《心理学报》审稿意见与作者回应

题目：行为可见增加利他偏好及其社会规范机制
作者：黄馨茹，李健，倪荫梅

第一轮

审稿人 1 意见：

本文考察了行为可见性对利他偏好的影响，研究较为规范，表述清楚，但有若干重要问题尚需作者进一步说明。

意见 1：本文的核心论点是行为可见性提升人们的利他行为，但行为可见性为利他的影响并不是一个新颖的课题，事实上早在上个世纪 70 年代就陆续有研究开始考察这一现象，如①Axelrod, R. (1984). The evolution of cooperation. New York: Basic Books. ②Fox, J., & Guyer, M. (1978). “Public” choice and cooperation in n-person prisoner’s dilemma. Journal of Conflict Resolution, 22(3), 469-481. ③Jerdee, T. H., & Rosen, B. (1974). Effects of opportunity to communicate and visibility of individual decisions on behavior in the common interest. Journal of Applied Psychology, 59(6), 712-716. 这些研究运用了实验法或计算机仿真探讨行为可见性对利他的影响，当然在不同的文献里，利他有时也被称为合作（cooperation）或亲社会（prosociality）等，但主旨没有太大区别。另外，Hardy, C., & Van Vugt, M. (2006). Giving for glory in social dilemmas: The competitive altruism hypothesis. Personality and Social Psychology Bulletin, 32, 1402-1413. 这项研究提出了“竞争性利他”的概念，虽然行为可见性并不是研究核心，但也考察了行为可见性对利他的影响。而在一项元分析中，Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality? Proceedings of the Royal Society B: Biological Sciences, 285(1875), 20180116，作者分析了 117 有关研究并提出行为可见性确实可以增加个体的利他行为，尽管效应量并不大，并且分析了上述效应的边界条件。其他相应的文献还有很多，恕不再这里一一列举，因此，但就本文的中心论点而言，我似乎看不出有重大的理论创新。

回应：感谢审稿人的宝贵意见。诚如审稿人所言，可见性（或称被观察）对于社会行为（如合作，亲社会行为，利他行为）的影响并非一个新课题。对此，我们的回应如下：

（1） 在前言部分更加明确地对本研究问题进行定义（见修改后的第 3 自然段）。参考 Bradley 等人(2018)的定义，我们将“可见性”更详细地分为“感知到被观察”“行为可见”和“行为与身份均可见”三类；此外，综合审稿人 2 的建议，我们区分了“对第三方可见”和“对接受方可见”。本研究中，我们量化地研究了“行为可见”以及“对接受方可见”，这一问题在前人文献中并未得到十分充分和明确的说明。

（2） 综合审稿人 1 和编委的意见，我们补充了实验 2，在实验 1 量化地揭示可见性对利他的影响基础之上，进一步考察社会规范对于“行为可见增加利他偏好”这一效应的影响与调节。具体内容请见文章正文。

（3） 此外，Hardy 和 Van Vugt (2006)在理论层面提出可见性对于利他的作用，因此我们在前言和讨论中删去了关于本研究对于利他理论的贡献的部分，并更改为较为符合本研究实际贡献的内容。在此十分感谢审稿人的提醒。

意见 2：作者在讨论部分也提出“本研究只从现象上验证了行为可见性对于利他偏好的影响，但未提出机制上的解释”，这一点我同意作者的看法。作为一个已经受到研究者关注超过半个世纪的课题，我作为一个读者更希望看到的不仅仅是对这个现象的重述，而是对这个现象的解释，如提出一种新的心理机制，或者考察这一效应新的边界条件，等等。因此，我觉得这一点并不能仅仅作为一个研究不足放在讨论部分，更可行的做法是增加一至两个研究来探讨上述现象的心理机制，比如，在研究 1 中作者发现在不同于以往研究的情境下也存在行为可见促进利他的效应，而在后续实验中逐一考察了上述效应潜在心理机制，只有这样才算得上一项完整的和有重要创新的研究。

回应：感谢审稿人的宝贵意见。研究 1 中，我们发现行为可见能够缩小不同反应类型间的偏好差异，这提示行为可见可能让人们的行为更加趋于某种规范。因此，在补充的研究 2 部分，我们对行为可见增加利他偏好这一现象的机制做出假设，即认为行为可见让人们更加遵从利他的社会规范，进而做出更多利他行为。实验 2 中，我们外显地操纵利他或非利他的社会规范，发现当存在利他规范时，行为可见增加利他偏好；而当环境社会规范为非利他时，这种效应减小甚至消失。诚然，这一实验并不能排除其他可能的机制，因此在讨论部分我们讨论了其他可能的机制，如声誉、共情和心理理论，后续研究或可进行进一步探索。

意见 3：本研究共包含一个实验，且涉及的被试量只有 38 个。当然实验数量少、样本量小本身并无过错，但考虑到现有心理学尤其是社会心理学的可重复性问题，这么小规模的一个研究似乎难以可靠地排除假阳性的可能。结合第二条意见，我仍然建议作者增加若干实验，拓展现有的研究结论，这样可以同时解决理论创新和方法论上的问题。

回应：感谢审稿人对于样本量的建议。为保证结果的稳健性，我们在实验 2 中招募了 53 名被试，并再次验证了行为可见增加利他偏好的结论。

意见 4：作者采用了两种方法来处理统计数据以增加结论的稳健性，这是本文严谨性的一个体现，值得肯定。但我个人觉得作者对计算建模方法的介绍和展示给予了过多的篇幅，如果作者能够从计算建模方法得出新的结论，增进读者对研究问题的新知识，那么大篇幅报告建模方法是有必要的，但从本文来看，两种数据处理方法得出了基本相似的结论，并没有增加读者对研究的新理解，因此计算建模方法仅仅作为一个稳健性检验所占篇幅实在过多，我建议适当删除这部分内容，只报告最重要的结果，把更多的篇幅留给理论探讨，这样才能提高本文的研究丰度和深度。

回应：感谢审稿人对于数据分析方法的建议。对此，我们有以下回应与修改：

（1） 在本研究中，我们确实花了较大的篇幅对计算建模进行描述以及对相应的结果进行解释。在修改版中，我们对行文进行了精简，减少了该部分的篇幅，但还是保留了相当的一部分。其原因在于计算建模结果是本研究讨论问题的核心部分，即对个体心理过程的量化刻画，该方法可以帮助我们对研究操纵引起的心理过程变化进行定性和定量的解释，同时尽可能排除其它混淆因素的影响。本研究中行为结果提示了可见性增加了个体利他，为我们后续的计算建模提供方向。然而，我们认为行为指标（如自我他人收益之差）只能粗略反映个体对优势与劣势不公平的容忍程度，但无法排除自我和他人收益值本身的影响。因此，本研究使用的计算建模除了提供稳健性检验之外，实际上更重要的是对个体的社会偏好变化进行定性和定量地刻画。由于原文的描述不够清晰，使得在不依赖模型的结果和依赖模型的结果两部分看起来是对应基本一致的内容。因此，在修改稿中，我们对两者进行了区分。实际上对利他偏好的刻画（AIA 和 DIA）是模型依赖的，而在不依赖模型的行为结果并不能对被试的社会偏好下定论。感谢审稿人的提问，使我们对两部分的结果阐述更为清

晰。

（2） 另一方面，从效用理论角度出发，“偏好”指个体内心对事物的喜好排序，而“行为”指个体外显的行为表现，二者有较多不同之处(Raiffa & Schlaifer, 1961)；不依赖模型的指标是对“行为”的直接分析，而基于模型的参数是对被试的心理过程即内在“偏好”的刻画，二者分别反应了不同的侧面，并不完全一一对应。我们在“方法：测量指标”部分对不依赖模型的指标和基于模型的指标进行了更为详细的区分，请审稿人审阅与指正。

（3） 此外，鉴于计算建模目前开始被广泛应用到社会心理学、发展心理学、临床心理学等心理学传统领域，我们认为详细地描述建模相关内容不仅能够清晰地说明本研究讨论的问题，同时也有助于新方法的引入，有助于和其他研究者进行交流，有助于后续研究者对方法的借鉴与拓展（我们已经将研究数据和相关分析的代码及结果上传到公共数据平台）。

