《心理学报》审稿意见与作者回应

题目：跨情境的刺激泛化在面孔信任形成中的作用：基于直接互动与观察学习的视角
作者：袁博；王晓萍；尹军；李伟强

第一轮

审稿人 1 意见：

《跨情境的刺激泛化在信任形成中的作用》通过 3 项实验考察了跨情境（公平—信任）的刺激泛化在信任形成中的作用。实验 1a 和实验 1b 分别从亲历者和观察者视角，发现了信任形成中跨情境的刺激泛化效应。实验 2 结果发现，对行为意图的感知参与了刺激泛化效应的产生；在无意图条件下，上述跨情境的刺激泛化效应消失。上述结果表明，个体采用联结学习机制将不同情境中习得的声誉信息泛化到新的互动情境中，进而指导其随后的信任决策。该研究设计新颖，数据处理方法恰当，前言的写作逻辑性强。有些问题尚需进一步修正。

意见 1：作者用泛化这一信任领域比较新颖的概念。这一概念与相似性有何区别？泛化一般是一个事物 A 上的特点，作用到其他相关联的事物 B 上。但是相似性更强调事物属性的类似之处。因此，泛化和相似性之间的关系，有待于进一步说明，研究到底是泛化还是相似性？它们对应的理论意义是否有所区别？

回应：感谢审稿专家的意见。刺激泛化(stimulus generalization)是指价值可以在感知上或概念上彼此相似的刺激之间传播和转移。因此，产生泛化的基础是基于两个刺激在某一属性上的相似性。比如，当陌生人与熟悉面孔共享某些属性时，熟悉面孔的的心理表征被自发激活，从而导致个体对陌生面孔产生与熟悉面孔相同的评价(Kraus et al., 2010)。Kocsor 等人(2017)研究表明，泛化效应的发生需要基于面孔物理特征的相似性，当陌生面孔与熟悉面孔相似时，个体通过加工面孔相似性信息，形成对陌生面孔的印象，进而导致个体对陌生面孔思想、情感和行为等的转变。在现实生活中，刺激很少以完全相同的形式出现，这种基于相似性的刺激泛化机制具有高度适应性。我们在修改稿的引言部分对刺激泛化的概念进行了一定的补充（*p2* 的引言部分）

意见 2：不公平相比于公平泛化更快，还是公平本身就是个天花板，两种情况下泛化的可上升空间本身就是不同的？

回应：感谢审稿专家的意见。本研究发现，对不公平分配者面孔的泛化强度高于对公平分配者面孔的泛化强度。从实验数据上看，在变形面孔与先前互动面孔相似度较低的情况下(23%, 34%)，被试选择变形面孔的比例大约 50%。这表明，在公平和不公平条件下，刺激泛化（上升或下降）的空间是等同的。此外，在变形面孔与公平分配者面孔相似度较高的情况下(78%)，被试选择变形面孔的比例大约 75%，仍具有可上升的空间（图 1 (A)）。因此，不公平相比于公平泛化更快，并非公平本身就是个天花板，导致没有刺激泛化产生的空间。

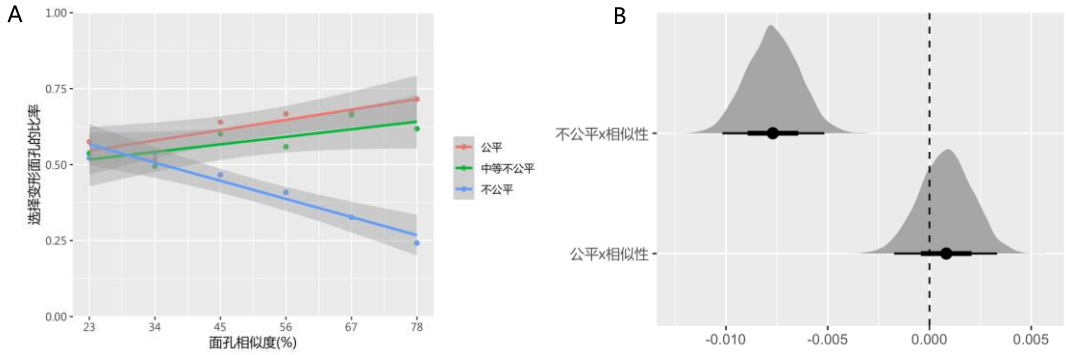

图 1 (A) 不同面孔联结类型下，面孔相似度对面孔选择比率的回归分析；(B) 公平/不公平面孔联结×面孔相似度回归系数的后验概率分布及相应的可信度区间 *CIs*。

我们认为，这种泛化的“不对称性”是一种进化上的适应机制。以往研究也发现，在令人厌恶的领域(aversive domains)存在更强的泛化，因为通常错误地将危险的刺激识别为安全的比将安全的刺激视为危险的代价更大(Bateson et al., 2011)。在对社会行为的观察学习中也普遍存在消极偏向(negativity bias)，即相比于积极刺激，人们更容易注意消极刺激，进而形成更加复杂的认知表征(Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001)。Feldmanhall 等人(2018)研究中发现，与值得信任的面孔相比，被试对不值得信任的面孔存在着更强的泛化，在令人厌恶的领域存在着更强的泛化。因此，在对不同公平程度的面孔泛化中也存在不对称效应，个体对不公平面孔的泛化程度高于对公平面孔的泛化程度。针对专家的问题，我们在讨论中，增加了对泛化的“不对称性”的解释（*p*23 的讨论部分）。

