《心理学报》审稿意见与作者回应

题目：自我–朋友冲突情境下智慧推理的文化差异及其机制

作者：魏新东　汪凤炎

第一轮

审稿人 1 意见：

《“怪异”的所罗门：自我-朋友冲突情境下智慧推理的文化差异及其机制》选题具有理论与现实意义。作者从不同文化背景下的自我类型结构出发，通过两个研究探讨了在不同自我结构背景下的所罗门悖论是一般性还是特殊性现象。结果发现，美国文化中自我类型对所罗门悖论基本无影响，但在中国文化下，高独立自我个体依然会出现所罗门悖论。不过，目前还有如下方面值得进一步商榷：

回应：非常感谢审稿专家的辛勤付出，阅读文稿，对选题意义的肯定与提出非常有价值的修改建议。您所指出的问题对我们进一步完善论文、提高论文质量非常重要，我们将一一予以改进，希望能达到您的要求。修改部分在论文中均已用红色字体标出。

意见 1：问题提出第一段中那么“所罗门们”，应为“所罗门”？

回应：谢谢您指出这一问题，我们修改了相应内容。

意见 2：作者在假设 1 中提到美国人面对朋友冲突时的智慧推理水平显著高于面对自身冲突的智慧推理水平，而中国人在两种情境下的智慧推理水平无显著差异。其中对于美国人的假设已有充分证据支持，美国人可以更好处理面对朋友冲突。但是对于中国人，作者在上文中提到了两种不同的可能，最后为何假设认为可能是没有显著差异？

回应：感谢您的评论。美国人面对朋友冲突时的智慧推理水平显著高于面对自身冲突的智慧推理水平，某种程度上可以看成是有关智慧推理的不等式“朋友冲突下 > 自我冲突下”成立。对于多持互依自我的中国人可能存在两种可能性，一是可能由于“对朋友的过度关心”阻碍了对朋友所遇到冲突的智慧推理，也就是常说的“关心则乱”，这就会造成上述不等式左边“朋友冲突下”智慧推理水平降低，第二个可能是由于“无我”从而提高对自己所遇到冲突的智慧推理，也就是持互依自我的个体更能以冲突方的视角考虑问题，一定程度上摆脱自我中心，这就是造成不等式右边“自我冲突下”智慧推理水平升高。无论是不等式左边的值降低还是右边的值升高，都会造成不等式两边的值相互接近，也就是所提出的两者没有显著差异的假设。我们在文章提出假设一段中对具体论述作了相应补充：“相对于持独立自我的美国人，持互依自我的中国人既可能由于“对朋友的过度关心”阻碍了对朋友所遇到冲突的智慧推理，也有可能由于“无我”从而提高对自己所遇到冲突的智慧推理，这就使得对朋友冲突的与自我冲突的智慧推理水平相接近，因此可以假设：……”。

意见 3：作者在前文提到在西方个体文化一般持有独立自我构想，而中国群体则主要是以互依自我为主。从这个角度来看，假设 1 和 2 是似乎没有区别？

回应：谢谢审稿人的评论。这里之所以提出两个假设主要出于以下考虑：虽然西方个体主义文化下的个体多持独立自我，而集体主义文化下的中国多持互依自我，但依据 Markus 与

Kitayama（2010）的观点，每个个体都同时包括独立自我与互依自我两种类型，而且二者之间彼此互动，依照情境具体表现，不同的情境会启动不同的自我类型；在群体层面上，独立自我与互依自我可以视为一个维度的两端，独立自我高则互依自我低（Twenge et al., 2010），不过在个体层面上独立自我与互依自我则应视为两个独立维度，独立自我高并不意味着互依自我低（Singelis, 1994），并且已有研究表明某些价值取向或行为模式，在文化内的个体间的变异是要大于文化间群体的变异（Fischer & Schwartz, 2011），表明文化内个体之间的差异不能够忽略，假设 1 更多聚焦于文化间的群体差异，而假设 2 更多地探究个体层面上的自我类型差异；另外考虑到各个文化间的交流与融合，尤其对于中国来说，随着改革开放与现代化，中国文化深受西方文化的影响，因此中国文化内个体可能会受多种文化的影响，例如中国香港社会中的“双文化个体”现象（Hong et al., 2000），这部分内容详见文章讨论部分。在假设部分我们突出强调了假设 1 主要针对群体层次与假设 2 主要聚焦个体层次。

参考文献

Fischer, R., & Schwartz, S. H. (2011). Whence differences in value priorities? Individual, cultural, or artifactual sources. *Journal of Cross-Cultural Psychology, 42*, 1127-1144.

Hong, Y., Morris, M. W., Chiu, C., & Benet-Mart ńez, V. (2000). Multicultural minds: A dynamic constructivist approach to culture and cognition. *American Psychologist, 55*(7), 709–20.

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

Twenge, J. M., Abebe, E. M., & Campbell, W. K. (2010). Fitting in or standing out: Trends in American parents' choices for children's names, 1880–2007. *Social Psychological & Personality Science, 1*, 19–25.

意见 4：对于研究 1 和 2 作者选取了类似研究的相关效应量可取的，但是应该补充进行的 prior power 分析所选择的统计检验（t 检验还是 F 检验），以及 Power 和显著性水平。

回应：感谢审稿人，我们在文中对相应部分进行了补充。

意见 5：研究 1 中美国的样本中有亚裔等不同族裔，不同族裔是否会影响研究结果？

回应：在研究过程中我们也曾考虑过是否需要删除亚裔美国人，考虑到虽然白人文化是主流，不过多种族是美国文化的一个主要特征，因此最终决定保留不同族裔的数据。此外，在数据上删除亚裔与否并不影响最终的结果，当删除亚裔数据时，简单效应分析结果显示美国文化中自我冲突下的智慧推理水平（M = 3.57，SD = .66）仍然低于朋友冲突下的智慧推理水平（M = 3.75，SD = .74），$F(1, 541) = 4.75$, $p = .030$, $\eta^2 = .009$（未删除前数据为：美国文化下“自我(M = 3.58, SD = .67)与朋友(M = 3.76, SD = .73)冲突条件下的智慧推理差异显著，$F(1, 567) = 4.62$, $p = .032$, $\eta^2 = .008$”），从数据结果上来看亚裔删除与否并不会影响研究结果。综合这两方面还是保持美国文化中原有的族裔多样性为好。

意见 6：2.1.2 中先汇报实验材料（工具）再汇报研究程序是否会更合适？

回应：感谢审稿专家指出这一问题，先介绍研究工具确实更适合读者的阅读习惯，我们调整了相应的顺序，并将 2.1.2 的标题也做了相应修改，改为“2.1.2 研究工具与程序”。

意见 7：2.1.2 研究程序与工具中“我们使用智慧推理量表（Brienza et al., 2018）评价被试面对该冲突时的智慧推理水平。”改为“本研究使用智慧推理量表（Brienza et al., 2018）来测量被试面对该冲突时的智慧推理水平。”是否更合适？

回应：感谢您的建议，我们将其改为“本研究……”。

意见 8：标点符号的细节问题，表 1 的注中的文字“下同”后缺少句号。2.2.2 对齐法内容中的参考文献引用“Muthén, & Asparouhov, 2018”&前面的逗号不需要。英文摘要 keyword 后缺少了“：”符号。

回应：感谢您的细心阅读，指出标点符号相关问题，我们修改了您点出的问题并对全文标点符号问题进行了详细的校对。

意见 9：2.2.2 对齐法内容中“智慧推理潜均值在各组间的比较来看，与前人关于所罗门悖论的研究结果相同”，使用相同一词来描述是否合适？

回应：感谢审稿专家指出这一问题，前人关于所罗门悖论的研究均是在西方样本中进行，主要采用的是方差分析来分析数据，比较各组之间平均数的差异，而本研究因为是跨文化的多组比较，需要先保证测量工具跨文化的等值性，因此首先采用的是对齐法考察其潜均值差异，使用“相同”确实不合适。这里更准确的表述应为“智慧推理潜均值在各组间的比较来看，在美国被试中，研究结果同样支持前人发现的所罗门悖论现象，即美国被试在朋友冲突组的智慧推理水平显著大于自我冲突组(4 > 3)”。

意见 10：注意到，在研究 1 的结果中美国和中国被试在独立自我和互依自我上没有显著差异，甚至得出中国被试独立自我要显著高于美国被试，尽管作者给出了相应的解释，并且不将两个文化下的自我类型合起来探讨，但是还让读者认为中国被试是以互依自我为主，而西方个体持独立自我的联系。

回应：感谢审稿人提出的问题。由于群体参照效应的存在(Heine et al., 2002)，即个体一般是和周围人相比来评价自己，不同的文化氛围会影响个体对自身的评价。例如自我建构量表中有一题“我乐意在许多方面与众不同”，如果在中国文化下，周围人的表现都相对而言比较追求与大家一致，那么个体只要稍微表现的与众不同一点就会给自己在这一题项上评高分；而在美国，可能由于文化鼓励大家追求独特，那么个体虽然客观而言比较“与众不同”，其在自我评价时也不会给自己打高分。这就可能造成在外显量表的得分上中国人的独立自我水平要高于美国人，但由于群体参照效应的存在并不能就此认为中国人的独立自我水平比美国人高，尤其是在群体层面上。不同于考察文化差异重点看群体层面上，研究 1 对自我类型的分析部分，主要还是聚焦在个体层面，重点探究中国文化下不同程度的独立自我与互依自我个体是否会对所罗门悖论产生影响。我们在文章的相应部分补充与突出了“个体层面上的自我类型”，而不是群体层面上的中国人或美国人。

参考文献

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference–group problem. *Journal of Personality and Social Psychology, 82*, 903–918.

意见 11：表 4 标题为“中国文化下自我类型对不同冲突类型的智慧推理影响”，表中的自我类型的变量只有独立自我？作者在结果中的文字还描述了美国文化下的结果，表格却只呈现了中国文化下的结果？

回应：谢谢审稿人提出的问题。由于量表所测得的独立自我与互依自我得分相关显著，我们采用与前人相同的处理方式，即将独立自我为因变量，互依自我为自变量作回归分析所得残差保存为表格中的“独立自我”指标，这一指标的具体含义为独立自我中互依自我无法解释的部分。前人在探究追寻美德的动机对所罗门悖论的影响中也是如此操作（Huynh et al.,

2017），由于“追寻美德的动机”与“追寻享乐的动机”两者之间相关显著，因此研究者为了得到较为“纯粹的”追寻美德的动机，也是将追寻享乐的动机预测追寻美德的动机所得的残差作为其指标。

虽然在表格中只呈现中国文化下的独立自我指标，不过在文中也同时呈现了中国文化下的互依自我情况（这里互依自我的指标也是通过回归分析所得的残差，方式与独立自我相同）、美国文化下独立自我与互依自我情况，不过它们与冲突类型的交互项对智慧推理的预测均不显著，在后面对这些差异进行了讨论，为了节省篇幅与方便读者看到最显著的结果，所以只将自我类型与冲突类型交互项显著的制成表格，将交互项不显著的，如美国文化下的结果放在文中呈现不将其制作成表格。

参考文献

Huynh, A. C., Oakes, H., Shay, G. R., & McGregor, I. (2017). The wisdom in virtue: Pursuit of virtue predicts wise reasoning about personal conflicts. *Psychological Science, 28*, 1848–1856.

