Investigating cross-cultural differences in reasoning, vision, and social cognition through replication

Alexandra Carstensen [1,*] **(abcarstensen@stanford.edu), Anjie Cao**[1,*] **(anjiecao@stanford.edu),**
Shan Gao[2] **(s.gao@berkeley.edu), and Michael C. Frank**[1] **(mcfrank@stanford.edu)**
[1]Department of Psychology, Stanford University, [2]Department of Psychology, UC Berkeley

Abstract

People perceive, think, and act in a multitude of different ways across cultures, and there is an extensive history of research documenting these differences. At the heart of much of this work is a contrast between Western and East Asian cultures that has inspired important efforts to document human psychology in populations outside of the WEIRD (Western, Educated, Industrial, Rich, Democratic) demographic, which is much overrepresented within psychological research (Henrich, Heine, & Norenzayan, 2010). Recent recommendations for measuring cultural distance (Muthukrishna et al., 2020) profile the US and China as focal points for cultural comparisons, but define cultural distance using explicit self-report measures. Here, we evaluate cross-cultural differences between the US and China using implicit and experimental measures. We attempt to reproduce and test extensions of prior work demonstrating cross-cultural differences in reasoning, vision, and social cognition with convenience snowball samples of university students. Few of these differences appeared in our sample.

Keywords: culture; replication; reasoning; attention; social cognition; variation; US; China

Introduction

Differences in values and behavior across cultures are apparent and large, and psychologists have devised measures of them for decades, resulting in a substantial literature. Comparisons between the United States and China have been especially well-researched, with differences attested in a range of cognitive domains, including visual attention (Ji, Peng, & Nisbett, 2000), language learning (Chan et al., 2011), executive function (Tan, 2020), relational reasoning (Carstensen et al., 2019), similarity judgments (Ji, Zhang, & Nisbett, 2004), values (Spencer-Rodgers, Williams, Hamilton, Peng, & Wang, 2007), preferences (Corriveau et al., 2017), and self-concepts (Spencer-Rodgers, Boucher, Mori, Wang, & Peng, 2009). As a result, the US and China are increasingly becoming major cultural poles in efforts to assess and measure differences (Muthukrishna et al., 2020) and correct for the pervasive bias in psychology research toward WEIRD samples (Henrich et al., 2010).

Despite a long empirical tradition of comparisons between these two cultures and an abundance of psychological accounts for observed differences, there is little consensus on how cultural constructs and measurements relate to each other. Additionally, research linking between cultural tasks is inherently correlational, and does not control for other factors, putting causal inference on shaky ground. These measures are difficult to aggregate and compare because of varying test populations, languages, and methods. Further, many of these reports predate some of the methodological issues that have been raised in the past 10 years (Open Science Collaboration, 2015) regarding the importance of limiting analytic flexibility in order to decrease false positives (Simmons, Nelson, & Simonsohn, 2011). As a result, replication of these results would provide important confirmation of the robustness of these effects. Accordingly, we report a systematic replication of a set of cross-cultural tasks in an online format.

There is already some empirical evidence suggesting issues in the reliability cross-cultural measurements. 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, often do 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) find a lack of coherence between tasks measuring social orientation and cognitive style, observing 8 correlations across 90 tests. The authors note that their findings imply the measures are orthogonal, and conclude that group-level differences between cultures are unlikely to relate to within-group individual differences. However, this study included only US participants and was therefore unable to reproduce cross-cultural differences, leaving open the alternative (and perhaps simpler) possibility that the task measurement properties are simply poor. Indeed, more recent work has raised concerns about low external validity and failed to replicate several related measures (Mercier, Yama, Kawasaki, Adachi, & Van der Henst, 2012; Mercier, Zhang, Qu, Lu, & Van der Henst, 2015; Zhou, Gotch, Zhou, & Liu, 2008).

We seek to gain traction on both theoretical and methodological issues in relating tasks by replicating and comparing several implicit and experimental measures in a single cross-cultural study. Our task selection process was shaped by an interest in relational reasoning and accounts explaining it with reference to cross-cultural differences in visual attention and social cognition (e.g., Kuwabara & Smith, 2012; Duffy, Toriyama, Itakura, & Kitayama, 2009). Specifically, we sought to examine whether greater attention to the visual context of objects or preferences for complex visual scenes facilitate relational reasoning. Within the social domain, we planned to test for self-focus, preferences for uniqueness, and sensitivity to personal motivations, all of which could indi-

cate attention to the self at the cost of broader social and contextual awareness; if this socialized neglect of context generalizes to broader informational biases against context then this could have a detrimental effect on relational reasoning. Additionally, we selected tasks that can 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 incorporated four desiderata in our task selection, preferentially choosing tasks that (1) have 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 linked to vision or social cognition accounts for relational reasoning. We further conducted a fairly extensive set of pilots to ensure that participants understood instructions and that tasks yielded interpretable data.