意见 3：讨论部分还要更细致和深刻的评述，整个讨论与前言和方法结果相比，还需要进一步打磨和提升。

回应：感谢审稿专家的建议。根据建议，在修改稿已对讨论部分进行了修改（*p21~p23* 的讨论部分）。

……

审稿人 2 意见：

这篇文章提出来的问题本身有一定意义，通过软件对不同面孔进行 morph 从而研究可能的面孔相关的社会行为/偏好的学习和应用，整体而言也是一个 well-posed 的科学问题。文章沿用了 Feldmanhall 的范式，只是把该文章第一阶段的 trust game 改成了 ultimatum game，认为这样改了之后可以说明刺激泛化可以跨情境，但这一点很牵强。具体意见如下：

意见 1：ultimatum game 的操作有问题，是不是泛化，有没有泛化，什么被泛化，没有说清楚。

回应：感谢审稿专家提出的问题。刺激泛化(stimulus generalization)是指价值(value)可以在感知上或概念上彼此相似的刺激之间传播和转移。产生泛化的基础是两个刺激在在某一属性上的相似性，比如，当陌生人与熟悉面孔共享某些属性时，熟悉面孔的的心理表征被自发激活，从而导致个体对陌生面孔产生与熟悉面孔相同的评价(Kraus et al., 2010)。在本研究中，我们旨在考察跨情境的刺激泛化在对陌生他人信任形成中的作用，即个体在前一个互动情境中形成的刺激与价值之间的联结，是否会泛化到随后互动情境中的信任决策中。因此，我们需要设置刺激联结(conditioning phases)和刺激泛化(generalization phases)两个阶段。

在刺激联结阶段（最后通牒博弈 ultimatum game），我们让被试与三个不同公平程度（公平、中等不公平、不公平）的分配者进行最后通牒博弈，通过接受到不同公平程度的分配提议，被试会形成对三个不同公平程度分配者面孔的刺激价值联结。为了检验联结的效果，我们对被试在不同公平程度下的接受率、分配方案引发的愉悦情绪以及对分配者（面孔）表现出的公平程度总体评分进行了分析。结果发现，不同条件下对分配者提议的接受率、愉悦情

绪以及分配者面孔的公平程度评分均存在差异显著。这表明，被试能够将分配者的面孔与不同公平程度之间形成刺激价值联结，最后通牒博弈操纵是有效的。

在随后的刺激泛化阶段，被试需要在两个面孔（变形 vs.匿名）之间选择其中一个作为信任游戏的搭档。我们通过操纵变形面孔与之前最后通牒博弈中分配者面孔的相似度（23%、34%、45%、56%、67%、78%），检测与先前互动中对不同公平程度面孔形成的刺激价值联结，是否会泛化到对不同任务情境下知觉相似的变形面孔的信任决策中。因此，在该任务中，对与之前互动对象面孔具有不同相似度面孔的信任选择是检验刺激泛化的指标。结果发现，相对于中等不公平面孔，随着与不公平面孔知觉相似性的增加，被试更少选择变形面孔进行接下来的信任游戏；相对于中等不公平，随着与原始公平面孔知觉相似性的增加，被试会更多的选择变形面孔进行接下来的信任游戏。

综上，最后通牒博弈操纵的目的是为了让不同分配者的面孔刺激与价值形成联结，然后考察这种联结是否会泛化到随后互动情境中对陌生人的信任决策中。我们的研究结果表明，个体采用联结学习机制将不同情境中习得的刺激价值联结泛化到新的互动情境中，进而指导其随后的信任决策。针对专家的问题，我们在引言最后一段，对刺激泛化的实验研究逻辑先进行了总结说明（*p5* 的引言部分）。

意见 2：文章写的啰嗦琐碎，详略不当，方法与结果混杂，每个实验以及相关数据分析缺少必要的假说和 justification，结果和结果之间也缺少必需的逻辑过度，很多时候是结果的罗列。像一个本科生的实验报告，而不是一个中国最好心理学刊物的论文。

回应：感谢审稿专家的意见。我们对结果和方法部分进行了调整，对实验以及相关数据分析的逻辑再次进行了说明和梳理。结果和结果之间也加上了逻辑过渡的解释（p9~p10、p15、p19~p20 的方法和结果部分）。

意见 3：为什么用 ultimatum game 加上 trust game？这种随机匹配下的 ultimatum game 里，接受者学习到的到底是啥？影响 trust 行为的因素有几个？那个和 ultimatum game 里的测量可能有关？除了泛化这个解释还有没有其他解释？其他解释是否可以排除？

