inclusion criteria for any specific task as specified in the preregistration.

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 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).

The sample size for each task after exclusion 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 online in English for the US sample and in Mandarin Chinese for the Chinese sample, with the exception of the Adult Causal Attribution task. The Adult Causal Attribution task was administered in English, and only Chinese participants who self-identified as being able to read English participated in this task. 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 includes a manipulation check that explicitly asks 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 repeated from Experiment 1. We replicated three tasks from Experiment 1 using identical procedures: Ambiguous RMTS, Picture Free Description, and Raven's 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 cultural difference 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 be reading news stories and answering 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 on a 7-point likert scale about the extent to which each was a likely cause of the murder. The statements included items that describe personal and situational factors. 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 or the background (e.g., a tractor in daylight with its lights on or off). 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 difference 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 task. Ji et al. (2004) showed that Chinese participants are more likely to categorize items based on thematic similarity, whereas US participants are more likely to categorize items based on taxonomic similarity. In this task, participants were presented with a list of word sets. Each set contained three words: a target word as prompt and two other words as options. The list included test sets and filler sets. 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 2AFC in which they chose one match for each cue item.

We used a subset of testing materials from Le (2021), including 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. Chinese participants are more likely to determine the referent of a name based on the description of the speaker (the descriptivist view) whereas U.S. participants are 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 testing 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 RMTS. 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 of 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.[2]

Picture Free Description. US participants were more likely to 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 found a main effect of country (β = 3.09, SE = 0.32, z = 9.61, $p < 0.01$). Our results replicate Experiment 1's finding for the first-mention measure with comparable effect size (standardized mean difference; Experiment 1: 1.48[1.24, 1.72]; Experiment 2: 1.57[1.34, 1.80]). [3]. These findings 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 identify 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 (β = 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 =

[2] 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.

[3] The comparable SMD suggests that the finding was not caused by the idiosyncrasy of our samples. As a result, we decided not to code the descriptive accounts for Experiment 2 that did not show differences in Experiment 1

4,257.15; Focal Object Changes: M = 10,646.54, SD = 4,816.10; β = 0.07, SE = 0.02, t = 3.45, $p < 0.01$). Chinese participants identified both types of change more quickly than U.S. participants (US: M = 10,689.49, SD = 4,406.73; CN: M = 9,875.67, SD = 4,733.57; β = 0.12, SE = 0.05, t = 2.27, $p < 0.05$). 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: U.S. 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, $p < 0.05$). In sum, we did not replicate the findings of Kitayama et al. (2009) (with Japanese participants) in either of our studies.

Adult Causal Attribution. We ran a mixed-effects linear regression predicting endorsement of each potential cause with country (U.S. 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 U.S. (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 extends the original findings by Morris and Peng (1994), and suggests that the measure of causal attribution in Experiment 1 (whcih 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 model 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 U.S. were more likely to choose taxonomic matches than participants in China (Proportion of taxonomic matches: US: M = 0.69; SD = 0.46; CN: M = 0.44; SD = 0.50), on both the main model (β = 2.02, SE = 0.89, t = 2.27, p < 0.05) . This finding replicates the findings of Ji et al. (2004) and Le, Frank, and Carstensen (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 U.S. participants made significantly more causal-historical choices than Chinese participants (Proportion of 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 a new sample of US and China adults.

Raven's Standard Progressive Matrices. We replicated the findings from Experiment 1. Chinese participants scored higher on Raven's Standard Progressive Matrices than U.S. 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).

Exploratory 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 (2014). Because there is no "correct" direction for all of the tasks except Raven's Matrices, we show the absolute value of effect size (Figure 2).

Across our two experiments, we saw consistent and generally large differences (SMD > 0.6) in Free Description, Raven's Matrices, Adult Causal Attribution, Semantic Intuition and Triads tasks. Aside from Raven's Matrices, all of these tasks had in common that they were 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