（4） 在文章讨论部分，我们增加了对研究结果理论探讨的篇幅。

意见 5：另外还有若干细节问题。

（1） 关于使用眼睛图片的问题。作者在引言部分提出了这个方法两个不足，但另一方面，作者在实验中同样采用了眼睛图片来作为实验操纵，这似乎有点前后自相矛盾，为什么作者认为本文采用眼睛图片就可以避免其他研究中眼睛图片的问题？仅仅是因为在本研究中伴随眼睛图片的还有在场的博弈对手（尽管事实上只是个虚拟被试）？如果是这样的话，我觉得更严谨的作法是应该增加一个操作有效性检验，也就是作者必须提供证据证明眼睛张开/闭合的图片的确能在被试心理上引发被观看/没被观看的感觉。我觉得这是本文实验设计的一个瑕疵。

回应：感谢审稿人对该问题的宝贵意见。对此，我们以下回应与修改：

① 在修改稿中，我们区分了“感知到被观察”和“行为实际被观察”的区别，并强调本研究只关注后者。本研究使用眼睛图片仅为了提醒被试此时对方正在看/不看他的选择。前人研究发现，使用眼睛图片诱发被观察感和较为底层的知觉与唤起有关(Hesslinger et al., 2017)，而本研究关注的“行为可见”可能涉及了更多高级的社会认知加工过程，因此我们认为使用眼睛图片的研究和本研究本质上关注的是不同的问题。非常抱歉，我们在原文中没有对这个问题的区分不够清晰，以至于产生混淆。非常感谢审稿人提出疑问，我们对这两方面的研究进行了更加细致的思考与辨别。修改稿中已进行明确的区分，具体请参见正文的前言部分。

② 非常感谢审稿人提出的操纵检验问题，在实验二中，我们补充了操纵检验的部分，即在每一个区块后让被试评价自己的感受有多符合描述“在刚才的游戏中，我的行为能够被其他一些人看到”，被试需要在 5 点李克特量表上打分，1 代表完全不符合，5 代表完全符合。结果发现，在可见条件下被试的评分总体较高(利他-可见条件下评分 $M = 3.86$，$SD = 1.23$，非利他-可见条件下评分 $M = 3.73$， $SD = 1.18$)，不可见条件下被试的评分总体较低(利他-不可见条件下评分 $M = 1.68$，$SD = 1.00$，非利他-不可见条件下评分 $M = 1.45$，$SD = .67$)。总的来说，可见条件下评分高于不可见条件，$F(1,52) = 155.15$，$p < .001$，$\eta_p^2 = .75$，说明我们对可见性的操纵确实让被试意识到自己的行为对他人可见/不可见。

（2）另一个问题和生态效度有关，作者在引言部分批评了现有相应研究缺乏生态效度，可是我看不出作者采用了什么措施来提高本文的生态效度。本文实验主要是一个独裁者博弈，而经济博弈的生态效度长期以来受到研究者，尤其是人类学家和心理学家的质疑（Brannen, J., & Coram, T. (Eds.). (1992). Mixing methods: Qualitative and quantitative research. Aldershot:

Routledge），且考虑到本文的样本量非常小，又是一个典型的实验室研究，无论从哪个角度来说似乎都没有理由认为本研究具有较高的生态效度。

回应：感谢审稿人的意见。对此我们的回应如下：第一，我们承认本研究采用经济学博弈范式，其生态效度有限；第二，我们在原文表述较为不清，事实上我们希望表达的是 Dana et al., 2006、Andreoni & Bernheim, 2009 等研究对“行为可见”的操纵不够直观，而并非批评前人研究不具有情境和被试样本的可推广性。原文中我们对“生态效度”这一词进行了误用，已在修改稿中进行相应调整。

（3）最后一个问题和被试的同质性有关，作者没有报告本研究采用了学生被试还是社会被试，从年龄上来看似乎包含着一定数量的社会被试，如果被试的异质性较大（例如有学生被试也有社会被试，或者年龄跨度较大），最好能简要分析下被试本身的特点对实验结果的影响（学生被试的话，控制变量应包括年龄、性别、专业，社会被试的话，最好还包括经济收入、教育水平、职业分类等），这样一方面能让研究结果更加严谨，另一方面可能帮助研究者发现意料之外的结果。

回应：感谢审稿人对于样本同质性的建议。本研究所招募的被试均为大学生被试，由于包含少数博士生，所以年龄跨度较大（已在修改稿中写明均为学生被试）。此外，我们针对被试年龄、性别、专业等因素建立了线性模型，分析被试特质对结果的影响，但未发现以上因素对结果的影响。

……

审稿人 2 意见：

本文通过使用修改版的独裁者游戏，考察了行为可见性对个体利他偏好（选择、评估）的影响。结合行为分析与计算建模，主要结果发现，行为可见情境下，个体对优势不公平的厌恶程度增加，且对劣势不公平的厌恶程度减小。本文所关注的主题（社会决策）与方法（认知计算建模）均紧紧贴合国际上的研究热点。同时本文行文流畅、内容可读性强，数据大体上支持研究结论并呼应研究假设。尽管如此，本文在行文结构和数据分析部分尚有一些不足。希望作者对下述建议进行考虑，并对文章进行改进。

意见 1：在介绍“行为可见性”时，希望作者纳入对于观众效应的讨论，如 Bradley et al 2018，Hamilton et al，2016；在探讨利他行为的动机和原因时，建议加入以下相关文献 Böckler et al，2016，Hu et al，2017。

回应：感谢审稿人对于相关文献的补充。我们已仔细阅读审稿人推荐的文献，回应如下：

（1） Bradley et al 2018 是一篇元分析，总结了观察性(observability)影响亲社会性的效应及其边界条件。我们对这篇文献进行了详细的阅读，并借鉴了文章中对不同程度的观察性的区分，即“感知到被观察”“行为可被观察”和“行为和身份均可被观察”。这篇文章的观点对我们对问题进行清晰定义起到了极大的帮助，十分感谢审稿人的补充。我们在修改稿中的前言第三段对其进行了引用。

（2） Hamilton et al，2016 是一篇关于观众效应(audience effect)的综述，总结了观众效应相关的行为、神经与发展研究以及观众效应在特殊人群中的研究。观众效应和本研究所关注的“行为可见对利他偏好的影响”有交叉（即都关注其他人的观察对个体行为的影响），但也有不同（观众效应范围更广，不只限于社会行为领域，还包括如物理性任务、记忆、问题解决等）。因此，我们在修改稿中并未对观众效应展开论述，而是在讨论部分简单提及。

（3） Böckler et al，2016 对各类经济学博弈中的亲社会动机进行了因子分析，区分出如

纯利他动机、规范性动机、策略性动机等因素。其中，策略性动机和本研究的出发点十分类似，均将利他视为一种策略性行为(在他人可见时做出更多利他行为，他人不可见时就更不利他)。为了控制篇幅，前言部分我们未对这篇研究的结果进行大面积引用，而是在讨论部分对其进行论述。

（4） Hu et al，2017 是一篇关于利他的实证研究，主要探究了“paying it forward”这一现象的神经基础，和利他相关，且该研究也使用了 Fehr(1999)的模型，因此在本研究中也引用。我们感谢审稿人对于稿件的仔细阅读和中肯意见。

意见 2：第三页，“[…]在实验界面中加入眼睛图案来代表正在观察的对方”，此处的描述可以在精确一些，即，本研究中的眼睛图案是二人独裁者游戏中另一方的眼睛，而不是第三方的眼睛。此外，我对这一实验设计有一些不同观点（见下文#5）。

回应：非常感谢审稿人指出可能的混淆点。在修改稿中，我们明确区分了“第三方观察”和“接受方观察”，并指出本研究仅关注后者（见修改稿前言部分）。

意见 3：第四页，这里介绍了 Fehr-Schmidt（FS）模型并解释了参数的意义。尽管这一模型广为研究者使用，但还是建议不要抛开公式讲参数。所以这里务必加上模型的公式，再对参数进行解释。

回应：感谢审稿人的建议。已在修改稿中加入公式及其描述。

意见 4：假设部分，“[…]但具体的差异方式有待实验验证。”这半部分作为研究假设欠妥。

回应：感谢审稿人的建议，已在修改稿中删去类似的描述。

意见 5：关于本研究中眼睛图案的使用，作者认为以往研究中的第三方眼睛效果劣于本研究中的游戏对方的眼睛；但我认为，个体行为被第三方可见时利他偏好的转变与本文关注的问题有必要进行区分。举例来说，在日常生活中，（A）如果地铁上有一个人没有抓紧栏杆马上要摔倒了，当地铁上有人多人的时候我可能更加倾向于伸手帮助，但此时那个要摔倒的人并不一定知道我要怎么做；（B）如果这个要摔倒的人在摔倒之前与我有眼神接触，那我大概率是会伸手帮助的。简而言之，本文关心的是 B 中的问题，即社会互动中当互动的对方能够知晓我要怎么做时，我会怎样做；而在文献综述和背景介绍时侧重的是 A，即，当我的行为（被第三方）可见时，我会怎样做。建议对 AB 进行区分，并对文章标题和摘要（以及英文摘要）进行微调。