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 recruited their friends and family members; Talhelm et al. (2014) report that this approach (in China) produces a broad cross-section of the country with representation across geographic regions and the 23 provinces. Participants in the US were compensated with $5 gift certificates (USD) and in China received ¥35 (CNY). We recruited 203 and 201 participants each from the US and China, respectively. The sample size is chosen to meet or exceed the sample sizes of the tasks in the literature.[1]

Our original preregistered exclusion plan was to exclude people from the whole experiment if they failed quality checks on any one task. However, due to the task demand associated with the Symbolic Self-Inflation task, this criterion would have led to the exclusion of 85 people (US: 59, CN: 26) from this task alone. As a result, we deviate from our pregistration in including people who only failed at the quality check for the sociogram task.

After exclusions (see SI), 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 M, 112 F, 1 NB, 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 has 110 US participants and 141 CN participants.

[1]Our study sample was substantially larger than six of the eight tasks we reproduce, comparable to one (Horizon Collage), and smaller only than the previous sample for our exploratory measure (Raven's SPM). While effect sizes may vary across populations, this sample should provide comparable power to detect a similarly robust cross-cultural difference as in each of these original studies.

Our sample covered a broad geographic cross-section of both countries, with US participants from 36 of the 50 US states and Chinese participants from 28 of the 34 Chinese provinces. Subjective SES ratings (using the MacArthur ladder; Adler, Epel, Castellazzo, & Ickovics, 2000) were similar across contexts, but slightly higher in the US sample (median = 6; CN median = 5). Participants in both samples had lived in the target country for most of their lives, with only 8.3% of the US sample living abroad for 2 years or longer and 6.0% of the CN sample. We included a question about international travel as a proxy of globalization in our samples and found that participants in the US sample generally had more experience traveling abroad (93.5% had traveled internationally, and the modal response was "6 or more" international travel experiences) than Chinese participants (39.5% had traveled internationally, with a mode of zero visits).

Procedure

Participants completed eight tasks and a brief demographics questionnaire, administered online 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 congratulates participants on being nearly done with the experiment. In total, the experiment took about 30 minutes to complete. Detailed methods are available at `https://osf.io/3hfwk/?view_only=04227ff032ad446fb126fa102ff056d6`

Measures

Ambiguous RMTS 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. Because (US) adults perform near ceiling on the causal relational match-to-sample task, 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, & Itakura (2013) found that children around the age of 6 showed cultural differences in describing pictures to others. Relative to the 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 this 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 Imada et al. (2013) found Japanese children are more susceptible to the Ebbinghaus Illusion – in which context alters the perceived size of a circle – than US children. In this task, we followed the original implementation, 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 measure accuracy of circle size judgments as a function of the actual size difference and flanker context, which is helpful when aligned with the correct answer and otherwise misleading.

Horizon Collage Senzaki, Masuda, & 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 larger 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 in a rectangular "canvas." They then labeled each circle for the person it represents. 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 others.

Uniqueness Preference Kim & Markus (1999) tested East Asians' and Americans' preferences for harmony or uniqueness by asking them to pick one gift pen from five provided options differing only in barrel colors, and found that European Americans were more likely to choose the uncommon color than East Asian participants. We adapted our task to better fit the format of our online experiment by showing a sticker book to measure progress through all tasks in our study. At the end of each task, participants received a sticker; for the uniqueness preference task, we let them select one of five dinosaur stickers, e.g. four blue dinosaurs and one yellow. Choice of the unique or repeated color was recorded.

Causal Attribution Previous work (Morris & Peng, 1994) has found that Chinese media is more likely than US media to attribute behaviors to situational contexts than to an actor's personal traits, and that participants from each context show corresponding preferences for situation (CN) or person (US) attributions. In a review of attribution studies comparing East Asian and Western styles, Choi, Nisbett, & Norenzayan (1999) note that cross-cultural differences in causal attribution emerge mainly in situations where there is evidence in favor of situational explanations. For this reason, we adapted our study from the deterministic situation condition in Seiver, Gopnik, & Goodman (2013), in which the corresponding behavior of two children who both engage in one activity and avoid another suggests 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 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, which was selected as the most difficult subset, and the one most dependent on true analogical reasoning (without alternative heuristic approaches like visual pattern completion).

