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

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

[3] [1] Department of Psychology, Stanford University

[4] [2] Department of Psychology, University of California, San Diego

[5] [3] FIXME

[6] Author Note

[11] Correspondence concerning this article should be addressed to Anjie Cao*, 450 Jane

[12] Stanford Way, Stanford, 94305. E-mail: anjiecao@stanford.edu

13   Abstract

14 Cultural differences between the US and China have been investigated using a broad array

15 of psychological tasks measuring differences between cognition, language, perception, and

16 reasoning. We conducted two large-scale replications of a selection of 12 tasks previously

17 reported to show cross-cultural differences, using online convenience samples of adults.

18 Five of these tasks showed robust cross-cultural differences, while six showed no difference

19 and one showed a small difference in the opposite direction. Tasks showing cross-cultural

20 differences tended to have multiple trials measuring high-level reasoning and language;

21 those that did not show cross-cultural difference included measures of attention/perception,

22 implicit social processes, or were initially designed to show differences in children. As in

23 prior work, cross-cultural differences in cognition (in those tasks showing differences) were

24 not strongly related to explicit measures of cultural identity and behavior.

25   *Keywords:* FIXME

26   Word count: FIXME

₂₇ Replicability of US-China differences in cognition and perception across 12 tasks

Introduction

₂₉ Cross-cultural differences are a striking part of the broader landscape of human
₃₀ variation. Differences in values and behavior across cultures are obvious to even a casual
₃₁ observer, and researchers have attempted to quantify these differences via a wide range of
₃₂ measures. Comparisons between the United States and China – often as exemplars of
₃₃ Western and East Asian cultures – have been especially well-researched, with differences
₃₄ attested in a wide range of cognitive domains, including visual attention (Chua, Boland, &
₃₅ Nisbett, 2005; Ji, Peng, & Nisbett, 2000; Waxman et al., 2016), executive function
₃₆ (Sabbagh, Xu, Carlson, Moses, & Lee, 2006; Tan, 2020), language learning (Chan et al.,
₃₇ 2011, 2011; Tardif, 1996; Waxman et al., 2016), relational reasoning (Carstensen et al.,
₃₈ 2019; Cheng, 2020; Richland, Chan, Morrison, & Au, 2010; Su, 2020), similarity judgments
₃₉ (Ji, Zhang, & Nisbett, 2004), values (Ji, Nisbett, & Su, 2001; Kwan, Bond, & Singelis,
₄₀ 1997; Spencer-Rodgers, Williams, Hamilton, Peng, & Wang, 2007), preferences (Corriveau
₄₁ et al., 2017; DiYanni, Corriveau, Kurkul, Nasrini, & Nini, 2015; Liang & He, 2012) and
₄₂ self-concepts (Spencer-Rodgers, Boucher, Mori, Wang, & Peng, 2009; Spencer-Rodgers,
₄₃ Boucher, Peng, & Wang, 2009). As a result, the US and China are increasingly becoming
₄₄ major cultural poles in efforts to assess and measure cultural differences (Muthukrishna et
₄₅ al., 2020) and correct for the pervasive bias in psychology research toward US and
₄₆ European samples (Arnett, 2016; Henrich, Heine, & Norenzayan, 2010; Nielsen, Haun,
₄₇ Kärtner, & Legare, 2017).

₄₈ Despite a long empirical tradition of comparisons between these two cultures and an
₄₉ abundance of psychological accounts for observed differences, there is little consensus on
₅₀ how measurements relate to each other. Reports of cultural differences are often difficult to
₅₁ compare quantitatively because of the varying samples, measures, and methods used in
₅₂ 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.

There is already some empirical evidence suggesting issues in the robustness of cross-cultural measurements. 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, 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) 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]. Further, more recent work has raised concerns about low external validity of some cross-cultural measures and 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

---

[1] The authors interpret their findings as imply the measures are orthogonal – indexing different constructs – and conclude 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. .

76 comparisons between US and either Chinese or Japanese participants). We made the

77 decision to pursue the strategy of gathering relatively large and heterogeneous convenience

78 samples using online recruitment, rather than recruiting smaller, more matched samples

79 using in-lab recruitment. Our reasoning was that the larger samples that we could access

80 using online recruitment would allow us to conduct highly-powered statistical tests,

81 allowing us to either reject or accept the null hypothesis of no cultural difference between

82 measures. Further, larger samples would afford the analysis of individual and demographic

83 differences within culture, a topic of considerable interest in this literature (e.g., Na et al.,

84 2010). Finally, the development of browser-based online versions of prominent

85 cross-cultural tasks would allow their inspection and reuse by other researchers, thus

86 promoting a more cumulative approach to the measurement of cultural differences.

87 The interpretation of any replication result is complex, given that disparate outcomes

88 between an initial study and a replication can occur for many reasons – including but not

89 limited to differences in experimental methods, sample or population differences, and

90 simple sampling variation in the outcomes Machery (2020). Our strategy of pursuing online

91 convenience samples limits the interpretation of our replication results: nearly all of the

92 tasks we selected had previously been administered in person, and the populations sampled

93 in previous reports varied but were largely convenience samples of either college students or

94 community members. More generally, our strategy of constructing a battery of replication

95 studies and administering them uniformly means that specific decisions about sampling

96 and administration will not be matched with the original studies (which were likely

97 heterogeneous in their samples and administration). Thus, our replication studies should

98 be taken as an assessment of whether a set of previously-reported cross-cultural differences

99 can be recovered in convenience populations recruited online, rather than as assessments of

100 the veracity of the original findings. Nevertheless, we believe that the field of cross-cultural

101 psychology can be advanced via the identification of tasks that yield cross-cultural

102 differences robustly across a variety of samples and administration formats and we hope our

103 work contributes to this aim. We return to this interpretive issue in the General Discussion.