回应：感谢审稿专家的意见。如在对您 Point #1 的回复中指出，刺激泛化是指价值(value)可以在感知上或概念上彼此相似的刺激之间传播和转移。为了考察跨情境的刺激泛化在对陌生他人信任形成中的作用。我们需要设置刺激联结(conditioning phases)和刺激泛化(generalization phases)两个阶段。在刺激联结阶段（或称联结学习阶段），我们旨在让被试与 3 名不同公平程度（公平、中等不公平、不公平）的分配者进行最后通牒博弈，形成对 3 名不同公平程度分配者面孔的刺激价值联结。在最后通牒博弈中，虽然是随机匹配的分配者，但被试要与这 3 名分配者重复进行最后通牒博弈。因此，被试通过与这 3 名不同公平程度分的配者互动，习得对面孔刺激与正性或负性价值（公平性）之间的联结。

影响 trust 行为的因素有很多，但这里我们旨在探讨在缺乏直接互动经验的情况下，是什么决定着个体的信任决策。因此，我们只关注基于面孔相似性的刺激泛化的作用。在刺激泛化阶段，被试需要在两个面孔（变形 vs.匿名）之间选择其中一个作为信任游戏的搭档。我们通过操纵其中变形面孔与之前最后通牒博弈中分配者面孔的相似度（23%、34%、45%、56%、67%、78%），检测与先前互动中对不同公平程度面孔形成的刺激价值联结，是否会泛化到对不同任务情境下知觉相似的变形面孔的信任决策中。在我们的分析中，也是将面孔联结类型（公平、中等不公平、不公平）和面孔相似度（23%、34%、45%、56%、67%、78%）作为被试是否选择变形面孔的预测变量进行分析。此外，在构建变形面孔刺激时，我们也对与原始面孔要进行 morph 的面孔的吸引力、可信度和领导力三个维度也进行了额外的控制。因此，最大可能地排除了其他可能的解释。

意见 4：不对称的”泛化“是和公平厌恶相关还是一个非社会的损失相关的现象？

回应：感谢审稿专家的提问。我们认为，这种泛化的“不对称性”是一种进化上的适应机制。以往研究也发现，在令人厌恶的领域(aversive domains)存在更强的泛化，因为通常错误地将危险的刺激识别为安全的比将安全的刺激视为危险的代价更大(Bateson et al., 2011)。在对社会行为的观察学习中也普遍存在消极偏向(negativity bias)，即相比于积极刺激，人们更容易注意消极刺激，进而形成更加复杂的认知表征(Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001)。Feldmanhall 等人(2018)研究中发现，与值得信任的面孔相比，被试对不值得信任的面孔存在着更强的泛化，在令人厌恶的领域存在着更强的泛化。因此，在对不同公平程度的面孔泛化中也存在不对称效应，个体对不公平面孔的泛化程度高于对公平面孔的泛化程度。

针对专家的问题，我们在讨论中，增加了对泛化的“不对称性”的解释（p22~p23 的讨论部分）。

意见 5：DDM 模型拟合好坏需不需要评估？DDM 中 drift rate 小，就说明个体更多积累不信任证据？？？

回应：感谢审稿专家的意见。DDM 模型拟合好坏需要进行评估，在初稿中由于字数的限制，我们没有过多地详细说明模型分析和拟合的细节。再修改稿中，我们增加了补充说明材料，对 DDM 模型拟合的结果进行了说明（见补充材料）。对于 DDM 中 drift rate (v)，其代表的是累积某一选择证据的速率，个体倾向于某一选项的偏好越强烈，信息向该选项积累的速度就越快。DDM 使用选择和反应时分布来描述被试如何累积证据做出选择，在本研究中，我们将被试选择信任编码为 1（上边界），选择不信任编码为 0（下边界）（见图 2）。因此，漂移率 v 可以量化被试通过加工面孔信息获得的有利于选择信任或不信任证据的强度，即对选择信任或不信任的价值权衡程度。

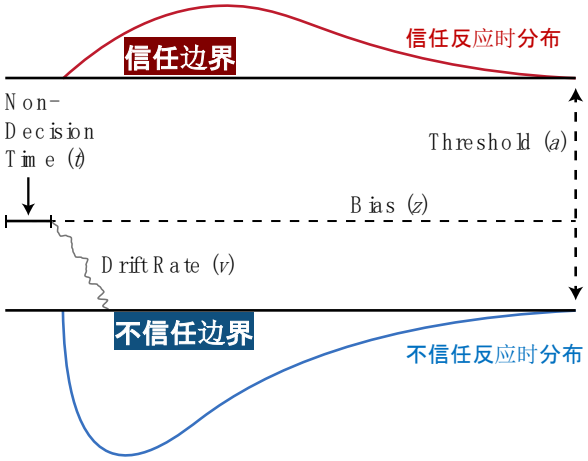

图 2 个体做出信任选择时，漂移扩散模型的示意图

在 DDM 分析中，我们使用分层贝叶斯参数估计(hierarchical Bayesian parameter estimation)的方法同时拟合个体(individual)与群体(group)层面的参数。结果发现，不公平与公平条件间漂移率 v 的差异存在较为可靠的证据(M = -0.51，95% HDI [-0.96, -0.05])，不公平条件下的漂移率 v 显著小于公平条件下的漂移率 v；不公平与中等不公平条件间漂移率 v 存在差异的证据并不明显(M = -0.31，95% HDI [-0.75, 0.13]) (如，见下图 3)。从图 3A 中也可以看出，不公平条件下的漂移率 v 大多分布在小于 0 的区间，而公平条件下的漂移率 v 大多分布在大于 0 的区间。因此，可以在一定程度上表明，在对与先前互动中不公平分配者面孔相似的陌生面孔进行信任决策时，个体更倾向累积不信任的证据。