回应：非常感谢审稿人的建议。我们非常赞成审稿人所说，这二者的影响在行为和心理机制上都可能存在差别，并且本研究中我们探讨的是“对方可见”情境。对此，我们进行了以下修改：

（1） 在前言部分明确指出第三方可见和接受者可见的区别，并指出本文关注后者。

（2） 在文献综述部分，删去有关第三方眼睛图片的研究，仅综述和本文关注的问题直接关联的研究(即 Andreoni & Bernheim, 2009)。

意见 6：在介绍 FS 模型时，请加入必要的细节。如，M_s, M_o 的意义，alpha、beta、lambda、b_0/1 等参数的取值范围。同时，本文使用了分层贝叶斯估计，建议加上所有参数的先验分布。最后，建议作者将数据和代码公开，上传到 github 或者 osf 等平台。

回应：感谢审稿人对建模方法的书写的建议。已在“方法：测量指标”部分加入关于模型的细节。所有的数据和代码已上传至公开平台，链接为：https://github.com/psych575/open-data-and-code-for-xb21-575.git

意见 7：我对于在不同行为类型（选择 vs 评分）下使用两套 FS 的参数这一做法并不完全认同。FS 模型是对选项之间公平判断/厌恶的一般性衡量，无论是选择、还是评分、甚至没有行为测量的情况下，我看到同样的一个公平选项和一个不公平选项，我的 utility 计算应该是一致的。但与此同时，如本文所呈现的，或许选择类型确实会对 utility 的计算产生影响。我建议的做法是，在当前模型的基础上在构建一个选择-评分联合模型，即选择和评分的情况下使用同样的 alpha 和 beta（当然，需要区分可见 vs 不可见），再与当前的模型进行比较，看哪一个模型胜出，再进行后续分析。

回应：感谢审稿人对模型设置的建议。确实如审稿人所言，utility 反映被试内在对不同选项的偏好，即使使用不同的行为指标，被试的选择和评分也应受到相同的 utility 的控制；当然，本研究提出行为指标本身的不同也会造成偏好的变化。我们补充了选择-评分联合模型，发现其 DIC 大于选择-评分分离模型，说明本文将选择和评分条件赋予不同参数值是较为合理的。我们已在实验一的“测量指标”和“结果”部分补充了模型比较的相关内容。

意见 8：对于公式 6，有必要加入可见/不可见这一条件作为自变量。因为 FS 模型进行了这一区分，在公式 6 中加入这一区分更有说服力。

回应：非常感谢审稿人对线性模型的建议。修改稿中，我们在公式 6 的线性模型中加入了 visibility 这一固定效应，模型估计结果和之前基本相同。

意见 9：关于图 3c-d，横轴是指他人减去自我？这里是正相关，所以意思是他人比自我的钱越多的时候，我越满意？请作者进行进一步的解释。如果我的理解错误，请指出。

回应：感谢审稿人提出疑惑。此处 “自我他人收益之差”指被试选中试次中，自我收益和他人收益的绝对值的均值(简称自我他人收益之差)。图 3c-d 对应劣势条件下的自我他人收益之差，即他人收益减去自我收益。图三呈现的是被试间的相关分析结果，即，将所有被试在选择条件下的自我他人收益之差，和在评分条件下的评分值，进行相关分析。图 3c-d 中正相关的意思是，如果一个被试在选择条件下选中试次中的他人收益越多于自我收益(说明越能接受他人收益多于自己)，那么他在评分条件下对劣势选项的评分值越高(说明对劣势不公平越满意)，这一结果说明同一个被试在选择和评分下的行为是正相关的，符合直觉。原文的图片横坐标没有写“绝对值”，可能容易造成误解。已对图片横坐标进行改正，并在测量指标部分对该因变量的含义与简写进行了更详细的说明。

意见 10：第 11 页，这里我认为更好的检验方法是方差分析而不是 t 检验，即 2（优势 vs 劣势）x2（可见 vs 不可见），且从图 4 可以看到潜在的交互作用。如果使用两个 t 检验的话，需要进行 p 值矫正。

回应：感谢审稿人对统计方法的建议。此处为不依赖模型分析部分，审稿人认为应该将优势/劣势也作为一个自变量进行分析。已按照审稿人的建议，将不公平类型(优势 vs. 劣势)作为一个自变量，用方差分析替代两个 t 检验。

意见 11：在阐释不依赖模型的结果与 AIA/DIA 的关系方面，有用结果解释机制（而非机制解释结果）之嫌。文章似乎在用得到的结果，来解释 AIA 升高且 DIA 降低；但其实是因为行为可见，AIA 升高 DIA 降低，才得到了行为结果。建议作者在这里修改一下逻辑和措辞。（个人观点：图 1 的理论框架很有意思也很有帮助，但是这个框架是来自于 FS 模型的；在阐释不依赖于模型的行为结果时，没有必要与 AIA/DIA 进行联系。这一联系可以在讨论部分进行阐释和深化。）

回应：非常感谢审稿人的建议。审稿人认为 AIA/DIA 属于心理机制层面，而不依赖模型的结果属于行为层面，对此我们十分认同。我们进行了以下修改：

（1） 在“测量指标”部分对不依赖模型的指标和基于模型的指标进行了更好的定义和区分；

（2） 在所有涉及不依赖模型指标的部分，不将其和 AIA/DIA 进行关联，而是仅将其解读为选择/评分的行为结果。

意见 12：Minor

（1）建议在第三页，“[…]通过决策行为反推偏好[…]和通过主观评价直接测量偏好[…]”，这里有必要解释什么是选择，什么是偏好。文章在后面确实对这两个概念进行了解释，但为了方便理解，在这里进行解释更好；尤其要强调选择的迫选性和离散型，以及偏好的连续性。

回应：感谢审稿人的建议，已在前言的对应部分加入了相应的描述。

（2）图 1，建议加上原点和坐标轴正负号

回应：感谢审稿人的建议，已添加坐标原点和坐标轴正负号。

（3）2.1 部分，在介绍样本计算时，烦请加上基于怎样的统计检验。

回应：感谢审稿人的建议，已在“方法：被试”部分加上了统计检验的类型。

（4）2.3 部分，“4-8 名被试”是指 4、6、8，还是 4-8 中的所有可能？我的理解是被试数需要是偶数。

回应：您的理解是正确的，确实是 4、6 或 8 人，已在文中说明。实际实验中，有的被试临时未能到场，由被试未知的实验助手假扮被试代替完成，因此被试总人数出现了奇数。

（5）英文摘要有若干语法错误：

-- “researches”，research 不可数

-- “between choice and rating condition”， → conditions

回应：感谢审稿人的建议，已对英文语法做出修改。

参考文献

Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality?. Proceedings of the Royal Society B: Biological Sciences, 285(1875), 20180116.

Böckler, A., Tusche, A., & Singer, T. (2016). The structure of human prosociality: Differentiating altruistically motivated, norm motivated, strategically motivated, and self-reported prosocial behavior. Social Psychological and Personality Science, 7(6), 530-541.

Hamilton, A. F. D. C., & Lind, F. (2016). Audience effects: what can they tell us about social neuroscience, theory of mind and autism? . Culture and Brain, 4(2), 159-177.

Hu, Y., He, L., Zhang, L., Wölk, T., Dreher, J. C., & Weber, B. (2018). Spreading inequality: neural computations underlying paying-it-forward reciprocity. Social cognitive and affective neuroscience, 13(6), 578-589.