104   Our task selection process was initially shaped by an interest in relational reasoning

105 and accounts explaining it with reference to cross-cultural differences in visual attention

106 and social cognition (Duffy, Toriyama, Itakura, & Kitayama, 2009; e.g., Kuwabara &

107 Smith, 2012; Moriguchi, Evans, Hiraki, Itakura, & Lee, 2012). Additionally, in Experiment

108 1, we selected tasks that could potentially be administered to young children as well as

109 adults, for use in future work addressing developmental questions about the relative time

110 course of cross-cultural differences across the visual, social, and cognitive domains. We

111 balanced four desiderata in our task selection, preferentially choosing tasks that (1) had

112 been theoretically or empirically implicated in relational reasoning, (2) were associated

113 with differential performance in US-China comparisons or related cultural contrasts (e.g.,

114 East Asian vs. Western cultures), (3) were relatively short, accessible tasks appropriate for

115 web administration, and (4) were vision or social cognition accounts for relational

116 reasoning. We further conducted a fairly extensive set of pilots to ensure that participants

117 understood instructions and that the tasks yielded interpretable data.

118   In Experiment 2, we selected a second set of tasks to investigate based in part on the

119 results of Experiment 1. In particular, we repeated a handful of tasks from Experiment 1

120 to collect further evidence (in some cases, varying task parameters). We then selected a

121 further set of tasks that probed both cross-cultural differences in higher-level cognition

122 (e.g., language and reasoning) and perception, again respecting the desideratum that the

123 tasks should be relatively short and amenable to administration in a web browser. The

124 final set of tasks included in each Experiment is listed in Table 1.

125   In addition to the goal of replicating individual tasks in which cross-cultural

126 differences in cognition and perception had been previously reported, our hope was that

127 the relatively large dataset that we collected could be used to explore the structure of

128 within- and across-cultural variation in cognition and perception more broadly. Towards

¹²⁹ this goal, we included a relatively extensive demographic questionnaire in both of our

¹³⁰ Experiments, with the aim of using these measures to explore variation within our samples.

¹³¹ In the final section of the paper, we report a series of exploratory analyses. The first of

¹³² these assess the reliability of individual tasks, aiming to gauge whether individual tasks are

¹³³ reliable enough from a psychometric point of view to support further individual differences

¹³⁴ analyses. We then report across-task correlations, aiming to discover covariation between

¹³⁵ tasks that might indicate that they load on the same construct. Finally, we turn to

¹³⁶ analyses of whether within-culture demographic variables predict variation in task

¹³⁷ performance. Overall, a number of tasks revealed acceptable levels of reliability, and tasks

¹³⁸ did not cluster together. We found relatively limited evidence for demographic predictors

¹³⁹ of within-culture variation, however.

¹⁴⁰ We make all code and data from our experiments available for further data collection

¹⁴¹ and analysis in hopes of promoting further cumulative work on measures and theories of

¹⁴² cross-cultural variation.

¹⁴³ **Experiment 1**

¹⁴⁴ In Experiment 1, our goal was to evaluate cross-cultural differences in a variety of

¹⁴⁵ constructs. We assembled a web-based battery of tasks and tested these on a snowball

¹⁴⁶ sample of US and Chinese participants.

¹⁴⁷ **Methods**

¹⁴⁸ **Participants.**   We recruited participants through snowball sampling seeded at large

¹⁴⁹ universities in the US and China, in which participants directly recruited by the researchers

¹⁵⁰ were encouraged to recruit their friends and family members through email forwarding and

¹⁵¹ social media sharing. Participants in the US were compensated with $5 gift certificates

¹⁵² (USD) and participants in China received ¥35 (CNY).

¹⁵³ We recruited 203 and 201 participants each from the US and China, respectively.

¹⁵⁴ Since we did not have strong a priori expectations about specific effect sizes, our overall

¹⁵⁵ preregistered sample size was chosen to meet or exceed the sample sizes used in prior

¹⁵⁶ reports in the literature from which our tasks were drawn.

¹⁵⁷ Our original preregistered exclusion plan was to exclude people from the full dataset

¹⁵⁸ if they failed quality checks on any one task. However, due to a task demand associated

¹⁵⁹ with the Symbolic Self-Inflation task, this criterion would have led to the exclusion of 85

¹⁶⁰ people (US: 59, CN: 26) due to this task alone. As a result, we deviate from our

¹⁶¹ preregistration and include participants in the broader dataset even if they failed the

¹⁶² quality check for the Symbolic Self-Inflation task.

¹⁶³ After exclusions, the US sample included 169 participants (44 Male, 114 Female, 9

¹⁶⁴ Non-binary, 2 Declined to answer), with a mean age of 21.79 years old, all of whom were

¹⁶⁵ native English speakers. The China sample included 167 participants (51 Male, 112

¹⁶⁶ Female, 1 Non-binary, 3 Declined to answer), with a mean age of 22.49 years old, who were

¹⁶⁷ all native speakers of Mandarin Chinese. This sample size is shared among all tasks except

¹⁶⁸ for the Symbolic Self-inflation task, which included 110 US participants and 141 CN

¹⁶⁹ participants.

¹⁷⁰ In addition to age, gender and linguistic background, we collected a range of

¹⁷¹ demographic information including subjective socioeconomic status measured using the

¹⁷² MacArthur Ladder (Adler, Epel, Castellazzo, & Ickovics, 2000), level of maternal

¹⁷³ education, the state or province the participant grew up in, residential mobility, and

¹⁷⁴ number of overseas experiences.

¹⁷⁵ **Procedure.**   Participants completed an online, browser-based sequence of eight

¹⁷⁶ tasks (see Table 1) and a brief demographic questionnaire. All tasks were implemented in a

¹⁷⁷ combination of jsPsych (De Leeuw, 2015) and custom HTML/JavaScript code. Tasks were

¹⁷⁸ 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. Because (US) adults perform near ceiling on the standard causal relational match-to-sample task, 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).

²³³   ***Symbolic Self-Inflation.***   Kitayama et al. (2009) found a difference between

²³⁴ Western and East Asian cultures in the size of circles participants drew to represent

²³⁵ themselves relative to other people in their social networks. Japanese participants drew

²³⁶ circles of similar sizes to represent themselves and others, while those from Western

²³⁷ countries (US, UK, Germany) tended to draw their "self" circles larger than those

²³⁸ representing others, indicating a symbolic self-inflation in the three western cultures

²³⁹ compared to Japan. We adapted this task, asking participants to draw themselves and the

²⁴⁰ family members they grew up with as circles by clicking and dragging the mouse on a

²⁴¹ rectangular "canvas" to draw circles of varying sizes. They then labeled each circle for the

²⁴² person it represented. We measured the diameter of each circle and calculated a percent

²⁴³ inflation score for each participant by dividing the diameter of the self circle by the average

²⁴⁴ diameter of circles for all others.