意见 12：3.2 结果中，“朋友冲突下的智慧推理显著高于自我冲突组”，请补充相关统计结果。
回应：感谢您指出这一问题。3.2 结果中补充了相关统计指标：“朋友冲突下的智慧推理(M = 3.94, SD = .57)显著高于自我冲突组(M = 3.83, SD = .65)”。

审稿人 2 意见：

这是一篇理论基础扎实、非常有文化心理学意义的优秀论文；但是目前的统计方法有误，实验方法需要改进，结果没有支持研究者的假设和结论。建议研究者补做实验，通过线下单独招募被试进行实验，在实验中加入“操纵检验”这一经典的要素，继续深入研究这一意义重大的主题。

回应：感谢审稿人对文章意义的肯定，给予进一步完善与修改论文的机会，对文章细致的阅读与辛勤付出，与提出非常具有启发性的问题。感谢审稿人提出“操纵检验”这一问题，我们按照审稿人的要求，添加了研究 3，补充了“操纵检验”，具体见回应意见与正文。不过，对于审稿专家指出的统计方法有误这一点，主要是交互效应不显著能否进行简单效应检验问题，我们想在回应意见 1 中尝试与审稿人进行讨论，希望我们可以论证清楚该问题，如果审稿人对于这一问题还有疑问，欢迎进一步的交流。另外，审稿人给出了非常有价值的评论，我们在没有删除与改动原文的情况下对审稿人的意见进行段落分类，希望我们归类整理出的下列意见段落能符合审稿人的原意。

意见 1：的确，“在交互不显著的情况下简单效应也可以显著(Tybout et al., 2001; Umesh et al., 1996)”；但是，交互作用不显著时能否直接进行简单效应检验？参见一篇被引约三千次的文章：Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594–604. doi: 10.1037/0003-066X.54.8.594 Add to Citavi project by DOI 第 599 页原文摘抄： “One of the most prevalent strategies psychologists use to handle multiplicity is to follow an ANOVA with pair-wise multiple-comparison tests. This approach is usually wrong for several reasons. First, pairwise methods such as Tukey's honestly significant difference procedure were designed to control a familywise error rate based on the sample size and number of comparisons. Preceding them with an omnibus F test in a stage wise testing procedure defeats this design, making it unnecessarily conservative. Second, researchers rarely need to compare all possible means to understand their results or assess their theory; by setting

their sights large, they sacrifice their power to see small. Third, the lattice of all possible pairs is a straight- jacket; forcing themselves to wear it often restricts researchers to uninteresting hypotheses and induces them to ignore more fruitful ones.” 此外，国内的一些心理统计学教材也明确指出，之所以要用方差分析而不是直接进行多重检验，就是要排除假阳性。当然，心理学工作者有时为了方便自己发表论文，可能会提出自己的想法，想要让自己不显著的交互作用也能够公开发表。但其实，本篇研究的结果就充分地说明了这种做法的不当之处：1. 研究一中，“在中国被试中，自我(M = 3.81, SD = .68)与朋友(M = 3.89, SD = .65)冲突条件下的智慧推理无显著差异，F(1, 567) = 1.11, p = .293, η^2 = .002；而在美国被试中，自我(M = 3.58, SD = .67)与朋友(M = 3.76, SD = .73)冲突条件下的智慧推理差异显著，F(1, 567) = 4.62, p = .032, η^2 = .008。”请注意，研究者做了两次差异检验（F 检验与 t 检验在两水平的对比条件下本质相同，可以互换），得到两个 p 值：0.293 与 0.032，事实上，如果非要直接进行多重比较，阿尔法水平必须校正（a 阈值改为 0.05/2=0.025），所以，两次差异检验的结果其实都不显著——而这是因为在这两次差异检验中样本量远不如合并数据进行检验时，故无法探测到如此小的效应量。类似的，在研究二中，研究者可以自己再重新审视下所谓“显著”的 p 值（0.043）及其对应的效果量（η^2 = .006）：“在启动独立自我条件下，朋友冲突组(*M* = 3.94, *SD* = .54)的智慧推理水平显著高于自我冲突组(*M* = 3.80, *SD* = .66)，*F*(1, 702) = 4.111, *p* = .043, η^2 = .006，而在启动互依自我条件下，朋友冲突组(M = 3.93, SD = .60)与自我冲突组(M = 3.85, SD = .64)的智慧推理无显著差异，F(1, 702) = 1.62, p = .204, η^2 = .002”。

回应：感谢审稿人分享的文献与提出有关交互作用显著是否可以进行简单效应检验的问题。

在讨论这一主要统计问题之前，我们需要澄清一下，此处的统计处理并不是为了让结果显著，或为了使用更多的样本来探测比较小的效应量，有针对性地将两个国家的样本合在一起做方差分析。实际上如果把两个文化下的样本分开做统计，某种程度上所得结果对我们更有利，如对于美国文化中的结果（共 280 个被试），控制性别、年龄等条件下，自我(*M* = 3.58, *SD* = .67)与朋友(*M* = 3.76, *SD* = .73)冲突条件下的智慧推理差异显著，*F*(1, 274) = 4.45, *p* = .036, η^2 = .016；中国文化下（共 295 个被试）自我(*M* = 3.81, *SD* = .68)与朋友(*M* = 3.89, *SD* = .65)冲突条件下的智慧推理无显著差异，*F*(1, 289) = 1.32, *p* = .252（按照学报投稿要求，将未出现显著结果的效应量删除）。也就是说如果我们把研究 1 拆分为两个单独的研究，即研究 1a 单考察美国样本中的情况，纯粹重复前人在美国文化下的研究，考察美国被试 280 个样本差异，其实效应量.008 增大到了.016；而研究 1b 考察中国文化，自我与朋友的差异并不显著，其实这也支持我们对于美国文化下自我与他人差异显著，而中国文化下自我与他人差异不显著的假设。在这里我们之所以将两个文化下的样本综合在一起，主要是出于两个方面的考虑：一是研究 1 在设计时就是主要考察两个群体的差异情况，也就是将文化作为一个自变量；二是为了研究的严谨，我们并没有采取在一个文化中收集完一批数据，分析出相关结果后，再去另一个文化中收集（这样一来其实就变成了两个独立研究），而是在两个文化下收集完数据后做统一处理。

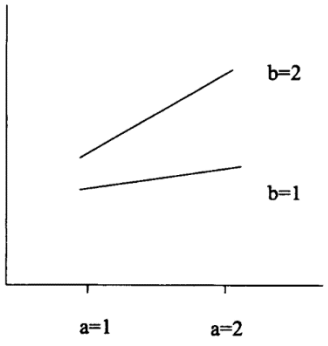
研究 2 的结果与研究 1 类似，即单纯只考虑一个启动条件下样本，如启动独立条件下的 345 个样本，在控制性别、年龄等条件下，自我(*M* = 3.80, *SD* = .66)与朋友(*M* = 3.94, *SD* = .54)的冲突条件下的智慧推理同样显著，*F*(1, 339) = 4.21, *p* = .041, η^2 = .012，此时的效应量同样大于两个样本合在一起做简单效应检验的情况（η^2 = .006），而在启动互依自我 365 个样本中，自我(*M* = 3.85, *SD* = .64)与朋友(*M* = 3.93, *SD* = .60)的冲突条件下的智慧推理同样显著，*F*(1, 359) = 1.60, *p* = .207。下面主要探讨交互效应不显著能否进行简单效应检验问题。

审稿人认为交互效应不显著情况下，不能进行简单效应检验，给出的理由是之所以进行方差分析而不直接进行多重检验主要的原因是为了排除假阳性。我们认为审稿人在这里可能是将简单效应检验和多重检验这两个概念等同了。两者虽然都属于事后检验（post hoc test,

Marascuilo & Levin, 1970），不过对于多重检验来说，它主要针对的是某个因素有多个水平的情况（两个以上水平），而简单效应检验则主要关注的是两个因素或多个因素之间的情况。国内心理学统计教材中所给出的要求一般为：对于多水平的，三个及三个以上的平均数进行差异研究，不能直接进行多重比较，需要先满足方差分析的 F 检验显著（张厚粲,徐建平, 2008, p.289）；对于两因素的，如果两个因素间交互作用显著则必须进行简单效应检验（张厚粲, 徐建平, 2008, p.391），因为交互效应显著表明主效应是一个没有考虑其他因素的一种检验，主效应的结果可能受到“歪曲”。在我们能够查阅到的国内主要统计心理学教材中，均没有提及“当交互作用不显著时，则不能进行简单效应检验”的相关论述（不过也必须要指出，查阅到的国内教材中也没有提及“当交互作用不显著时，也可以进行简单效应检验”的正面证据）。不过我们发现在一本英文统计教材中确实比较明确的指出（某位研究心理统计老师推荐）：如果只关注简单效应情况的话，交互作用其实压根就不需要检验，英文原文为：“A case can be made that the interaction need not be tested at all if a set of comparisons such as these do in fact turn out as expected(Hayes, 2005, p.447)”，原书作者还认为交互作用之所以必要，主要在于它可以提供比简单效应检验更多的额外信息，而不是简单效应检验的前提。具体内容以及有关这本教材的作者介绍，与这句话的上下文，可以参见这条回应下面所列出的参考文献。我们将这本教材作为一个重要参考文献补充在文中。

对于某个因素有两个以上水平间的比较，即多个平均数间的比较，方差分析通过 F 检验讨论组间变异在总变异中的作用，借以对两组以上的平均数进行差异检验，得到一个整体性的检验结果，而通过类似 t 检验的方法两两进行比较，做统计结论时就会增大犯 α 错误的概率，所以诚如审稿人所言，对于两组以上平均数的比较确实不能直接做多重比较，而要做方差检验，方差检验不显著情况下，多重比较结果并没有意义（张厚粲,徐建平, 2008, p.289）。审稿人所分享的文献涉及的也主要是多重比较，并没有涉及“interaction”（交互）以及“simple effects”（简单效应）的论述。

对于交互效应与简单效应检验而言：以最简单的情况，也就是本研究 2（A_1，A_2）×2（B_1，B_2）情况为例，交互效应表示的是 A 因素对因变量的影响在 B 因素两个水平上是否相同，如果将在 B_1 水平上的差值，即 M（B_1A_1）与 M（B_1A_2）差值记为 M_1，将 B_2 水平上的差值，即 M(B_2A_1)与 M(B_2A_2)的差值记为 M_2，那么交互效应显著就是指 M_1 与 M_2 的差异显著；而简单效应关注的只是其中的一个 A 因素对因变量的影响在某个水平，如 B1 水平的差异，从数值上来说也就是 M_1 与“0”是否差异显著，并不是如交互效应那样比较 M_1 与 M_2。如果用图来表示上述情况将更为直观，b = 1 直线的斜率表示的是在 b_1 水平上 A 因素的简单效应，而 b = 2 直线的斜率表示的是在 b_2 水平上 A 因素的简单效应，简单效应显著表示的是直线的斜率与水平线的斜率，即斜率为“0”差异显著，而交互效应显著则表示的是 b =1 直线的斜率与 b=2 直线的斜率差异显著（Marascuilo & Levin, 1970; Tybout et al., 2001; Umesh et al., 1996）。



总的来看，从逻辑条件来说，多因素交互作用显著与简单效应检验的关系相当于是充分不必要的关系，即如果交互作用显著则必须要进行简单效应检验，但是进行简单效应检验并不意味着一定要满足交互作用显著这一前提条件。而多水平的方差分析与事后比较则是充分且必要关系，即方差分析的结果显著，就需要事后检验确定具体那两个水平之间存在差异；进行多水平间的事后检验也一定要首先满足方差分析显著这一前提。

回到本研究的问题，我们发现的是交互作用不显著，也就是美国文化下自我——朋友的差异与中国文化下自我——朋友的差异比较无显著差异，不过美国文化下的简单效应显著，就是说明在美国文化下自我——朋友之间的差异与“0”相比存在显著差异，而中国文化下的简单效应不显著，说明在中国文化下自我——朋友之间的差异与“0”相比不存在显著差异。用两个式子比较形象的表达就是：0 ＜ 美国（自我与朋友差异）和 0 ≈ 中国（自我与朋友差异） ≈ 美国（自我与朋友差异），这里的约等于号“≈”表示差异不显著，而且“≈”不具备传递性，即 a ≈ b ≈ c，并不能说 a ≈ c。也就是说中国文化下自我与朋友差异介于“不显著”与“显著”之间，进一步探究发现，即在小节“2.2.4 自我类型不同对所罗门悖论影响”，发现中国文化下独立自我与冲突类型的交互作用（调节本质上就是交互）显著，在独立自我高条件下自我与朋友差异显著，而在独立自我低条件下自我与朋友差异不显著，此时用式子比较形象的表达就是，在中国文化下：0 ≈ 独立自我低（自我与朋友差异）< 独立自我高（自我与朋友差异）。