第二轮

审稿人 1 意见：

作者增加了一个实验，对之前的问题做了一些修改和回应，总体来说，文章的创新性和严谨性有了一定提升。不过我还是有若干问题希望和作者作进一步探讨。

意见 1：实验 2 在实验 1 基础上探讨了社会规范的作用，这个想法本身是很有价值的。但我

最大的疑问也是关于实验 2：

（1）在社会心理学中，社会规范通常被区分为描述性规范和命令性规范（Cialdini et al., 1991），而实验 2 中的社会规范实际上只是涵盖了描述性规范这一分类，因此我建议使用描述性规范来取代社会规范，从而使表述更为精准。

回应：感谢审稿人的建议，已在修改稿的前言部分区分描述性规范和命令性规范，并强调本文所探究的是描述性规范。感谢审稿人帮助我们对概念进行更加清晰的定义与区分。

（2）实验 2 的基本操作本质上属于社会规范方法（social norms approach, SNA），即通过向个体展示某种规范性信息来改变个体的态度或行为，这是规范心理学中普遍应用的一种干预实践，但 SNA 的使用是基于一定的前提条件的。Dempsey 等（2018）以及 Blanton 等（2008）提出了实施 SNA 的若干基本前提：①接受 SNA 干预的个体存在规范错觉，即被试错估了实际存在的规范，只有在这种情况下 SNA 才能取得良好的效果，最近发表在美国科学院院报 PNAS 的一篇论文也证实了这种观点（Palacios et al., 2022）。对比而言，本文实验 2 并没有将被试是否存在规范错觉这一前提条件置于分析之内，从而使操作和结论略显粗糙。②更为重要的是，在 SNA 中，实验者所提供的规范信息必须可靠，使被试相信某种行为的普遍程度确实高于自己的估计，从而调整自己对规范的估计和相应的行为，这是 SNA 产生作用的基本机制，举例而言，在我前面提到的研究中（Palacios et al., 2022），作者首先开展了一个较大规模的调查，调查了某个行为在人群中的普及程度，然后用这个数据作为真实规范信息的操作定义。而在本文实验 2 中，规范信息完全来自作者的操纵，很难说在多大程度上是真实的。更让我困惑的是，实验 2 采用了被试内设计，如果我没有理解错的话，被试会在短时间内接受两种完全相反的规范信息：大部分人都表现出了利他/不利他的行为。那么被试在短时间内接受的这两个相互矛盾的信息如何让被试相信这是真实可靠的？或者说哪一个信息才是真实可靠的？如何排除这种疑惑对实验结果的影响？

回应：感谢审稿人对社会规范操纵的建议，对此，我们有如下回复与修订：

① 我们仔细阅读了审稿人提供的关于“社会规范方法(SNA)”的文献，并对 SNA 的起源、发展与应用有了大致的了解。我们发现 SNA 主要针对的是<u>现实生活中的行为干预</u>(如酗酒、吸毒、安全驾驶、消费等)，而非针对实验室研究。我们认为，SNA 应用的两个前提(存在规范错觉、传达的是真实的规范)主要是针对现实情境提出的：

i) 在现实情境中，人们可以通过各种渠道获得关于社会规范的信息(如 2022 年的 PNAS paper 期待通过社会规范改变人们的消费行为，而人们可以通过观察邻居的消费行为推断社区中的普遍消费水平)，因此在 SNA 实施过程中最好使用<u>真实的社会规范</u>，否则被试容易看出这一规范是虚假的。而在本研究的实验室情境中，社会规范描述的是“人们在这一实验的分配任务中如何选择”，除了主试提供的前人选择信息，被试没有其他渠道获得“大多数人如何选择”的信息，因此被试只能相信主试提供的是真实的社会规范。因此，“传达的规范真实可靠”这一前提只适用于现实情境，而不适用于实验室情境。事实上，在 Asch(1953)的经典线段实验中，几名假被试明显给出了错误的线段长度判断(明显错误的社会规范)，而真被试仍然受到了这一错误规范的影响。由此可见，在实验室研究中，社会规范是否是现实中大多数人的普遍行为并不是必要前提。

ii) “存在规范错觉”这一前提和“规范信息真实可靠”这一前提是相辅相成的。由于 SNA 要使用真实的社会规范来改变人们的行为，因此只有对规范存在错误估计的人群才会受到社会规范操纵的影响。事实上，“存在规范错觉”这一前提的本质要求是“提供的社会规范信息要和被试自身的行为选择有一定差距，这样被试才会根据新的规范信息来调整行为”。在本研究的实验二中，提供的利他规范

和非利他规范都是使用较为极端的 AIA 和 DIA 值来生成，其利他/非利他程度显著有别于大多数被试自身的偏好，因此被试获取新的规范信息后倾向于调整自己的选择行为，说明社会规范操纵是有效的。

② 事实上，在较多实验室研究中，尤其是探究社会规范如何影响利他决策的研究中，对社会规范的操纵和本研究类似，即主试直接向被试提供实验情境下的大多数人如何决策的信息(e.g., Fathi et al., 2014; Raihani & McAuliffe, 2014; Oda & Ichihashi, 2016; Agerström et al., 2016; Kawamura & Kusumi, 2017) (可在正文的参考文献列表中找到引用的文献)。如，Kawamura & Kusumi(2017)探究社会规范对分配行为的影响，向被试呈现前人分配数额的均值来操纵社会规范。相比于现实情境，实验室研究的优势在于可以探究特定情境下人们的行为倾向，这种特定的情境可能并不是现实生活中常见的情境，而是主试构建和操纵的情境。因此，实验室研究中操纵的社会规范可能是现实中不常见的规范(如 Asch 的实验)，但不代表这种操纵是错误的。

③ 对于被试需要在短时间内接受两种规范，我们在实验设计阶段也考虑到了这一问题。由于社会偏好的个体差异较大，本研究采取被试内设计是更好的选择。为了让两种社会规范的呈现显得合理，我们告诉被试在之前已经有很多人参与了这一实验，前人的实验结果被上传到两个服务器上，分别为服务器 1 和服务器 2，被试在实验中将依次登录这两个服务器，并看到这两个服务器上的前人选择；实验结束后，被试的选择也将被上传到两个服务器上，并呈现给之后来做实验的被试。这种类似游戏服务器的设置引入了两个不同的平台，那么两个平台上的人们有不同的选择也是较为合理的。我们在“3.1.3 实验流程”部分详细介绍了社会规范的操纵方式，请审稿专家批评指正。

④ 最后，数据结果表明我们对“利他规范/非利他规范”的操纵确实带来了被试选择结果和利他偏好的差异，结合不依赖模型和基于模型的结果，相比于非利他规范，在利他规范下被试的 AIA 更高，DIA 更低，表现出更大的利他偏好，说明被试的行为确实受到了我们所操纵的社会规范的影响。

（3）在上述疑问的基础上，我对实验 2 中社会规范的操纵持保留意见，而作者也并未进行操纵有效性检验，这让实验 2 结果的说服力不足。

回应：感谢审稿专家的批评，事实上我们对实验二的社会规范操纵进行了一系列的操纵检验。在每一个区块的实验(共 4 个区块)结束后，被试需要对三句陈述进行评价，并进行 1-5 打分，1 表示“完全不认同”，5 表示“完全认同”。三句陈述分别为“1.在刚才的分配中，我认为之前参与游戏的玩家比较友善”“2.在刚才的分配中，我认为之前参与游戏的玩家比较自私”“3.在刚才的游戏中，我的行为受到了之前游戏参与者的影响”。得到四个条件下被试对这三句陈述的评分后，对其进行统计检验。结果发现，对于第一句陈述，在利他规范条件下被试的评分显著高于非利他规范条件，$F(1,52) = 32.18$，$p < .001$，$\eta_p^2 = .54$；对于第二句陈述，在利他规范下被试的评分显著低于非利他规范条件，$F(1,52) = 37.32$，$p < .001$，$\eta_p^2 = .57$。这一结果说明，在利他规范下被试认为之前的参与者更加友善，在非利他规范下被试认为之前的参与者更加自私，说明我们对的操纵成功地让被试形成了之前大多数人在实验情境下的行为倾向性的认知，即对社会规范的认知。对于第三句陈述，结果发现，四个条件下被试的评分都显著高于 2.5 分，说明被试认为自己的行为受到了社会规范的影响，表明我们的社会规范操纵成功影响了被试的分配行为。

意见 2：关于“接受者结果可见性”我也存在一点疑问：作者在回应中指出，本文研究的是“行为可见”以及“对接受方可见”，这里的“行为可见”在文章中也可以说是“行为结果可见”，那么通俗地说：本文研究的是在行为结果对接受者来说可见/不可见时，个体的行为/偏好有何

