

《心理学报》审稿意见与作者回应

题目：说话人观念与任务要求对反语理解中观点采择的影响

作者：杜雅楠，徐文苑，葛枫岭，孙凤娇，夏天，宗文惠，周茂楠

第一轮

审稿人 1 意见：

本研究通过两个实验系统探讨了说话人观念（经典刻板、反向刻板、中性）与任务要求（外显判断 vs. 自然阅读）对反语理解中观点采择的影响，并结合个体差异变量进行了深入分析。研究设计严谨，方法先进，数据分析详实，具有较强的创新性。研究结果支持基于约束的并行加工模型，对理解社会认知与语言理解的交互机制具有一定贡献。然而，文章存在以下问题，导致文章的可读性和阅读体验较差。

意见 1：

文章的逻辑结构不够清晰，例如如何引进个体差异变量以及实验任务等。

回应：

感谢审稿人的建议。我们非常认同审稿人的意见，初稿在引入实验变量（尤其是实验任务和个体差异）时，理论论证确实不够充分，导致逻辑不够连贯。针对这一问题，我们对引言部分进行了系统性修订与逻辑重组，主要修改如下：

首先，我们重新梳理了反语理解中关于“序列加工（优先从自身视角出发）”与“基于约束的并行加工（自身视角和观点采择并行加工）”的理论争论（见修改稿第 4 页最后一段至第五页第三段，“序列加工与基于约束的并行加工的理论争论也体现在反语理解中的观点采择研究中。……，表现出典型的序列加工模式”，修改稿中已用红色字体标识）。通过回顾 Akimoto 等(2012)、Deliens 等(2017)、Weng 等(2023)和 Ronderos 等(2024)研究，我们明确了反语理解中观点采择的时程是实验任务（外显识别反语意图的任务和自然阅读伴随“是/否”理解问题作为注意力监控）与语境线索强度交互影响的观点。这为本研究对比实验 1（反语等级评定任务）与实验 2（自然阅读，“是/否”理解问题作为注意力监控）的设计提供了直接的理论依据，即考察不同任务需求如何调节反语理解中观点采择的过程。

其次，进一步强化了说话人的性别观念和读者的内隐和外显的性别刻板印象的引入逻辑（见修改稿第 5 页第四段和第五段，“值得注意的是，……，“自我视角”即个体所持有的内隐或外显性别刻板印象”，修改稿中已用红色字体标识）。我们指出，以往研究多聚焦于实验提供的临时性外部预期，而性别刻板印象作为一种长期内化、具有高度认知稳定性的认知先验，其整合机制可能与临时线索存在本质差异。通过这种对比，我们论证了探讨稳定社会信念对反语加工中观点采择影响的独特价值与必要性。而读者的内隐和外显性别刻板印象作为他们自己内化的观念，是本研究中重要的“自身视角”的体现。

再次，关于一般认知资源和社会认知特质两个维度的个体差异，我们补充了理论支持（见修改稿第 5 页最后一段和第 6 页第一段，“其次，为阐明社会信念在观点采择中的动态整合过程，……，这说明了成功解读反语需要多维度认知资源的协同”，修改稿中已用红色字体标识）。执行功能（尤其是 Stroop 任务表现）与语言理解中观点采择的关系是序列加工和基于约束的并行加工之间的理论争论之一。而共情和语用等社会认知特质对于反语理解有重要作用。

最后，我们对本研究各变量之间的关系提出了明确假设（见修改稿第 6 页第 3 段，“基于以往研究及本实验设计，……，从而促进观点采择”，修改稿中已用红色字体标识）。我们预期语境强度与实验任务存在交互作用，即强预期线索在两类任务中均能促进反语意图的并行整合，而冲突语境则会诱发对自身视角的依赖，且该效应在实验 2 中更明显。同时，个体差异会调节上述效应，表现为执行功能较强者优先更能抑制自身视角，更倾向进行观点采择，而社会认知特质较高者则更易在加工早期整合刻板印象线索以促进观点采择。

通过上述修订，我们不仅理顺了从理论争论到实验设计（任务选择）的逻辑，也强化了从个体先验认知（性别刻板印象）到个体特质差异（执行功能与社会认知）的论证链条，使整项研究的逻辑结构更加严谨清晰。

意见 2:

实验一图 1 的实验流程，为什么是这样的一个流程？在阅读实验之前进行内隐联想测验是否有什么特殊的实验目的，为何是内隐联想测验在阅读实验之前进行，其他的任务在阅读实验之后？

回应:

感谢审稿人对于实验流程设计逻辑的关注。针对实验流程安排的必要性及潜在的干扰因素，我们已经在实验流程部分进行了详细补充与完善（见第 9 页，2.1.5 实验流程 第一、二段“实验流程如图 1B 所示，……，以缓解被试疲劳”，修改稿中已用红色字体标识）。

具体而言，本研究采用 IAT——阅读实验——认知能力测验——特质问卷的顺序，主要基于对启动效应、认知资源损耗及测量稳定性的权衡。首先，将 IAT 置于首位是为了获取未受启动影响的内隐态度，避免阅读实验中性别刻板印象信息对内隐态度产生干扰。其次，为确保核心任务在被试认知资源相对充沛的状态下完成，我们将阅读实验安排在认知能力测验之前。最后鉴于 AQ、IRI 及 BSRI 所测特质具有较高的重测信度和稳定性，我们将特质量表任务安排在最后。

意见 3:

从实验 1 到实验 2 的逻辑是实验 1 采用了外显的意图判断任务，实验 2 取消这个任务，要求被试在自然阅读之后回答可理解问题。此处很难读懂，实验 1 也是采用可理解性问题，外显意图判断任务是指实验 1 中的反讽程度评定吗？

回应:

感谢审稿人的意见。实验 1 中的“外显意图判断任务”确实是指反语程度等级评定任务。在实验 1 中，被试在阅读完句子后需对说话人的反语程度进行外显评定，这一过程会诱发被试主动评估说话人意图的加工策略。而实验 2 则取消了该评定环节，仅在部分试次后设置简单的“是/否”理解问题（如询问故事场景细节），其目的仅在于监控被试的注意力，而非引导其进行意图识别。因此，实验 2 更接近自然阅读情境，能够考察在无外显任务驱动时，个体差异和语境线索如何自发地影响观点采择。