²⁴⁵   ***Uniqueness Preference.***   Kim and Markus (1999) tested East Asians' and

²⁴⁶ Americans' preferences for harmony or uniqueness by asking them to pick one gift pen

²⁴⁷ from five options. In the condition that we replicated, the options differed only in the

²⁴⁸ barrel colors – four were the same and one was unique. They found that European

²⁴⁹ Americans were more likely to choose the unique colored one than East Asian participants.

²⁵⁰ We adapted our task to better fit the format of our online experiment by showing a virtual

²⁵¹ "sticker book" to measure progress through all tasks in our study. At the end of each task,

²⁵² participants received a virtual sticker. For the uniqueness preference task, we let them

²⁵³ select one of five dinosaur stickers: four blue dinosaurs and one yellow. Choice of the

²⁵⁴ unique vs. repeated color was recorded.

²⁵⁵   ***Causal Attribution.***   Previous work has shown that participants from South

²⁵⁶ Korea and the U.S. attribute behaviors differently in situations where there is evidence in

²⁵⁷ favor of situational explanations (Choi, Nisbett, & Norenzayan, 1999). Similarly, Chinese

²⁵⁸ media is more likely than U.S. media to attribute a person's behaviors to situational

²⁵⁹ context as opposed to individual traits (Morris, Nisbett, & Peng, 1995; Morris & Peng,

260 1994). To create a child-friendly version of the causal attribution task, we adapted our

261 study from the deterministic situation condition in Seiver, Gopnik, and Goodman (2013),

262 in which two children both engage in one activity and avoid another, suggesting that

263 situational constraints (e.g., the latter activity being dangerous) may be guiding their

264 decisions. Participants watched a series of four short, animated vignettes in which two

265 children both played in a pool and neither child played on a bicycle. We then asked

266 participants to explain in text why each child did not play on the bicycle, making for two

267 test trials per participant. We used the prompt question from Seiver et al. (2013), which

268 explicitly pits person attributions against situational ones: "Why didn't Sally play on the

269 bicycle? Is it because she's the kind of person who gets scared, or because the bicycle is

270 dangerous to play on?" We coded each response for per-trial count of (a) person and (b)

271 situation attributions.

272 ***Raven's Standard Progressive Matrices.*** As an additional attention check as

273 well as an exploratory measure of relational reasoning assessing performance rather than

274 preference, we included the 12 questions from Set E of Raven's Standard Progressive

275 Matrices. Su (2020) found cross-cultural differences between adults in the US and China in

276 performance on this set. This set of questions was selected because it was the most difficult

277 subset and also the one most dependent on true analogical reasoning (without alternative

278 heuristic approaches like visual pattern completion).

279 **Analyitic approach.** Our sample size, methods, and main analyses were

280 pre-registered and are available at https:// aspredicted.org/37y6a.pdf. Data and analysis

281 scripts are available at FIXME

282 The specific papers that we drew on for our tasks used a heterogeneous set of analytic

283 methods. Rather than planning to replicate these specific analyses, we instead attempted

284 to follow current best practices by using linear mixed effects models with maximal random

285 effect structure as a unified analytic framework (Barr, Levy, Scheepers, & Tily, 2013). We

286 fit a separate model to each task. In case of convergence failure, we followed standard

287 operating procedure of pruning random slopes first and then random intercepts, always

288 maintaining random intercepts by participant. We report p-values derived from

289 approximating t-scores from z-scores, which is appropriate for relatively large samples

290 (Blouin & Riopelle, 2004). Our key tests of interest were typically either the coefficient for

291 a main effect of country (US/China) or an interaction of country and condition.

292   While our main analyses used our preregistered models, described above, we also

293 computed Bayes Factors (BFs) for each model to evaluate evidence for null hypotheses

294 relative to test hypotheses. In each case, we fit a Bayesian linear mixed effects model with

295 the maximal random effect structure and default priors using the brms package in R and

296 evaluated evidence for this model as compared to one without the key culture term (either

297 main effect or interaction) using the bridge sampling method (Bürkner, 2017). We adopt a

298 conventional threshold of >3 of $<.3$ for interpreting the BF ratio as evidence for the test or

299 null hypothesis, respectively.

Results

301   FIXME: The majority of results are visualized in Figure 1, except for the Ebbinghaus

302 Illusion data, in Figure 2. Below we discuss the results of each task in turn.

303   **Ambiguous cRMTS.** To examine whether adults in the US and China show

304 differing preferences for object-based or relational solutions, we ran a mixed-effects logistic

305 regression predicting response choice (object or relation) with country (US or China) as a

306 fixed effect. There was no main effect of country on response choice (object or relation; US:

307 $M = 0.39$, $SD = 0.48$; CN: $M = 0.37$, $SD = 0.47$; $\beta = 0.14$, $SE = 0.89$, $z = 0.16$, $p =$

308 0.87). The Bayes Factor analysis suggested that the evidence was in favor of the null

309 hypothesis ($BF = 0.17$). The preference for object-based solutions seen in US preschoolers

310 and the corresponding preference for relational solutions observed in China in an

311 ambiguous context did not extend to adults in our samples.

312   Our US results replicate findings by Goddu and Walker (2018), who reported that US

313 adults are at chance in this paradigm. It seems likely that adults in both groups of our

314 study are aware of the ambiguous evidence and their near-chance selections reflect

315 (reasonable) uncertainty.