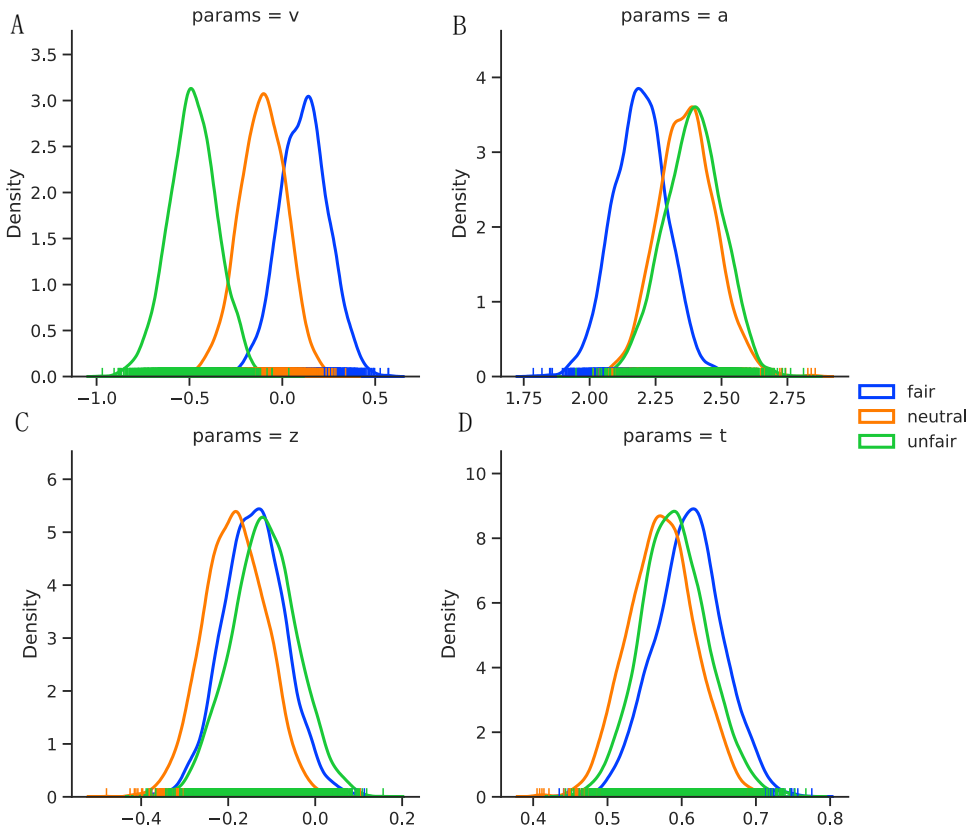

图 3 不同面孔联结类型下，DDM 估计出的 4 个参数之间的差异比较。图 A 代表表漂移率 v；图 B 代表边界起始点偏差 z；图 C 代表边界高度 α；图 D 代表非决策时间 τ。

……

审稿人 3 意见：

“跨情境的刺激泛化在信任形成中的作用”基于联结学习理论，通过 3 个实验考察了刺激泛化在信任形成中的作用。研究结果表明，个体对不公平分配者的面孔产生泛化，影响其在信任博弈任务中的行为决策；刺激泛化效应会受到行为意图信息的影响。研究选题具有一定的理论价值和实践意义，但仍然存在以下一些问题有待进一步斟酌。

意见 1：本研究中涉及变量较多，因此希望引言部分能够将内容的逻辑性和层级结构进一步完善。

回应：感谢审稿专家的建议。在修改稿我们对引言部分进行了修改，进一步完善了逻辑性和层级结构（*p1~p5* 的引言部分）。

意见 2：本研究强调的“跨情境间（公平-信任）的刺激泛化在信任形成中的作用”，其中“跨情境”这一界定有待斟酌。在刺激联结和泛化阶段，测试均使用信任博弈任务考察个体的投资决策，加之在这一任务范式中，公平是信任建立的基础。因此本研究结果是否可以反映刺激泛化在不同情境中的作用这一问题有待商榷。

回应：感谢审稿专家的意见。在本研究中，为了考察跨情境的刺激泛化在对陌生他人信任形成中的作用，我们设置了刺激联结和刺激泛化两个阶段。在刺激联结阶段，我们采用的是最后通牒博弈任务，让被试与三个不同公平程度（公平、中等不公平、不公平）的分配者进行最后通牒博弈，通过接受到不同公平程度的分配提议，被试会形成对三个不同公平程度分配

者面孔的刺激联结。在随后的刺激泛化阶段，我们采用是信任博弈任务，让被试在两个面孔（变形 vs.匿名）之间选择其中一个作为信任游戏的搭档。通过操纵其中一个面孔与之前最后通牒博弈中分配者面孔的相似度（23%、34%、45%、56%、67%、78%），检测与先前互动中对不同公平程度面孔形成的联结，是否会泛化到对不同任务情境下知觉相似的变形面孔的信任决策中。