差异。但如果行为结果连接受者都不可见，那么该行为到底对哪一方产生了影响？该行为的意义又是什么？作者在引言中补充的例子“分配者知道每一试次是否由自己决定，但接受者只知道分配由分配者决定的概率”，也很难算是操纵行为可见性的研究，因为接受者知道分配结果，只是不知道这个结果是否是分配者做出的。作者自己也提到“这项研究以概率操纵可见性，较不直观，与现实社会的情境相差较大”。那么本文中接受者不知道分配的结果，这种情况是否更是一种脱离现实而只存在于实验室中的情境？而且我想知道实验指导语中对接受者和分配者的报酬是如何说明的？因为分配者有很多次是不知道自己被分配的金额，那么如何说明她/他最终报酬与实验结果之间的关系？

回应：谢谢审稿专家提出的疑问。我们的回复如下：

(1) 本文探究的“可见性”确切来说应该是“在分配者视角下的可见性”，即分配者认为他/她的行为结果能否被其他人看到，以及能否被接受者看到。我们希望探究这种对可见性的认知对分配者行为的影响，即分配者认为自己的行为能够/不能够被接受者看到时，分配者将如何分配。

(2) 在补充的“分配者知道每一试次是否由自己决定，但接受者只知道分配由分配者决定的概率”的例子中，当接受者认为只有极小概率分配是由分配者做出时，在分配者的视角下，相当于是分配者认为接受者不知道这个分配是由分配者做出的，那么分配者就能够做出更加自私的决定而不会损害自己的声誉；这一研究的情境和本研究希望探究的情境是十分类似的，都操纵了分配者对于接受者能否看到自己行为的认知。

(3) 在本实验中，有可见和不可见两种情境，在可见条件下，(被试作为分配者认为)接受者能够实时地看到自己被分配了多少钱；而在不可见条件下，接受者无法实时看到这一轮的分配数额。最终无论是分配者还是接受者，每一轮得到的分数会累加起来，并以一定比例转换为实验报酬。主试会在实验结束后告知被试 ta 一共获得了多少分数，并支付被试费。也就是说，在分配者的视角看来，接受者在一些条件下能够知道这一轮得到多少分数，在一些条件下不知道，但是最终都会知道一个总的分数(但是接受者无法根据总分数来逆向推断不可见条件下被分配了多少分数)。当然，事实上本实验中并没有接受者存在，所有被试都会成为分配者，可见性的操纵实际上只是操纵了被试对于实验情境的认知。在修改稿中，我们对于实验结果和最终报酬的关系进行了更加详细的说明。

(4) 另外，我们认为“接受者不可见”的情境在现实生活中也是较为常见的，譬如匿名捐赠中，被捐赠者只会获得一个总的数额，但不知道具体每一笔钱来自于哪一个捐赠者。生活中许多涉及多人福利的决策都面临是否公开透明的问题，因此研究可见/不可见具有一定现实意义。

意见 3：关于被试身份的匿名问题。作者指出，4-8 名被试同时来到实验室，且通过大屏幕介绍了实验规则，这就意味着在该阶段被试之间是知晓彼此身份的。也就是说对于每一位分配者而言，潜在接受者的身份并非完全匿名，且始终限定在 3 人/5 人/7 人的小范围内；对于接受者而言也是一样。因此即便告知被试游戏是匿名进行的，也很难达到预期效果。

回应：感谢审稿人对于“身份匿名”的疑问。被试在同一实验室参与实验，是为了增加实验的真实性，让被试相信和自己互动的是真实玩家(而实际上本实验中的接受者并不存在，因此需要进行一系列操作来增加真实性)。在实验过程中，被试被告知一半的人是分配者，另一半是接受者，每一轮游戏都需要一名分配者和一名接受者配对。每一轮游戏中，配对关系是不固定的，因此分配者和接受者都无法唯一地定位现在是谁在和自己配对，搭档的性别、外貌等信息无法用于指导决策，因此决策过程可以视作匿名进行，这种匿名处理的方式在社会决策领域较为常见(e.g., Gao et al., 2018)。此外，本实验在招募被试时保证了同时进行实验的被试之间相互不认识，被试不知道一起做实验的其他人的身份信息，保证了被试之间的匿

名性。

意见 4：新稿标题为“行为可见性与社会规范对利他偏好的影响”，这给人感觉行为可见性和社会规范并行地影响了利他偏好，但实际上社会规范在实验 2 中更接近调节变量的角色，即行为可见性对利他偏好的影响在一定程度上取决于存在何种性质的社会规范，我觉得标题可以适当修改以体现出社会规范的这一作用。
回应：谢谢审稿人的建议，我们认同行为可见和社会规范在研究中的作用是不完全对等的，社会规范更接近于是解释“行为可见增加利他偏好”的机制。我们拟将题目改为《行为可见增加利他偏好及其社会规范机制》，请审稿人批评指正。

意见 5：关于前言的论述，有一些可以适当修改的地方。
（1）在第一段中，作者提出了利他偏好的定义，但作者没有说明这个定义是参考现有文献还是自己提出的定义，如果是前者应补充上相应文献；如果是后者，应说明和现有定义相比这一定义的优点。
回应：谢谢审稿人的建议，已在前言部分补充对利他偏好定义的说明。利他偏好的概念源自于社会价值取向理论，将利他偏好定义为“不在意自我收益，只希望最大化他人收益”(Murphy et al., 2011)。然而在较多情况下，利他者并非完全不考虑自我收益(West et al., 2007; Pfattheicher et al., 2022)，因此，参照 Sáez 等人(2015)，本研究将利他偏好被定义为“对他人收益有正向的考虑”。

（2）在第四段中，作者引用“第三方惩罚”来说明第三方观察对个体利他行为的影响。第三方惩罚也被称为利他性惩罚，但尽管被冠以“利他”二字，第三方惩罚的利他属性（即利他性惩罚是否真的利他？）目前仍然是一个有争议的话题（陈思静, 杨莎莎, 2020; Rodrigues et al., 2020），因此我建议这里最好引用通常意义上的利他行为（如帮助或捐赠）而非第三方惩罚来说明自己的观点，从而避免无谓的争议。
回应：感谢审稿人对于引用第三方惩罚是否合适的建议。第三方惩罚的研究有两个关注点：一是第三方在观察到不道德行为时的惩罚行为(第三方的行为)；二是当存在第三方且第三方具有惩罚可能时，分配者将如何根据此信息调整自己的行为(分配者的行为)。本文希望引用的是第二点，即希望说明“第三方的存在对分配者利他行为的影响”，而不是探究“第三方的行为是否属于利他范畴”。抱歉我们在原文中叙述得有些模糊，可能导致读者产生误解，已在修改稿中进行说明。

（3）假设 2 的提出方式我觉得不够妥当，在提出假设前，作者主要在阐述评分和选择这两种测量方式的差异，但假设却是：行为可见增加利他偏好在选择和评分两种反应类型中均成立。前后的逻辑衔接不够顺畅。
回应：感谢审稿人的建议。本文使用选择和评分两种方式，主要目的是探究行为可见对利他偏好的影响是否在多种测量方式下均成立，因此在修改稿中，我们在前言部分删去了讨论选择和评分的差异的内容，改为在讨论部分再详细讨论两种测量方式的差别。

（4）在引入社会规范这一概念时应对社会规范的定义、分类和如何测量做一个简单的介绍，因为并不是每个读者都了解这块内容。
回应：感谢审稿人的建议，已在修改稿的前言部分补充对社会规范的介绍。

意见 6：在被试量的计算中作者应报告所参考的效应量。

回应：感谢审稿人对于被试量计算的建议，已做出相应修改。

意见 7：作者区分了利他行为和利他偏好，并用了两种方式来处理数据，这给人以启发，但关于这一做法我仍然有若干疑问。

（1）作者用两种指标来测度利他，即利他行为和利他偏好，这是本文的核心内容之一，但作者主要论述了利他行为/偏好在方法论上的区别，而没有充分讨论它们在概念上的差异。换言之，作者是否认为利他是一个多面向的概念（multi-faceted construct），而这两个指标分别体现了利他这个概念的不同面向？我希望作者能更深入讨论这个话题。

回应：谢谢审稿人提出的疑问。利他行为和利他偏好的差异可以对应于我们在前言第 8 段所叙述的“行为”与“偏好”的差异：利他行为指人们外显表现出的倾向选择增加他人收益的行为，利他偏好指人们内隐的希望增加他人收益的倾向性。在本研究(以及多数决策研究)中，引入偏好是为了揭示行为背后的认知计算机制，通过计算建模的方式推导出人们在大脑中进行认知加工的“算法”。诚然，这种对偏好与行为的区分确实没有为“利他”的概念本身提供更多信息，或许在后续研究中可对此进行深入探索。