316   **Picture Free Description.**   Based on Imada et al (2013), we expected Chinese

317 participants would be more likely to mention background objects first and provide more

318 descriptive accounts for background objects relative to focal objects, in comparison with

319 US participants. Our results extend previous findings with the former metric (first

320 mention; Proportion of Relational Match choice: US: $M = 0.90$, $SD = 0.17$; CN: $M = 0.56$,

321 $SD = 0.30$) but not the latter (number of descriptive accounts; For focal objects: US: $M =$

322 1.06, $SD = 0.51$, ; CN: $M = 0.88$, $SD = 0.44$; For background objects: US: $M = 1.31$, SD

323 $= 0.94$; CN: $M = 0.94$, $SD = 0.72$).

324   For first mention, we ran a mixed-effects logistic regression predicting the type of first

325 mention (object or relation) with country (US or China) as a fixed effect. We found a main

326 effect of country ($\beta = 3.36$, $SE = 0.34$, $z = 9.94$, $p < 0.01$). For descriptive accounts, we

327 ran a mixed-effect Poisson regression model predicting the number of descriptive accounts,

328 with description type (focal or background), country (US or China), and their interaction

329 as fixed effects. There was a significant main effect of culture (with US participants

330 providing more descriptions overall: $\beta = 0.36$, $SE = 0.13$, $t = 2.68$, $p < 0.01$). The culture

331 effect interacted with the description types, but the effect was in the opposite direction,

332 with U.S participants provided more background descriptions than focal descriptions,

333 relative to Chinese participants ($\beta = -0.16$, $SE = 0.07$, $t = -2.16$, $p < 0.05$). The Bayes

334 Factor analysis was consistent with the frequentist models (For first mention: $BF =$

335 6,918.24; For description type: $BF = 12.03$). The mixed results between the first mention

336 and descriptive accounts measures suggest that there is some complexity in linking broader

337 theoretical accounts to specific measures; we interpret this result with caution and include

338 the task in Experiment 2 to follow up further.

339   **Ebbinghaus Illusion.**   To test whether perception of the Ebbinghaus illusion

340   varied across populations in our sample, we ran a mixed-effects logistic regression

341   predicting accuracy on each trial, with country (US or China), context (No Context or

342   Illusion Context), and circle size difference (the percent of difference in diameters) as fixed

343   effects, along with their interactions. We found main effects of context (with worse

344   performance in the Illusion Context; $\beta = 4.95$, $SE = 0.29$, $z = 17.03$, $p < 0.01$) and circle

345   size difference (worse performance for smaller differences; $\beta = 0.34$, $SE = 0.01$, $z = 27.33$,

346    $p < 0.01$). There was a marginally significant main effect of country at the opposite of the

347   predicted direction (US participants performed worse: $\beta = 0.52$, $SE = 0.26$, $z = 1.95$, $p = $

348   0.05) but no interactions with country (All $\beta < 0.01$; All $p > 0.05$). The Bayes Factor

349   suggested that the results were extremely in favor of the null hypothesis ($BF = 0$).

350   In sum, we failed to replicate cultural differences found between Western (US/UK)

351   and Japanese participants in susceptibility to the Ebbinghaus illusion in our sample of US

352   and Chinese adults.

353   **Horizon Collage.**   In the Horizon Collage task, three key measurements are

354   calculated from the "collage" participants created: the height of the horizon (height in

355   proportion to the height of the frame), the number of stickers, and the total area of the

356   stickers covered (following the original analysis, we added up the area occupied by each

357   individual sticker); Japanese children tend to put the horizon higher and include more

358   stickers that cover more area in their collage, compared with Canadian children. We ran a

359   fixed effect linear model with culture as the main predictor for each of the measurements.

360   Culture did not significantly predict any of the three measurements (Sticker height: US: M

361    $= 0.57$, $SD = 0.15$; CN: $M = 0.54$, $SD = 0.20$; Sticker number: US: $M = 11.51$, $SD = 5.81$;

362   CN: $M = 11.77$, $SD = 5.80$; Sticker area: US: $M = 16.98$, $SD = 8.36$; CN: $M = 17.43$, SD

363    $= 8.60$; All $\beta < 0.03$; All $p > 0.1$). All bayes factors suggest that there are only anecdotal

364   evidence supporting the culture effect (Sticker height: $BF = 1.02$; Sticker number: $BF = $

365   1.70; Sticker area: $BF = 2.36$)

366   Our experiment contrasted Chinese and US adults, rather than Japanese and

367 Canadian children. Although Senzaki et al. (2014) found that the cultural differences were

368 more salient in older children than younger children, suggesting that cultural differences

369 might increase with development, interpretation of our failure to replicate is still qualified

370 by differences in culture and medium of administration.

371   **Symbolic Self-Inflation.**   To test whether US adults have a larger symbolic self

372 than Chinese adults, we ran a linear regression predicting percent inflation score (calculated

373 by dividing the diameter of the self circle by the average diameter of circles for others) with

374 country (US or China) as a fixed effect. No difference was found in the degree of symbolic

375 self-inflation between US and China adults based on percent inflation scores (US: $M =$

376 0.95, $SD = 0.26$; CN: $M = 0.95$, $SD = 0.55$; $\beta = 0.36$, $SE = 0.13$, $t = 2.68$, $p < 0.01$). The

377 Bayes Factor shows moderate evidence in favor of the null hypothesis ($BF = 0.14$)

378   One possible explanation for our null results is that we adopted a different task

379 design from Kitayama et al. (2009). Instead of asking participants to draw their social

380 network, our design asked participants to draw themselves and the family members they

381 grew up with. During the coding process, we noticed that people from both cultures

382 tended to draw older people, e.g., their parents, into larger circles, which might have

383 resulted in overall larger circles for other people than the self-circles in our task for both

384 cultures, masking any US-China difference in the degree of self-inflation. It is possible that

385 there are also cultural differences between Japan and China in self concept; Japanese

386 samples typically demonstrate characteristics previously associated with East Asian

387 cultures in general, with Chinese samples deviating from these characteristics at times

388 (Bailey, Chen, & Dou, 1997; Church et al., 2012, 2014).

389   **Uniqueness Preference.**   We examined cross-cultural preferences for uniqueness

390 by running a simple logistic regression predicting each participant's single choice (minority

391 or majority color) with country (US or China) as a fixed effect; we used logistic regression

392 rather than mixed effects logistic regression due to the absence of repeated observations.