因此，无论在亲历者还是在观察者视角下，刺激联结阶段和刺激泛化阶段所使用都的是不同的实验任务。在刺激联结阶段，个体处于公平/不公平情境；在刺激泛化阶段，我让被试进行的是信任决策。虽然公平和信任都是人类社会的重要互动情境，但是这两类行为的有着明显区别。公平强调社会互动中资源分配的平等性(equality)，而信任建立在对他人的意向或行为的积极预期基础上，敢于托付(愿意承受风险)的一种意愿(Rousseau & Camerer, 1998)。在现实生活中，社会环境的不断变化，刺激很少在完全相同的情境中出现，对已知面孔的学习情境通常不同于随后做出信任决策的情境。因此，本研究中我们强调“跨情境的刺激泛化在信任形成中的作用”，这也是对以往研究的重要扩展。

意见 3：实验 1a 和实验 1b 分别从亲历者和观察者角度展开探讨，但作者在文中并未对两个实验的异同性及目的做出明确阐述。虽然引言中论述了直接经历和观察学习都会使个体对他人形成声誉印象，从而影响随后的行为决策。但本研究中作者预期两种视角下的效应是一致还是有差异？为什么要同时从两种视角进行探讨？这个内容在讨论中也并未涉及，希望作者在文中可以做出进一步的说明。

回应：感谢审稿专家的意见。本研究从亲历者和观察者两个视角进行探讨刺激泛化的作用，主要是考虑到在日常生活中，我们不仅会亲身经历不公平的事件，更多的时候还作为第三方观察到不公平事件。间接的观察学习在自然界中广泛存在，对个体适应复杂的社会环境以及优化社会决策有着重要的意义(Mineka & Ohman, 2002; Olsson, Nearing, & Phelps, 2007)。个体通过观察他人行动的结果，能够间接的习得刺激的效价，进而指导个体未来的行为。此外，以往研究也发现，当个体作为第三方观察到不公平的行为时，也会产生经历不公平时的情绪体验，并且愿意牺牲自己的利益以惩罚不公平的实施者(Buckholtz et al., 2008)。通过观察他人在社会交互中的行为表现，人们能够形成对他人的声誉表征(Milinski, 2016; Milinski, Semmann, Bakker, & Krambeck, 2001; Wedekind & Milinski, 2000)，并在随后的互动中根据他人声誉做出相应调整行为(Milinski, 2016)。因此，本研究中，我们预期在亲历者和观察者视角下刺激泛化的效应是一致的。根据专家的建议，我们在讨论中对两种视角下的结果做了进一步的说明（p21~p22 讨论部分）。

意见 4：文中多处提到形成情感联结后刺激才会泛化的观点。例如，在引言（P7）中写道，“我们推测，个体会通过直接互动或间接观察，对社会互动中表现出不同公平程度的互动对象形成相应的情感联结，并对他们的面孔产生刺激泛化，用以指导在随后的互动中对陌生他人的信任决策。”“我们推测，行为意图在刺激泛化效应的产生中起到调节作用，在有意图的条件下，对先前互动中面孔形成的情感联结才会泛化到随后的信任决策中。”在 2.3 实验流程中，也同样写道“刺激联结阶段采用最后通牒博弈任务（改编自 Xiang 等人(2013)研究），让被试形成对 3 个分配者面孔的情感联结”。

研究中通过将 3 张面孔与不同公平程度的分配行为进行匹配，让被试形成的是面孔-行为间的联结。虽然研究测试了被试对不同公平程度的分配者的积极情绪体验水平，并不能简单归结为个体与交往对象间形成了情感联结，并且本研究中对于刺激泛化阶段的结果分析中也再未考虑之前情绪体验的影响。因此，本研究实际上并未涉及“情感联结”的问题。

回应：感谢审稿专家的意见。“情感联结”是联结学习领域常用的一种表述。比如，让被试对

积极、中性、消极的行为描述和面孔进行匹配学习形成情感联结，然后对与其相似的陌生面孔进行特质评价，结果发现当熟悉面孔具有积极特质时，与其相似的陌生面孔评价也是积极的；而熟悉面孔具有消极特质时，与其相似的陌生面孔评价也更消极，即发生了情感学习的泛化(affective learning generalization)。但确如审稿人所言，使用“情感联结”未必完全准确，因为情感是一个含义范围比较广泛的概念。本研究中我们提到的“情感联结”是指被试能否将三张分配者面孔与不同公平程度之间形成积极和消极的价值(value)联结，主要通过被试对不同条件下分配者提议的接受率、以及评价分配者面孔的总体公平程度进行进行操作性检验。由于刺激泛化(stimulus generalization)是指价值(value)可以在感知上或概念上彼此相似的刺激之间传播和转移。综合考虑之后，我们觉得用价值联结应该会更准确些，因此，将文中相应部分的“情感联结”改为了“价值联结”。

意见 5：在实验设计中“中等不公平”条件起到怎样的作用？在回归分析中，面孔联结类型是作为哑变量处理的吗？在各项回归分析中，似乎都是以“中等不公平”条件对参照，那么所有的结果都是相对性结果吗？中等不公平条件下，被试的刺激泛化又具有怎样的特征呢？如果变化参照条件，结果会发生改变吗？考虑到本研究中数据统计分析和结果相对较多，建议增加对多元回归和混合逻辑回归中变量的编码、放入方法、以及对主效应和交互效应的明确解释。