对于出现 0 ＜ 美国（自我与朋友差异）和 0 ≈ 中国（自我与朋友差异） ≈ 美国（自我与朋友差异）这一结果，它其实是符合假设 1，即认为美国文化下自我——朋友差异显著，即 0 ＜ 美国（自我与朋友差异），而中国文化下差异不显著，即 0 ≈ 中国（自我与朋友差异）；虽然假设中没有涉及交互作用的情况，但是出现中国（自我与朋友差异） ≈ 美国（自我与朋友差异）这一结果，需要对此作出进一步解释，我们在讨论中，从全球化的西方文化对中国文化的影响与现代化经济发展带来的个体主义上升两个角度，来解释中国文化的自我与朋友差异的这一“折中状态”，详见讨论部分内容。考虑到文化变迁，可能交互作用不显著，简单效应显著这一模式更符合相关理论，因为对于中国文化而言，文化变迁向着个体主义不断增强的方向也就是与美国文化相接近（即交互作用不显著，中国（自我与朋友差异） ≈ 美国（自我与朋友差异）），但同时也保持着自身的特色（即与美国文化中的简单效应检验显著不同 0 ＜ 美国（自我与朋友差异），中国文化中的简单效应检验也不显著 0 ≈ 中国（自我与朋友差异））

我们后续开展的一个相关研究中就涉及多重检验的问题：在研究中我们假设在美国文化下，个体面对自我与朋友冲突时的智慧推理有显著差异，不过个体在面对朋友与陌生人冲突时的智慧推理无显著差异，即自我 < 朋友 ≈ 陌生人；而在中国文化下，个体面对自我与朋友冲突时的智慧推理无显著差异，不过个体在面对朋友与陌生人冲突时的智慧推理差异显著，即自我 ≈ 朋友 < 陌生人。我们计划开展两个研究，研究 1a 在美国文化下进行，研究 1b 在中国文化下进行，因为分为两个研究其实就根本不涉及文化（中美）与冲突类型（自我、朋友、陌生人）的交互作用是否显著这一问题，我们只是在不同文化下进行冲突类型差异的检验，某种程度上也就相当于直接做的简单效应检验，不过此时的简单效应检验并不能直接按照假设所预期的那样直接进行自我与朋友间的 t 或 F 检验，朋友与陌生人间的 t 或 F 检验，也就是多重比较检验，必须要在三个水平整体 F 检验显著基础上才能进行下一步的事后检验。目前这一研究计划正在预注册程序中。

以下讨论交互作用与简单效应分析具体关系的文献：

Hayes, A. F. (2005). *Statistical methods for communication science*. New York, NY: Routledge.

作者 Andrew F. Hayes 在谷歌学术中的被引次数为 121083（截止 2020/11/18），统计分

析中常用的 process 程序就是由其开发。在这本英文教材的第 447 页中，作者讨论了交互作用与简单效应之间的关系问题，附英文原文以及上下文：[“One can sensibly ask why the test for interaction is even necessary. Berger predicted that the population trend information should reduce perceived vulnerability among men but not among women. **Why not just compare the two simple effects with a series of t tests? What information is gained by testing the interaction first? A case can be made that the interaction need not be tested at all if a set of comparisons such as these do in fact turn out as expected.** But you need to recognize that what is left out from this strategy is an explicit test of whether the difference between these differences is statistically different. The interaction tests the significance of the difference between the differences.”]作者在这里指出了之所以做交互作用检验，是因为它比简单效应分析提供额外信息，简单效应只是分析差异是否显著，而交互作用则是差异和差异之间的差异是否显著（也就是上面引述英文的最后一句：“The interaction tests the significance of the difference between the differences”），可见交互作用并不是简单效应的前提，它的必要性体现在比简单效应提供更多的信息。

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal, 7*, 397–421. doi:10.3102/00028312007003397（EDUCATION & EDUCATIONAL RESEARCH 5/263 Q1）

Tybout, A., Sternthal, B., Keppel, G., Verducci, J., Meyers-Levy, J., Barnes, J., … Maxwell, S. (2001). Analysis of variance. *Journal of Consumer Psychology, 10*(1–2), 5–35. doi:10.1207/S15327663JCP1001&2_03（PSYCHOLOGY, APPLIED 19/84 Q1），这个参考文献包括众多作者针对统计方法的讨论其中第一篇题目就是 Tybout 与 Sternthal 讨论的“Can I test for simple effects in the presence of an insignificant interaction?”(“我可以在交互作用不显著情况下进行简单效应分析吗？”)

Umesh, U. N., Peterson, R. A., McCann-Nelson, M., & Vaidyanathan, R. (1996). Type IV error in marketing research: The investigation of ANOVA interactions. *Journal of the Academy of Marketing Science, 24*(1), 17–26. doi:10.1007/BF02893934（BUSINESS 8/152 Q1）

张厚粲, 徐建平. (2008). *现代心理与教育统计学*. 北京师范大学出版社.

意见 2：其实，综合来看，这两个研究比较一致的结果是：自我类型的效应不显著，而中国人和美国人都存在所罗门效应（虽然很微弱）。具体分析如下：研究一：“文化的主效应显著，F(1, 567) = 6.01, p = .015, η2 = .010，中国被试的智慧推理水平显著高于美国被试的智慧推理水平；冲突类型的主效应显著，F(1, 567) = 5.17, p = .023, η^2 = .009”“在独立自我分量表总分上，美国被试(M = 29.87, SD = 5.71)显著低于中国被试(M = 30.91, SD = 5.17)，t(575) = 2.30, p = .022, 95% CI = [.153, 1.933]；在互依自我分量表总分上美国被试与中国被试无显著差异，”结合起来说：美国被试独立自我得分更低，根据研究者在前言中的推论，美国被试应该有更高的智慧推理，但是研究结果与之刚好相反。研究二：“自我类型的主效应不显著，F(1, 702) = .36, p = .549，η^2 = .001；冲突类型的主效应显著，F(1, 702) = 5.506, p = .019, η^2 = .008，” 因此，这两个研究的结果与研究前言中的“因为不同于独立自我将自我视为一个实体，具有明确的边界，互依自我的边界有弹性，强调人与人之间的联系，某种程度上可视为“无我”（即无小己）的一种形态，这可能会使持互依自我个体在面对涉及自身冲突时更容易摆脱自我中心式的偏见”有矛盾。

回应：感谢审稿人提出的问题。审稿人通过研究 1 与研究 2 两个结果指出自我类型对智慧推理影响的结果，与我们前面的假设推论相矛盾。我们这个推论的主要内容是互依自我比独立自我在处理涉及自身的人际冲突时智慧推理的水平更高。而研究 1 中在自评量表中出现中国人的独立自我水平高于美国人独立自我水平，但是却是美国人的智慧推理水平更低，此为矛盾 1；在研究 2 中自我类型的主效应不显著，也就是说并未发现启动互依自我的个体智慧推理水平要高于启动独立自我的个体，此为矛盾 2，希望我们这里的概括符合审稿人的原意。

对于矛盾 1，在研究 1 中关于为何在自我构念的自评量表中，出现与一般认知不符的情况，即美国人独立自我反而低于中国人独立自我得分，前人对于这一现象的解释是由于群体参照效应的存在，即个体一般是和周围人相比来评价自己，不同的文化氛围会影响个体对自身的评价（Heine et al., 2002）。例如自我建构量表中有一题“我乐意在许多方面与众不同”，如果在中国集体主义文化下，周围人的表现都相对而言比较追求与大家一致，那么个体只要稍微表现的与众不同一点就会给自己在这一题项上评高分；而在美国，可能由于文化鼓励大家追求独特，周围人普遍都比较“独特”，那么个体即使比较“与众不同”，在自我评价时可能也不会给自己打高分。这就如同一个身高 170cm 的人在平均身高只有 165cm 群体中，会自评认为自身的身高较高，而身高 175cm 的人在平均身高 180cm 群体中，会自评认为自身的身高较矮，但从绝对值上来说 175 是要大于 170 的。因此不能依据这里的美国人独立自我自评得分比中国人独立自我自评得分低，就得出美国人的独立自我水平低于中国人的结论，所以矛盾 1 的前提条件并不成立。

另外探究不同文化下自我类型对智慧推理的影响，主要关注的是个体层面上的自我类型，并不是两个群体层面间的相互比较，我们在假设 2 中增加了聚焦个体层面。在群体层面上一般将互依自我与独立自我视为一个维度的两端，独立自我高则互依自我低（Twenge et al., 2010），但依据 Markus 与 Kitayama（2010）的观点，每个个体都同时包括独立自我与互依自我两种类型，而且二者之间彼此互动，依照情境具体表现，因此在个体层面上则不能将互依自我与独立自我视为一个维度的两端，而应视为两个独立维度，独立自我高并不意味着互依自我低（Singelis, 1994），有可能出现在同一个体身上独立自我与互依自我都很高的情况。基于上述理由我们除了要分开探究两个文化中，还要把两种自我类型也分开探究。在中国文化下，互依自我程度可以显著预测智慧推理水平：$\beta = .18$, 95% CI = [.047, .199], $t(288) = 3.17$, $p = .002$。在美国文化下互依自我程度也可以显著预测智慧推理水平：$\beta = .33$, 95% CI = [.157, .315], $t(273) = 5.87$, $p < .001$ 这某种程度上部分支持我们的推论，即互依自我程度越高，智慧推理水平越高。不过需要指出的是在研究 1 中互依自我高并不意味着独立自我低，所以这些发现不能说完全支持原本的推论，即“互依自我比独立自我在处理涉及自身的人际冲突时智慧推理的水平更高”。

研究 2 的结果是自我类型的主效应不显著，我们采用启动法，确实是将独立自我与互依自我视为同一维度，也确实未发现启动互依自我的个体智慧推理水平显著高于启动独立自我个体。不过这些结果虽然未支持“互依自我比独立自我在处理涉及自身的人际冲突时智慧推理的水平更高”这一推论，但并不意味结果就与这一推论相矛盾，因为数据呈现出来的趋势是与这一推论吻合的，考虑到我们补充的研究 3 发现自我类型的主效应虽然不显著，但 p 值某种程度上达到了“边缘显著”的状态（在正文中还是采用无显著差异说法，并未使用边缘显著一说），$F(1, 524) = 3.53$, $p = .061$，这一趋势更为明显。

总结来看，研究 1 实验数据是部分支持该推论，研究 2 或研究 3 的数据没有支持该推论（但并不矛盾），造成这一结果可能有多种原因，例如在研究 1 中使用自评量表测量自我类型，将独立自我与互依自我视为两个不同的维度，而研究 2 和 3 使用启动个体不同的自我类型，则是将独立自我与互依自我视为同一维度的两端。我们在讨论中指出，“一方面可能学界普遍支持的观点，即互依自我与独立自我视为两个不同的维度更为合理；另一方面，独立

自我可以很好地代表西方人或受西方文化影响较深的中国，不过互依自我可能只是学者以西方独立自我为参照所构造出的对应概念(Fiske, 2002; 黄光国, 2012)，在中国本土学者观点中，中国人的互依自我至少存在依然以个体为中心的差序格局式自我与以个体与他人关系为中心的关系型自我两种形态(费孝通, 2011; 李抗, 汪凤炎, 2019; 许烺光, 1989)，而差序格式互依自我并不能满足智慧推理中蕴含的对个体去自我中心式的“无我”要求(Grossmann, 2017)。”

参考文献

Fischer, R., & Schwartz, S. H. (2011). Whence differences in value priorities? Individual, cultural, or artifactual sources. *Journal of Cross-Cultural Psychology, 42*, 1127-1144.