393   There was not a large cross-cultural difference in the probability of choosing the uniquely

394   colored sticker (US: $M = 0.57$, $SD = 0.50$; CN: $M = 0.63$, $SD = 0.48$; $\beta = $ -0.23, $SE = $

395   0.22, $z = $ -1.02, $p = 0.31$). Our Bayes Factor analysis suggested that we have no evidence

396   supporting the test hypothesis that culture is a meaningful predictor in participants' choice

397   ($BF = 0.95$).

398   The difference between our result and that of the original study by Kim and Markus

399   (1999) might be related to the use of online format in our study. In the original study,

400   participants were asked to pick a gift pen from five physical pens with different barrel

401   colors. It could be that Asian American participants in the previous study chose the more

402   common color because they wanted the next person to also have room for decision making

403   in the face of resource scarcity, or because they were expressing values or identities

404   influenced by East Asian cultural mandates favoring interpersonal harmony and similarity.

405   Our finding is also consistent with previous work demonstrating that tendencies toward

406   conformity in East Asian samples are linked to reputation management (Yamagishi,

407   Hashimoto, & Schug, 2008); it may be that our online experiment did not establish a

408   sufficient social context to motivate participant concern about reputation, and accordingly

409   failed to motivate reputation management in the form of a conformity preference.

410   **Causal Attribution.**   To test whether Chinese participants tended to make more

411   situational attributions, and US adults more personal attributions, we ran a mixed-effects

412   Poisson regression predicting the number of attributions included in each explanation, with

413   attribution type (situation or person), country (US or CN), and their interaction as fixed

414   effects. We found a main effect of attribution type (Situation attribution: US: $M = 0.65$,

415    $SD = 0.61$; CN: $M = 0.65$, $SD = 0.52$; Person attribution: US: $M = 0.33$, $SD = 0.55$; CN:

416    $M = 0.51$, $SD = 0.52$; $\beta = 0.24$, $SE = 0.10$, $z = 2.37$, $p < 0.05$). Neither the interaction

417   nor the main effect of culture was significant (Both $\beta < 0.3$; $p > 0.05$). The Bayes Factor

418   analysis shows strong support for the null hypothesis ($BF = 0$)

419   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: $M = 0.84$, $SD = 0.17$; β = -1.31, $SE = 0.23$, z = -5.64, $p < 0.01$). The bayes factor is consistent with the frequentist model, showing strong support for the test hypothesis ($BF = 88{,}944.40$). This finding replicates Su (2020) in finding an advantage on Raven's Matrices. In our context, we interpret the relatively high scores we observed as evidence that participants were engaging fully with our tasks.

Discussion

We did not observe cross-cultural differences in the majority of 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 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). Nevertheless, our failure to find robust Western vs. East Asian cultural differences was dispiriting.

We designed Experiment 2 to extend Experiment 1 by recruiting a different sample

and identifying followup or replacement tasks that we hoped would yield a broader set of cross-cultural differences.

Experiment 2

Experiment 2 was designed to follow up on Experiment 1 and further evaluate cross-cultural differences across a battery of tasks. 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).

470 **Methods**

471   **Participants.**   We recruited participants through online crowdsourcing websites.

472 For the US, we used Prolific and applied the following screening criteria: a) U.S.

473 nationality; b) born in the U.S. and c) currently reside in the U.S. For China, we used

474 Naodao (www.naodao.com), a platform designed for conducting online experiments in

475 mainland China. Participants in U.S. were compensated at the rate of \$12.25 per

476 submission and in China ¥35 per submission. We recruited 304 participants from the U.S.

477 and 185 participants from China.

478   10 participants were excluded because they did not meet our demographic inclusion

479 criteria. Following our preregistration (available at https://osf.io/u7mzg), we applied a

480 task-based exclusion procedure in which we excluded a participant's responses in a

481 particular task if they a) showed a response bias in the tasks, b) had missing data on more

482 than 25% of trials or c) failed to meet the inclusion criteria for any specific task as specified

483 in the preregistration.

484   Similar to Experiment 1, we collected demographic information from participants,

485 including subjective socioeconomic status, the state or province the participant grew up in

486 and the one they currently reside in, residential mobility, number of international

487 experiences, education, and undergraduate area of study (STEM or non-STEM). We also

488 administered scales to collect explicit measures of participants' cultural identities and

489 behaviors (Cleveland & Laroche, 2007; Cleveland, Laroche, & Takahashi, 2015).

490   The sample size for each task after exclusion and the descriptive statistics for each

491 demographic question are reported in TABLE X FIXME.

492   **Procedure.**   Similar to Experiment 1, participants completed eight tasks and a

493 brief demographics questionnaire online. The experiment was administered online in

494 English for the US sample and in Mandarin Chinese for the Chinese sample, with the

495 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.

Identical Tasks. 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

549 similarly.

550   We coded difference descriptions to exclude trials in which participants did not

551 identify the change, and checked agreement on focal objects across cultures. We measured

552 how quickly participants identified the difference on trials in which they reported the

553 difference correctly.

554   ***Taxonomic/thematic similarity task.***   Ji et al. (2004) showed that Chinese

555 participants are more likely to categorize items based on thematic similarity, whereas US

556 participants are more likely to categorize items based on taxonomic similarity. In this task,

557 participants were presented with a list of word sets. Each set contained three words: a

558 target word as prompt and two other words as options. The list included test sets and filler

559 sets. In each test set, one option was a taxonomic match (e.g. monkey - elephant) and the

560 other a thematic match (e.g. monkey - banana). In each filler set, the cue item and the

561 options were broadly similar, thematically and taxonomically, making for a more

562 ambiguous decision (e.g. monkey - elephant, tiger). Participants completed a 2AFC in

563 which they chose one match for each cue item.