回应：感谢审稿专家的意见。实验设计中，我们将“中等不公平”作为参照水平，将不公平和公平条件的泛化效应与其进行对比。在回归分析中，我们也将面孔联结类型作了哑变量处理的，将“中等不公平”作为参照水平。对于中等不公平条件下刺激泛化的特征，单独将这个条件下数据进行回归分析发现，在研究实验 1a 和实验 2 中，面孔相似度的预测效应不显著，实验 1a：$t(184) = 1.71, p = 0.087$；实验 2：$t(178) = -0.90, p = 0.370$；在研究实验 1b 中，$t(178) = 2.32, p = 0.020$，面孔相似度的预测效应显著的。对于审稿人提出的变化参照条件分析，因为本研究的主要目的是考察不公平和公平条件下的刺激泛化效应，将“中等不公平”作为参照水平是合理的；如果改用其他条件（如，公平条件）作为参照，数据分析上可以进行的，但是数据解释结果的解释上就比较困难。此外，根据您的建议，我们在方法部分增加了对多元回归和混合逻辑回归中变量的编码、放入方法、以及对主效应和交互效应的解释（p9~p10 数据分析处理部分）。

意见 6：对结果部分的几个困惑：

（1）在“2.4.2 刺激泛化阶段”结果分析中，“以面孔联结的类型（公平、中等不公平、不公平）、面孔相似度（23%、34%、45%、56%、67%、78%）为预测变量，以被试选择变形面孔的比率为响应变量进行多元线性回归分析。随着与原始公平面孔知觉相似性的增加，被试更多选择变形面孔进行接下来的信任游戏，但与中等不公平条件下的差异并未到达显著，$t = 0.44, p = 0.662$。”通过表 1 回归结果显示，公平的主效应是不显著的，那么上述表达中“随着与原始公平面孔知觉相似性的增加，被试更多选择变形面孔进行接下来的信任游戏”，这里表述是否有出入？

（2）4.3.2 刺激泛化阶段的混合逻辑回归分析中，“此外，相对于中等不公平，随着与原始公平面孔知觉相似性的增加，被试可能表现出一定的刺激泛化效应，$t = 2.12, p = 0.034$。为什么是“可能表现出一定的刺激泛化效应”，差异显著不能明确是否存在泛化效应吗？

（3）多元回归分析和混合逻辑回归如果分析目的一致，建议选择最佳检验方法，不需要重复使用多种方法进行相同的差异检验，而且两种统计分析方法在结果上并不完全一致，作者如何权衡和解释统计分析中的结果差异性。

（4）第 24 页（讨论上面最后一段），将实验 1a 和实验 2 结果进行综合分析中。在无意图条

件下的分析结果为什么不同于 4.3.2 中的分析结果（23 页第一段）。

回应：感谢审稿专家的提问。

（1）确如专家所言，表 1 回归结果显示，公平的主效应是不显著的，公平×知觉相似性也是不显著的。但从回归分析中斜率（见下图 3A）可以看出，随着与公平面孔知觉相似性的增加，被试会更多的选择变形面孔进行接下来的信任游戏，只是与中等不公平条件下的差异并未到达显著。因此，我们将文中的表述修改为：随着变形面孔与原始公平面孔知觉相似性的增加，被试选择变形面孔比例有所提高，但与中等不公平条件下的差异并未到达显著，t(550) = 0.64, p = 0.524（p11 结果部分）。

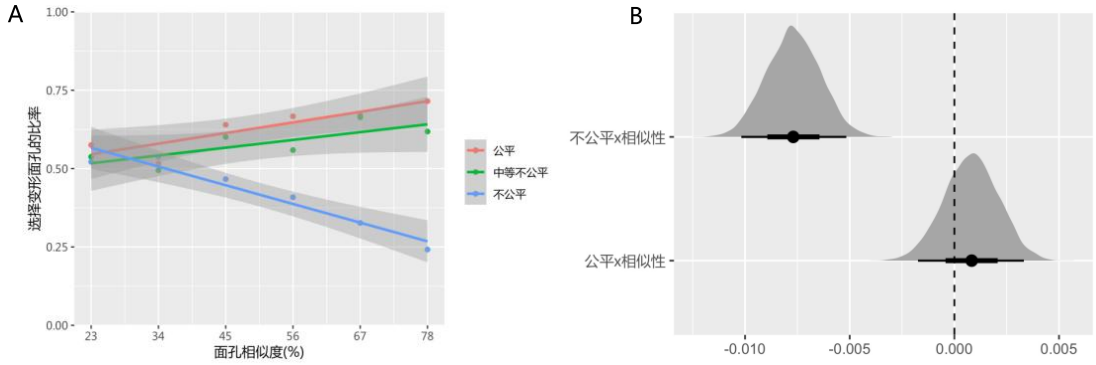

图 3 (A) 不同面孔联结类型下，面孔相似度对面孔选择比率的回归分析；(B) 公平/不公平面孔联结×面孔相似度回归系数的后验概率分布及相应的可信度区间 *CIs*。

