《心理科学进展》审稿意见与作者回应

题目：“惩前毖后”与“率先垂范”：第三方干预行为的影响效应

作者：郭禹辰　刘艳彬　程 远

第一轮

审稿人 1 意见：

本文献综述旨在探讨第三方干预行为对社会规范的影响和作用机制。分别探究了第三方惩罚和第三方补偿两种干预行为在维护社会规范上的作用，并讨论了威慑和规范信号两种社会效应作为第三方干预行为的有效性的理论机制。整体表述清晰，结构和论点合理。一些意见建议如下：

意见 1：文章的标题给人的阅读感受是第三方干预的作用要在“惩前毖后”和“率先垂范”之间进行二选一，但实际上文章在第四部分最后的论述中表达的是威慑和垂范都有可能是第三方干预的作用机制，而并不是二者之中孰优孰劣。建议标题与文章内容的主旨观点保持一致。

回应：非常感谢您的提醒。“惩前毖后”和“率先垂范”作为第三方干预促进规范遵从的不同作用机制，二者确实应当是并列关系。为保持标题与文章主旨观点的一致性，我们修改了文章标题，修改后的标题为《“惩前毖后”与“率先垂范”：第三方干预行为的影响效应》。

意见 2：在引言的第二段，作者对于“惩罚”对于维护社会规范的研究现状的综述可能不准确。例如在提及“惩罚无法有效促进公平合作”的观点中，引的研究表达的是“缺乏正当性的第三方惩罚”（来源于后文中对这次文章的再次引用），可能并不是所有的惩罚都无法促进公平合作。因此惩罚对于维护社会规范的作用上存在争议的原因，可能是不同的应用视角或者不同性质的惩罚动机与行为造成的。文章 3.2 小节第二段的论述的开头部分存在类似的问题。

回应：非常感谢您的建议。我们再次梳理了当前有关惩罚性干预在维护社会规范方面的研究，发现争议主要体现在两方面：首先，惩罚作为一种外部激励，会破坏人际信任并排挤人们遵守规范的内部动机，导致惩罚在维护社会规范方面的效果表现出“有益于当下，有损于未来”的特点。例如 Mulder 等人(2006)的研究利用“解除惩罚”范式发现，参与者虽然在有惩罚阶段表现出较高的合作水平，但在惩罚解除后其合作水平便迅速下降，甚至远低于未经历过惩罚的参与者。这表明惩罚破坏了人际信任，排挤了合作的内部动机，未能真正起到维护规范的作用。Rand 等人(2009)以及陈思静等人(2015)的研究也发现，惩罚撤销之后人们的合作水平骤然降低，远低于无惩罚情况下。

其次，惩罚的正当性也会影响其干预效果。只有正当、合理的惩罚干预才能有效维护社会规范，而缺乏正当性的惩罚不仅无法促进规范遵从，反而可能适得其反，引起怨恨和对抗

(Baldassarri & Grossman, 2011)。例如，Xiao 和 Tan(2014)的研究发现，具备恰当理由的第三方惩罚有着更好的规范维护效果；Herrmann 等人(2008)以及 Fatas 和 Mateu (2015)的研究则发现，缺乏正当性的反社会惩罚（即惩罚者的合作水平低于被惩罚者）会损害正当惩罚的积极效果，群体中的反社会惩罚越高，群体成员的合作水平就越低。

综上，我们对正文第 1 页 1 小节第 2 段的表述修改如下：

“近年来，有研究者提出了不同观点，认为惩罚性干预作为外部激励虽然能够在短期内促进规范遵从，但是也会破坏人际信任、排挤人们遵守规范的内部动机，致使惩罚撤销后不良行为迅速反弹(Mulder et al., 2006; 陈思静 等, 2015)，无法真正起到促进社会规范的作用。此外，惩罚性干预的效果也取决于其正当性，当惩罚被滥用时（如实施反社会惩罚），该行为不仅无法有效维护规范，反而会损害正当惩罚的积极效果，导致群体合作水平下降(Herrmann et al., 2008; Fatas & Mateu, 2015)。”