564   We used a subset of testing materials from Le (2021), including 15 test triads, 15 filler

565 triads, and 2 attention check questions. The order of the triads was randomized between

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

567   ***Semantic Intuition.***   Li, Liu, Chalmers, and Snedeker (2018) found cultural

568 differences in semantic intuitions about ambiguous referents in Chinese and US

569 participants. Chinese participants are more likely to determine the referent of a name

570 based on the description of the speaker (the descriptivist view) whereas U.S. participants

571 are more likely to determine the referent based on the original usage (the causal-historical

572 view). In the study, participants read five separate stories and judged the correctness of

573 statements referring to a character after each story. Two comprehension check questions

574 were included for each story. We followed the original testing procedure closely and used

575 the same materials. We measured participants' semantic intuition as their judgment on the

576 correctness of statements referring to the critical characters.

Results

578   **Ambiguous RMTS.** Our analysis was identical to that in Experiment 1. We did

579 not observe a main effect of country on participants' preference for object vs relational

580 matches (Proportion of relational match: US: $M = 0.41$, $SD = 0.44$; CN: $M = 0.41$, $SD = $

581 0.42; β = -0.01, $SE = 0.48$, z = -0.03, $p = 0.98$). This analysis again supports the null

582 hypothesis ($BF = 0$). As in Experiment 1, we did not find evidence that the differential

583 preferences observed in preschoolers extend to adults. It seems likely that adults in both

584 populations are aware of the mixed evidence for the relational and object solution and that

585 their responses reflect sensitivity to this ambiguous design. [ˆp]

586   [ˆp] Interestingly, our reliability analysis shows that adults expressed this uncertainty

587 only at the population level: individuals tended to be consistent in choosing the same

588 solution type across all four test trials, with ambiguity expressed as disagreement between

589 participants.

590   **Picture Free Description.** US participants were more likely to mention the focal

591 objects than the background objects (First mention: US: $M = 0.94$, $SD = 0.14$; CN: $M = $

592 0.69, $SD = 0.26$). We used the same regression analysis as in Experiment 1 and found a

593 main effect of country ($\beta = 3.09$, $SE = 0.32$, $z = 9.61$, $p < 0.01$). The result of our Bayes

594 Factor analysis is consistent with this model ($BF = 797{,}161.24$). Our results replicate

595 Experiment 1's finding for the first-mention measure with comparable effect size

596 (standardized mean difference; Experiment 1: 1.48[1.24, 1.72]; Experiment 2: 1.57[1.34,

597 1.80]). This suggests that the finding was not caused by the idiosyncrasy of our samples.

598 As a result, we decided not to code the descriptive accounts for Experiment 2 that did not

599 show differences in Experiment 1. In summary, we partially extended Imada et al.'s (2013)

600 findings to Chinese adults.

601   **Change Detection.**   We ran a linear mixed-effects model predicting the reaction

602   time to correctly identify changes in the pictures, with country (U.S. or China) and type of

603   change detected (focal or background) as main effects, as well as their interaction. We did

604   not find evidence for an interaction between culture and type of change detected ($\beta = 0.04$,

605    $SE = 0.03$, $z = 1.40$, $p = 0.16$). Participants in both countries identified changes to the

606   context faster than changes to focal objects (Context Changes: $M = 10{,}101.87$, $SD =$

607    $4{,}257.15$; Focal Object Changes: $M = 10{,}646.54$, $SD = 4{,}816.10$; $\beta = 0.07$, $SE = 0.02$, $t =$

608   3.45, $p < 0.01$). Chinese participants identified both types of change more quickly than

609   U.S. participants (US: $M = 10{,}689.49$, $SD = 4{,}406.73$; CN: $M = 9{,}875.67$, $SD = 4{,}733.57$;

610    $\beta = 0.12$, $SE = 0.05$, $t = 2.27$, $p < 0.05$). The result of our Bayes Factor analysis is

611   consistent with this result ($BF = 0$). We did not extend the findings from Masuda and

612   Nisbett (2006) to the current study.

613   **Symbolic Self-Inflation.**   In Experiment 1, we did not find a significant difference

614   in the degree of symbolic self-inflation between adults in the US and China. Here, we

615   observed a pattern contrary to the prediction: U.S. adults showed less self-inflation than

616   Chinese adults (US: $M = 1.30$, $SD = 0.51$; CN: $M = 1.45$, $SD = 0.65$; $\beta = $ -0.15, $SE =$

617   0.06, $t = $ -2.56, $p < 0.05$). However, the Bayes Factor analysis shows support for null

618   hypothesis ($BF = $). We did not extend the findings of Kitayama et al. (2009) with

619   Japanese participants to our comparison with Chinese participants.

620   **Adult Causal Attribution.**   We ran a mixed-effects linear regression predicting

621   endorsement of each potential cause with country (U.S. or China) and attribution type

622   (personal or situational) as fixed effects, as well as their interaction. We found an

623   interaction in the predicted direction: Chinese participants endorsed situational

624   attributions to a greater extent than their counterparts in the U.S. (Situational ratings:

625   US: $M = 1.71$, $SD = 0.80$; CN: $M = 3.17$, $SD = 0.89$; Personal ratings: US: $M = 3.12$, SD

626    $= 1.10$; CN: $M = 3.14$, $SD = 1.07$; $\beta = $ -1.39, $SE = 0.14$, $t = $ -9.71, $p < 0.01$). The Bayes

627   Factor analysis was consistent with this finding ($BF = 685969773091435776$). This result

628 extends the original findings by Morris and Peng (1994), and suggests that our Experiment

629 1 measure of causal attribution (designed for use with child participants) may not be

630 appropriate for measuring cross-cultural differences in causal attribution among adults.

631   **Taxonomic-Thematic Similarity.**   We used a mixed-effects logistic regression

632 model predicting response (taxonomic or thematic match) with country (US or China) as a

633 fixed effect. There was a significant effect in the predicted direction: participants in the

634 U.S. were more likely to choose taxonomic matches than participants in China (Proportion

635 of taxonomic matches: US: $M = 0.69$; $SD = 0.46$; CN: $M = 0.44$; $SD = 0.50$), on both the

636 main model ($\beta = 2.02$, $SE = 0.89$, $t = 2.27$, $p < 0.05$) and the Bayesian analysis ($BF =$

637 3,766,960.49). This finding replicates those of Ji et al. (2004) and Khuyen, Frank, and

638 Carstensen (2021).

639   **Semantic Intuition.**

640   **Raven's Standard Progressive Matrices.**   We replicated the findings from

641 Experiment 1. Chinese participants scored higher on Raven's Standard Progressive

642 Matrices than U.S. participants (US: $M = 0.49$, $SD = 0.27$; CN: $M = 0.73$, $SD = 0.23$; β

643 $= -1.82$, $SE = 0.25$, $z = -7.39$, $p < 0.01$). The Bayes Factor analysis is consistent with our

644 model findings ($BF = 1,093,530,161.67$).