（2）感谢审稿专家的提问。这里混合逻辑回归分析结果表明，公平×知觉相似性的回归是显著的，t = 2.12, p = 0.034，但是多元回归分析的结果发现，公平×知觉相似性的回归并未达到显著，t = 1.29, p = 0.198。因此，综合两个结果，我们当时写的是：可能表现出一定的刺激泛化效应。在修改稿中，根据专家的建议，我们只保留了多元回归分析的结果。

（3）感谢审稿专家的建议。总体而言，本研究中多元线性回归和混合逻辑回归分析的结果基本一致。混合逻辑回归分析考虑了随机效应，并且将被试每个试次的 binary 选择纳入分析，具有更高的统计检验力。因此，在某些条件下（如，知觉相似性的主效应）发现显著性的结果($p < 0.05$)，而多元线性回归分析没有发现显著性的结果。但考虑到多元线性回归分析对结果更好的解读性以及结果呈现上的直观性，我们将两个结果都进行了报告。在修改稿中，根据专家的建议，我们在多元线性回归中增加了随机效应，同时也增加了对多元线性回归的贝叶斯分析，以便更好地呈现不同公平程度条件下回归系数的差异，不再报告混合逻辑回归分析的结果（*p9* 数据分析处理部分）。

（4）感谢审稿专家的提问。这里结果并没有不同，只是在综合分析中，我们报告的是多元回归分析的结果。在无意图条件下，多元线性回归和混合逻辑回归分析的结果略有差异。修改稿中，已经根据新的数据分析方法对这一部分内容进行了更新（*p20~p21* 结果部分）。

意见 7：格式问题：表 1-表 6 的表头“是否选择变形面孔 ＝β_0＋β_1 面孔联结类型 ×β_2 面孔相似度 ＋ε”命名不规范，建议修改。另外，表 4 中有显著性***的标注，但不规范，其他表格中均未标注，建议将全文格式统一。

回应：感谢审稿专家的建议。根据建议，已对修改稿中的表格格式进行统一。

第二轮

审稿人 1 意见：

作者已回复了我所有问题，手稿已经有了很大进步，同意发表。

回应：感谢审稿专家对我们修改稿的肯定。

审稿人 3 意见：作者针对上一轮评审意见进行了详细说明和认真的修改完善，但针对本研究的研究问题和理论意义或价值，仍存在需要进一步深入思考的地方：

意见 1：本研究“跨情境的刺激泛化在信任形成中的作用”相较于 Feldmanhall 等人(2018)所提出的信任的学习或建立机制，有哪些进一步的推进或贡献需要明确阐述。作者在前言中提到“Feldmanhall 等人(2018)的研究仅探讨了同一任务情境下（联结和泛化阶段均为信任博弈任务）的刺激泛化现象。”本研究将联结阶段的信任博弈任务改为最后通牒，认为可以验证跨情境的泛化效应是不恰当的，至少需要增加不同的情境（如与积极消极效价行为）加以验证。本研究最终仍然在考察“刺激泛化在信任形成中的作用”。

回应：感谢审稿专家的意见。Feldmanhall 等人(2018)的研究仅探讨了同一任务中直接互动情境下（联结和泛化阶段均为信任博弈任务）的刺激泛化现象。然而，在现实生活中，随着社会环境的不断变化，刺激很少会在完全相同的情境中出现，对已知面孔的学习情境通常与随后进行信任决策的情境是不同的。例如，人们可能在先前的互动情境中与他人进行资源分配，在随后的情境中对他人做出信任决策。此外，Feldmanhall 等人(2018)的研究仅探讨了直接互动学习下的刺激泛化现象，除了直接互动学习，人们是否能够通过观察学习形成类似的刺激泛化效应，目前尚不清楚。最后，在 Feldmanhall 等人(2018)的研究基础上，本研究还探讨了信任形成中的刺激泛化仅需要刺激与行为结果之间的简单联结，还是需要基于对他人行为意图的感知？我们在修改稿中对上述推进或贡献进行了进一步的补充说明（p3 的引言部分）。

针对专家提出的「需要增加不同的情境（如与积极消极效价行为）加以验证」，实际上，在本研究中包含了不同的效价（公平、中等不公平、不公平）情境，对应于 Feldmanhall 等人(2018)研究中的可信、中等不可信、不可信的三种效价情境。我们在刺激联结阶段（最后通牒博弈 ultimatum game），让被试与三个不同公平程度（公平、中等不公平、不公平）的分配者进行最后通牒博弈，通过接受到不同公平程度的分配提议，被试形成对三个不同公平程度分配者面孔的价值联结。综上，本研究所探讨的跨情境的刺激泛化，是指个体在先前联结学习互动中的情境与随后刺激泛化的信任决策情境是不同的，两个情境并非都是在信任互动情境下。

意见 2：本研究与 Feldmanhall 等人(2018)的研究的差异还表现在实验 2 探讨意图在刺激泛化效应中的调节作用。但实验 2 中对于无意图条件的操纵是“通过计算机进行分配”，此时，计算机的分配结果无论公平与否，都与交往对象无关。如果要考察行为意图是否会影响刺激泛化，设置由交往对象有意或无意，进而引发公平或不公平行为更为恰当。当交往对象/面孔与行为间没有关系时，无法建立刺激与价值间的联结，没有联结就谈不上泛化。