对正文第 4 页 3.2 小节第 2 段的表述修改如下：

“然而，也有研究发现第三方惩罚并不总是能够促进规范遵从，特别是在第三方惩罚未能持续存在，或者第三方惩罚缺乏正当性的时候，其维护规范的效果会明显减弱。首先，第三方惩罚在促进规范遵从方面的效果经常表现出“有益于当下，有损于未来”的特点，即人们迫于第三方惩罚的压力而遵守规范，却在惩罚撤销后立刻原形毕露，甚至变本加厉。这是因为惩罚作为一种外部激励，虽然能立即改变人们行为，却也会排挤人们遵守规范的内部动机。研究发现，人们虽然在有惩罚阶段表现出较高的合作水平，但在惩罚解除后其合作水平便迅速下降，甚至远低于未经历惩罚者(Rand et al., 2009; 陈思静 等, 2015)。有研究者认为，人际信任遭到破坏以及内部动机削弱可能是导致该现象的重要原因(Mulder et al., 2006; Xiao, Houser, 2011)。进化动力学研究也发现，虽然第三方惩罚可以促使发展中社会更快地向高度合作的社会过渡，但是当社会已经进入合作状态后，过度重视惩罚会导致更多社会损失(Yu et al., 2016)。此外，当第三方惩罚缺乏正当性时，也无法有效维护社会规范。”

意见 3：本文列举出了第三方惩罚和第三方补偿分别对于维护社会规范各自独特的作用与缺陷，补偿对于维护社会规范具有很大的促进作用，并且引起的负面影响较小。但是相较于惩罚，补偿行为要求第三方对于受害方进行实际的帮助，而惩罚更多的是通过干预使违规者付出代价，因此对于第三方来说，补偿是一种相对来说成本更高的干预方式。不同的干预成本和干预场景也是干预效果的影响因素。

回应：非常感谢您的建议。首先，就第三方惩罚与第三方补偿的成本孰高孰低，我们认为此问题或许没有定论，而是取决于干预的具体场景。例如在大多数实验室研究中，研究者经常以金钱惩罚或金钱补偿作为第三方干预的手段，并且特别控制第三方惩罚与补偿成本保持一致以提高研究内部效度。而在田野实验或现实生活中，第三方惩罚可能是代价微小的社会排斥，也可能是代价高昂的身体对抗；第三方补偿可能是代价微小的言语安慰，也可能是代价高昂的经济援助。总之，第三方惩罚与第三方补偿的成本可能具有高度的情境特异性。当然，除了源于情境的成本外，第三方惩罚相比补偿更可能附带隐性成本，比如可能招致被惩罚者的报复和怨恨。

其次，第三方干预维护社会规范的影响效果确实受到很多因素的调节，正如您所提到的，第三方干预的场景和成本是干预效果的重要影响因素。研究显示在涉及资源分配的场景中，人们普遍认为应该对不公平的分配者给予惩罚，甚至未惩罚的人会因此而受到高阶惩罚(Martin et al., 2019)；然而，人们进行合作互动时却并不赞同对未合作者实施惩罚(Sutter et al., 2010)。这意味着，人们可能对于不同场景下惩罚的合理性与正当性有着自己的判断和认识，

并且会根据惩罚发生的场景推断惩罚者动机，而不是对所有的第三方惩罚行为都盲目认可。而许多研究显示，只有被认为具备正当性的惩罚才能起到规范维护功能。因此可以推论，第三方惩罚的具体场景，特别是在该场景下是否应当实施干预，可能影响其维护规范的效果。

此外，第三方干预的成本也可能影响干预效果。一方面，第三方干预的成本越低，干预出现的频率就越高(Guala, 2012)。这意味着在低干预成本时，违规者和受害者得到应得惩罚或帮助的可能性更高，因此干预行为所产生的威慑效力会更强、传递规范信号的频率也更高，进而得以更好地促使他人遵守社会规范。然而，也有研究发现了不一致的结果，比如有成本的惩罚和奖励比无成本时更能促进合作(Balliet et al., 2011)；特别是对于第三方惩罚而言，有代价的惩罚比无代价的惩罚更有效地维持了合作(Kuwabara & Yu, 2017)。一个可能的解释是，无成本或低成本的惩罚行为可能引发观察者对其道德合法性的质疑（如惩罚者可能出于竞争而非利他做出惩罚），而高成本的惩罚则更能彰显惩罚者的大公无私(Raihani & Bshary, 2015)，因此后者可能判断为更具正当性，并传递出无私的、愿意维护社会规范的强烈信号。总之，第三方干预成本对干预效果的影响似乎颇为复杂，二者之间的确切关系还有待未来研究深入探讨。

综上，根据您的建议，我们在研究展望部分增加了对此问题的讨论。我们对正文第 10 页 5 小节第 3 段的表述补充如下：

“第三方惩罚发生的场景可能会影响人们对惩罚行为道德合法性的判断，如在涉及资源分配的场景中，人们普遍认为应该对不公平的分配者给予惩罚(Martin et al., 2019)，而在涉及合作互动的场景中，人们却并不赞同对未合作者施加惩罚(Sutter et al., 2010)。这意味着，人们并非对所有第三方惩罚行为都盲目认可，而是会对不同场景下惩罚行为的合理性加以判断，并根据情境背景推断惩罚者动机。那么，人们对于第三方惩罚动机与合理性的感知会影响到惩罚效果吗？感知利他的第三方惩罚是否比感知利己的惩罚更好地澄清了社会规范？不同来源及场景下的第三方惩罚是否在传递规范信号的效果上也有所不同？这些问题有待未来研究深入探讨。”