Exploratory analysis

Mini meta-analysis
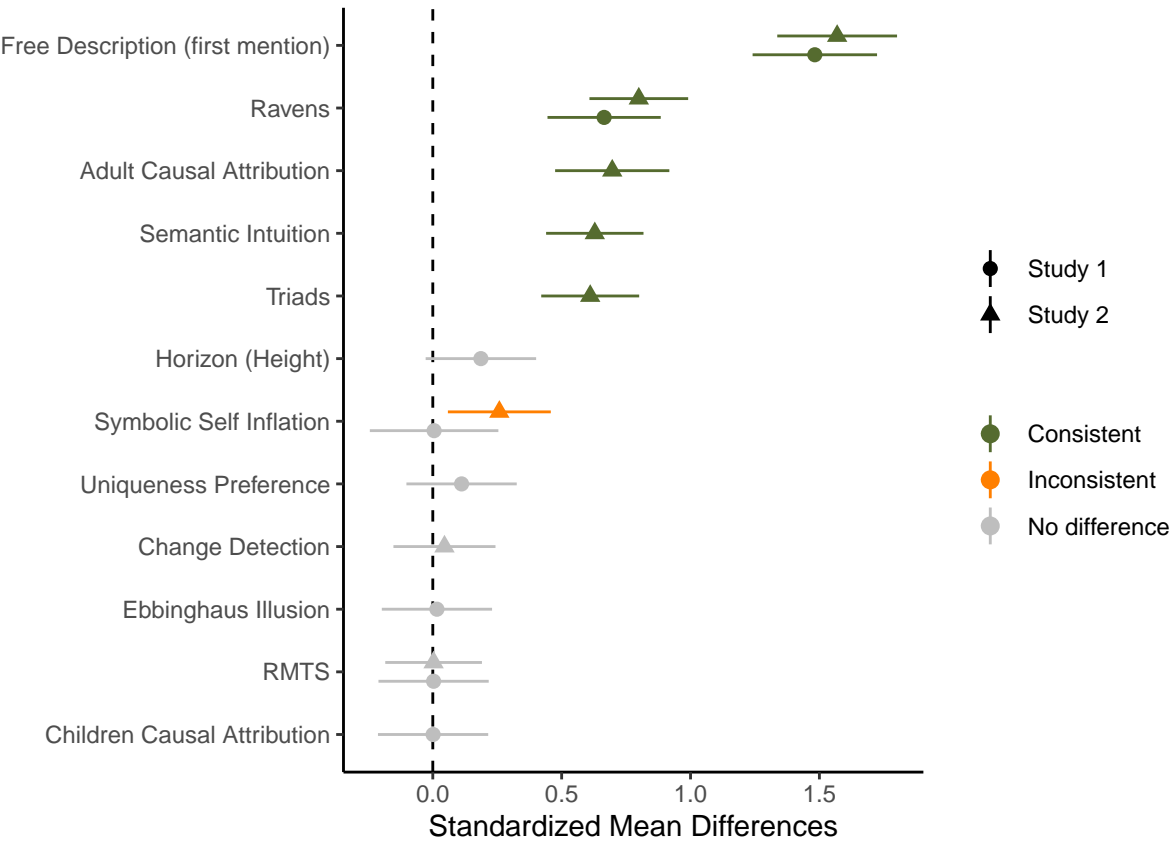
As our first exploratory analysis, we identified the key effect of interest from our

pre-registration (usually a main effect of culture or an interaction of culture, depending on

task) and converted the coefficient into a standardized measure of effect size (standardized

mean difference; SMD) via the method described by (2014). Because there is no "correct"

direction for all of the tasks except Raven's Matrices, we show the absolute value of the

effect size. Figure FIXME shows these effect sizes.

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.

658 In contrast, we observed effect sizes close to zero for our more aesthetic and perceptual

659 tasks (Change Detection, Ebbinghaus Illusion, and Horizon). We also observed little

660 consistent difference in four other tasks (RMTS, Symbolic Self-Inflation, Uniqueness