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

Twenge, J. M., Abebe, E. M., & Campbell, W. K. (2010). Fitting in or standing out: Trends in American parents' choices for children's names, 1880–2007. *Social Psychological & Personality Science, 1*, 19–25.

意见 3：而且，研究结果不支持本论文的主要结论：怪异的“所罗门”。那么，这两个研究的结果是否表明“普适”的“所罗门”呢？本人也持怀疑态度。

回应：感谢审稿人的评论。审稿人认为本文的研究结果不支持文章主要结论，即“怪异的所罗门”。审稿人之所以持这一观点的主要原因还是在于认为我们上述的统计方法有误，也就是交互作用不显著的情况不能做简单效应分析，这一问题我们已经在回应 1 中回答，论证了这一处理方法是可行的。另外，我们除了做方差分析外，还使用了跨文化研究中常用的在测量等值基础上的“对齐法”，两个统计方法的结果其实都在某种程度上支持我们的假设。

另外，研究 1 中国文化下自我与朋友差异不显著，而研究 2 中国文化下自我与朋友差异显著（冲突类型主效应显著），这两个结果的不同可能来自于样本量，研究 1 中国文化样本量为 295（与前人研究以及本研究的美国文化中样本量相当），而研究 2 样本量为 710，也就是说中国文化下当样本量足够大的时候也会发现所罗门悖论的现象。不过结合研究 1 中，中国文化下独立自我与冲突类型的交互作用显著，研究 2 中也是启动独立自我条件下才发现自我与朋友差异显著的情况，可以说，所罗门可能“怪异”在独立自我高的情况，并不能笼统地说是美国文化与中国文化的区别。因为随着全球化与现代化的进程，中国文化中个体主义与独立自我出现上升的趋势（黄梓航等, 2018; 蔡华俭等, 2020）。

由此，我们意识到原来的结论可能并不准确，给审稿人造成了误解，原结论为：“(1)相对于面对自己的人际冲突，美国文化下的个体在面对朋友的人际冲突时的智慧推理水平更高，即存在所罗门悖论的现象，而中国文化下的个体在面对自我与朋友的人际冲突时的智慧推理水平无显著差异。(2)中国文化下独立自我程度高的个体同样会表现出所罗门悖论现象，而独立自我程度低的个体未出现所罗门悖论现象。”我们根据上述结果将原结论中的“中国文化下的个体在面对自我与朋友的人际冲突时的智慧推理水平无显著差异。”删除，并将两者合并为一个结论：“相对于面对自己的人际冲突，美国文化下的个体在面对朋友的人际冲突时的智慧推理水平更高，即存在所罗门悖论的现象；中国文化下独立自我程度高的个体同样会表现出所罗门悖论现象，而独立自我程度低的个体未出现所罗门悖论现象，在中国文化下，通过启动不同自我类型也发现相同模式，可见所罗门悖论现象并不具备普适性。”这样的结论相对于之前的可能更为严谨。

参考文献

蔡华俭, 黄梓航, 林莉, 张明杨, 王潇欧, 朱慧珺, 谢怡萍, 杨盈, 杨紫嫣, 敬一鸣. (2020). 半个世纪来中国人的心理与行为变化——心理学视野下的实证研究. *心理科学进展, 28*(10), 1599–1618.

黄梓航, 敬一鸣, 喻丰, 古若雷, 周欣悦, 张建新, 蔡华俭. (2018). 个人主义上升,集体主义式微?——全球文化变迁与民众心理变化. *心理科学进展, 26*, 2068–2080.

意见 4：因为：1．研究一中的被试取样可能无法代表典型的美国人和中国人（“最终腾讯问卷收集到610份数据，MTurk平台收集到594份数据……有效样本分别为中国295人(男144，女 151；平均年龄 23.22±4.34)与美国 282 人”）。所使用的外显自我类型量表可能受社会赞许的影响；此外，研究者本人也有睿智的洞见：“外显自评量表施测于不同文化群体上存在群体参照效应(Heine et al., 2002)”。而研究二中使用的启动法是否成功地操纵了被试的自我类型呢？这需要“操纵检验”，比如通过自我类型量表的测量来说明两组人在短暂的写作任务后确实存在自我类型上的显著差异——尤其是在网络实验的情境下，存在大量的不认真做实验的被试，操纵检验是必不可少的。

回应：感谢审稿人对我们研究设想的支持并提出具有价值的意见与建议。

对于网上收集的样本不具备代表性问题，我们在本研究的局限性中补充了该观点，不过其实与线下招募的被试（一般都为大学生、研究生群体）相比，线上招募样本的代表性还是相对而言更强(Buhrmester et al., 2011)。因此在补充的包含操纵检验的研究 3 中，我们认为招募线上被试比较合理一点。审稿人所说的线上被试不认真情况确实存在，例如我们研究 1 中共收集到 610 份中国被试数据，和 594 份美国被试数据，而经过测谎题与两道自评回答认真程度的题项筛选后分别还剩有效数据 295 份与 282 份，有效率分别为 48.36%和 47.47%，可见认真回答的被试不到一半；而研究 2 共收集到 1124 份数据，经过同样的标准筛选后还剩有效被试 710 份，有效率 63.17%。这里有效率虽然大幅提升但依然存在大量不认真作答的情况，研究 2 与研究 1 中有效率不同的一个重要原因可能在于研究 2 的问卷调查的长度与题项远低于研究 1。可以看出在线上调查的过程中，设置测谎题与一些筛选标准非常重要，操纵检验也同样重要。在补充的研究 3 中我们将测谎题安插在智慧推理量表中，与被试自评认真作答的筛选题分开，使其更为隐蔽，研究 3 共收集数据 1108 份，通过三个筛选标准后还剩 537 份，有效率 48.47%，研究 3 因为添加了情绪测量与自尊量表，题项与长度也高于研究 2。另外，之所以采用线上调查的一个主要原因也是 2017 年发表在 psychological science 上的探讨所罗门悖论的论文(Huynh et al., 2017)也是采用线上调查的方式，不过作者将冲突类型作为被试内变量，我们认为这可能会使得被试猜测到实验研究目的，因此将冲突类型改为被试间变量。

对于“操纵检验”：为了检验研究 2 是否操纵了个体的自我类型，我们补充了研究 3，在研究 2 启动研究的基础上添加了一个检验题项，即自我包含他人量表(Inclusion of the Other in the Self Scale, IOS; Aron et al. 1991; Aron et al. 1992)，审稿人建议的自评式自我构念量表虽然也可以检验被试是否被启动了相应的自我类型，不过我们担心过多的文字表述可能会对后续造成干扰，因此我们选择这一以图形为主的测量方式来检验启动效果。IOS 量表包含 7 组双圆组成的图形，用来评估个体在多大程度上感到自我与他人之间的相互联系，或在多大程度上感受到他人是自我的一部分(Aron et al. 1992)。本研究要求被试从重合度不同的七个双圆图形中选择一个分别代表他们和好友的关系。得分越高，表示两个圆圈的重合度越高，表示与好友的关系越亲密。研究结果表明本研究的启动是有效的，启动互依自我条件下的个体认为自己与好友的重合度（M = 4.67, SD = 1.43）显著高于启动独立自我条件下（M = 4.38, SD = 1.58）的个体，t = -2.26, p = .024, 95% CI [-.549, -.038]，具体见文中补充的研究 3。另外对于研究 3 经过与一位研究者讨论与在其建议下，我们参照以往文献进一步控制了被试在重构冲突情境中的情绪（Grossmann et al., 2019）与个体的自尊情况(Huynh et al., 2017)。希望这

样所得的结果更加严谨，具体见正文部分。研究 3 同样发现两者的交互作用不显著，$F(1, 524) = .29$, $p = .590$，不过简单效应检验，对不同自我类型下的自我与朋友冲突中智慧推理进行比较，发现在启动独立自我条件下，朋友冲突组($M = 3.97$, $SD = .66$)的智慧推理水平显著高于自我冲突组($M = 3.82$, $SD = .80$)，$F(1, 524) = 4.67$, $p = .031$, $\eta^2 = .009$，而在启动互依自我条件下，朋友冲突组($M = 4.05$, $SD = .67$)与自我冲突组($M = 4.00$, $SD = .67$)的智慧推理无显著差异，$F(1, 524) = 2.30$, $p = .130$。同样如果只考虑不同条件下的样本的话，如启动独立自我条件下的 253 个样本，自我与朋友间的差异同样显著，$F(1, 245) = 4.08$, $p = .044$, $\eta^2 = .016$，这里的效应量也增大；而启动互依自我条件下的 281 个样本，自我与朋友间的差异不显著，$F(1, 273) = 2.90$, $p = .090$。

参考文献

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self-scale and the structure of interpersonal closeness. *Journal of Personality & Social Psychology, 63*, 596–612.

Aron, A., Aron, E. N., Tudor, M., & Nelson, G. (1991). Close relationships as including other in the self. *Journal of Personality & Social Psychology, 60*, 241–253.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5.

Grossmann, I., Oakes, H., & Santos, H. C. (2019). Wise reasoning benefits from emodiversity, irrespective of emotional intensity. *Journal of Experimental Psychology: General, 148*(5), 805-823.

Huynh, A. C., Oakes, H., Shay, G. R., & McGregor, I. (2017). The wisdom in virtue: Pursuit of virtue predicts wise reasoning about personal conflicts. *Psychological Science, 28*, 1848–1856.

第二轮

审稿人 1 意见：

作者较好的回答了审稿人第一轮提出的问题，较第一轮的有了一定的改进。但论文现在仍然存在以下问题，需要作者回复。

回应：非常感谢审稿专家再一次细致的审阅文稿，很抱歉上一次回应过程中一些问题未能清晰阐明，对于本次意见我们将一一予以改进，希望能够达到您的要求，修改部分在论文中均已用蓝色字体标出。

意见 1：对于作者在第一轮的意见 2 中回应中有以下两点建议和疑问。1）作者从智慧推理的不等式“朋友冲突下 > 自我冲突下”背景下回应了审稿人关于“中国人在两种情境下的智慧推理水平无显著差异”的疑问，并且在手稿中补充了相应内容，考虑到如果缺乏智慧不等式的背景，读者可能会对于中国人在两种情境下的智慧推理水平无显著差异存在疑问。建议作者在相应地方对假设的依据进行更清晰的论述。2）作者这部分的回应是否有相关参考文献为依据？

回应：感谢审稿人认真审查本部分内容！之前的回应确实未能从读者角度考虑问题。我们在文中结合以往研究补充了这一不等式描述：“当个体面对涉及自身的冲突或困境时，可能由于更容易沉浸在自身的观点或情绪之中，忽视他人观点，偏向于坚定自己所认定的立场，从而会抑制其智慧推理能力(Huynh et al., 2017; Kross & Grossmann, 2012; McGregor et al., 2001)，而当面对朋友等他人冲突或困境时，则能够采纳不同观点同时避免如对自身缺点认

识不足的认知偏差，从而会促进智慧推理(Grossmann & Kross, 2014; Pronin et al., 2008)，形成个体在朋友冲突下智慧推理大于自我冲突下智慧推理的不等式。”

在提出假设部分进一步详细阐明中国人在两种情境下智慧推理水平无差异的依据，并补充了有关“自我抽离”(李天然等, 2015)、“自我中心”(Self-centeredness)与“无我”(selflessness)的相关文献(Dambrun & Ricard, 2011)：“以上述朋友冲突下智慧推理水平大于自我冲突下智慧推理水平的不等式为基础（即朋友冲突 > 自我冲突），相对于持独立自我的美国人，持互依自我的中国人既可能由于与朋友的心理距离较近，更容易卷入朋友冲突之中，无法抽离自身，一定程度上阻碍了对朋友所遇到冲突的智慧推理(李天然 等, 2015)，使得不等式左边的朋友冲突下智慧推理变低，也有可能由于“无我”从而提高对自己所遇到冲突的智慧推理，使得不等式右边的自我冲突下智慧推理变高(魏新东, 汪凤炎, 2020; Dambrun & Ricard, 2011)，不等式左边变低而右边变高从而使得不等式两边接近，即对朋友冲突的与自我冲突的智慧推理水平相接近”。

参考文献

李天然, 李晶, 俞国良. (2015). 自我抽离:一种适应性的自我反省视角. *心理科学进展, 23*(6), 1052–1060.