我们对正文第 10 页 5 小节第 4 段的表述补充如下：“*此外，第三方干预的成本也可能影响干预效果。一方面，第三方干预的成本越低，干预出现的频率就越高(Guala, 2012)。这意味着在低干预成本时，违规者和受害者得到应得惩罚或帮助的可能性更高，进而干预行为所产生的威慑力会更强、传递规范信号的频率也更高，从而应当在促进规范遵从方面起到更好效果。然而，也有研究发现了不一致的结果：有成本的第三方惩罚比无成本时更能维持和促进合作(Balliet et al., 2011; Kuwabara & Yu, 2017)。这是因为无成本或低成本的惩罚行为可能引发观察者对其道德合法性的质疑（如惩罚者可能出于竞争而非利他做出惩罚），而高成本的惩罚则更能彰显惩罚者的大公无私(Raihani & Bshary, 2015)。后者因此可能被认为更具正当性，并传递出惩罚者无私的、愿意维护社会规范的强烈信号。总之，第三方干预成本对其维护社会规范效果的影响似乎颇为复杂，二者之间的确切关系还有待进一步探究。”*

意见 4：文章中存在一些语义不清晰或者不通顺的表达，建议作者优化较为不通顺或不易理解的语句。例如文章 4.2 小节第三段中的“并且预期未传递规范信息的惩罚将无法抑制违规行为”，这样的表达降低了可读性，其他类似地方建议改进。

回应：非常感谢您的建议。我们再次对全文语句进行了细致地修正，调整了部分语句的表达方式，提升了文章语言表达的清晰性与可读性。**我们对文章语句的主要修改如下（仅罗列部分主要修改，其他细微修改请见正文部分）：**

正文第 1 页 1 小节第 1 段第 2 句修改为：“*在漫长的社会发展进程中，人类逐渐形成*

了对彼此行为方式的期望与承诺.....”

正文第 1 页 1 小节第 1 段第 4 句修改为：“*因此在发现他人违背规范时，即使事不关己，甚至代价高昂，人们也会自发地进行干预来维护社会规范。”*

正文第 1 页 1 小节第 2 段第 1 句修改为：“*.....然而在理论层面上还缺少对第三方干预作用机制的系统梳理和总结。”*

正文第 2 页 2 小节第 3 段第 2 句修改为：“*尤其当第三方具备较高的特质性共情时，他们会更倾向于补偿受害者而非惩罚违规者(Hu et al., 2015; Leliveld et al., 2012)。”*

正文第 3 页 3 小节第 1 段第 4 句修改为：“*在无法依赖于法律规章等正式系统制裁的情况下，社会规范之所以能够得以长期维持，离不开第三方干预的力量(Schroeder et al., 2003; Tomasello & Vaish, 2013)。”*

正文第 4 页 3.2 小节第 1 段第 2 句修改为：“*.....即人们做出惩罚是为了震慑潜在违规者，从而避免违规行为再次发生，因此它也被称为功利主义动机(Akers, 1990; Tan & Xiao, 2018)。”*

正文第 5 页 3.2 小节第 3 段第 5 句修改为：“*.....研究者发现相比实施惩罚，恩加人更愿意通过弥补已造成的伤害来维持公平，而这种恢复性措施对维持社会秩序和良好人际关系都具有重要作用(Wiessner , 2020)。”*

正文第 8 页 4.2 小节第 3 段第 5 句修改为：“*此外，人们在实施第三方惩罚时也更偏好那些传递了规范信息的惩罚，并且认为那些未能传递规范信息的惩罚方式无法阻止违规行为再次发生(Marshall et al., 2021)。”*

……

审稿人 2 意见：

该论文综述了第三方干预研究的已有进展，并提出了第三方干预行为维护社会规范的可能作用机制。论文围绕一个核心的科学问题，提出了理论构想，并层层深入地阐述了相关作用机制。论文的主要创新观点是：第三方干预不仅通过惩罚威慑促进规范遵从，也通过传递规范信息调整人们对社会规范的感知，进而促使其行为改变。总的来说，论文框架合理，逻辑严密，论证具有理论深度，并且所提出的理论较好地弥合了现有研究发现中的一些冲突。以下是一些小的修改意见（minor concerns）。