Results

Analytic approach

The sample size, methods, and main analyses were pre-registered and are available at `https://aspredicted.org/37y6a.pdf`. Data and analy-

sis scripts are available at https://osf.io/65hwd/?viewonly=04227ff032ad446fb126fa102ff056d6. Departing from the heterogeneous approaches used by previous authors, we attempted to follow current best practices by using linear mixed effects models with maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013), fit 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). The majority of results are visualized in Figure 1, except for the Ebbinghaus Illusion data, in Figure 2.

While our main analyses followed our preregistered tests, described above, in response to reviewer feedback, we included Bayes Factor (BF) analyses to evaluate evidence for null hypotheses relative to test hypotheses. In each case, we evaluated evidence for the full model specified in our preregistration (unless otherwise noted). We adopt a conventional threshold of >3 of $<.3$ for interpreting the BF ratio as evidence for the test or null hypothesis, respectively.

Individual Task Results

The majority of tasks results are visualized in Figure 1, except for the Ebbinghaus Illusion data, in Figure 2.

Ambiguous RMTS 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. We included an intercept per subject as a random effect, as well as by-subject random slopes for trial number to model order effects (as participants may be more likely to consider ambiguity in the task and switch response strategies with time). This model did not converge; following lab standard operating procedures, we pruned the random slopes for trial number. There was no main effect of country on response choice (object or relation; US: $M = 0.39$; CN: $M = 0.37$; $\beta = 0.14$, $SE = 0.89$, $z = 0.16$, $p = 0.87$). Due to the run time of the Bayes Factor analysis on the full model, we ran the analysis on the pruned model. Our results suggested that the evidence was in favor of the null hypothesis ($BF = 0.003$).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.

Picture Free Description We selected this task to investigate whether there are cultural differences in visual attention, as measured by the content of picture descriptions. 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: $M = 0.9$; CN: $M = 0.56$) but not the latter (number of descriptive accounts; For focal objects: US: $M = 1.06$; CN: $M = 0.88$; For background objects: US: $M = 1.31$; CN: $M = 0.94$). 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 and subject as random 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 the interaction between description type (focal or background) and country (US or China) as fixed effect, description type and subject as random effect, and by-picture random slope for country. 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 result of our Bayes Factor analyses is consistent with our model ($BF = 174743.6$). These mixed results underscore the importance of metric, and they should be interpreted with caution.

Ebbinghaus Illusion This task was included as a second measurement for cultural differences in visual attention. Imada et al. (2013) found that Japanese children are more susceptible to Ebbinghaus illusion than US children. 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, with their interactions. As random effects, we included intercepts for subjects, as well as by-subject random slopes for the effect of context. We used this model instead of the full model due to its lack of stable estimates. 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 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 $p > 0.05$). Due to the runtime of the Bayes Factor analysis on the full model, we ran the analysis on the pruned model instead. The Bayes Factor suggested that the results were in favor of the null hypothesis ($BF = 26646.4$).

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). Senzaki et al. (2014) found that Japanese children tend to put the horizon higher and include more stickers that cover more area in their collage, compared to the Canadian children. We ran