我们意识到原稿中关于实验设计的逻辑阐述不够清晰，主要原因有两点：一是任务需求变量的引入逻辑不够顺畅。为此，我们在修改稿第 4 页末尾至第 5 页详细阐述了这一逻辑，明确了从实验 1 到实验 2 的转变是为了考察不同任务需求对观点采择加工时程的调节作用，从而使两个实验的递进关系更加透明。二是原稿中使用“外显任务”和“内隐任务”表述不够精准。为此，在修订稿中，我们将涉及实验 1 任务的表述统一修订为“反语程度等级评定任务”；将实验 2 的任务统称为“自然阅读任务”。

意见 4:

实验 2 取消反讽程度的评定，作者如何确定读者确实有进行反语的加工还是按照字面意思理解的？有没有后测确认被试确实有加工句子中反语意图。

回应:

感谢审稿人对实验 2 设计的深入思考与建设性意见。我们完全理解审稿人的担忧：在没有外显意图任务的情况下，如何确认读者确实对反语进行了加工，而非停留在字面理解层面。为此，我们在实验 2 的讨论部分专门补充了一段说明（见 22 页，3.3 讨论的第二段，“值得注意的是，……该交互模式更倾向于支持读者确实进行了反语意图的推断”，修改稿中已用红色字体标识），对这一问题进行澄清和讨论，以便读者更全面地理解本研究的结论基础与方法选择。

关于未设置后测，我们当时的考虑是：如果在实验结束后让被试重新阅读所有句子并逐条进行反语等级评定，这种回溯性任务可能较难准确反映在线加工过程中的实时理解状态，其对于自定步速阅读数据的解释效力也存在一定局限。

不过，我们认为实验 2 的在线加工数据模式可以为这一问题提供了证据。具体而言，在溢出区段，说话人观念与话语类型呈现出显著的交互作用：在中性观念条件下，反语句的阅读时间显著长于字面语句；而在经典刻板与反向刻板观念条件下，这一差异不再显著。这一模式与心理语言学文献中常被讨论的“反语加工成本”现象较为吻合——当字面意义与语境发生冲突时，读者往往需要额外的时间进行反语意图的推理。更重要的是，这种加工成本并非在所有条件下普遍出现，而是仅在说话人观念未能提供明确预期（即中性条件）时存在；当说话人观念能够形成对反语的合理预期（刻板条件）时，加工成本随之减弱甚至消失。这种条件性的加工差异，或许可以在一定程度上说明被试并非仅停留于字面水平的机械加工，而是能够实时整合说话人的社会身份信息，并据此调节对反语意图的理解过程。此外，本实验中的理解问题回答正确率整体较高，也在一定程度上表明被试对句子语义进行了有效加工。

审稿人 2 意见:

本研究采用 2 个实验探讨了说话人观念与任务要求如何影响反语理解中的社会观点采纳。实验 1 和 2 分别从外显和内隐层面探讨了上述问题。本研究的选题新颖，研究有一定的理论价值。不过，本文尚有部分问题与作者探讨。

意见 1:

实验 1-2 中 Stroop 任务不建议翻译为中文，直接用英文名称就好。图 1 字体较小，较难辨识。类似的情况也出现在其他图片上，不再一一指出。

回应:

感谢审稿人敏锐细致地审阅。关于术语翻译问题，我们完全赞同您的观点，Stroop 任务作为经典范式，直接使用英文原名更符合学术惯例，且能避免翻译可能带来的理解偏差，因此我们已遵照建议将全文统一修改为英文专有名词。

针对文稿中图片辨识度不佳的问题，我们已对全部图片完成了系统性优化。具体而言，一方面统一增大了图中标注的文字字号，保障其在正常浏览时清晰可读；另一方面，将图片统一转换为高分辨率.tif 格式，确保图表在放大查看时仍能保持线条与文字的清晰度。

意见 2:

实验 1-2 的实验任务较多，总的实验时长多久？

回应：

感谢审稿人的建议。针对实验 1-2 时长及任务安排的问题，我们已经在修改稿“实验流程”部分补充了详细的时间与任务排序逻辑。

实验 1 的总时长约为 70-75 分钟，实验 2 为 85-90 分钟，包含了任务执行与必要的休息时间。

实验 1 的具体安排如下：实验首先进行 IAT 测验（约 4 分钟）；随后进入核心的阅读实验（约 33 分钟，分为两组进行，每组的 15 分钟，组间休息约 3 分钟）；之后进行认知能力测评，包括 Stroop 任务（约 3 分钟）和阅读广度任务（约 15 分钟）。最后，被试需完成三份个体特质问卷（AQ、IRI、BSRI，共计约 10 分钟）。为了缓解被试的疲劳感并保障数据质量，各任务模块之间均设置了 1-3 分钟的休息时间。这种排序方式在确保核心任务（阅读实验）获得充足认知资源的同时，也兼顾了内隐态度的测量效度与个体特质的稳定性（见第 9 页，2.1.5 实验流程 第一、二段“实验流程如图 1B 所示，……，以缓解被试疲劳”修改稿中已用红色字体标识）。

为更全面评估工作记忆容量，实验 2 在问卷测量结束后，增加操作广度任务，用时约为 18 分钟（见第 19 页 3.1.3 实验程序与实验流程标红部分）。

意见 3：

实验 1-2 中为什么量表总分与分量表相关高，就要用分量表进行统计分析？量表总分不应该与分量表相关高吗？

回应：

非常感谢审稿人的意见。我们意识到初稿中的表述不够严谨，容易引起误解。实际上，我们在个体差异统计分析的方案中，遵循的是研究的理论驱动、量表构念维度的独立性和统计的多重共线性三个核心原则。修改稿中，我们重写了相关部分（见修改稿第 12 页最后一段到 13 页第一段红色字体部分以及第 13 页 2.2.3 变量间的相关 的红色字体部分）。