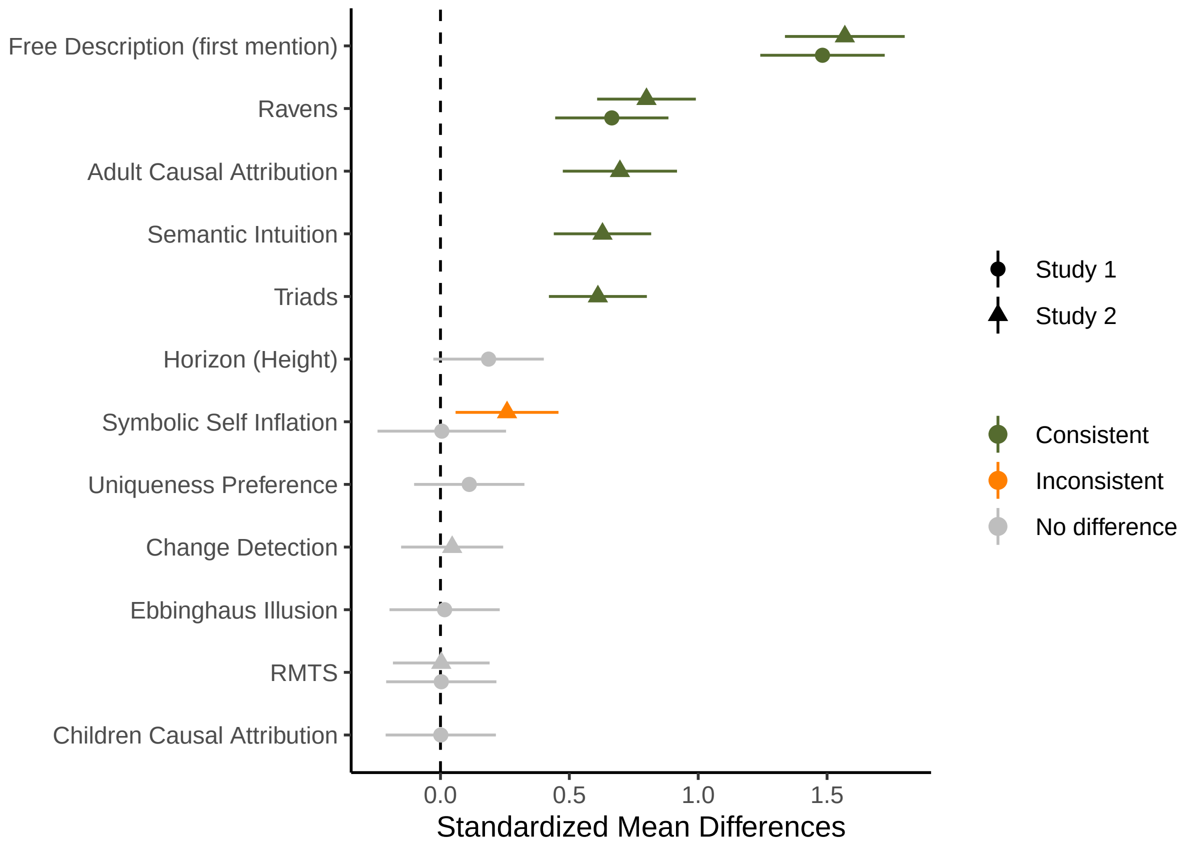

Figure 2. Forest plot of effect sizes (standardized mean difference) for each task across both experiments. Point shape shows experiment number and color provides a guide to whether effects were consistent with prior literature.

tasks (Change Detection, Ebbinghaus Illusion, and Horizon). We also observed little consistent difference in four other tasks (RMTS, Symbolic Self-Inflation, Uniqueness Preference, and Children's Causal Attribution), perhaps for reasons idiosyncratic to each. We return to the broader question of generalization across task types in the General Discussion.