Dambrun, M., & Ricard, M. (2011). Self-centeredness and selflessness: A theory of self-based psychological functioning and its consequences for happiness. *Review of General Psychology, 15*, 138–157.

Grossmann, I., & Kross, E. (2014). Exploring “Solomon’s paradox”: Self-distancing eliminates the self-other asymmetry in wise reasoning about close relations in younger and older adults. *Psychological Science, 25*, 1571–1580.

Huynh, A. C., Oakes, H., Shay, G. R., & McGregor, I. (2017). The wisdom in virtue: Pursuit of virtue predicts wise reasoning about personal conflicts. *Psychological Science, 28*, 1848–1856.

Kross, E., & Grossmann, I. (2012). Boosting wisdom: Distance from the self-enhances wise reasoning, attitudes, and behavior. *Journal of Experimental Psychology: General, 141*, 43–48.

McGregor, I., Zanna, M. P., Holmes, J. G., & Spencer, S. J. (2001). Compensatory conviction in the face of personal uncertainty: Going to extremes and being oneself. *Journal of Personality and Social Psychology, 80*, 472–488.

Pronin, E., Olivola, C. Y., & Kennedy, K. A. (2008). Doing unto future selves as you would do unto others: Psychological distance and decision making. *Personality and Social Psychology Bulletin, 34*, 224–236.

意见 2：对于作者对于第一轮意见 4 的回应中有以下疑问：作者是采取的 A Priori 还是 Post Hoc 或其它的 power 分析类型？请明确说明。此外，F 检验中包括 ANOVA, MANOVA 和 Linear Multiple Regression 等，请作者明确写明是具体统计检验。最后，建议作者参考相关文献完善这部分的内容。

回应：谢谢审稿人指出这一问题！研究主要采用的是 G*Power 中 A Priori，目前主要有两篇与所罗门悖论直接相关的文献(Grossmann & Kross, 2014; Huynh et al., 2017)，本文在研究方法与研究工具上更接近 2017 年文献，因此我们将研究 1 与 2 统一改为采用 Huynh 等（2017）研究的效应量作为估计样本量的指标。参照《学报》最新发表的一篇文献(王慧媛 等, 2021)，将这一部分内容改为：

采用 G*Power 软件计算研究样本量(Faul et al., 2007)，采用 2(文化：中国与美国) × 2(冲突类型：自我冲突与朋友冲突)被试间设计，并使用方差分析进行统计检验，依据以往类似研究效应量(f^2 = .117, Huynh et al., 2017)，Ⅰ类错误的概率 α err prob 为.05, 检验效能 Power(1−β err prob)为.80 计算样本量，每个文化下需要 288 人(Huynh et al., 2017)。

鉴于研究 2 与研究 3 都采用相同的 2×2 被试间设计，并且考虑网上作答的有效率与所设置的筛选标准，研究 2 与 3 会相对多招募被试，所以此部分内容在研究 2 与 3 中略去。

对于最新补充的研究 4，考虑到线下被试的同质性较高，对于冲突情境也改为标准化的假设情境（泄密情境），在疫情背景下招募被试比较困难，因此我们每个实验条件招募 50 人，共 200 人，灵敏度功效分析(sensitivity power analysis; 假设 $\alpha = 0.05$, power = 0.80), 结果发现，根据 200 样本量，我们有能力检测到的最小效应量为 $f = 0.20$，符合小效应的标准。

参考文献：

王慧媛, 陈艾睿, 张明. (2021). 意义关联的注意定向效应：基于空间位置的抑制和捕获. *心理学报, 53*, 113-127.

Brienza, J. P., Kung, F., Santos, H. C., Bobocel, D. R, & Grossmann, I. (2018). Wisdom, bias, and balance: Toward a process–sensitive measurement of wisdom-related cognition. *Journal of Personality and Social Psychology, 115*, 1093–1126.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

意见 3：作者结果中表 1 的(n = 577)，表 2 和表 3 的等值检验被试数量(n = 575)与表 1 不一致？

回应：感谢审稿人指出这点！我们整理数据后发现，研究 1 通过测谎题等筛选标准后，在中美两国分别招募被试中国 295 人(男 144，女 151)与美国 282 人(男 155，女 125，未报告 2 人)，初稿中被试信息部分已有此说明。由于有两个性别是缺失，所有我们将这两个样本删除，最终是 575 人，因此表 1 使用的也是 575。表 1 的 $n = 577$ 可能是后面添加过程中未将两个未报告性别被试排除，我们在文中修改了这一信息。

意见 4：在“2.2.4 自我类型不同对所罗门悖论影响”中以及研究 3 的 4.2 结果中 t 检验显著的结果请补充相关效应量大小。

回应：感谢审稿人指出这点，研究 3 的 Cohen's d 为.20（通过效应量在线计算软件得到，https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php）。我们在文中补充了这一信息。

意见 5：“2.2.4 自我类型不同对所罗门悖论影响”的简单斜率检验发现在中国文化下不同水平的独立自我在自我冲突和朋友冲突上表现出了交互作用，作者认为该结果支持了假设 2。作者原文假设 2 是“(2)聚焦于个体层次，独立自我个体在面对朋友冲突时的智慧推理水平显著高于面对自身冲突时的，而互依自我的个体在两种情境下的智慧推理水平无显著差异。”上述结果仅是中国文化下得到的交互，假设中似乎没有区分，而美国文化下独立自我与冲突类型交互作用不显著。

回应：谢谢审稿人的意见！考虑到美国文化下独立自我与冲突类型交互作用不显著，这里“支持假设 2”的论述确实不准确，我们在文中将其改为“部分支持假设 2”。不过，在作出假设 2 时确实是从“个体层面”考量，即认为这一假设既适合中国文化下个体，也适合美国文化下个体，这里美国文化下未显著，在假设中对其进一步区分可能并不合适。我们希望能在讨论中对这一现象进行探讨，结合文化心理学有关“怪异”样本的相关问题，在讨论中指出：“研究 1 中发现与中国样本不同，美国样本中独立自我与冲突类型的交互项无法显著预测智慧推理，而这可能由于美国“怪异”样本总体上的独立自我程度较高(Henrich et al., 2010)，或因为作为文化输出国的美国，独立自我文化某种程度上影响了中国人自我，而中国文化中

的自我类型对美国人自我的影响甚微，总之若只探究美国的“怪异”样本，自我类型影响所罗门悖论的机制可能无法在经验数据中呈现。”

参考文献：

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral & Brain Sciences, 33*, 61–83.

意见 6：研究 2 的 3.2 结果当中“其中潜均值有显著差异的为 2 组与 1 组，某种程度上该结果支持假设 2。”某种程度上该结果支持假设 2，这样的描述是否合适，请作者斟酌。

回应：感谢审稿人指出这一问题！这一表述应改为支持假设 2，考虑到研究 1 采用“对齐法”比较各组的潜均值差异主要是由于在两个不同的文化背景下，而且一个采用的英文量表，一个是中文量表，有可能在文化、语言上造成理解的差异，很有必要进行测量等值性检验（梁觉, 周帆, 2010）。研究 2、研究 3 以及研究 4 均是在中国文化背景下采用中文量表，在研究 1 等值性检验的基础上，出于对文章的篇幅考量，将研究 2 对齐法检验潜均值的部分删除，只保留平均值的方差检验可能比较合适。附研究 2 的对齐法检验结果：

采用对齐法比较四组条件下的智慧推理潜均值差异，表 S1 表明四组条件下的所有截距和因子载荷都近似等值，可以进行各组间智慧推理潜均值比较，表 S2 表明四组智慧推理潜均值大小排序为启动独立自我的朋友冲突组(2)、启动互依自我的朋友冲突组(4)、启动互依自我的自我冲突组(3)、启动独立自我的自我冲突组(1)，其中潜均值有显著差异的为 2 组与 1 组，该结果支持假设 2。

表 S1 不同自我类型与冲突条件下截距与因子载荷近似测量等值性(n = 710)

智慧推理五维度	截距等值组别	因子载荷等值组别
对方视角	1, 2, 3, 4	1, 2, 3, 4
变化/多种结果	1, 2, 3, 4	1, 2, 3, 4
局限/理智的谦逊	1, 2, 3, 4	1, 2, 3, 4
局外人视角	1, 2, 3, 4	1, 2, 3, 4
妥协/解决方案	1, 2, 3, 4	1, 2, 3, 4

注：1 代表启动独立自我的自我冲突组；2 代表启动独立自我的朋友冲突组；3 代表启动互依自我的自我冲突组；4 代表启动互依自我的朋友冲突组，下同。

表 S2 各组智慧推理潜均值比较(n = 710)

排序	组别代码	潜均值	潜均值显著小于该组
1	2	.221	1
2	4	.196	
3	3	.030	
4	1	.000	

参考文献：

梁觉, 周帆. (2010). 跨文化研究方法的回顾及展望. *心理学报, 42*(01), 41-47.

意见 7：作者根据第一轮的意见 6 将 2.1.2 的标题也做了相应修改，改为“2.1.2 研究工具与程序”。研究 2 和 3 的这个部分的小标题描述以及具体介绍流程最好与研究 1 一致。

回应：谢谢审稿人指出这点，我们在文中调整了这一顺序。

审稿人 2 意见：

非常感谢作者们的积极回应！文章有所提高，但本人仍感到内心矛盾：一方面，本人认可你们的理论与假设，从情感上也希望学生能够早日发表这篇论文；但是从理智上，本人认为该论文目前仍不适合在国内最权威的杂志《心理学报》上发表，否则对于国内的许多学生都会是误导，甚至影响你们这个领域今后的发展。本人的担忧，不仅是研究者的统计检验方法，也是其行为实验的开展方式：通过网络平台迅速收集大量样本，最终以不到一半的有效被试来检验假设的做法。这两个弊病可能是你们最终在 3 个研究中都没有发现显著交互作用的原因。而你们的主要理论假设，即怪异的“所罗门”——文化对冲突类型效应的调节作用，如你们所说，本质上就是一个交互作用。

回应：感谢审稿人对我们的督促，非常用心的给出意见，同时比较快的回应给予我们比较充足的时间收集线下被试。我们认同与接受审稿人提出的需要线下被试进一步完善论文，使研究更严谨的观点，补充以大学生为主的线下被试的研究 4：“研究 2 与研究 3 均采用在线调查的方式，并且冲突情境都来自于个体回忆最近发生在自身或朋友身上，虽然大都为日常生活冲突，但冲突对象、冲突的严重程度等也各不相同，因此研究 4 在研究 2 与 3 基础上，采用线下招募在校生被试的方式，并使用该线下群体较为常见的标准化冲突情境（朋友泄密情境），来进一步检验研究假设。”不过考虑到线下被试的同质性较高，对于冲突情境也改为标准化的假设情境（泄密情境），在疫情背景下招募比较困难，因此我们每个实验条件招募 50 人，共 200 人，女生 148，男生 52 名，平均年龄 20.17±1.56，18-27 岁。灵敏度功效分析 (sensitivity power analysis; 假设 $\alpha = 0.05$, power = 0.80), 结果发现，根据 200 样本量，我们有能力检测到的最小效应量为 $f = 0.20$，符合小效应的标准。此外，确实如审稿人所预想发现了交互作用（见文章研究 4 结果部分），$F(1, 193) = 6.66$, $p = .011$, $\eta^2 = .033$，进一步简单效应检验，对不同自我类型下的自我与朋友冲突中智慧推理进行比较，在启动独立自我条件下，朋友冲突组($M = 4.08$, $SD = .50$)的智慧推理水平显著高于自我冲突组($M = 3.63$, $SD = .58$)，$F(1, 193) = 14.37$, $p < .001$, $\eta^2 = .069$，而在启动互依自我条件下，朋友冲突组($M = 3.91$, $SD = .68$)与自我冲突组($M = 3.89$, $SD = .62$)的智慧推理无显著差异，$F(1, 193) = .02$, $p = .886$，该结果支持假设 2。