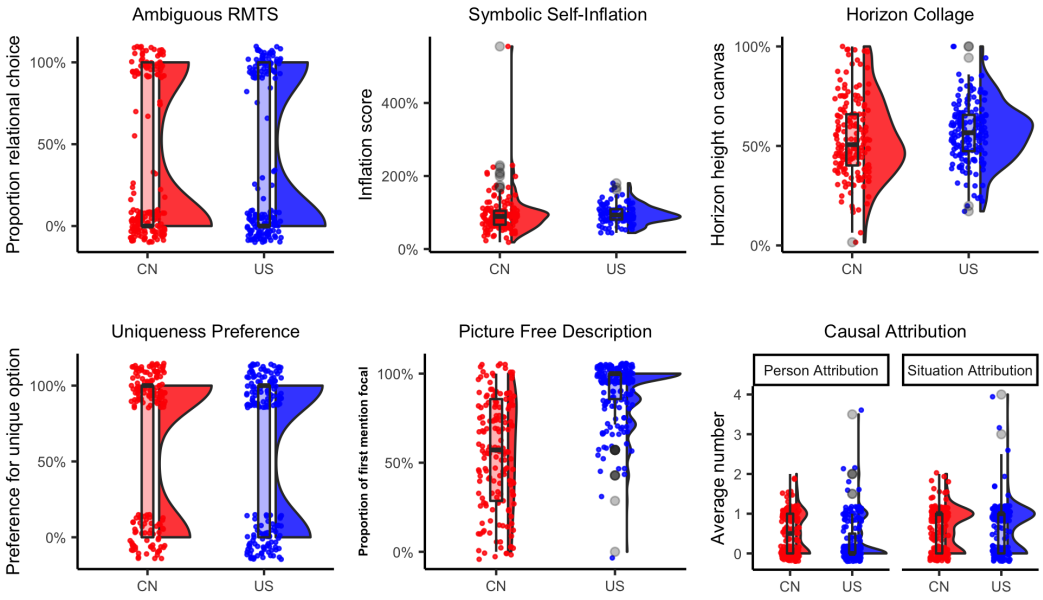

Figure 1: Results from each task. Results from the CN sample are plotted in red (N = 167, N = 141 for Symbolic Self-Inflation (SSI) Task), and the US in blue (N = 169, N = 110 for SSI). Only first-mentioned objects in Picture Free Descriptions showed cross-cultural differences.

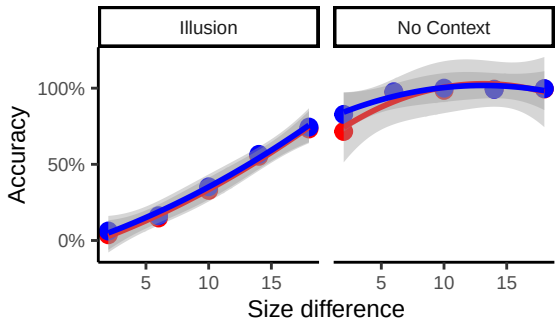

Figure 2: Ebbinghaus Illusion. The magnitude of differences between the two circles is plotted on the x-axis, accuracy on the y-axis. The left panel shows accuracy in the Illusion block, and right in the No Context block. The red (China) and blue (US) lines show that accuracy varies as a function of trial and condition.

a fixed effect linear model with culture as the main predictor for each of the measurements. We found that culture did not significantly predict any of the three measurements (Sticker height: US: M = 0.57; CN: M = 0.54; Sticker number: US: M = 11.51; CN: M = 11.77; Sticker area: US: M = 16.98; CN: M = 17.43; All $\beta < 0.03$; All $p > 0.1$). Bayes Factor analyses suggest that we have equal evidence for the test and null hypotheses regarding horizon height (BF = 1.00). For the number of stickers and area covered by stickers, we observe slightly stronger evidence for test hypotheses, but the results were still inconclusive

Symbolic Self-Inflation To test whether US adults show a greater tendency toward symbolic self inflation 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; CN M = 0.95; $\beta < 0.01$, SE = 0.06, t = 0.04, p = 0.97). The Bayes Factor also suggests that the current data are in favor of the null hypothesis (BF = 0.14). In order to test whether this null finding resulted from the different ways of calculating inflation scores between our study and Kitayama et al.'s (2009) original study, we then tried to do a replication of the original analyses, using subtraction-based inflation scores (which may be subject to baseline effects). The degree of self-inflation still did not differ significantly between US and Chinese adults (US M = -6.11; CN M = -13.58; β = 7.48, SE = 4.77, t = 1.57, p = 0.12). The Bayes Factor for this alternative analysis suggested that the evidence is in favor of the test hypothesis (BF = 40.91). The mixed results underscored the importance of metrics; they should be interpreted with cautions.

Uniqueness Preference As the second task measuring social cognition, this task was included to examine cross-cultural preferences for uniqueness. We ran a simple logistic regression predicting each participant's single choice (minority or majority color) with country (US or China) as a fixed effect. No cross-cultural difference was found in the likelihood of choosing the uniquely colored sticker (US: M = 0.57; CN: M = 0.63; β = -0.23, SE = 0.22, z = -1.02, p = 0.31).Our

Bayes Factor analysis suggested that we have equal evidence for the null hypothesis and the test hypothesis (BF = 0.95). Therefore, we did not find support for the hypothesis that US adults are more likely to show a preference for uniqueness.

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 and subject and trial as random effects, with by-subject random slopes for attribution type and by-trial random slopes for country. This model failed to converge. Following the standard pruning procedure, we pruned the model into a fixed effect modle with attribution type, country, and their interaction.We found an interaction effect between description type and country, but in the opposite direction as previous results (β = 0.44, SE = 0.16, z = 2.84, $p < 0.01$). US participants provided more situation attributions relative to personal attributions, compared to Chinese participants (Situation attribution: US: M = 0.65; CN: M = 0.65; Person attribution: US: M = 0.33; CN: M = 0.51). Our Bayes Factor analysis was consistent with our models' results (BF = 12918.32[2])

Raven's SPM To test for differences in our exploratory measure of relational reasoning performance in adults, 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; CN: M = 0.84; β = -1.31, SE = 0.23, z = -5.64, $p < 0.01$). This finding replicates Su (2020) in finding an advantage on this measure. In our context, we interpret the relatively high scores we observed as evidence that participants were engaging fully with our tasks.