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

Reliability assessment

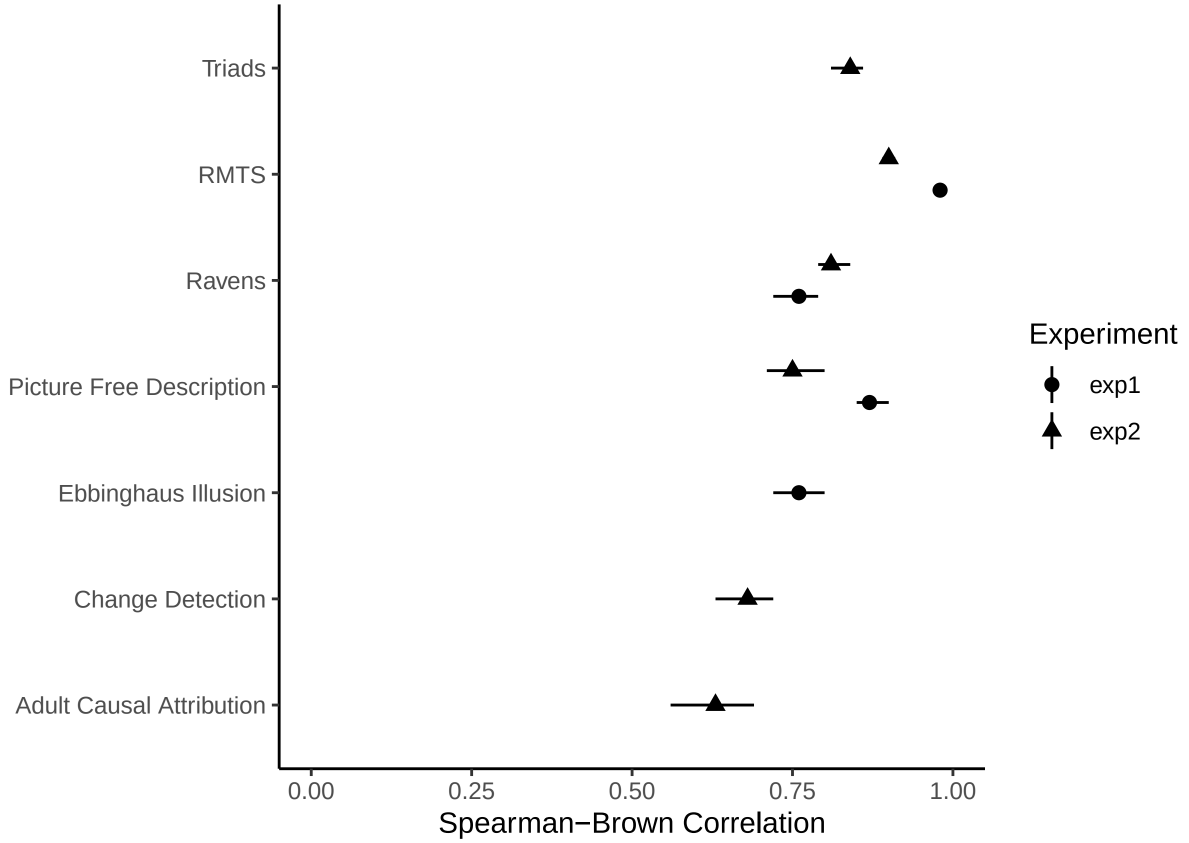

Figure 3. Spearman-Brown adjusted reliabilities for tasks with more than four trials. Point shape shows experiment number. Error bars show 95% confidence intervals.

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 the measurement of 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. Thus we used a split-half approach, asking whether participants' answers on individual questions related to one another. 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 split-half approach is only suitable for tasks with multiple trials, we removed tasks with less than four trials from the analysis. For tasks with more than one condition, we focused on the conditions that were predicted to show cultural differences (i.e. Illusion context condition for Ebbinghaus task; Situational judgements for Adult Causal Attribution task; Context condition for Change Detection task).

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 = -0.31), Free Description in Study 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 the 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 the hypothesis of a single factor explanation. Across both Experiments 1 and 2, the largest absolute magnitude of correlations observed were -0.29 (Triads and Adult Causal Attribution in study 2) and -0.28 (Free Description and Ravens in Study 2), and -0.24 (Adult Causal Attribution and Free Description in Study 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.

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 across-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 identity and consumption patterns), 2) geographic information about where participants grew up Markus & Conner (2014), and 3) a range of demographic factors, including age, gender identities, residential mobility, number of international experiences, maternal education level, and subjective socioeconomic status as measured by the MacArthur Ladder (Adler et al., 2000).

Task-Global identity relationships. We fit models predicting task scores based on culture and its interaction with global-local identity for tasks in Experiment 2 (we did not collect these scales in Experiment 1). We include the coefficients for all models in Supplementary Table FIXME Two of these relationships were statistically significant at .01 < p < .05 (Adult Causal Attribution: $p = 0.05$; Triads task: $p = 0.04$) but neither of these relationships survived Bonferroni correction for multiple comparisons across the family of coefficients for the models across all tasks.

Task-Geographic origin relationships. We next considered whether regions within each country were meaningful predictors of task performance. We fit models predicting task scores based on the categories of regions the participants reported grew up in. For China, provinces were categorized as rice-cultivating regions and wheat-cultivating regions based on Talhelm et al. (2014). For U.S., 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). We fit the region models for each task in each study separately, and coefficients for all models were included in the Supplementary Table FIXME.