661 Preference, and Children's Causal Attribution), perhaps for reasons idiosyncratic to each.

662 We return to the broader question of generalization across task types in the General

663 Discussion.

664 **Reliability assessment**

665
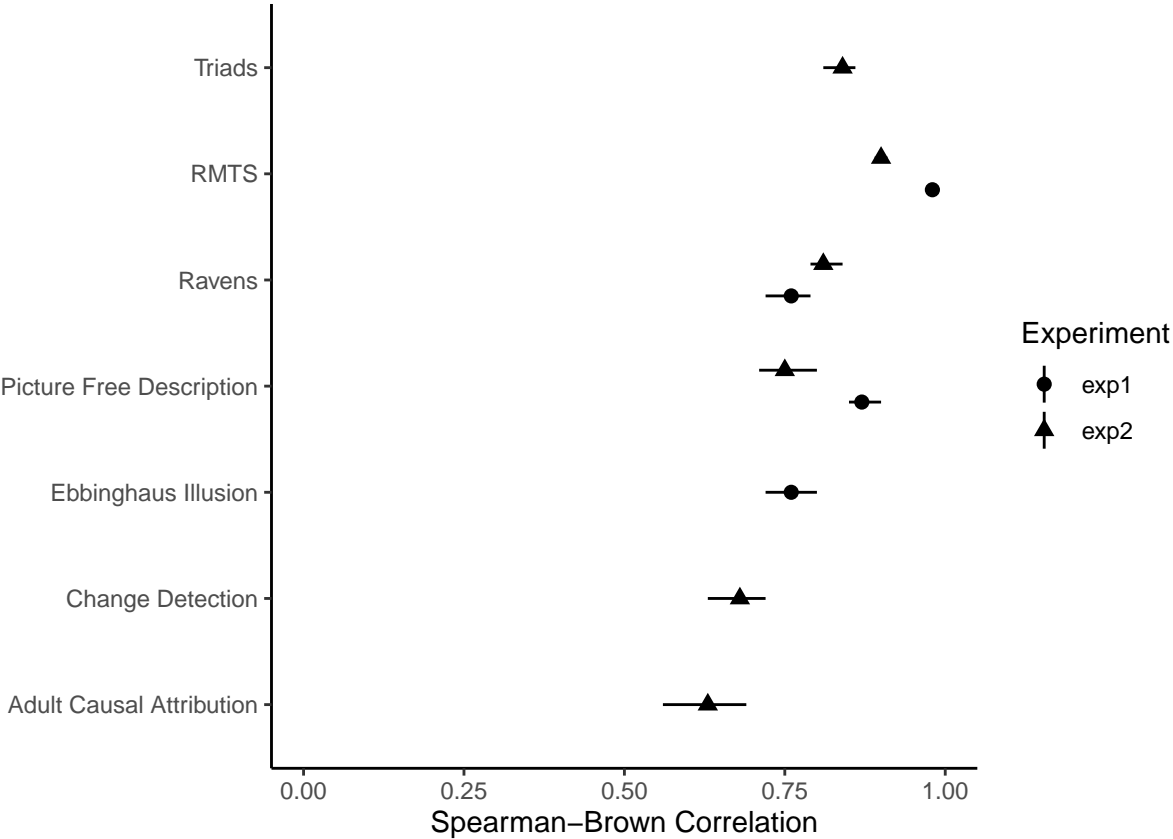
666   One question motivating our work was whether the individual tasks we used were

667 reliable enough – had low enough measurement error – to be used for further investigation

668 of individual differences. The gold standard for the measurement of whether a task yields

669 stable within-person measurements is test-retest reliability (simply because test-retest gives

670 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 thus

dropped the 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 FIXME shows the corrected split-half reliabilities for all tasks in both of our

experiments. Overall the reliabilities are high (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 = -0.19), 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 cross-cultural variation in our data. We treated

the missing data using two approaches: listwise deletion and imputation with means. Two approaches yielded comparable results with the correlation values, so here we reported correlations from listwise deletion.

Correlations between task scores were quite low on average, however, suggesting limited support for this hypothesis. 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. We did not pursue this direction further though we welcome further explorations of this type using our dataset.

Demographic variation and explicit measures of cultural identity

Does demographic or variation in cultural identity predict responding in our tasks? Our approach to this question 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 (specifically, regions that have historically farmed primarily rice or wheat in China, following Talhelm et al. (2014), and culturally distinct regions of the US following Markus and Conner (2014)), and 3) a range of demographic factors, including age, gender identities, residential mobility, number of international experiences, maternal education level, subjective socioeconomic status as measured by MacArthur Ladder (Adler et al., 2000).

₇₂₂ **Task-Global identity relationships.**   We fit models predicting task scores based

₇₂₃ on culture and its interaction with global-local identity for tasks in Experiment 2 (we did

₇₂₄ not collect these scales in Experiment 1). We include the coefficients for all models in

₇₂₅ Supplementary Table FIXME Two of these relationships were statistically significant at .01

₇₂₆ $< \text{p} < .05$ (Adult Causal Attribution: $p = 0.05$; Triads task: $p = 0.04$) but neither of these

₇₂₇ relationships survived Bonferroni correction for multiple comparisons across the family of

₇₂₈ coefficients for the models across all tasks.

₇₂₉ **Task-Geographic origin relationships.**   We next considered whether regions

₇₃₀ within each country were meaningful predictors of task performance. We fit models

₇₃₁ predicting task scores based on the categories of regions the participants reported grew up

₇₃₂ in. For China, provinces were categorized as rice-cultivating regions and wheat-cultivating

₇₃₃ regions based on Talhelm et al. (2014). For U.S., states were categorized based on either

₇₃₄ the coastal locations (West Coast, East Coast and Inland) or broad geographic locations

₇₃₅ (West, South, Northeast, Midwest), following the categorization reported in Carstensen,

₇₃₆ Saponaro, Frank, and Walker (2022). We fit the region models for each task in each study

₇₃₇ separately, and coefficients for all models were included in the Supplementary Table

₇₃₈ FIXME.

₇₃₉ 5 out of 48 models we ran showed a statistically significant relationships between

₇₄₀ regions and task performance. In Study 1, coastal location was a significant predictor for

₇₄₁ Free Description task. Participants who grew up in Inland regions or East Coast were more

₇₄₂ likely to mention the focal object first when describing the pictures (Inland: $p = 0.02$; East

₇₄₃ Coast: $p = 0.05$). In Study 2, both coastal location and broad geographic location were

₇₄₄ significant predictors for Ravens, with participants from Inland and East Coast scoring

₇₄₅ higher than participants from West Coast (Inland: $p = 0.00$; East Coast: $p = 0.05$), and

₇₄₆ participants from Midwest and South scoring higher than participants from the West

₇₄₇ (Midwest: 0.00; South: 0.04). In addition, the two categories also predicted performance in

₇₄₈ Change Detection. East Coast participants took longer to respond than West Coast

749 participants ($p = 0.02$), and Northeastern participants took longer to respond than

750 participants grew up in the West ($p = 0.01$). However, none of these relationship survived

751 Bonferroni correction.

752 **Basic demographic effects.** We fit 192 to see if basic demographic factors could

753 predict task performance. The demographic factors we explored were age, gender

754 identities, residential mobility, number of international experiences, maternal education

755 level, subjective socioeconomic status as measured by MacArthur Ladder (Adler et al.,

756 2000). 24 were statistically significant, but only one model survived Bonferroni correction.

757 Change detection was predicted by age in the U.S. sample, with older participants took

758 longer to respond than younger participants (Adjusted $p < 0.0001$).

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.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 1–28.

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.

Khuyen, L., 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.

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.

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., & 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*.