Conclusion

A rich literature investigates differences in cognition between the US and China, with this comparison serving as a case study for broader cross-cultural differences. Yet the robustness of this literature is difficult to ascertain due to limited direct replications and widely varying methods and analyses. To begin to address this situation, we planned a structured replication of eight cross-cultural tasks with online convenience samples of US and Chinese university students.

We did not observe cross-cultural differences in the majority of the tasks we tested. The only exceptions were in picture description and our exploratory measure of reasoning performance (Raven's matrices). Many of our tasks do 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. We discuss other limitations of our study below, but – to a first approximation – the low level of replication we saw was dispiriting. Despite the self-evident cultural differences between the US and China, measuring differences in cognition based on the previous literature may be non-trivial.

In Picture Free Description, the only one of our primary tasks that showed a difference, US participants were more likely to reference focal elements first when describing the scene. This finding is consistent with the "culture task" account, by which practice with behaviors that are uniquely relevant in a given cultural context can produce culture-specific psychological tendencies (Kitayama et al., 2009). Specifically, China has been characterized as a high-context culture, where people are expected to pay close attention to context, and integrate this information to interpret reltively subtle social cues. In contrast, the US has been characterized as a low-context culture, where communication tends to rely on explicit verbal information, contextual cues carry less weight, and people are therefore less likely to prioritize background information. However, we did not find cross cultural differences in the Horizon Collage task and Ebbinghuas Illusion task, suggesting that not every task induces differential focus on contextual elements for Chinese participants. Perhaps the specific linguistic focus of picture description leads to cross-cultural differences; other task comparisons would be necessary to probe this finding further.

None of the three social cognition measures we used showed cross-cultural differences, though in each case we made some modifications to the original task to adapt it for our study. In the Symbolic Self-Inflation task, we asked participants to draw themselves and the family members they grew up with instead of their "social network" (as in the original task). It is possible that size symbolism (e.g. drawing larger parent circles) washed out cultural differences to some extent, though this result is qualified by the suggestion that some difference might be present on an alternative analysis. Adapting the Uniqueness Preference task to an online format also may have trivialized the choice and undermined incentives for unique or harmonious behavior as the expression of cultural values (consistent with the random responding we observed). The null results in the Causal Attribution task are perhaps more surprising, but it may be that situational explanations are too salient for adults in our version of the task producing a ceiling effect. Nonetheless, our findings do provide qualifying evidence on the robustness of cross-cultural differences in these tasks.

It seems likely that a combination of factors is responsible for the lack of robust differences observed in these tasks, including both cultural convergence as a result of globalization (increasing cultural similarity between the US and China), and cultural heterogeneity within East Asia broadly or mainland China more specifically (limiting the generality of previous findings among other populations in East Asia). In the

[2]The test and null models had divergent transitions that were not resolved by adjusting model parameters so the estimates may be unstable. However, this analysis is consistent with our main regression in finding a main effect of culture and interaction which, taken together, are inconsistent with the original predictions.

latter case, these findings may serve as a cautionary note on generalizations about East Asian culture based on a single sample within East Asia.

Our study has a variety of limitations that should be noted in the interpretation. First, as discussed above, our tasks and analyses are typically adaptations of the original literature rather than direct replications. Task or analytic differences could in principle play a role in explaining our failures. Second, although our sample size is larger than many of the original studies, our design is underpowered to detect small cultural differences. Third, the young, well-educated, and relatively worldly student populations in our study do not provide the most representative samples of their countries, especially given that 20 or more years have elapsed since several of the original reports. It may be the case that Chinese college students who were not born at the time of the original studies have somewhat different cultural cognition than the original participants. Further work with other samples would be critical to assessing the generality of our effect measurements.

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.

Acknowledgements

We are grateful for Toshie Imada, Daniel Casasanto, and Shikun Su for materials, and to members of Language and Cognition Lab and Culture Collab group at Stanford University for their valuable feedback. This work was funded in part by awards from the McDonnell Foundation and the Center for the Study of Language and Information at Stanford University supporting A. Carstensen.

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.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem & Lang*, *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.

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.

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

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.

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.

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., 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.

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.

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.

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.

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

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., 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.

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.