5 out of 48 models we ran showed a statistically significant relationships between regions and task performance. In Study 1, coastal location was a significant predictor for Free Description task. Participants who grew up in Inland regions or East Coast were more likely to mention the focal object first when describing the pictures (Inland: $p = 0.02$; East Coast: $p = 0.05$). In Study 2, both coastal location and broad geographic location were significant predictors for Ravens, with participants from Inland and East Coast scoring higher than participants from West Coast (Inland: $p = 0.00$; East Coast: $p = 0.05$), and participants from Midwest and South scoring higher than participants from the West (Midwest: 0.00; South: 0.04). In addition, the two categories also predicted performance in Change Detection. East Coast participants took longer to respond than West Coast participants ($p = 0.02$), and Northeastern participants took longer to respond than participants grew up in the West ($p = 0.01$). However, none of these relationship 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 identities, residential mobility, number of international experiences, maternal education level, subjective socioeconomic status as measured by MacArthur Ladder (Adler et al., 2000). 24 were statistically significant, but only one model survived Bonferroni correction. Change detection was predicted by age in the U.S. sample, with older participants taking longer to respond than younger participants (Adjusted $p < 0.0001$).

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 particular case study of 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 range 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 US and Chinese participants. 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. Finally, we turn to broader interpretation of our results including our exploratory analyses.

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 overall these effects have tended to be small in US samples (Buhrmester, Kwang, & Gosling, 2016). Contra this concern, however, participants did perform relatively accurately on those tasks that had correct answers (e.g., Raven's Matrices, Ebbinghaus Illusion), and in our exploratory analysis, we found relatively high reliabilities on all tasks. Further, our pre-registered exclusion criteria removed participants who performed poorly. Thus, we do not believe that participants were inattentive overall.

Another limitation of our estimates of US-China differences comes from differences in 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 U.S. 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; Ravens; and RMTS). 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 and found our results were qualitatively identical. 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.

Another difference between our experiments and previous work was the lack of an experimenter, and some of our tasks may be particularly sensitive to the presence of an experimenter. In a web experiment, participants are often isolated in front of their own computer. In contrast, when participating in an in-person experiment, participants need to interact with and perform the task in front of the experimenters who are often from the same social group. Indeed, in the uniqueness preference pen choice task, cross-cultural differences are dependent on the presence of an experimenter (Yamagishi et al., 2008). Our null results, obtained in the absence of an experimenter, can be seen as a conceptual replication of this work.

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.

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 choices in this task 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 RMTS 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 and 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, cultural variation within the broader 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 children and adults 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, and broad construals of the relevant cultural factors in previous work (e.g., Ji et al., 2004). Nonetheless, this variation could reflect similar psychological tendencies that are expressed differently as a result of distinct sociocultural contexts and traditions across differing regions and countries. As another example, responding in the Horizon Collage task 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 typically be expressed through distinct visual techniques.

Conclusion

We conducted two sets of 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 cultural variations. In addition, we also explored the effect 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.

References

Adler, N. E., Epel, E. S., Castellazzo, G., & Ickovics, J. R. (2000). Relationship of subjective and objective social status with psychological and physiological functioning. *Health Psychology*, *19*(6), 586.

Arnett, J. J. (2016). The neglected 95%: Why american psychology needs to become less american.

Bailey, J. R., Chen, C. C., & Dou, S.-G. (1997). Conceptions of self and performance-related feedback in the US, japan and china. *J Int Bus Stud*, *28*(3), 605–625.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Blouin, D. C., & Riopelle, A. J. (2004). The difference between t and z and the difference it makes. *J of Gen Psych*, *131*(1), 77–84.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?

Carstensen, A., Saponaro, C., Frank, M. C., & Walker, C. M. (2022). Bridging cultural and cognitive perspectives on similarity reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Carstensen, A., Zhang, J., Heyman, G. D., Fu, G., Lee, K., & Walker, C. M. (2019). Context shapes early diversity in abstract thought. *PNAS*, *116*(28), 13891–13896.

Chan, C. C., Tardif, T., Chen, J., Pulverman, R. B., Zhu, L., & Meng, X. (2011). English-and chinese-learning infants map novel labels to objects and actions differently. *Dev Psy*, *47*(5), 1459.

Cheng, L. (2020). *The development of cognitive styles among american and chinese children* (PhD thesis).

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures. *Psy Bull*, *125*(1), 47.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *PNAS*, *102*(35), 12629–12633.