具体说，本研究中的多维量表包括 Bem 性别角色倾向（BSRI）和人际关系指数（IRI）两个。对于 BSRI 而言，本研究将其做为测量听者推导语义时的一种整体性“自我视角”干扰源，旨在考察个体性别刻板印象的宏观强度如何影响反语理解中的观点采择，而非区分特定性别角色的细分维度。从维度独立性来看，BSRI 各分量表之间相关较高（0.58 到 0.82），这种高相关说明各维度在本样本中测量的是同一核心构念，缺乏足够的区分效度。因此，为了规避潜在的严重多重共线性问题并保持模型的简洁性，我们最终采用 BSRI 量表总分作为衡量个体外显性别刻板印象的指标。

对于 IRI，正如我们在修改稿第 6 页第一段强调的那样，共情作为多维特质，其各成分在反语加工中具有明确的功能差异与不同的神经机制（Valles-Capetillo et al., 2026）。已有研究表明，“观点采择”能预测反语识别准确性并关联视角整合脑区（如右侧缘上回、双侧岛叶），“幻想”亚成分与加工效率有关，而“共情关怀”与“个人痛苦”则参与调节情绪模拟过程中的神经活动（Valles-Capetillo et al., 2026; Zhu & Filik, 2023）。从维度独立性来看，IRI 四个分量表两两相关较低（0.09 到 0.47），表明其认知与情感成分在经验水平上具有良好的独立性。在统计共线性层面，这种低相关保证了在回归模型中保留分量表不会引发共线性问题，能更精准地捕捉“观点采择”等特定维度对反语理解的特异性预测作用，避免总分掩盖了具有重要理论价值的维度差异。

意见 4：

实验 1-2 采用阅读时间不太合理，因为词语太多了，而且不同故事间字数可能也不同，阅读时间中的混淆因素太多。如果对过程进行分析，建议用眼动仪，通过眼动指标研究阅读

过程，更加科学。

回应：

感谢审稿人的建议。眼动技术能够通过注视时长、回视频率等指标精细捕捉词汇层面的加工过程，有助于区分不同的加工阶段。然而，针对本研究的研究问题，采用自定步速阅读范式下的整句呈现方式，更能有效回答我们所关注的科学问题。如我们在引言中所述，自定步速阅读范式是本研究领域内探讨类似问题时广泛使用的实验范式，采用该范式有助于将本研究的结果与已有研究（Akimoto et al., 2012; Deliens et al., 2017; Ronderos et al., 2024; Weng et al., 2023）进行比较，从而增强研究发现的可比性和可推广性。

此外，采用自定步速阅读范式还可以有效规避高频功能词带来的“跳读”干扰。本研究的目标区段均为 5 到 8 个词构成的简单句（如“这工作可真好！”），其中包含一些短小且高频的词汇（如“这”“可”“真”）。在眼动实验中，这类词汇的跳读率通常较高，容易导致关键语用标记词的注视数据缺失。相比之下，自定步速阅读通过“按键触发”的物理操作，强制被试将认知资源分配到每一个句子区段。这种强制性加工确保了被试对包含语用线索的每一个语素都进行了处理，从而提供了更完整的整体加工耗时数据。

我们充分理解审稿人对语境长度的顾虑：相比于孤立句子，完整故事文本的篇幅更长，潜在的干扰确实可能增加。针对这一潜在风险，我们在实验材料编制阶段已经进行了严格的匹配和平衡。针对审稿人担心的词数差异及故事间差异，我们在设计中均已严格控制。对同一种说话人观念条件，同一套材料的背景区段、观念区段、目标区段、溢出区段和结尾区段完全一致；唯一的微小差异仅存在于非关键的事件区段（1 至 2 个词之差，例如表 1 中“保安”和“保姆”）。我们也意识到，不同说话人观念条件虽然在字数和理解难度上总体相近，但较长的故事语境仍可能消耗不同的认知资源，构成潜在威胁。因此，我们的假设仅聚焦于不同说话人观念的字面意思与反语之间的加工差异，并在统计分析时采用有计划的比较（planned pairwise），即只比较不同说话人观念下，字面意思条件和反语条件下加工时长的差异。

除此之外，针对可能的干扰，我们也在统计中进行了控制。我们使用线性混合效应模型（LMM）进行数据分析，将“被试”和“项目”同时设为随机效应，以有效控制个体差异与故事间变异。同时，我们将语境的自然程度、事件区段的字数等作为协变量纳入模型。这意味着，即便存在细微的混淆因素，其影响也已在回归模型中得到估计和控制，从而确保最终观察到的显著差异确实源于自变量的操纵。

综上所述，尽管眼动技术具有更高的空间分辨率，但在本研究严格的材料控制与统计建模下，自定步速阅读能够提供足够科学、稳健且具有解释力的实验证据。

意见 5：

实验 1-2 中反语评分的反应时间也不建议进行反应时分析，因为里面涉及的加工过程较多。单纯进行反语等级分析是比较合理的。

回应：

感谢审稿人的建议。我们意识到，反语等级评定任务的反应时确实包含了记忆整合、决策等认知加工过程，不能反映反语理解的即时加工。我们已经删除了相关的数据分析和讨论。

意见 6：

执行功能包括很多子成分，仅仅用 Stroop 任务作为判断执行任务的指标是否合适？并且，实验 2 将该任务每个词的呈现时间设置为 700 ms，以方便与以往研究比较，但是与实验 1 就没可比性了。

回应：

感谢审稿人的建议。执行功能确实是一个包含抑制、转换和更新等核心子成分的多维复合概念。我们在本研究中选择 Stroop 任务作为主要指标，并非试图以此涵盖执行功能的所有侧面，而是基于理论假设和先前研究作出的针对性选择。在修改稿中，我们重新梳理并重写了关于个体差异变量选取的理论依据（详见修改稿第五页第 5-6 段）。

观点采择的核心在于解决“自我-他人”的视角冲突。根据 Brown-Schmidt (2009) 的研究，个体在语言理解中需要实时抑制“自我中心”的干扰，才能成功采择他人视角。该研究发现言语 Stroop 任务表现与观点采择任务表现之间表现出了显著的正相关，而 go-nogo 任务与观点采择之间没有显著相关。此后，围绕观点采择加工中自我视角抑制问题的研究，多集中在 Stroop 任务上，如 Brown-Schmidt 和 Fraundorf (2015)、Luce 和 Almor (2023) 等。