意见 1：“Fehr 及其合作者们最先在实验室中观察到了第三方惩罚的存在”，这句话的表述可能可以完善。第三方惩罚作为一类亲社会行为，在一些社会心理学研究中很早就有，只是 Fehr 等采用了专门衡量第三方惩罚力度的实验范式，从而被观察到并被显性地提出来了。

回应：非常感谢您的提醒。我们对此句的表述方式进行了完善，**修改后的表述如下：***“Fehr 及其合作者们开发了基于博弈任务的第三方观察者范式，率先在实验室内对第三方惩罚行为进行了定量研究。”*

意见 2：“违规行为背离规范的程度越高、受害者遭受的损失越大，第三方做出惩罚或补偿的力度就越大”。此处，是否可以进一步补充有关惩罚或补偿力度的相关数据。以及，惩罚或补偿的不同力度是否可以反应个体的“报应主义”或“结果主义”动机差异？

回应：非常感谢您的建议。首先，违背规范的严重程度与第三方惩罚和补偿力度的相关性已经得到许多研究结果的支持，例如 Fabbri 和 Carbonara(2017)的研究发现，随着独裁者自私程度的增加，第三方对独裁者的惩罚力度也随着增加：当独裁者从接受者那里拿走 10 个代币时（共 30 个代币），第三方惩罚者对独裁者的平均惩罚为扣除其 9.6 个代币；当独裁者拿走 20 个代币时，第三方惩罚者对独裁者的平均惩罚为扣除其 16.5 个代币；而当独裁者拿走全部 30 个代币时，第三方惩罚者对独裁者的平均惩罚为扣除其 23.2 个代币。Rodrigues 等人(2020)的研究同样发现独裁者公平程度正向预测第三方惩罚与补偿力度：当独裁者做出相对不公平的分配时（共 8 个代币，分给接受者 2 个，留给自己 6 个），独裁者会被扣除约 21%的收益，接受者会被补偿约 35%的收益；当独裁者做出非常不公平的分配时（共 8 个代币，分给接受者 0 个，留给自己 8 个），独裁者会被扣除约 29%的收益，接受者会被补偿约 44%的收益。

综上，我们对正文第 3 页 3.1 小节第 2 段的表述补充如下：“*例如，Rodrigues 等人(2020)的研究显示，独裁者的公平程度正向预测第三方惩罚与补偿的力度：当独裁者做出相对不公平的分配时，独裁者会被扣除约 21%的收益，接受者会被补偿约 35%的收益；而当独裁者做出非常不公平的分配时，独裁者会被扣除约 29%的收益，接受者会被补偿约 44%的收益。*”

其次，关于“惩罚或补偿的不同力度是否可以反应个体的“报应主义”或“结果主义”动机差异？”这是一个好的问题，但遗憾的是我们暂时没有发现能够支持“干预力度反映第三方动机差异”的相关研究，这可能是未来一个重要的研究问题。

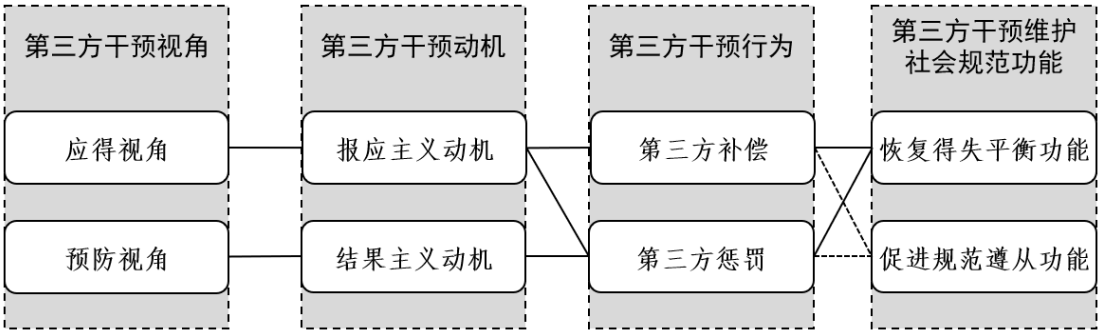
就已有研究而言，首先由于第三方补偿不具备威慑力，在第三方干预者的认知里应当无法抑制未来违规行为发生，因此补偿行为往往受到“报应主义”动机的驱动，与“结果主义”动机的关系可能相对薄弱。其次，有一些研究曾试图区分第三方惩罚的两种不同动机，但没有发现惩罚力度与“报应主义”和“结果主义”动机的关系。例如，Tan 和 Xiao(2018)的研究通过操作第三方惩罚是在合作行动前还是行动后做出，从而对惩罚者的“报应主义”和“结果主义”动机加以区分，发现个人惩罚者在行动前后所做出的惩罚力度和频率均没有显著差异，表明惩罚的力度与动机取向并没有明显关联。Marshall 等人(2021)的研究则通过操作能否传递错误行为信息来区分惩罚者的“报应主义”和“结果主义”动机，发现儿童作为惩罚者时同时具备“报应主义”和“结果主义”两种动机，并且两种动机的强度没有显著差异。此外，Carlsmith 等人(2002)的研究通过情境材料操纵违规行为的惩罚应得性（对应于报应主义动机）与威慑性（对应于结果主义动机），发现人们在定性层面同等支持两种惩罚动机，但在实际做出惩罚决策时，则更多出于“报应主义”的考虑，即依据违规者的错误程度给予应有的惩罚。总之，现有研究结果显示，第三方干预的力度更多取决于干预成本以及错误行为的严重程度，暂时没有研究表明惩罚或补偿的力度可以反应干预者“报应主义”或“结果主义”的动机差异。