Church, A. T., Alvarez, J. M., Katigbak, M. S., Mastor, K. A., Cabrera, H. F., Tanaka-Matsumi, J., et al.others. (2012). Self-concept consistency and short-term stability in eight cultures. *J Res Pers*, *46*(5), 556–570.

Church, A. T., Katigbak, M. S., Ibáñez-Reyes, J., Jesús Vargas-Flores, J. de, Curtis, G. J., Tanaka-Matsumi, J., et al.others. (2014). Relating self-concept consistency to hedonic and eudaimonic well-being in eight cultures. *J Cross Cult Psy*, *45*(5), 695–712.

Cleveland, M., & Laroche, M. (2007). Acculturaton to the global consumer culture: Scale development and research paradigm. *Journal of Business Research*, *60*(3), 249–259.

Cleveland, M., Laroche, M., & Takahashi, I. (2015). The intersection of global consumer culture and national identity and the effect on japanese consumer behavior. *Journal of International Consumer Marketing*, *27*(5), 364–387.

Corriveau, K. H., DiYanni, C. J., Clegg, J. M., Min, G., Chin, J., & Nasrini, J. (2017). Cultural differences in the imitation and transmission of inefficient actions. *J Exp Child Psy*, *161*, 1–18.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

DiYanni, C. J., Corriveau, K. H., Kurkul, K., Nasrini, J., & Nini, D. (2015). The role of consensus and culture in children's imitation of inefficient actions. *J Exp Child Psy*, *137*, 99–110.

Doherty, M. J., Tsuji, H., & Phillips, W. A. (2008). The context sensitivity of visual size perception varies across cultures. *Perception*, *37*(9), 1426–1433.

Duffy, S., Toriyama, R., Itakura, S., & Kitayama, S. (2009). Development of cultural strategies of attention in North American and Japanese children. *J Exp Child Psy*, *102*(3), 351–359.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Goddu, M., & Walker, C. M. (2018). Toddlers and adults simultaneously track multiple hypotheses in a causal learning task. In *CogSci*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behav Brain Sci*, *33*(2-3), 61–83.

Imada, T., Carlson, S. M., & Itakura, S. (2013). East–west cultural differences in context-sensitivity are evident in early childhood. *Dev Sci*, *16*(2), 198–208.

Ji, L.-J., Nisbett, R. E., & Su, Y. (2001). Culture, change, and prediction. *Psych Sci*, *12*(6), 450–456.

Ji, L.-J., Peng, K., & Nisbett, R. E. (2000). Culture, control, and perception of relationships in the environment. *JPSP*, *78*(5), 943.

Ji, L.-J., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? *JPSP*, *87*(1), 57.

Jurkat, S., Iza Simba, N. B., Hernández Chacón, L., Itakura, S., & Kärtner, J. (2022). Cultural similarities and differences in explaining others' behavior in 4-to 9-year-old children from three cultural contexts. *Journal of Cross-Cultural Psychology*, 00220221221098423.

Kim, H., & Markus, H. R. (1999). Deviance or uniqueness, harmony or conformity? A cultural analysis. *JPSP*, *77*(4), 785.

Kitayama, S., Park, H., Sevincer, A. T., Karasawa, M., & Uskul, A. K. (2009). A cultural task analysis of implicit independence. *JPSP*, *97*(2), 236.

Kuwabara, M., & Smith, L. B. (2012). Cross-cultural differences in cognitive development: Attention to relations and objects. *J Exp Child Psy*, *113*(1), 20–35.

Kwan, V. S., Bond, M. H., & Singelis, T. M. (1997). Pancultural explanations for life satisfaction: Adding relationship harmony to self-esteem. *JPSP*, *73*(5), 1038.

Le, K., Frank, M., & Carstensen, Alex. (2021). *Is it language or is it culture? Re-examining cross-cultural similarity judgments using lexical co-occurrence* (Undergraduate Honors Theses). Stanford University.

Li, J., Liu, L., Chalmers, E., & Snedeker, J. (2018). What is in a name?: The development of cross-cultural differences in referential intuitions. *Cognition*, *171*, 108–111.

Liang, B., & He, Y. (2012). The effect of culture on consumer choice: The need for conformity vs. The need for uniqueness. *Int J of Consum Stu*, *36*(3), 352–359.

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567.

Markus, H. R., & Conner, A. (2014). *Clash!: How to thrive in a multicultural world*. Penguin.

Markus, H. R., & Kitayama, S. (1992). The what, why and how of cultural psychology: A review of shweder's thinking through cultures. *Psychological Inquiry*, *3*(4), 357–364.