不过，我们想指出研究 1 主要理论假设并没有说“文化对冲突类型效应的调节作用”，而是为“美国人面对朋友冲突时的智慧推理水平显著高于面对自身冲突时的智慧推理水平，而中国人在两种情境下的智慧推理水平无显著差异”，从事前假设而言，这一假设的发生概率为两个独立事件的乘积.05*.95 = .0475 满足小概率事件（这里按审稿人在下文中认为一个显著一个不显著的概率为 0.095，其实包含了两种情况，即“美国显著、中国不显著”与“美国不显著、中国显著”，因此需要将.095÷2 = .0475），而且这一假设并不一定表明文化与冲突类型的交互作用显著。由于不存在交互作用，我们在后文讨论与结论中也一直在避免下“中国文化不存在所罗门悖论”这一结论，而是相对而言更确切的认为“中国文化下独立自我程度高的个体同样会表现出所罗门悖论现象，而独立自我程度低的个体未出现所罗门悖论现象”。“所罗门”的“怪异”主要体现在只存在于独立自我高的人身上。同时我们在讨论部分也从文化变迁的角度对出现这种情况进行一定程度的解释。

其实研究 1 数据呈现的交互作用不显著，但是满足我们假设的数据结果，如审稿人所言可能存在线上收集数据方法学上的问题，但是在统计方法无误的基础上（即交互作用不显著也可以进行简单效应分析），连续两次研究都出现相同模式，其实从理论上讨论这一现象的成因可能更有价值，我们认为这一现象比较支持目前中国文化不断变迁的观点，即随着现代化与全球化时代到来，中国文化中的个体主义不断增强，越来越接近美国等西方文化（交互作用不显著），但同时保持着自身文化特色的一种“折中状态”，详见文章讨论部分。

意见 1：检验交互作用的意义何在？不应该把 Hayes, A. F. 著作中的单句话（“A case can be made that the interaction need not be tested at all if a set of comparisons such as these do in fact turn out as expected”）单独拿出来，而应放回原文联系上下文去理解：“One can sensibly ask why the test for interaction is even necessary. Berger predicted that the population trend information should reduce perceived vulnerability among men but not among women. Why not just compare the two simple effects with a series of t tests? What information is gained by testing the interaction first? （首先进行交互作用能得到什么信息呢？）A case can be made that the interaction need not be tested at all if a set of comparisons such as these do in fact turn out as expected. But （但是）you need to recognize that what is left out from this strategy is an explicit test of whether the difference between these differences is statistically different. The interaction tests the significance of the difference between the differences.” 正如你们所说：“交互效应表示的是 A 因素对因变量的影响在 B 因素两个水平上是否相同，如果将在 B1 水平上的差值，即 M（B1A1）与 M（B1A2）差值记为 M1，将 B2 水平上的差值，即 M(B2A1)与 M(B2A2)的差值记为 M2，那么交互效应显著就是指 M1 与 M2 的差异显著” 所以，交互作用不显著，就说明了所罗门效应在中美之间、不同自我类型之间无显著差异；也就是说你们三个网络研究的结果没有支持你们的假设（即：所罗门效应是“怪异”的，存在文化差异和自我类型差异）的确，有些学者像你们一样关注的是交互作用，于是他们不检验交互作用，而是分别计算出差值 M1 和 M2（仅限于 A 因素是被试内因素时），然后只做一次 t 检验或 F 检验，比较 M1 和 M2 是否有显著差异；这样避免了 multiple-comparison tests 所造成的假阳性。否则，按照你们的做法，任何一个因素（比如年龄、性别），都可以成为所谓的“调节变量”。比如，年龄与文化间的交互作用不显著，但是如果你把 18-21 岁的所有被试分成四组，而阿尔法水平又不做相应地调节，大概率（概率是 0.1854875）就会发现，在某个年龄上，文化差异显著，而在某个年龄上，差异又不显著。但是，这样的结果不能成为交互作用显著的证据；类似的，中国人中发现差异显著，而美国人中发现差异不显著，也不能成为某个效应存在文化差异的证据。具体数学推导证明如下：假设 a 水平为 0.05,在效应不存在的情况下，2 次检验同时发现差异显著的概率是 0.05 的二次方 0.0025；同时发现差异不显著的概率是 0.9025；所以，有的检验显著、有的不显著的概率是 1-0.0025-0.9025=0.095，远大于心理学上约定的 0.05 的小概率事件的概率；而在效应存在的情况下，假设研究范式有 0.8 的统计效力，那么有检验显著、有检验不显著的概率是 1- 0.64- 0.04= 0.32！大概三分之一的概率，会得到一次 p 值显著，一次不显著。更关键的是：你们的研究范式是否有 0.8 的统计效力？以研究 3 为例，所罗门的总体效果量是 η2 = 0.01，假设你们的四组人数是平均分布，每组有 537/4=134.25!!——假设是 135 人每组，根据 Gpower3.1 的测算，简单效应检验的统计效力只有 0.37！也就是说，按照你们研究 3 的范式，只有 37%的概率可以检验到这个微弱的效应！那么，你们得到一个效应显著，一个效应不显著的概率是多大呢？1-0.1369-0.3969=0.4662。也就是说，在所罗门效应普遍存在的情况下，仍然有将近一半的概率会得到这样的结果：一次 p 值显著，一次不显著。 “What information is gained by testing the interaction first? （首先进行交互作用能得到什么信息呢？）A case can be made that the interaction need not be tested at all if a set of comparisons such as these do in fact turn out as expected. But（但是） you need to recognize that what is left out from this strategy is an explicit test of whether the difference between these differences is statistically different. The interaction tests the significance of the difference between the differences.” 综上所述，不能以一个 p 值显著、一个 p 值不显著来说明某个效应受到某一因素（比如：文化）的调节。

回应：感谢审稿人的评论！首先想澄清一下引用 Hayes 著作中的观点并不是想否认检验交互作用的意义，只想为在交互作用不显著的情况下也可以进行简单效应检验提供支撑论据。

我们认同审稿人所说的美国 *p* 值显著、中国 *p* 值不显著不能说明两个自变量之间存在交互作用的观点，不过正如前面所言，我们两个主要假设其实一直都没有说文化或自我类型与冲突类型存在交互作用，而是“(1)美国人面对朋友冲突时的智慧推理水平显著高于面对自身冲突时的智慧推理水平，而中国人在两种情境下的智慧推理水平无显著差异；(2)聚焦于个体层次，独立自我个体在面对朋友冲突时的智慧推理水平显著高于面对自身冲突时的，而互依自我的个体在两种情境下的智慧推理水平无显著差异。”

审稿人认为我们的假设不满足小概率事件，给出的论证是，从假设效应不存在的事前角度而言，“有的检验显著、有的不显著的概率是 1-0.0025-0.9025=0.095，远大于心理学上约定的 0.05 的小概率事件的概率”，这里存在一点小瑕疵，一个检验显著、一个检验不显著的概率确实为.095，但是这一概率事件包含了两种情况，（a）中国文化下显著、美国文化下不显著与（b）中国文化下不显著、美国文化下显著，而我们的假设只是其中的一种情况，所以这一准确的概率为.095/2 = .0475 < .05，因此按审稿人逻辑的话，也是符合小概率事件要求，不存在假阳性问题。

不过诚如审稿人所言，当考虑到所罗门效应是普遍存在的情况，事后的统计效力和效应量等指标后，修正审稿人给出的 0.4662 概率也有.4662/2 = .2331，显然是不满足小概率事件要求，这或许也是审稿人需要我们展开线下，增加研究效应量的一个重要原因。统计效力与效应量等主要是在心理学可重复性危机的背景下“登上舞台”（胡传鹏等, 2016），因此我们认为最好方式就是进行重复研究，当然审稿人所说线下实验也是重复的一种，不过从重复研究的角度考虑，还是以接近原研究的方法为好，这也是我们第一轮补充研究 3 依然采用线上的一个原因。如果把三个研究放一起综合考察，因为三个研究的模式是相同的（概率上也相差不大，可以都按第三个研究为标准），三个概率为.2331 的事件发生的概率为 0.2331*0.2331*0.2331 = 0.012，如果不把第一个研究算上（因为它主要采用测量自我类型方式，而不是启动自我类型），那么两个事件的概率也为 0.2331*0.2331 = 0.054，某种程度上从事后效应量角度而言，“启动独立自我显著，启动互依自我不显著”虽然在线上研究中的效应量不高，但是这一情况连续发生两到三次，并且呈现的模式都相同而且效应量都不高的话，有理由认为，这一结果至少在线上研究中有很好的鲁棒性（robustness）。

参考文献：

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题：从危机到契机. 心理科学进展, 24(9), 1504-1518.

意见 2：网络研究是否是普适的？以上推导也说明了：效应量和统计效力非常关键。一个检验不显著，也可能是因为统计效力不够，无法检测到微弱的效应。特别是，你们的三个网络研究的效应量都很低。诚然，网络研究因其便利、高效而得到国内外许多学者的偏爱；特别是，它能够低成本地收集不同人群的数据（比如公司白领），提高研究的外推效度；但是你们的三个网络研究的有效样本率普遍低于 50%。所以，它可能并不适用于研究中国网民的智慧，因其特别需要认真投入和脑力。我也下载、查阅了你们设定样本量时参考的文献（你们说：“之所以采用线上调查的一个主要原因也是 2017 年发表在 psychological science 上的探讨所罗门悖论的论文(Huynh et al., 2017)也是采用线上调查的方式”），Huynh et al 文中的原话是：“Based on effect sizes from prior research on Solomon’s Paradox (Grossmann & Kross, 2014) a G*power analysis suggested a total sample size of approximately 275 participants....Participants were brought into the lab as part of a larger study on personality and

motivation.” 他们设定 275 人所参考的效应量是依据前人的实验室研究，而他们所做的是传统的实验室研究，排除了环境干扰，被试更加认真投入，所以才有较大的效应量。他们只在 study2 中使用了 Mtrurk，而且，有效样本率是 356/393=90%.此外，我也查阅了他们参考的最早的所罗门效应研究报告(Grossmann & Kross, 2014)，其中： Study1，线下，自我 VS 他人 η^2 = .25 Study2，线下，重复研究一结果得到的效应量 η^2 = .12； Study3，线上，重复研究一结果得到的效应量 η^2 = .05； 可见，国外学者均是将实验室研究当作基础，网络研究当作扩展。而且，网络研究的效应量远低于实验室研究。实验室研究样本同质性高，这是局限，也可以是优点——可以排除额外变量（如职业在智慧上的效应）的干扰。

回应：感谢审稿人的细致，当前直接探究所罗门悖论的主要有两篇文献，即审稿人提及的 2014 年与 2017 年，开始我们理解两篇文章的线上研究，认为它们是在完善前面研究基础上做的，相对于前面线下研究是比较成熟与相对完善的设计，因此前面回应时会认为“前人也是采用线上收集”。不过经过审稿人点出后，我们意识到线下研究确实是为基础，从线下到线上的研究设计并不是简单的递进关系。我们在研究 4 开展的线下实验室研究中发现的效应量为：交互作用效应量为 η^2 = .03；简单效应分析中启动独立自我条件下自我——朋友差异的效应量为 η^2 = .07。这里需要指出 Grossmann 与 Kross（2014）年中使用的标准化冲突情境为伴侣（朋友伴侣）出轨，而我们这里考虑到中国大学生被试对此情境可能相对并不熟悉，因此改为比较熟悉的朋友或同学泄密情境。