（2）利他行为和利他偏好这两个指标之间是什么关系？比如，我们是否可以合理地认为利他偏好在一定程度影响了利他行为？抑或是相反？又或者两者是独立的？增加这方面的论述可以让读者更好地理解采用两套指标的必要性。

回应：感谢审稿人的提问。参见上一个问题的回答，我们认为偏好反映了行为背后的认知加工机制，利他偏好和利他行为是决策过程中的上下游关系；然而，根据定义，利他行为是主试可以测量的变量，而我们认为偏好是引发行为的隐变量，无法直接测量，故此通过建模方式对偏好进行估计。在测量指标上，行为指标和偏好指标体现为高度相关的关系。我们认同审稿人所述，利他偏好很大程度影响和导向了利他行为。我们已在修改稿的前言和方法部分进行一系列语句的调整，以帮助读者更加理解行为和偏好的关系。

（3）作者用选择和评分两种方式在实验 1 中测量利他行为，事实上评分测量的应该是态度而非行为，具体而言是对某个特定方案的满意度。如果从概念上讲，评分的测量更接近偏好而不是行为。

回应：十分感谢审稿人的建议。我们认同评分反映的是个体对特定方案的满意程度，然而在本研究中，我们依然希望将评分本值划分为行为指标。态度一旦体现在评分中，便成为了外显的行为指标，而不再是内隐的偏好。在决策研究中，对行为和偏好的区分主要在于“外显、离散”或是“内隐、连续”，无论是评分还是选择，都只是进行了局部的行为采样，只有通过建模的方式，才能得到内隐的、连续的偏好。

参考文献

陈思静, 杨莎莎. (2020). 利他性惩罚的动机. 心理科学进展, 28(11), 1901–1910.

Blanton, H., K?blitz, A., & McCaul, K. D. (2008). Misperceptions about norm misperceptions: Descriptive, injunctive, and affective ‘social norming’ efforts to change health behaviors. Social and Personality Psychology Compass, 2(3), 1379–1399.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. Advances in Experimental Social Psychology, 24, 201–234.

Dempsey, R. C., McAlaney, J., & Bewick, B. M. (2018). A critical appraisal of the social norms approach as an interventional strategy for health-related behavior and attitude change. Frontiers in Psychology, 9, 2180.

Palacios, J., Fan, Y., Yoeli, E., Wang, J., Chai, Y., Sun, W., ... & Zheng, S. (2022). Encouraging the resumption of economic activity after COVID-19: Evidence from a large scale-field experiment in China. Proceedings of the National Academy of Sciences of the United States of America, 119(5), e2100719119.

Rodrigues, J., Liesner, M., Reutter, M., Mussel, P., & Hewig, J. (2020). It's costly punishment, not altruistic: Low midfrontal theta and state anger predict punishment. Psychophysiology, 57(8), e13557.

审稿人 2 意见：

作者极大程度上回答了我先前的问题，同时实验 2 的增加使得本研究更有说服力。此外感谢作者将数据与分析代码上传到 OSF。 不过我还有一些问题，希望作者予以考虑。

意见 1：实验 1 对于评分的模型，公式 5/6 都用到了 e(psilon)，但在 jags 代码中并未体现。我认为完整起见，e 也需要作为一个自由参数被估计。当前的 sample statement: rating ~ normal (R, 1)，假设正态分布的 SD=1，在评分数据有潜在的 outlier 时可能并不能进行很好的拟合。

回应：感谢审稿专家对于实验 1 建模方式的建议。我们已按照审稿人的建议进行修改，将正态分布的标准差也作为模型中的自由参数，发现拟合出的 SD 在 1~1.5 左右。此外，我们还对 HBM 模型进行了如下修改：①假设选择条件下的温度参数 λ、评分条件下的 b0、b1、e 在可见/不可见条件下不同，这一模型对数据的拟合效果更好(原模型 DIC = 24158.16，现模型 DIC = 21240.21)，说明可见/不可见除了影响 α、β，也会影响其他一系列决策参数 ②将群体分布的均值的先验分布从均匀分布改为正态分布，这样参数的先验取值不会被局限在一个小的区间，而是可以取实数域上的任意值，这样的设置对参数取值有更强的包容性，在本研究中更加合理。进行以上修改后，结果和之前大致相同，但发现反应类型(选择 vs 评分)在可见条件下对 AIA 的影响从不显著变为显著，所以我们删去了讨论部分之前对于不显著结果的阐释。

意见 2：实验 2 的参数估计使用了最大似然估计（LME）；但是既然在实验 1 中使用了贝叶斯分层估计，为何在实验 2 使用 LME？这使得前后文的一致性大大降低，建议实验 2 的模型分析也采用贝叶斯分层估计。

回应：感谢审稿人的建议，修改稿中我们已将实验二参数拟合的方式改为了分层贝叶斯估计，模型结果和不依赖模型的结果类似，详情请见修改后稿件。

意见 3：对于实验 2 模型参数的报告，除了图 6d，最好将原始的参数进行报告，而不仅是做差。同时，为了更好地解释实验数据并探究认知计算机制，可以考虑调节分析，即社会规范（利他与否）作为调节变量，来考察 model extracted trial-by-trial utility（自变量）与被试利他选择（因变量）之间的关系。

回应：感谢审稿人的建议，修改稿中我们在图 7 报告了 HBM 模型的参数的原始取值。此外，我们按照审稿人的建议，建立了如下线性模型：

$$choice \sim utility + type + norm + norm \times utility$$

其中，choice 为是否选择公平选项(选择公平选项取 1，选项不公平选项取 0)，utility 为根据建模结果的 AIA、DIA 和 Ms、Mo 计算出的效用(公平选项 – 不公平选项)，type 代表不公平类型(优势不公平取 1，劣势不公平取 0)，norm 代表社会规范条件(利他规范取 1，非利他规范取 0)。结果发现，在控制了 utility 后，norm 的主效应(p = .285)和 norm 与 utility 的交互作用(p = .063)都不显著。对此我们认为，由于在 utility 的计算中使用了不同条件下的 AIA、DIA 参数，而在建模中的参数估计值在很大程度上已经捕获（或体现）了不同社会规范下的

决策差异，所以控制 utility 之后，norm 对 choice 不再有影响，norm 也不调节 utility 对 choice 的作用。换言之，norm 对于 choice 的影响已经体现在改变的 AIA 和 DIA 参数中，也体现在由 AIA 和 DIA 决定的选项 utility 中。

第三轮

审稿人 1 意见：

作者回复了审稿意见，对我在上一稿中的疑问做了回答，但以下三个问题我觉得作者的解释并不是很充分。

意见 1：关于实验 2 的操纵检验。上一稿原件中我并没有看到任何有关操作检验的内容，但作者却在回复中告诉我“事实上我们对实验二的社会规范操纵进行了一系列的操纵检验”，那为什么不在原稿中报告这一部分内容呢？等外审指出这一问题后再提供数据这一做法恕我无法接受，因为这无法有效保证数据的真实性。

回应：感谢审稿专家对于操纵检验的质疑，请允许我们对未在原稿中报告该部分结果进行解释。在本研究中，我们操纵社会规范的方式并不是独创，而是借鉴已有研究(e.g., Raihani & McAuliffe, 2014; Kawamura & Kusumi, 2017)，前人研究已充分证明操纵描述性规范能够改变被试行为，并采用类似的事后报告作为操纵检验。因此在我们的研究中操纵检验并不是必须的，只是作为辅助验证。此外，由于篇幅限制，我们并未在上一稿中报告操纵检验结果，但该部分数据确实是在实验过程中收集的真实数据。

意见 2：关于实验 2 的设计。作者认为由于社会偏好的个体差异较大，所以采用了被试内设计，但这个理由无法让人信服。如果被试的个体差异较大，我们可以采用增加被试数量或配对等方式来降低个体差异的影响。当然，如果能有效排除学习的影响，那被试内设计确实是更好的选项。但问题是本文的被试内设计并没有很好地控制学习的影响，即我在上一稿意见所提及的如何排除“接受了一种规范信息”对“接受另一种完全不同的规范信息”的影响？作者在回复中提到，在操纵检验时，被试需要对三句陈述进行评价，并进行 1-5 打分，1 表示“完全不认同”，5 表示“完全认同”，其中一个陈述是“在刚才的游戏中，我的行为受到了之前游戏参与者的影响”。作者给出的检验结果表明被试确实受到了规范信息的影响，那我们假设被试首先接受利他规范信息，接着又接受非利他规范信息，那么被试在接受第二种规范信息时（非利他）如何排除第一种规范信息（利他）的影响？反之亦然。