关于 Stroop 任务参数的设置，在实验 1 中，考虑到中文加工的复杂性，我们采取了“语音触发即消失”的宽松设计，以确保被试有充足的加工时间。但这也导致了天花板效应（正确率 0.97 ± 0.04 ），且无法与 Brown-Schmidt (2009) 等采用固定时长的研究直接比较。因此，实验 2 将呈现时间统一设为 700 ms。实验结果显示，时间压力设置确实对被试构成了挑战：实验 2 的平均正确率（ 0.83 ± 0.16 ）下降，且表现出更大的个体差异。我们完全认同审稿人的观点，两实验因参数、难度不同，不宜进行直接统计比较，故在正文中避免对 Stroop 任务的正确率等进行跨实验比较。

不过，在难度跨度较大的两种实验条件下——实验 1 的宽松加工与实验 2 的时间压力——Stroop 任务表现与阅读时间的交互作用均未达到统计显著水平，这表明该研究发现并非特定参数设置下的偶然结果。

.....

审稿人 3 意见：

本研究探讨了说话人社会观念（性别刻板印象）与任务要求如何影响反语理解中的观点采择过程，通过两个实验（外显任务 vs. 内隐任务）系统考察了这一问题。研究问题具有一定的理论价值，实验设计相对完整，但在方法学、数据分析和结果解释等方面存在若干重要问题，需要作者进行实质性修改。

意见 1：

本研究将反语等级判断任务的反应时作为反语理解加工的重要指标之一。然而，从实验流程来看（图 1A），反语等级判断任务出现在目标区段之后，还需经历溢出区段和结尾区段后才进行评分。这意味着：

- a 反语等级判断任务与目标区段之间间隔了两个区段的阅读
- b 这并非真正意义上的在线（on-line）加工测量，判断任务的反应时更多反映的是离线决策过程，而非即时的反语理解过程
- c 这一时间延迟会极大降低反语评分任务反应时指标对目标区段加工过程的敏感度

回应：

感谢审稿人的建议和细致分析。我们认同，由于反语等级判断任务与目标区段之间间隔了两个区段的阅读，其反应时确实已难以作为对目标区段即时加工过程的有效指标，而更多反映了任务后期的离线整合与决策过程。因此，将这一反应时数据解释为即时反语理解过程的测量是不准确的。

基于此，我们决定在修改稿中删除所有关于反语等级判断任务反应时的分析与相关论述，将分析重点集中于更能反映即时加工过程的指标（如自定步速阅读任务中的数据）。

意见 2:

论文将自定步速阅读时间与多个个体差异量表指标进行了回归分析,通过分析斜率等指标间接探讨个体差异因素对反语理解的影响。

这一分析策略存在以下问题:

方法学局限: a 这种分析本质上只能揭示相关性,无法探讨深层次的认知机制,无法建立因果关系(如内隐态度"导致"加工困难; b 多个个体差异变量之间可能存在复杂的交互作用,但模型难以完全捕捉; c. 论文多处使用因果性语言解释相关性结果,如"内隐态度可能干扰意图推断".基于回归斜率的三重交互作用解释过于复杂,难以重复验证。

回应:

非常感谢审稿人的意见。针对审稿专家提出的关于回归分析方法的局限性,我们进行了深刻反思。首先,关于因果推断的严谨性,我们认同横截面数据的回归分析在推导因果链条上的天然限制。为此,我们已对全稿进行了语言规范化处理,将原先带有因果导向的表述(如“导致”、“驱动”)修正为更为客观的关联性描述(如“预测”、“与……相关”),以准确反映变量间的统计关系。我们重写了有关个体差异的讨论,谨慎推论,避免对结果进行过度解释。并且我们在总讨论中也强调了混合效应模型分析本质属于相关性探讨,无法推导出因果关系(详见 24 页的倒数第二段红色字体部分)。

其次,我们认同多个个体差异变量之间可能存在复杂的交互作用,模型难以完全捕捉,所以我们在进行数据分析时,只考虑个体差异变量和自变量之间的交互作用,不考虑个体差异之间的交互作用。同时,为了防止模型因为多重共线性无法捕获复杂交互作用,我们对模型进行了共线性检验,发现所有 VIF 值都小于 5,确保模型估计的稳健性。

针对审稿人关于三重交互作用不可重复性的担忧。需要向审稿人坦诚说明的是,本研究纳入的多种个体差异变量(如内隐联想测验 D 分数、观点采择得分等)虽均有前期理论与实证研究支撑,但在整体设计上仍带有一定探索性质。我们在报告结果时,既希望与学界分享更多的研究发现,也试图在复杂的交互作用中梳理出有理论价值的规律。目前的好消息是,尽管两个实验都包含了多种个体差异,但两个实验中具有显著调节作用的变量相对一致,是内隐联想测验的 D 分数和观点采择得分。这一跨实验的一致性在一定程度上缓解了对于多重交互作用“偶然显著”或“统计假象”的担忧,为我们的论点——即读者自身视角与社会认知中的观点采择能力影响反语理解过程中的观点采择——提供了初步证据。

意见 3:

本研究测量了大量个体差异变量,涵盖工作记忆(阅读广度任务、操作广度任务)、认知控制(Stroop 任务)、社会认知能力(IRI 量表的观点采择、共情关注、个人悲伤、想象四个分量表)、内隐与外显性别态度(IAT、Bem 量表)以及语用能力(AQ 量表)等多个维度,并报告了大量涉及这些变量的统计分析结果。然而,这些指标的选取依据及其与核心研究问题的理论关联性不够清晰,存在理论驱动不足:论文的核心问题是“稳定的社会观念如何作为认知约束调节反语理解中的观点采择”。从这一问题出发,各个体差异变量应服务于特定的理论预测:共情关注(EC)、个人悲伤(PD)、想象(FS)这些 IRI 分量表与反语理解中观点采择的关系是什么?论文缺乏明确的理论预测。工作记忆与认知控制:作为一般性认知资源,其作用机制是什么?是支持信息整合,还是支持冲突抑制?论文未予明确。