另外这里可能需要解释一下为何 Huynh et al(2017)中同样采用 Mturk 收集数据，最后的有效率高达 90%，远远高于我们这里的有效率。主要因为一般而言在 Mturk 中收集，网站自身会有一个类似的筛选标准，也就是你可以选择把这一任务投放给以往有效作答率超过 80%或 90%的人群，这一选择本身其实没有问题，因为在问卷中设置测谎题实质上也是在筛选有效被试，这个设置相当于网站提前帮助研究者筛选了一下，一般情况下也会选择一个指标（有的研究甚至会选择 Mturk 职业答题者以提高有效率），而由于我们做文化比较研究，相应的中文网站并无此功能，因此在 Mturk 中我们也就没对此进行设置，为与中国网站的设置保持一致，可能造成了样本有效率的不同。

意见 3：三个研究都是采用被试间设计，需要报告每组人数。有些组人数少，可能是使某些简单效应不显著的原因。此外，研究 2 和 3 中统计结果是否报告错误？研究 2 有效样本 710，为什么独立我与互依我启动条件下都是“F(1, 702)”？各组人数具体是多少？类似的，研究 3 也要注意。

回应：谢谢审稿人指出这一问题，我们在文中对相关人数均作了补充。其中研究 1 中国文化下自我冲突 153 人，朋友冲突 142 人；美国文化下自我冲突 138 人，朋友冲突 142 人。研究 2 启动独立自我条件下自我冲突 180 人，朋友冲突 165 人；启动互依自我条件下自我冲突 171 人，朋友冲突 194 人。研究 3 启动独立自我条件下自我冲突 126 人，朋友冲突 129 人；启动互依自我条件下自我冲突 141 人，朋友冲突 141 人。“F (1, 702)”中的 *df* 是 2×2 方差分析中的误差 *df*，也就是两个条件下一起考量的总误差 *df*，所以都是 702。若分开比较的话研究 2 独立自我条件下误差 *df* 为 339，互依自我条件下误差 *df* 为 359，当然分开之后的结果也与假设相符：启动独立自我下自我与朋友冲突情境比较，$F(1, 339) = 4.21$, $p = .041$, $\eta^2 = .012$；启动互依自我的比较，$F(1, 359) = 1.60$, $p = .207$。

意见 4：本研究中的行为实验特别需要线下实施。你们的线上研究中，流失的那一半被试可能在独立我水平上显著高于留下来的那一半被试；而在线下招募中，保证这一部分被试流失率不超过 10%的话，你们可能发现与之前研究不同的结果。总之，本人衷心希望你们早日发表这篇论文。这篇论文的发表，不仅对你们领域，对于国内心理学专业的其它学生也将是

很好的案例，以此说明网络研究的效应量可能远比实验室研究的小，不能单纯依靠网络平台去进行所有的行为实验。西方有其特有的文化背景（劳动力价格高、被试招募困难）而选择网络研究，而我们也有我们的文化背景要求我们不盲目跟风。希望 90 后、00 后的学生仍能继承中国实验心理学前辈们勤劳诚恳、精益求精的传统。

回应：感谢审稿人非常用心的意见！之前我们一直把重点放在关注研究 1 与 2 线上结果是否可重复，对线下研究的重要性认识不足，采用审稿人宝贵意见后确实出现了预期结果。对于本研究而言，线下研究的效应量也有一定提高，本次研究让我们意识到线上与线下的结合确实很有必要。

第三轮

审稿人 1 意见：没有意见了。

回应：再次感谢评审老师在审阅过程中提出的修改建议，很好地帮助我们提升文稿质量。谢谢！

审稿人 2 意见：作者接纳了审稿人的意见，补充了研究 4 作为线下实验，使研究结果与最初的理论假设更加吻合，论文基本达到《心理学报》的发表水准。但仍有以下问题，期待作者进一步解释与修改：

回应：非常感谢审稿专家再一次审阅文稿，对文章的认可，您的意见对于我们进一步完善论文非常重要，对于本次意见我们将一一予以改进，希望能够达到您的要求，修改部分在论文中均已用紫色字体标出。

意见 1：“自我建构量表由 Singelis(1994)开发，原量表共 24 题，包括 12 道独立自我题项与 12 道互依自我题项，Na 等人(2010)在使用时将其中与年龄密切相关的题项删除形成 20 题，两种自我类型各 10 题。潘黎和吕巍(2013)修订中文版成人自我建构量表，最终包含 10 道互依自我题项与 6 道独立自我题项。本研究在此基础上选取 6 道独立自我题项与 6 道互依自我题项”，作者为什么要舍弃本量表中的 4 个题目？

回应：感谢审稿人的意见！研究 1 舍弃修订后中文版量表中 4 个题项，主要是为了保持量表在研究 1 跨文化使用中的一致性。潘黎和吕巍(2013)修订的量表主要为了满足相关统计学指标，并未考虑年龄因素，10 道互依自我题项中就包括与年龄等额外因素相关的题项，如：“年轻人在制定教育/职业规划时，应该考虑父母的建议”。这就使得 Na 等(2010)使用的互依自我题项与中文版修订的互依自我题项并不一致，而中文版量表中独立自我的 6 题均包括在 Na 等(2010)使用的 10 道独立自我题项中。提取两个量表的公共题项(Na et al., 2010; 潘黎, 吕巍, 2013)，会得到 7 道互依自我题项与 6 道独立自我题项。

在研究 1 后面的统计分析中，采用回归分析生成相应残差分别作为“纯粹的”互依自我与独立自我指标，需要两个分量表题项间的平衡，该方法与前人在探究追寻美德的动机对所罗门悖论的影响中操作相同(Huynh et al., 2017)，由于“追寻美德的动机”与“追寻享乐的动机”两者之间相关显著（本研究中在两个文化下独立自我与互依自我相关均显著），研究者为了得到“纯粹的”追寻美德的动机，也是将追寻享乐的动机预测追寻美德的动机所得的残差作为其指标，“追寻美德的动机”与“追寻享乐的动机”分量表各包含 5 道题项。为了使得互依自我与独立自我题项平衡，需要删除一道互依自我题项，删除的一道为：“如果我所在的群体需要我，即使我呆得不开心，我也会留在那里”，主要因为该题项在潘黎和吕巍

(2013)修订研究中的因子载荷（样本 1）低于本研究所选的 6 道互依自我题项。我们在文中相应部分简要补充了相关论述。

参考文献：

潘黎, 吕巍. (2013). 自我建构量表在成人中的应用和修订. *中国健康心理学杂志, 21*(5), 710−712.

Huynh, A. C., Oakes, H., Shay, G. R., & McGregor, I. (2017). The wisdom in virtue: Pursuit of virtue predicts wise reasoning about personal conflicts. *Psychological Science, 28*, 1848–1856.

Na, J., Grossmann, I., Varnum, M. E. W., Kitayama, S., Gonzalez, R., & Nisbett, R. E. (2010). Cultural differences are not always reducible to individual differences. *Proceedings of the National Academy of Sciences, 107*, 6192–6197.

意见 2：对于第一轮的意见 2，尽管作者针对“两个矛盾”进行了补充说明，但审稿人仍有以下问题，望作者可以解答。针对矛盾 1：作者指出可能存在参照群体效应从而导致中国人倾向于报告比实际更高的独立自我。如果是这样，互依自我上是否也有"参照群体效应"?——美国人平均的互依我更低，是否会让他们倾向于报告比实际更高的互依我？——可是文中结果显示互依我并无显著差异，怎么理解？

回应：感谢审稿人的意见！与审稿人的推理一致，有研究者在提前启动“参照群体效应”，即明确与参照群体对比情况下，确实发现了西方被试在互依自我得分上高于东方被试的情况，具体为在加拿大（日本）被试群体回答自陈量表的题项前加入：“与大多数加拿大人（日本人）相比，……”，结果发现加拿大被试在互依自我量表上的得分显著高于日本被试，也高于未启动，即使用原量表题项的加拿大被试的互依自我得分(Heine et al., 2002)。

在本研究中，中国被试在独立自我量表上得分高于美国被试，可以用“参照群体效应”解释，但在互依自我得分上并没有出现预想中的结果，一个可能的解释为：被试可能在填写互依自我题项时，没有达到填写独立自我题项时“参照群体”的程度，即诱发的“参照群体效应”较小。“参照群体效应”本质上来自于社会比较(social comparison, Festinger, 1954)，社会比较是个体进行自我评估、形成自我概念的重要途径(刘艳, 2011)，但并不是唯一途径，很多情况下个体也可以依据内省等方法对自身情况进行判断与评估，例如在使用关于自尊的自陈式量表进行跨文化的研究中“参照群体效应”就很小(Heine & Lehman, 2004)。

本研究中独立自我题项比互依自我题项可能更容易让被试进行社会比较，继而产生更大的“参照群体效应”，这一现象某种程度上可以从文化变迁的角度来解释。当今世界的一个总体趋势是个体主义上升，而集体主义式微，个体主义中包含的追求独立，优先个人目标等价值取向，目前在东西方文化中均可算作比较主流(Varnum & Grossmann, 2017; 黄梓航 等, 2018)；而且对于当今中国人来说，长期被忽视的、甚至被压抑的独立自我，逐渐在日常生活中得到培养、发展与强调(Lu & Gilmour, 2007)，相对而言可能更容易引起关注与社会比较。

参考文献：

黄梓航, 敬一鸣, 喻丰, 古若雷, 周欣悦, 张建新, 蔡华俭. (2018). 个人主义上升,集体主义式微?——全球文化变迁与民众心理变化. *心理科学进展, 26*, 2068−2080.

刘艳. (2011). 自我建构研究的现状与展望. *心理科学进展, 19*, 427–439

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117–140.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What’s wrong with cross-cultural comparisons of subjective Likert scales? The reference–group problem. *Journal of Personality and Social Psychology, 82*, 903–918.

Heine, S. J., & Lehman, D. R. (2004). Move the body, change the self: Acculturative effects on the self-concept. In

M. Schaller & C. Crandall (Eds.), *Psychological foundations of culture.* Mahwah, NJ: Erlbaum.

Lu, L., & Gilmour, R. (2007) Developing a new measure of independent and interdependent views of the self. *Journal of Research in Personality, 41*, 249–257.

Varnum, M. E., & Grossmann, I. (2017). Cultural Change: The How and the Why. *Perspectives on Psychological Science, 12*, 956−972.

意见 3：作者在回复意见的第三段说："因此在个体层面上则不能将互依自我与独立自我视为一个维度的两端，而应视为两个独立维度，独立自我高并不意味着互依自我低（Singelis, 1994）"。但是在后文中又说："研究 2 与 3 的启动研究将独立自我与互依自我视为同一维度的两端" 前后似乎有矛盾？如何理解？

回应：感谢审稿人指出这一点！

虽然独立自我与互依自我的提出者与自我建构量表的开发者，都强调应该将独立自我与互依自我视为两个独立维度(Markus & Kitayama, 2010; Singelis, 1994)，但在一些应用自我建构量表的研究中，研究者会采用将互依自我得分减去独立自我得分所得的"自我建构指数"作为判别独立自我与互依自我程度的指标(曾世强 等, 2016; 朱振中 等, 2018)。这样一来就使得原本包含独立自我与互依自我的二维，通过统计处理变为了"自我建构指数"一维，"自我建构指数"高的就倾向互依自我，低的就倾向独立自我，这种处理方式可能不是很妥，可能导致独立自我与互依自我得分均高的，与得分均低的在"自我建构指数"上相同。

原本我们认为，通过实验启动的不同自我类型与上述的处理方式类似，也是将二维变为一维，但经过审稿人点出后发现，二者其实并不类似，相减后所得的"自我建构指数"是一个连续变量，而使用实验启动两种不同的自我类型背后主要是一种"二元式"思维，通过启动得到的不同自我类型是相对立的二分变量。所以后文中"视为同一维度的两端"表述确实不妥，更确切的来说应该是"研究 2 与 3 的启动研究将独立自我与互依自我视为相互对立的二分变量"。

参考文献

曾世强, 陈健, 吕巍, 潘黎. (2016). 独立自我"嗇于人"，相依自我"嗇于己"——为自己消费还是为他人消费与自我建构对储蓄和消费选择的影响. *管理评论, 28*(6), 119–130.