回应：感谢审稿专家再次对实验设计提出疑问，我们更加了解了您的问题的核心，请允许我们进一步阐释使用被试内设计的理由，以及如何处理学习的影响：

(1) 我们当前研究中使用的与社会规范相关的被试内操纵在近年的研究中都多有使用(e.g., Beltzer et al., 2019; Deffner et al., 2020; Aberg et al., 2021)，这些研究都提示被试在不同的环境中会根据环境的变化适应性的调整自己的行为。以我们当前的利他规范为例，当被试从一种社会规范情境切换到另一种社会规范情境中时，在适应新的社会规范的过程，被试会在上一规范的影响的基础之上根据当前规范的信息调整自己的偏好，直到被试主观认为适应当前社会规范情境。如果前后两种规范具有足够的区分度，我们就有可能观测到两种规范下的偏好差异。

(2) 同时，在实验设计上，对于每个被试，我们尽量对两种社会规范情境进行区分,以降低前一社会规范情境对后一情境的影响。i)我们在实验中营造了两种完全不同的环境，即不

同的游戏服务器; ii)两种环境之间使用区块设计; iii) 我们也分别使用绿色和橙色的文字来代表不同的服务器，以此区分不同的规范条件。

(3) 此外，不同社会规范区块的呈现顺序在被试间进行了随机化处理，即一部分被试先接受利他规范、再接受非利他规范；另一部分被试先接受非利他规范、再接受利他规范。根据实验心理学的方法，这种随机呈现能够平衡不同条件出现顺序的影响。这也能够直接回答审稿专家担心的问题，即社会规范之间的顺序效应能够通过被试间的随机化消除。

(4) 从被试表现来看，我们的操纵性检验通过被试的主观汇报证明被试确实受到了实验情境操纵的影响，而基于行为的指标和计算建模的指标都表明这种被试内社会规范的操纵是有效的。

意见 3：关于态度与行为的问题。我在上一稿意见提到，作者所测量的“行为”实际上是“态度”，作者认为“态度一旦体现在评分中，便成为了外显的行为指标，而不再是内隐的偏好”。这个理由恕我无法接受。首先，态度——至少是外显态度——多采用评分的方式，为什么“态度出现在评分中，便成为了外显的行为指标”？如果这样的话，那为什么心理学研究还要测量行为意向和实际行为？其次，事实上，态度-意向-行为问题（attitude-intention-behavior gap 或 attitude-intention-behavior relations）是心理行为科学中的一个经典问题，这三者间虽然存在一定关联，但绝不是同一个东西，如广为接受的计划行为理论明确区分了这三者。关于这一点，我的意见很明确：我们无法通过询问“对…表示满意/不满意”来测量行为。再者，作者在回复中讨论了偏好和行为之间的关系，认为“利他偏好和利他行为是决策过程中的上下游关系”，但如果作者测量仅仅是态度而非行为，那么作者所说的关系还存在吗？换言之，作者应讨论态度和偏好的关系，而不是行为和偏好。

回应：感谢审稿人的疑问，事实上本文的写法和审稿人所列举的理论并不矛盾。我们认同在传统心理行为科学中，“态度”和“行为”有着明显的区分。而在当前研究中，我们延续了大多数计算建模研究的习惯，将指标分为“行为指标”和“模型指标”。此处的“行为”暗含着非模型指标的意思，更加广义，指代被试在实验中的诸多外显反应；而态度-意向-行为中的“行为”更加具体；二者在操作性定义上或许有所不同。我们已对文章进行修改，尽量用“行为指标”来指代选择和评分，而避免将评分称为“行为”。本文中，将评分称为行为指标具有合理性，且也具有偏好和行为在决策过程中的上下游关系：被试通过内在的偏好，决定对选项的外显评分。 态度和行为受到可见性的影响是否具有一致性确实是我们在研究中探讨的问题之一，我们在文中将之称为反应类型差异，态度对应评分，而行为对应选择。文中的态度或者说满意程度评分确实更接近偏好的概念，它是个体对偏好的一种外显表现，从我们的设计上说它是从不满意到满意几个离散值。我们将之称为“行为指标”，即被试外显表型出来的指标。通过计算建模，我们可以利用满意度对被试在有利和不利不公平条件下的利他程度进行更为细致地连续刻画，并且使得其在不同情境和范式之间具有可比性。 评分（态度）和行为（选择）的关系，也是我们研究试图回答的问题。正如我们在前言中所述，评分和行为两种方式在以往研究中都被广泛用于利他研究，但是对于两种测量方式所揭示的利他程度会有怎样的差异，以及受到情境影响的变化是否具有一致性都还没有研究进行过探讨。在本研究中，我们的效用建模的形式使得两种情境具下的利他偏好具有可比性，我们在图 4 中呈现了结果，并在正文中进行了汇报，两者确有共性也有各自的特性。简而言之，行为可见增加利他偏好，在评分和选择间具有一致性；并且，相比于评分，在选择时被试更加在意分配的效率。在讨论部分，我们也对其差异背后的原因进了探讨（4 讨论，第四段）。

参考文献

Aberg, K. C., Toren, I., & Paz, R. (2021;2022;). A neural and behavioral trade-off between value and uncertainty

underlies exploratory decisions in normative anxiety.*Molecular Psychiatry, 27*(3), 1573-1587. https://doi.org/10.1038/s41380-021-01363-z

Beltzer, M. L., Adams, S., Beling, P. A., & Teachman, B. A. (2019). Social anxiety and dynamic social reinforcement learning in a volatile environment. *Clinical Psychological Science, 7*(6), 1372-1388. https://doi.org/10.1177/2167702619858425

Deffner, D., Kleinow, V., & McElreath, R. (2020). Dynamic social learning in temporally and spatially variable environments: Dynamic social learning. *Royal Society Open Science, 7*(12)https://doi.org/10.1098/rsos.200734

Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P., & Robinson, O. J. (2017). Modelling avoidance in mood and anxiety disorders using reinforcement-learning. *Biological Psychiatry (1969), 82*(7), 532-539. https://doi.org/10.1016/j.biopsych.2017.01.017

……

审稿人 2 意见：

再次感谢作者采纳之前的审稿意见，文章的质量有了很大提高，数据结果也更加支持研究假设与实验结论。我没有额外的问题，建议发表，并恭喜作者。

回应：十分感谢审稿专家从始至终对我们的帮助与建议，您的宝贵建议极大帮助我们提高文章质量，精炼分析方法，梳理文章逻辑。我们已将更新后的数据和程序代码上传到 github，链接同前：https://github.com/psych575/open-data-and-code-for-xb21-575.git。再次致以我们真挚的感谢。

编委意见：作者们好，感谢你们对研究和文章做出的诸多修改。修改后的文章和你们对审稿人意见的回复我都仔细看了，觉得研究在添加一个实验后，丰富了很多，也对先前实验的结果进行了重复，还拓展了研究范围，增加了对社会规范的讨论，对文章整体质量的提高帮助明显。虽然如此，审稿人 1 还有一些质疑，尤其是第二和第三点，希望你们能对文章进一步修改，体现对这些质疑的回复。另外，文章现在过长，应该删减一些不重要的内容。比如说，研究的亮点是模型和对模型参数的比较，因此可以不用对行为结果汇报过多，尤其是实验一。最后，文章的书写质量还可以进一步提高，尤其是新增的内容。重要问题是太过琐碎，使得重点不够突出，建议进一步修改、精简。祝好，责任编辑

回应：十分感谢编委老师的建议，我们已对审稿人 1 的建议进行回复以及做出相应修改。此外，我们已按照您的建议，对实验一的行为结果进行了精简，并对文章总体的书写进行修改，合并、删除了一些重复内容，并对语句进行精炼，使重点更加突出。

第四轮

审稿人 3 意见：

本文探讨了行为可见性对利他倾向的影响，发现在行为可见时人们更可能利他，并检验了该效应的社会规范机制，即存在不利他社会规范时，行为可见对利他的促进作用减弱。研究亮点在于运用认知计算建模及比较模型参数来回答核心问题，行文逻辑流畅，严谨规范。作者对前几轮的审稿意见作出了较充分的回应。但是，本文的弱点是理论创新有限，尚有一些理论问题需要作者再考虑。