回应:

感谢审稿人提出的宝贵意见,您指出的个体差异变量理论驱动不足、与核心问题关联不清晰的问题,对我们完善研究的逻辑框架至关重要。我们认同您的观点,变量的选取应直接服务于“稳定的社会观念如何作为认知约束调节反语理解中的观点采择”这一核心问题的理论预测。

基于您的建议，我们在修改稿中对前言部分进行了重点重写，旨在明确每一项个体差异变量的理论角色与纳入依据。核心思路是：将内化的社会信念（性别刻板印象）视为需要被克服或整合的“自我视角”基础，而一般认知资源与社会认知能力则分别从“加工能力”和“加工倾向”两个层面，调节个体在冲突情境下超越自我、采纳他人观点（即说话人观念）的动态过程。

具体而言，我们重新审视了理论框架与既有文献，决定将社会认知能力的分析焦点集中于认知观点采择特质。针对 IRI 多维共情量表，基于最新的神经科学证据（如 Valles-Capetillo et al., 2026），我们将共情视为个体克服认知约束、实现视角整合的关键特质。具体而言，观点采择（PT）代表了个体在认知上推断他人心理状态的稳定倾向；幻想（FS）反映了个体对虚拟语境的卷入能力，直接调节反语加工的效率；而共情关怀（EC）与个人痛苦（PD）则通过情绪模拟机制，辅助个体在社会认知冲突中对说话人的意图进行“情感定性”。这种多维度的考察能够更完整地揭示社会认知能力如何作为一种正向调节力量，帮助个体在面对违反社会信念的语境时，依然能够灵敏地捕捉并采纳他人视角。

同时，修改稿已明确区分工作记忆与抑制控制这两项一般认知资源的不同作用机制。工作记忆主要为支持信息整合提供容量基础，而抑制控制则主要用于在视角冲突时支持对自我中心反应的抑制。这一区分直接对应于我们的研究假设，即高执行功能的个体在冲突语境下，能更好地整合信息并抑制干扰，从而更有效地完成观点采择。

意见 4:

实验 2 的一个关键发现是反语效应从目标区段延迟至溢出区段出现。作者将此归因于内隐任务条件下加工过程的自然延展。然而，这一效应也可能源于可理解性问题的设置特点——若被试在溢出区段仍需整合信息以应对后续提问，则该区段的阅读时间延长可能反映的是任务驱动的信息保持与整合需求，而非反语理解本身的认知特征。换言之，实验 2 溢出区段的效应可能与问题设置的任务需求有关，而非单纯反映“外显 vs. 内隐”任务类型的差异。

回应:

非常感谢这一具有启发性的建议。我们认真考虑了审稿人提出的这一替代解释。的确，仅从实验 2 内部看，溢出区段的效应既可能反映内隐加工的延迟效应，也可能源于被试为应对理解性问题而进行的策略性信息保持。为厘清这一问题，我们依托实验 1 与实验 2 的对比设计进行分析。

两个实验采用了完全相同的是/否理解性问题（频率、位置、内容均一致），唯一的系统性差异在于实验 1 要求被试在阅读后完成外显的反语等级评定，而实验 2 无此要求。若理解性问题本身足以驱动溢出区段的策略性整合，那么在任务负荷更高、被试对问题更警觉的实验 1（反语等级评定 + 理解性问题），理应出现同等甚至更明显的溢出区段效应。然而，实验 1 的反语效应严格限定于目标区段，未延伸至溢出区段；实验 2 则在目标区段无效应，溢出区段出现显著效应。这一反向关联模式强烈表明：效应位置的移动并非由理解性问题驱动，而是由外显评价任务的移除所导致。

此外，实验 2 目标区段的阅读速度模式也为这一推论提供了证据。无论经典刻板还是反向刻板条件，其阅读速度均显著快于中性条件。这说明在自然阅读中，明确的社会预期起到了认知支架的作用，促进了初步语义整合；同时也暗示，在自然阅读中，语言理解系统对语义冲突的监控阈限相对较低——反语带来的冲突并未在目标区段立即触发高强度的重新分析，而是被暂存或延缓处理，直至溢出区段才完成修正。这种加工的异步性特征，而非对特定问题的策略性应对，恰好印证了我们的推论。

最后，从实验 2 的效应时程分布来看，也为上述解释提供了支持。若溢出区段的阅读时间延长确实是被试为应对后续理解性问题而主动保持信息的结果，那么这种策略性保持应当

具有持续性——即从冲突信息出现之时（目标/溢出区段）直至问题呈现之前，加工负荷均应维持在较高水平。然而，我们在故事结果区段并未观测到稳定的反语效应。换言之，反语效应集中于溢出区段，恰是语言理解系统在自然阅读条件下异步修正的典型表现，而非对特定问题任务的策略性应答。

我们在修改稿的总讨论中补充了相关论述（见第 24 页 4.2 任务需求对社会预期效应及时间进程的影响 第二段红色字体部分）。

意见 5:

边缘显著结果过多，论文报告了大量边缘显著结果（ p 值在 0.05-0.10 之间）。过多报告边缘显著结果会增加一类错误风险，降低了研究结论的可靠性。

回应:

感谢审稿人的宝贵意见。我们承认，由于本研究设计涉及多个阅读区段、自变量、协变量及个体差异指标，客观上导致统计检验数量较多，这确实增加了出现边缘显著结果的概率。我们仔细复核全文后发现，大多数边缘显著的结果出现在与个体差异相关的三重交互作用之后进行的简单斜率分析中。为尽可能全面、透明地呈现研究发现，我们在文中详细报告了所有显著和边缘显著的简单效应分析的结果（包含 β 值、 p 值、95%置信区间等），但在结果讨论与核心结论的提炼上，严格依据经过多重比较校正后仍达到统计显著水平的效应展开，不以边缘显著作为推论依据。

意见 6:

统计结果汇报不完整。以事件区段为例：

"话语意图主效应不显著，二者的交互作用边缘显著。简单效应分析显示，三种说话人观念条件下，读者阅读字面意思和反语条件的时间均无显著差异($ps > 0.12$)。"

回应:

感谢审稿人的建议。在初稿中，我们没有完整报告出所有简单效应的详细结果，而是详细写出显著和边缘显著的成对比较结果，并简要总结不显著的结果。我们认同完整呈现简单效应统计参数对于读者全面评估数据的重要性，但在正文中呈现所有结果所占篇幅过大。为此，我们决定在附录中补全所有模型交互作用显著之后，基于 R 的 `emmeans`、`emtrends` 包进行的简单效应估计的结果（包括具体的估计值、标准误、 t 值及 p 值），并在正文相应位置增加了查阅指引。

需要说明的是，我们呈现结果的初衷是为了平衡信息的详略、避免正文过于冗长，从而突出核心研究发现。为此，我们在表格中汇总了所有阅读区段的主效应和交互作用，同时在附录中提供了混合效应模型的完整汇总表（`summary`），既方便对比结果差异，又保整体模型估计的透明度。

意见 7:

论文中术语使用前后不统一，例如："话语意图"与"话语类型"混用：有时使用"话语意图主效应"，有时使用"话语类型主效应"，指代同一自变量。

回应:

非常感谢审稿人的细致观察。我们已注意到术语混用可能带来困扰，因此在修改稿（含正文和图表）中，将相关表述统一为“话语类型”（修改处已用红色字体标识）。需要说明的是，虽然反语在语用层面包含讽刺、批评、幽默等多种具体功能，但本研究的核心目的在于考察反语表征与字面表征在认知加工上的差异，而不对具体的细分意图进行区分，因此我们认为使用话语类型作为条件名称更能客观描述实验处理。

此外，我们通读全文，对术语进行了全面梳理，将涉及可理解性问题的表述统一修改为“是/否”理解性问题，以进一步提升全文术语的一致性。同时，需要使用简称时，我们将 Bem 性别角色量表统一写为 BSRI，以与其他量表（如 IRI、AQ）的表达方式保持一致。

第二轮

审稿人 1 意见：

本文经过修改，论文质量得到进一步提升。尚有部分问题与作者探讨，具体如下：

意见 1：

在进行材料评定时是否评定了材料故事的合理性？特别是经典刻板、反向刻板和中性故事间的合理性应不存在显著差异。

回应：

非常感谢审稿人就材料合理性提出的建议。在先前版本中，我们仅评定了材料的反语程度和自然程度（指导语为“日常生活中，人们有多大可能会使用目标句来评价故事中所描述的事件”）。我们非常认同审稿人的观点，即“合理性匹配”对实验效度至关重要。因此，在修改版中，我们补充了合理性评定。

我们重新招募了 24 名被试，要求他们对实验材料的合理性进行 5 点评分（1-5, 1 表示非常不合理，5 表示非常合理）。评定结果与相关的分析见修改稿第 10 页第 1 段和第 9 页表 2 绿色字体部分。

具体说，本次评定的结果显示，字面意思条件下，三种说话人观念条件下，故事的合理性均较高（ $M_{\text{经典刻板}} = 4.39$, $SD_{\text{经典刻板}} = 0.57$; $M_{\text{反向刻板}} = 4.15$, $SD_{\text{经典刻板}} = 0.72$; $M_{\text{中性}} = 4.69$, $SD_{\text{中性}} = 0.34$ ），说明故事情境的合理性良好。相比于字面意思条件，反语条件下所有说话人观念条件下的合理性评分都较低（ $M_{\text{经典刻板}} = 2.10$, $SD_{\text{经典刻板}} = 0.65$; $M_{\text{反向刻板}} = 2.58$, $SD_{\text{经典刻板}} = 0.84$; $M_{\text{中性}} = 1.70$, $SD_{\text{中性}} = 0.53$ ）。合理性评分与反语等级评分呈极显著负相关（ $r = -0.90$, $p < 0.001$ ）。这表明读者对合理性的理解高度依赖于话语字面语义与语境事实的一致性，即合理性评分在很大程度上反映了反语操纵本身所带来的语义冲突，其较低的合理性评分验证了反语操纵的有效性。

但是，三种观念在故事间合理性上并不完全相等，字面意思条件下，反向刻板条件下的合理性略低于另外两种条件；反语条件下，反向刻板条件下的合理性略高于另外两种条件。累计链接混合效应模型（CLMM）分析显示，说话人观念主效应不显著（ $\chi^2(2) = 1.61$, $p = 0.448$ ），话语类型主效应显著（ $\chi^2(1) = 82.89$, $p < 0.001$ ），二者之间的交互作用显著（ $\chi^2(1) = 45.39$, $p < 0.001$ ）。进一步分析发现，字面意思条件下，反向刻板条件的合理性显著低于中性条件（ $\beta = 1.39$, $SE = 0.36$, $95\%CI = [0.54, 2.24]$, $z = 3.92$, $p < 0.001$ ），且低于经典刻板条件（边缘显著， $\beta = -0.89$, $SE = 0.38$, $95\%CI = [-1.80, 0.02]$, $z = -2.35$, $p = 0.057$ ）；经典刻板与中性条件之间无显著差异（ $\beta = 0.50$, $SE = 0.43$, $95\%CI = [-0.53, 1.53]$, $z = 1.16$, $p = 0.73$ ）。反语条件下，反向刻板条件的合理性显著高于中性条件（ $\beta = -1.97$, $SE = 0.34$, $95\%CI = [-2.79, -1.16]$, $z = -5.81$, $p < 0.001$ ），且高于经典刻板条件（ $\beta = 0.10$, $SE = 0.24$, $95\%CI = [0.42, 1.58]$, $z = 4.10$, $p < 0.001$ ）；中性条件低于经典刻板条件（ $\beta = -0.98$, $SE = 0.33$, $95\%CI = [-1.76, -0.19]$, $z = -2.98$, $p = 0.009$ ）。