朱振中, 李晓君, 刘福, Haipeng (Allan) Chen. (2020). 外观新颖性对消费者购买意愿的影响：自我建构与产品类型的调节效应. *心理学报, 52*(11), 1352–1364.

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

Singelis, T. M. (1994). The measurement of independent and interdependent self-construals. *Personality & Social Psychology Bulletin, 20*(5), 580–591.

意见 4：研究 3 需要具体（或简要）说明为什么要控制情绪和自尊。结果部分，4.2 部分协变量效应均显著，且效果量都远远高于观察变量的效果量，需要进一步说明协变量对因变量的具体影响，并解释缘由。

回应：感谢审稿人指出这一点！控制情绪主要有两个原因，一是情绪有可能会"干扰"个体冲突事件的重构或回忆，例如有研究表明相对于比较中性的事件或经验，个体更容易回忆起一些情绪特征较为明显的事件或经验(Levine & Edelstein, 2009)。二是情绪与冲突情境下智慧推理的关系(Kunzmann & Glück, 2019)，从情绪的功能主义视角来看(Campos et al., 1989; Keltner & Gross, 1999)，情绪在某种程度上可以为个体提供关于情境的有价值信息或信号，

继而可以让个体更好的把握冲突事件的发展趋势。在具体研究方面，情绪与智慧推理的关系还未取得共识，例如有研究者发现智慧推理与积极情绪正相关(Grossmann, Gerlach, & Denissen, 2016)，而另一些研究发现积极情绪与智慧推理无关或呈现出较低的负相关(Kunzmann & Baltes, 2003; Mickler & Staudinger, 2008)。还有一些研究表明情绪与智慧推理的关系要视具体情境而定，例如在阻止朋友自杀的假设情境中，个体的悲伤水平与其智慧推理呈显著正相关(Hu et al., 2018)，而在与本研究相似的涉及自身人际冲突情境中，研究者发现丰富的情绪体验有助于提高个体的智慧推理水平(Grossmann et al., 2019)，本研究结果发现，积极情绪(B = .17, SE = .02, t = 8.06, p < .001)与消极情绪(B = .09, SE = .03, t = 2.69, p = .007)均正向预测智慧推理，在某种程度上也支持该观点。此外由于研究 3 并不是使用标准化冲突情境，而是被试回忆重建的自身亲历的冲突事件，情绪也可以在一定程度上作为不同冲突事件性质与强度的一个指标。

对于自尊而言，依据自尊的社会计量器理论，自尊是个体社会关系好坏的一个内在标准，与个体在一个群体或一段关系中被认可与接纳的程度密切相关(sociometer theory, Leary & Baumeister, 2000)。当个体的人际关系出现问题时，自尊就会作为社会计量器就会发出一种信号，使得个体的自尊感下降并引起个体焦虑，沮丧等情绪反应，进而促使个体必须采取某种行为去获得，维持和恢复人际关系的和谐感(张林, 李元元, 2009)，而智慧推理中采取更广泛视角，整合不同观点与意见以及优先考虑让步等都有助于促进人际冲突的妥善解决与恢复人际关系的和谐(Grossmann et al., 2012)。最近在中国老年人群体做的一项调查也发现自尊水平与智慧呈正相关(Chen et al., 2021)，本研究的结果也发现自尊可以正向预测人际冲突中的智慧推理(B = .20, SE = .05, t = 4.41, p < .001)。

根据上述分析，我们在文中相应部分补充了情绪与自尊作为协变量的相关说明。

参考文献

张林, 李元元. (2009). 自尊社会计量器理论的研究述评. *心理科学进展, 17*(4), 852–856

Campos, J. J., Campos, R. G., & Barrett, K. C. (1989). Emergent themes in the study of emotional development and emotion regulation. *Developmental Psychology, 25*, 394–402.

Chen, Z., Zhu, M., Zheng, L., & Xie, X. (2021). Personal wisdom and quality of life among Chinese older adults. *Journal of Health Psychology*.

Grossmann, I., Gerlach, T. M., & Denissen, J. J. A. (2016). Wise reasoning in the face of everyday life challenges. *Social Psychological and Personality Science, 7*, 611–622.

Grossmann, I., Karasawa, M., Izumi, S., Na, J., Varnum, M. E. W., Kitayama, S., & Nisbett, R. E. (2012). Aging and wisdom: Culture matters. *Psychological Science, 23*, 1059–1066.

Grossmann, I., Oakes, H., & Santos, H. C. (2019). Wise reasoning benefits from emodiversity, irrespective of emotional intensity. *Journal of Experimental Psychology. General, 148*, 805–823.

Hu, C. S., Huang, J., Ferrari, M., Wang, Q., Xie, D., & Zhang, H. (2018). Sadder but wiser: Emotional reactions and wisdom in a simulated suicide intervention. *International Journal of Psychology, 54*(6), 1–9.

Keltner, D., & Gross, J. J. (1999). Functional accounts of emotion. *Cognition and Emotion, 13*, 467–480.

Kunzmann, U., & Baltes, P. B. (2003). Wisdom-related knowledge: Affective, motivational, and interpersonal correlates. *Personality and Social Psychology Bulletin, 29*, 1104–1119.

Kunzmann, U., & Glück, J. (2019). Wisdom and emotion. In R. J. Sternberg & J. Glück (Eds.), *The Cambridge handbook of wisdom* (pp. 575–601). United Kingdom: Cambridge University Press.

Leary, M. R., & Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. *Advances in experimental social psychology, 32*, 1–62.

Levine, L. J., & Edelstein, R. S. (2009). Emotion and memory narrowing: A review and goal-relevance approach. *Cognition and Emotion, 23*, 833–875.

Mickler, C., & Staudinger, U. M. (2008). Personal wisdom: Validation and age-related differences of a performance measure. *Psychology and Aging, 23*, 787–799.

意见 5：5.2 部分。建议作者报告下协变量的效应。

回应：感谢审稿人指出这一点！我们在文中相应地方补充了协变量的效应情况。“协变量的效应均不显著，性别：$F(1, 193) = .05$, $p = .825$，年龄：$F(1, 193) = .48$, $p = .488$，社会阶层：$F(1, 193) = .01$, $p = .922$，”。

意见 6：摘要部分。“研究表明，一方面，所罗门悖论可能只存在独立自我高的人身上，并不具有普适性；另一方面，研究心理学除了关注“怪异”样本外，还需关注使用单一样本但却默认结论具有文化普适性的“怪异”研究者。” “另一方面”的表述是通过本研究得出的结论？还是作者提出的建议？

回应：感谢审稿人的意见！“怪异”样本的问题主要由 Henrich 等于 2010 年指出，该文一经发表就引起学界的广泛讨论，并且绝大多数研究者都认为应该重视心理学界的“怪异”样本问题。不过与这种讨论的热烈情况相比，研究者在实际研究工作中却并没有对此特别重视，例如有研究者于 2018 年在 *PNAS* 上发表文章就指出，发表在 2014 年与 2017 年心理学旗舰期刊 *psychological science* 上的实证论文，不仅大部分样本依然来自西方世界，而且这些研究者中的大部分在论文中也并未对样本局限性进行讨论(Rad et al. 2018)。比较巧合的是，目前已有的与“所罗门悖论”直接相关的论文就是分别发表在 2014 与 2017 年的 *psychological science* 期刊上(Huynh et al., 2017; Grossmann et al., 2014)，并且他们也未在文章局限性部分尝试对所罗门悖论的文化普适性问题进行探讨，这种联系与巧合也是我们探究所罗门悖论是否具有文化普适性的一个重要动机。

此外，相对于当前心理学界对“可重复性”问题的重视，例如诸多心理学期刊都针对“可重复性”问题设置了很多数据、方法以及结果汇报方面等硬性要求(Greenfield, 2017)，样本的“多样性”问题还基本处于依靠研究者自觉阶段（当然样本问题更多地受到研究者所处的现实情境限制），结合这两个方面的原因，总体而言这里更多的是基于研究结果提出的一个建议。摘要中的表述改为这样可能更合理：“研究表明所罗门悖论可能只存在独立自我高的人身上，并不具有普适性。可见心理学研究除了关注“怪异”样本外，更需关注使用单一样本但却默认结论具有文化普适性的“怪异”研究者。”

参考文献

Grossmann, I., & Kross, E. (2014). Exploring “Solomon’s paradox”: Self-distancing eliminates the self-other asymmetry in wise reasoning about close relations in younger and older adults. *Psychological Science, 25*, 1571–1580.

Huynh, A. C., Oakes, H., Shay, G. R., & McGregor, I. (2017). The wisdom in virtue: Pursuit of virtue predicts wise reasoning about personal conflicts. *Psychological Science, 28*, 1848–1856.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences, 115*(45), 11401–11405.

Greenfield, P. M. (2017). Cultural change over time: why replicability should not be the gold standard in psychological science. *Perspectives on Psychological Science, 12*(5), 762–771.

意见 7：不知道“两因素方差分析”是否是“协方差分析”？其对应的英文术语是？
回应：感谢审稿人指出这一点！协方差分析确实更为恰当，协方差分析对应的英文为：“analysis of covariance (ANCOVA)”(Hayes et al., 2005, p.408)。协方差分析又称带有协变量的方差分析，是一种将回归分析与方差分析结合起来使用的一种分析方法，这里区别于自变量为一个单因素协方差分析(陈敏琼, 2014)，称为两因素协方差分析可能更恰当。本研究协变量的处理类似于阶层回归分析中加入的第一步控制变量，第二步就是在第一步控制基础上，采用比较常规的两个自变量对因变量的两因素方差分析。鉴于本文主要目的在于方差分析，出于文章行文简洁考量，除了第一次出现外，后文中将其统称为方差分析。

参考文献

陈敏琼. (2014). 单因素协方差分析的一种回归算法. *统计与决策, 15*, 73-75.

Hayes, A. F. (2005). *Statistical methods for communication science.* Routledge.

第四轮

审稿人 2 意见：论文经过几轮修改后已经有很大提高，建议立即发表！
回应：非常感谢审稿专家对论文的肯定！再次感谢审稿专家在审阅过程中提出的建议，帮助我们提升文稿质量。谢谢！

编委复审意见：经过多轮修改，该稿已有很大提升，同意发表。
回应：感谢编委专家对修改稿的肯定！

主编终审意见：本文经过多轮评审和修改，质量答复提高，建议文章继续做几处小的修改：
回应：感谢主编对修改稿的审核与意见，帮助我们减少了文章歧义以及进一步方便读者理解。文中相关修改均以绿色字体标注。

意见 1：文章的“问题提出”标题改为“引言”
回应：“问题提出”已改为“引言”。

意见 2：中文题目中的“怪异所罗门”和英文题目中的“Solomon is weird" 不合适，建议拿掉：Solomon 现象不是西方独有的，中国高独立自我的被试也有。
回应：感谢审稿人意见！中英文题目中的“怪异的所罗门”（Solomon is weird）均已删除。

意见 3：建议不要用“怪异”来描述样本，直接说西方大学样本：“怪异”让人觉得很奇怪，没有读过那个文章的人会觉得莫名其妙。“WEIRD”的直接翻译是“怪异”，但是其本意就是西方大学生，好的文章应该能让没有太多的背景的人也很容易读懂，学报不鼓励故弄玄虚。
回应：感谢审稿人意见！我们将摘要与正文涉及“怪异”样本描述的相关部分均以“以西方大学生为主的样本”、“西方人”、“西方被试”等替换。仅在直接引用“WEIRD”的文献时，出现

WEIRD 一词，但在文中作了解释，即它是“西方的(Western)、受过教育的(Educated)、工业化的(Industrialized)、富有的(Rich)及民主的(Democratic)”五个英文单词首字母的合写，是一种修辞手法，这样既能让读者一眼就明白“WEIRD”在文中的准确含义，又能表达对英文原文的尊重。