Markus, H. R., & Kitayama, S. (2010). Cultures and selves: A cycle of mutual constitution. *Perspectives on Psychological Science*, *5*(4), 420–430.

Masuda, T., Ishii, K., & Kimura, J. (2016). When does the culturally dominant mode of attention appear or disappear? Comparing patterns of eye movement during the visual flicker task between european canadians and japanese. *Journal of Cross-Cultural Psychology*, *47*(7), 997–1014.

Masuda, T., & Nisbett, R. E. (2006). Culture and change blindness. *Cognitive Science*, *30*(2), 381–399.

Mercier, H., Yama, H., Kawasaki, Y., Adachi, K., & Van der Henst, J.-B. (2012). Is the use of averaging in advice taking modulated by culture? *J Cog & Culture*, *12*(1-2), 1–16.

Mercier, H., Zhang, J., Qu, Y., Lu, P., & Van der Henst, J.-B. (2015). Do easterners and westerners treat contradiction differently? *J Cog & Culture*, *15*(1-2), 45–63.

Moriguchi, Y., Evans, A. D., Hiraki, K., Itakura, S., & Lee, K. (2012). Cultural differences in the development of cognitive shifting. *J Exp Child Psy*, *111*(2), 156–163.

Morris, M. W., Nisbett, R. E., & Peng, K. (1995). Causal attribution across domains and cultures.

Morris, M. W., & Peng, K. (1994). Culture and cause: American and chinese attributions for social and physical events. *JPSP*, *67*(6), 949.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond WEIRD psychology: Measuring and mapping scales of cultural and psychological distance. *Psych Sci*, *31*(6), 678–701.

Na, J., Grossmann, I., Varnum, M. E., Kitayama, S., Gonzalez, R., & Nisbett, R. E. (2010). Cultural differences are not always reducible to individual differences. *PNAS*, *107*(14), 6192–6197.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *J Exp Child Psy*, *162*, 31–38.

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Parsons, S. (2021). Splithalf: Robust estimates of split half reliability. *Journal of Open Source Software*, *6*(60), 3041.

Richland, L. E., Chan, T.-K., Morrison, R. G., & Au, T. K.-F. (2010). Young children's analogical reasoning across cultures: Similarities and differences. *J*

Exp Child Psy, *105*(1-2), 146–153.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of chinese and US preschoolers. *Psych Sci*, *17*(1), 74–81.

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? *Child Dev*, *84*(2), 443–454.

Senzaki, S., Masuda, T., & Nand, K. (2014). Holistic versus analytic expressions in artworks: Cross-cultural differences and similarities in drawings and collages by Canadian and Japanese school-age children. *J Cross Cult Psy*, *45*(8), 1297–1316.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psych Sci*, *22*(11), 1359–1366.

Spencer-Rodgers, J., Boucher, H. C., Mori, S. C., Wang, L., & Peng, K. (2009). The dialectical self-concept. *Pers Soc Psy B*, *35*(1), 29–44.

Spencer-Rodgers, J., Boucher, H. C., Peng, K., & Wang, L. (2009). Cultural differences in self-verification. *J Exp Soc Psy*, *45*(4), 860–866.

Spencer-Rodgers, J., Williams, M. J., Hamilton, D. L., Peng, K., & Wang, L. (2007). Culture and group perception: Dispositional and stereotypic inferences about novel and national groups. *JPSP*, *93*(4), 525.

Su, S. (2020). *Analogical reasoning in Chinese and US adults* (Master's thesis). Cornell University.

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., & Kitayama, S. (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science*, *344*(6184), 603–608.

Tan, B. (2020). *Chinese and US young children's executive function and its sociocultural antecedents* (PhD thesis). The University of Memphis.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. *Dev Psy*, *32*(3), 492.

Waxman, S. R., Fu, X., Ferguson, B., Geraghty, K., Leddon, E., Liang, J., & Zhao, M.-F. (2016). How early is infants' attention to objects and actions shaped by culture? New evidence from 24-month-olds raised in the US and china. *Front Psy*, *7*, 97.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020.

Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences versus strategies as explanations for culture-specific behavior. *Psychological Science*, *19*(6), 579–584.

Zhou, J., Gotch, C., Zhou, Y., & Liu, Z. (2008). Perceiving an object in its context—is the context cultural or perceptual? *Journal of Vision*, *8*(12), 2–2.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.