正如审稿人指出的，在故事的合理性方面，经典刻板、反向刻板和中性故事间的合理性应不存在显著差异。这样才能保证后续的阅读结果的差异不是由于故事本身合理性的差异所致。客观而言，本研究材料评定确实未能实现三类故事在合理性上的完全等价（尤其体现在话语类型与故事类型的交互上）。

作为补充性的稳健性检验,我们尝试在实验 1 和实验 2 阅读时间相关模型中加入合理性评分作为协变量后发现,无论是实验 1 的目标区段和实验 2 的溢出区段,说话人观念与句子类型的交互作用依然显著(实验 1: $\chi^2(2) = 11.97, p = 0.003$; 实验 2: $\chi^2(2) = 7.19, p = 0.027$)。这说明,在剔除背景故事间合理性差异造成的干扰后,说话人身份对反语加工的调节作用依然存在,合理性差异对阅读时间的影响不足以构成对原有发现的有效竞争性解释。然而,考虑到从测量构念上看,故事的合理性与话语类型(字面/反语)存在不可分割的概念重叠——反语的本质正是字面意义的“不合理”。将合理性评分作为协变量纳入主模型,会在统计上过度控制掉反语的核心特征。因此,我们不将其作为正式分析中的协变量。我们已将上述讨论补充在讨论部分(第 27 页, 4.3 验材料合理程度的影响, 已用绿色字体标明)。考虑到合理性评分和反语在构念上高度重叠,若直接将合理性作为协变量纳入主模型,可能会掩盖真实的加工效应。因此,我们的主要结果仍然使用原分析策略。

意见 2:

10 个题目的 AQ 问卷信度只有 0.54,有些太低了,其结果是否可靠?类似的情况也出现在 IRI 的观点采择维度?但是 IRI 的信度在其他研究中报告的都不错,本研究比较低,是否意味着数据质量不够稳定?

回应:

感谢审稿人的敏锐观察。针对个别量表的信度偏低的问题,我们进行了以下分析,认为这更多反映了量表本身的结构特性和题目设计问题,而非数据采集质量的不稳定。

首先,我们推测,AQ 量表的沟通维度和 IRI 的观点采择分量表(PT)信度较低的原因可能是反向题目过多。AQ 量表的沟通维度 10 道题中包含 4 道反向题,IRI 量表 PT 维度 7 道题目中包含 4 道反向题。心理测量学研究指出,加入反向题通常会降低项目与总分之间的相关,进而导致量表内部一致性(如 Cronbach's α)降低(Vigil-Colet et al., 2020, Suárez-Alvarez et al., 2018)。相比之下,反向题较少的维度(如 FS 和 EC)在本研究中表现出了较好的信度(0.76 和 0.62)。

其次,无论是 AQ 量表,还是 IRI 量表,本研究中发现的信度模式和原量表中较为一致。虽然 AQ 量表和 IRI 量表都是在相关领域中被广泛使用的量表,但最初量表开发者报告的信度并不高(尤其是观点采择维度的建构信度)。具体的说,Baron-Cohen 等(2001)报告的 AQ 沟通维度的 alpha 系数为 0.65。戎幸等(2010)没有报告 alpha 系数,而是报告的四个维度的建构信度(Construct Reliability)和重测信度:FS、EC、PT 和 PD 四个分量表的建构信度依次为 0.75、0.60、0.59、0.69;间隔两周后的重测信度为 0.75、0.59、0.64 和 0.78。本研究发现的 AQ 沟通维度的 alpha 系数为 0.54。在本研究中,FS、EC、PT 和 PD 的 alpha 系数分别 0.76、0.62、0.30、0.60。本研究的结果与前人研究报告的模式较为契合。

尽管 alpha 系数较低,但本研究两个独立实验的数据显示出了较高的模式一致性。当前数据中,我们报告的是整个研究中(包含两个实验)的信度。拆开看,实验 1 的 AQ 信度系数为 0.58 ($SD = 0.09$)、实验 2 为 0.33 ($SD = 0.12$);实验 1 的 IRI 的四个分维度 FS、EC、PT 和 PD 的 alpha 系数分别 0.64 (0.06)、0.51 (0.12)、0.33 (0.17) 和 0.61 (0.08);实验 2 的分别为 0.74 (0.04)、0.71 (0.05)、0.26 (0.16) 和 0.60 (0.12)。可见,AQ 量表和 IRI 量表的 PT 维度在两个研究中信度是稳定的。IRI 的其他维度在本研究中表现良好,说明被试在作答时是认真的,数据稳定。

最后,我们认为较低的信度可能影响该变量作为调节因子的解释效力,尤其是 PT 维度。在两个实验中,我们都发现了 PT 维度与说话人观念和话语类型之间的三重交互作用,但是当前版本对于低信度对这一结果的分析不足。鉴于此,我们在每个实验的讨论部分和总讨论部分,都说明低信度可能表明测量可靠性较低。因此,谨慎看待文章中得出的与 PT 相关结

果（见第 20 页第三段最后两句、第 25 页第三段最后一句和第 28 页第一句，已用绿色字体标明）。此外，我们删除了有关 PT 的确定性结论。

参考文献

Vigil-Colet, A., Navarro-González, D., & Morales-Vives, F. (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema*, 32(1), 108–114. <https://doi.org/10.7334/psicothema2019.286>

Suárez-Álvarez, J., Pedrosa, I., Lozano, L., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, 30(2), 149 – 158. <https://doi.org/10.7334/psicothema2018.33>

意见 3:

请补充伦理审查信息，以及报告实验 2 的被试是否参加过实验 1。

回应:

感谢审稿人的宝贵建议。我们在实验 2 中补充了伦理审查信息。伦理审查文件已按要求发送至编辑部，如文章被录用，我们将补充包括伦理审批号等在内的完整信息。

我们确保了实验 2 的被试与实验 1 相互独立，避免潜在的练习效应或对实验范式的熟悉感。尽管有 2 名被试曾参加过实验 1，但其数据已从后续所有分析中剔除。因此，最终纳入实验 2 分析的 72 名被试均为首次参加本系列实验，未参与过实验 1 或者任何材料评定任务。