回应：感谢审稿专家的意见。在实验 2 中，为了探讨刺激泛化的产生是仅需要刺激与效价之间的简单联结，还是需要基于对他人行为意图的感知？在与实验 1 分配设置完全保持一致的情况下，我们告知被试分配方案由计算机提出。这样的操作可以有效排除互动对象的意图，同时不引入其他可能的无关变量。以往一些关于意图的研究，例如，Sun 等人(2020)关于意图对广义互惠行为的影响的研究，Cushman 等人(2013)对意图对道德判断的影响进行研究，均是使用计算机分配来对无意图条件进行操纵。虽然，在这种操作下，计算机的分配结果与交往对象的行为意图无关；但交往对象的面孔依然与不同效价的分配结果匹配呈现，只是这些结果并非由交往对象有意导致。按照传统联结主义的观点，在这种情境下也应该会出现联结学习。因此，在保持其他变量不变的情况下，这样的实验操纵可以考察意图在刺激泛化中

的作用，验证认知因素在社会情境下联结学习中的作用。我们在修改稿中对上述实验操纵的逻辑进行了补充说明（p17 的实验 2 实验材料与任务部分）。

意见 3：建议根据本研究的重点或突出贡献进一步修改和完善题目。本研究的刺激泛化仅限于面孔刺激，目前题目过于宽泛。
回应：感谢审稿专家的意见。如前说述，本研究在以往研究基础上，从直接学习和观察学习视角，探讨跨情境间的刺激泛化在信任形成中的作用。因此，根据本研究的重点或突出贡献，我们将题目修改为："跨情境的刺激泛化在面孔信任形成中的作用：基于直接与观察学习的视角"。

第三轮

审稿人 3 意见：

作者已回复了我所有问题，手稿已经有了很大进步，同意发表。
回应：感谢审稿专家对我们修改稿的肯定。

编委意见：

意见 1：修改后的文章我看了，和之前相比有了较大的提高。但现在的文章篇幅过长，图表过多（图就有 12 个），真的有必要吗？写作的一个重要目的是把重要发现以最简单明了的方式传递给读者，而不是把七七八八的内容都放在文章里，让读者去猜哪个更重要。所以，希望你们能自己再思考一下文章的结果，缩减文章篇幅和图表。
回应：感谢编委专家的意见。根据意见，我们对文章方法、结果部分以及图表进行了缩减。去掉了解释不同条件下分钱方案分布的示意图，以及实验 1b 和实验 2 的实验流程图（与实验 1a 的流程图大致相同）。在结果部分，考虑到图更能直观的呈现研究结果，我们选择对表格进行删减，保留重要的能够直观展示研究结果的图。请编委专家审查。

意见 2：关于文章内容，个人觉得前言过长，讨论中有过多的对结果的复述。研究的主要发现是人对面孔信任的泛化。针对这个结果，你们主要强调泛化的积极作用。但应该指出，过分泛化可能是刻板印象和偏见形成的机制之一，也有潜在的消极作用。另外，泛化的发生是有意识还是无意识的？你们在文章最后说："而在我们的实验中，被试并没有意识到面孔是经过变形而来的"，这个的证据从何而来？如果是的话，泛化可否被认为是内隐学习的一种？
回应：感谢编委专家的意见。根据意见，我们对文章引言和讨论部分进行了精简，去掉了其中一些可能重复的论证以及讨论中对结果的复述。对于过分泛化潜在的消极作用，我们也在讨论中增加了相关的论述（见 p19 的讨论部分）。此外，关于文章最后的论述"而在我们的实验中，被试并没有意识到面孔是经过变形而来的"，这个证据来自于实验后被试的主观报告。实验后被试主观报告，他们没有意识到这些面孔是由联结阶段的面孔变形而来的，他们相信每一张变形面孔是一个真实的信任合作伙伴。在这种内隐情形下，个体对变形面孔的信任仍然受到过去经验的影响，表明面孔信任中的刺激泛化具有一定的内隐性。我们对讨论中的这一观点进行了补充说明（见 p21 的讨论部分）。

意见 3：最后，图 6，9，12 中的文字都是英文，包括貌似给自己看的"params"这个缩写，Panel B 和 C 在图示中也搞错了，参数的字母也不一致（英文和希腊字母）。类似的低级错误希望在下一稿中不要出现。

回应：感谢编委专家的提醒。我们修改了图示中的表述错误，并再次对文章中的内容进行核对检查，避免类似的低级错误。

第四轮

编委意见：

作者们好，感谢你们对我上次意见的回复。虽然文章现在已经简洁了许多，但还是过长，加上 references，有近 2 万字。我还是希望你们能进一步删减，去掉不必要的描述（可以从方法部分入手）和 references，最好能删到 17000 字左右。责任编辑

回应：感谢编委专家的意见。根据意见，我们对文章的方法和 *references* 部分进行了一定删减，在保证文章可读性的前提下，去掉了一些不那么重要的表述（如，实验具体指导语），将其放在补充材料中，供感兴趣的读者查阅。目前文章总字数（包括参考文献）在 18000 字以内。

主编意见：该文经过多轮评审和修改，已经达到学报发表要求。同意发表！谢谢作者们和审稿专家的辛勤付出！