意见 1：作者提出本文的贡献之一在于探讨“行为对接受者可见”，区别于以往研究的“行为对第三方可见”，但没有具体说明理论上“行为对接受者可见”有何独特性，本研究如何帮助理解不同观察者的差异。因此，本文聚焦于“行为对接受者可见”的理论意义显得有些薄弱，作者需要对此作更具体的说明。也许“行为对接受者可见”的理论特殊性在于：第一，接受者卷入度高，比第三方旁观者更在意分配结果，而分配者也能意识到这一点。第二，分配者若选择优势不公平方案，直接损害接收者的利益，但不会直接损害第三方旁观者的利益。这两个特点可能放大分配者对优势不公平的厌恶。

回应：非常感谢审稿专家关于文章理论价值的一系列建议，您的建议对于我们梳理和突出文章的理论意义具有非常大的帮助。本文章对“接受者可见”有何理论独特性的说明确实不够充分，感谢审稿专家指出以及提供了十分有效的建议，我们已在前言部分对其进行补充。

意见 2：作者在本研究问题中将观察者身份限定为分配方案的接受者，同样存在上述问题，与第三方视角相比，研究接受者视角带来了哪些结果差异以及理论创新？建议作者对上述问题进行思考，这有助于展现研究的理论贡献。

回应：非常感谢审稿专家的建议。诚如审稿专家所言，相比于第三方，接受者与分配者利益相关，因此接受者观察相比于第三方观察，对于分配者行为的驱使可能不同。作为补充，我们在讨论部分引用了 Bradley 等人(2018)的研究结果，对两种观察者身份带来的可能差异进行了更多的阐释，请审稿专家批评指正。

意见 3：在引言中，作者认为利他具有信号功能，人们通过利他塑造自己的良好形象，因此利他行为可见是其发挥信号功能的前提之一。然而，本研究中，即使在“行为可见”时，游戏也是匿名进行，观察者不会知道是谁作出了利他行为，此时被试应当没有必要发出利他信号。如果按作者所说，利他是为了传递信号，如何理解被试即使自知匿名，也在被观察时更利他？

回应：谢谢审稿专家提出的疑问。在涉及社会互动的实验中，保持被试间的匿名性主要是为了将实验和被试日常生活中的身份分隔开，使得被试的行为动机尽可能仅来源于实验因素，因此本研究采用了匿名操作。在 Hardy 和 Van Vugt(2006)的研究中，实验同样匿名进行，被试不知晓彼此的身份，但被试间存在多次互动。而本研究中，每一试次都会重新随机配对，因此分配者和接受者只存在单次互动。本研究结果正好能够说明，即使只存在单次互动，被试仍然会在可见条件下做出更多利他行为，说明在单次互动中人们同样在意自己的形象、愿意发送利他信号。在利他的信号假说中，匿名与否并不是最重要的因素，因为即使实验匿名进行，被试在实验中依然是一个独立的、主动的行为主体，接受者也能够知晓行为来自于这一行为主体，那么被试就能够感知到为自己行为负责、维护形象的需求。通俗而言，即使某人处于全是陌生人的环境，当有人向其寻求帮助，该人也会考虑到维护自身的良好形象而给予帮助。对于利他作为信号这一假说，我们认为更为重要的因素是行为是否可见，当行为可见时，人们发送利他信号才有意义。

意见 4：作者认为“利他是社会规范”导致人在行为可见时利他，并将“利他以维护声誉”作为竞争假设。两者从理论上如何区分？人也可能依据社会规范采取利他行为，以表明自己遵守规范，从而塑造良好形象、维护声誉。此外，作者在引言中提出的利他的信号功能也较符合“维护声誉”的假设。 总体而言，不论是“社会规范”解释，还是“声誉”解释，似乎都是社会赞许性的体现，即人倾向于按社会所期望的方式行动，并且行为可见时这种倾向更强。作者需要考虑两种假设有何本质区别，在理论推导部分对社会规范、利他的信号功能、维护声誉等概念之间的关系进行更清晰的阐述。

回应：感谢审稿人对于本研究心理机制的讨论和建议。我们在讨论部分介绍了声誉对行为可

见促进利他的可能作用方式。请允许我们在这里说明，我们的本意并不在于将声誉作为与社会规范对立的竞争性假设，在我们研究中也没有通过实验操纵的方式对两者进行讨论，我们已经对相应的行文阐述进行修改。同时，我们非常认同您的观点，即社会规范和声誉似乎都是社会赞许性的体现。不管是社会规范本身还是声誉，其结果都使个体更多的做出了符合社会期许的行为，向社会传递了自己的亲社会性，体现了利他的信号功能，只是在研究社会规范和声誉的影响时两者通常对应不同的实验操纵形式(e.g., Piazza & Bering, 2008)。我们在讨论部分将社会规范、利他的信号功能和维护声誉等概念之间的关系进行了更清晰的阐述。

意见 5：关于计算建模“行为指标”的表述问题，作者可以考虑用脚注注明本研究中“行为指标”的意义，方便读者理解。
回应：十分感谢审稿专家的建议。为避免混淆，我们在正文中统一使用“不依赖模型的指标”来替代“行为指标”。

意见 6：研究 1 为 2 (行为可见性：可见 vs. 不可见) × 2 (反应类型：选择 vs. 评分)被试内设计。但根据研究流程与结果分析，不公平类型似乎也是想要探讨的自变量之一，因此实验设计是否应为 2 (不公平类型：优势 vs. 劣势) × 2 (行为可见性：可见 vs. 不可见) × 2 (反应类型：选择 vs. 评分)被试内设计？
回应：感谢审稿专家的建议，我们确实在实验一的结果分析部分将不公平类型也当作一个因子进行多因素方差分析，已按照审稿人的建议进行修改。

意见 7：研究 1 实验设计与流程中，作者提到“被试作出反应后，进入反馈阶段。可见条件下……不可见条件下……”。此描述容易产生歧义，似乎是被试作出选择和评分后，才知道自己处于可见还是不可见条件。

回应：感谢审稿专家指出这一漏洞。实际实验过程中，被试在每一个区块开始前就会被提示这一区块是“可见”还是“不可见”条件，而不是做出选择后才知晓。已在方法部分补充这一细节。

意见 8：根据结果分析，研究 2 用分配点数表示利他偏好，但相关点数分配的内容并未在研究流程中体现，建议补充。
回应：感谢审稿人提出疑问。研究 2 的因变量之一“平均分给他人的点数”，即指代二选独裁者游戏中，被试选中试次中他人的平均收益，这一指标仍然是通过二项选择来体现，并非其他单独的测量。抱歉我们在原文中的模糊叙述导致理解有误，已在“3.1.3 测量指标”部分进行补充和说明。

意见 9：结果图表下显著性标注 p 需斜体。
回应：十分感谢审稿专家指出错误，已对所有图片下方的 p 值标注进行修改。

意见 10：章节 3.2.1 第 2 段 3 行，“社会规范的主效应显著”与“F”之间应有逗号连接。
回应：感谢审稿专家勘误，已修正。

第五轮

审稿人 3 意见：

作者较好进行了修改，突出了研究的理论意义。本文已达到《心理学报》发表要求，建议发表。

回应：十分感谢审稿专家的宝贵建议及对我们文章的支持！

编委意见：作者们好，感谢对之前评审意见和审稿人 3 的意见的回复。文章在之前基础上又修改了不少，尤其是在理论方面得到了显著的加强，总体已经达到发表的水平。我唯一的意见是希望在对文章做进一步的精简，现在的篇幅还是过长，对阅读的挑战比较大，与文章并不复杂的 message 不相符。

回应：十分感谢编辑老师，我们已对文章的篇幅进行再次缩减，尤其对方法、结果和讨论部分进行了精简，删去了一部分与文章主线关联较弱的、较不重要的内容，并对全文的语言进行精炼，简化长句，以提升阅读的流畅性、降低阅读难度。当前文章正文字数在 14000-15000 字左右，各部分的篇幅均符合杂志要求。

主编意见：

本文通过使用修改版的独裁者游戏，考察了行为可见性对个体利他偏好（选择、评分）的影响。结合行为分析与计算建模、实验方法，主要结果发现，行为可见情境下，个体对优势不公平的厌恶程度增加，且对劣势不公平的厌恶程度减小。此外，研究引入社会规范作为调节，发现行为可见增加利他偏好的作用依赖于利他的社会规范，在利他社会规范下，当行为对接受者可见时，人们将表现出更多利他偏好。该研究为行为对接受者可见时，利他决策的研究提供了新观点和证据，有一定的理论贡献。经过多轮的修改和完善，达到心理学报的发表要求，同意发表。