意见 3：图 1 描述了“第三方惩罚和补偿”均出现在报应主义动机和结果主义动机中，这本身没有错。不过，是否可以在图中或者正文中适当描述“第三方惩罚”和“第三方补偿”在两类动机中出现的倾向性，甚至是可能的概率。

回应：非常感谢您的建议。考虑到第三方补偿虽然弥补了受害者的损失，却没有改变违

规者的收益矩阵，难以起到威慑作用，因此第三方干预者不会预期通过补偿来阻止未来违规行为再次发生。故而第三方补偿更多受到“报应主义”动机驱动，与“结果主义”动机的关系较为微弱。为了更清晰、准确地展示第三方干预行为与动机的关系，我们重新绘制了图 1，并且在正文中增加了有关“报应主义”动机驱动下第三方干预者做出惩罚及补偿行动倾向的阐述。

我们对正文第 5 页图 1 的修改如下：



我们对正文第 3 页 3.1 小节第 1 段的表述补充如下：

“研究发现，在不受限制的情况下，人们往往同时实施第三方惩罚以及第三方补偿来恢复正义(Lotz et al., 2011; Van Doorn et al., 2018)；而当只能选择一种干预方式时，人们整体上更倾向于对受害者做出补偿(Dhaliwal et al., 2021)，并且其实际行为决策经常受到自身人格特质(Leliveld et al., 2012; Hu et al., 2015)和得失情境框架(Liu et al., 2019)的影响。”

意见 4：在研究展望中，是否可以进一步探讨第三方补偿的新的可能不同形式，以及除了惩罚和补偿外还有哪些第三方干预措施。

回应：非常感谢您的建议，我们在研究展望部分增加例举了对数字化时代下第三方补偿的新形式，并阐释了第三方奖励在维护社会规范方面的可能价值。

我们对正文第 9 页 5 小节第 1 段的表述补充如下：

“从广义上来讲，任何针对受害者的恢复性措施都属于第三方补偿的范畴，其形式非常多样。在数字化时代下，第三方补偿也衍生出许多低成本、高参与的新形式，如互联网慈善日捐、在线公益课程分享、信息披露与法律援助虚拟社区、社交媒体舆论声援支持等等。由于补偿行为具有纯粹的利他性，也不会如同惩罚一般带来高额附加成本，故而有必要探讨第三方补偿能否成为惩罚的替代性措施，在非正式规范维护系统内发挥同样的作用。”

我们对正文第 9 页 5 小节第 2 段的表述补充如下：

“除了惩罚与补偿外，奖励可能也是一种重要的第三方干预方式。一项元分析研究发现，与惩罚相类似，奖励同样对社会合作具有中等程度的正向影响作用(Balliet et al., 2011)。如果说第三方惩罚与补偿反映了干预者对不良行为的反对态度，那么第三方奖励则反映了干预者对积极行为的肯定与认可。因此第三方奖励不仅是一种外部激励，也可能具有信号作用。未来的研究可以探讨第三方奖励维护社会规范的作用机制，比如奖励是否同样具有传递规范信息、促进规范遵从的功能？以及奖励作为一种外部激励，是否也存在排挤内部动机等潜在的负面效果？”

第二轮

审稿人 1 意见：作者针对审稿意见做了较多改动，已经没有进一步的意见，谢谢。

审稿人 2 意见：作者针对上一轮审稿意见，详实、有效地修改了文本，改善了论文质量。没有新的问题了。建议录用。

编委 1 意见：同意发表。

编委 2 意见：同意发表。

主编意见： 根据编委和审稿专家的意见，建议发表。