此外，所有研究阶段的被试样本均保持独立：参与过三次材料评定任务或两个正式行为实验（实验 1、实验 2）中任一实验的被试，均被排除在其他实验之外。我们在文中补充了这一信息（见第 21 页，3.1.1 被试部分的绿色字体部分）。

审稿人 2 意见:

本文经过作者修改后，内容表达上有大幅提升，逻辑也更加清晰。但文章仍然存在一些表述或逻辑不清的地方，需要进一步改进。

具体修改意见:

意见 1:

P5，引言，第二段中“这些研究主要聚焦于对物理层面共同知识(如“谁看到什么”)的在线整合,尚不足以揭示真实语言理解作为社会认知活动的核心——即推断他人内在的观念与态度(“谁相信什么”)。”这里的两类研究讲的不清楚。

回应:

感谢审稿人的指正。原稿中关于“物理层面”与“观念层面”观点采择的表述未能充分揭示两者的本质差异。针对这一问题，我们在修订稿中从信息基础与神经机制两个维度进行了重新梳理（详见修改稿中第 4 页最后三段绿色字体部分“既往研究常聚焦于视觉观点采择 (Visual Perspective Taking), ……从而无法完全采纳说话人的心理视角(Akimoto et al., 2012; Galinsky & Ku, 2004; 吴梦慧 等, 2022)。”）。

简言之，基于感知的视觉观点采择依赖当前物理环境中的直接线索（如遮挡、空间朝向、视线方向），判断他人“能看见什么”或“如何看见”，在神经层面主要表现为颞顶联合区（TPJ）的参与。基于社会知识的意图推断则依赖长期内化的社会知识（如刻板印象），推测他人的意图、信念和欲望，在神经层面更多涉及内侧前额叶（mPFC）。

意见 2:

P5 倒数第 5 行,“即使采用多种观点采择任务, Stroop 任务表现和工作记忆广度均与观点采择表现之间仍无稳定相关(Luce & Almor, 2023)。”那么本文为何要进行 Stroop 任务和工作记忆广度的测量。

回应:

感谢审稿人的宝贵意见。关于执行功能与观点采择关系的相关论述,我们补充了研究结果之间不一致的可能原因,即 Luce 和 Almor(2023)指出的观点采择在理解中可能并非单一能力,在不同任务中可能涉及不同的加工机制,需要区分任务类型。鉴于他们的研究侧重于视觉或语言结构层面的视角转换,而本研究关注的是性别刻板印象驱动的语用推理,其对执行功能的依赖程度可能有所不同。因此,本研究将 Stroop 任务和工作记忆广度测量作为探索性指标,考察执行功能在此类社会信念加工中是否表现出不同的关联模式(相关修改见第五页最后一句到第六页第一段绿色字体部分“现有证据关于执行功能与观点采择的关系并不一致……以探测执行功能在社会信念驱动的语用推理中是否呈现不同的关联模式”)。

意见 3:

P6, 第二段中“无论在外显反语意图加工任务中还是自然阅读中,均能在早期被自动整合。”这里对比的不是两种任务(一个是实验任务,一个是阅读方式),并且与上文表达内容不一致。请核查。

回应:

感谢审稿人的指正。在上一版修改中,我们使用“自然阅读”替换了“内隐任务”。如您所述,“自然阅读”是阅读方式,这种表述容易引起误会。相比于反语等级评定任务,仅用理解性问题监控注意力的任务的阅读只是更接近自然阅读,但不能说是自然阅读。实际上实验 2 使用“是/否”理解性问题监控注意力。因此,在这一版本的修改中,我们将文章中所有的“自然阅读”替换成“理解性问题任务”。

意见 4:

P6 最后一段,从共情,到共情量表的五个维度(认知层面和情感层面),缺乏直接的说明。

回应:

感谢审稿人的指正。我们增加了共情分维度的说明“共情包含认知和情感两个层面。其中,认知层面包含观点采择(Perspective Taking, PT)与幻想(Fantasy, FS)两个亚成分,情感层面包含共情关怀(Empathic Concern, EC)与个人痛苦(Personal Distress, PD)两个亚成分,这些亚成分通过不同的路径共同参与反语加工。”相关修改见第六页第二段绿色字体部分。

意见 5:

方法部分:从表 1 的实验材料来看,目标区段在经典刻板条件和反向刻板条件都可以表达反讽和字面意思,取决于事件区段的职业名词是否与观念一致,如果我的理解对的话,

内隐联想测验测的是性别与学科的内隐关联强度,研究者的实验材料中的刻板印象操纵的是性别-职业刻板印象,为什么不用性别-职业内隐联想测验?

回应:

感谢审稿人的宝贵意见。如您所指出的,我们通过“事件区段所描述的事件是否符合性别刻板印象(例如,母亲从事保姆或保安工作)”与说话人观念的一致性,来操纵话语类型变量。

在此需要说明的是,表 1 仅作为材料示例,其中使用的是性别-职业刻板印象。实际上,

如我们在第 7 页“2.1.3 实验材料与材料核证”第一段所述，本研究涉及多种类型的性别刻板印象材料，涵盖职业（如保姆、保安）、服饰（如领带、短裙）、专业（如英语、物理）和爱好（如化妆、抽烟）等。

经典文献（如 Nosek et al., 2007）中测量性别内隐偏见的范式通常有两种：一是性别-职业（家庭）IAT，二是性别-学科 IAT。性别-职业（家庭）IAT 通常操作化为“男性-职业、女性-家庭”的联结，而本研究所操纵的是不同具体工种之间的职业分工（例如保姆 vs. 保安），两者概念并不完全匹配。此外，Nosek et al.(2007)基于超过 250 万名被试的数据发现，性别-职业（家庭）IAT 存在显著的性别差异，即女性的内隐刻板效应强于男性；而性别-学科 IAT（测量女性与文科、男性与理科的联结）则无显著性别差异，具有良好稳定性，且普遍存在于各类人群中。因此，本研究采用性别-学科 IAT 作为内隐性别刻板印象的测量工具。

我们已经将上述内容，补充进 2.1.5 实验流程的内隐联想测验部分，以阐明选择性别-学科 IAT 的原因及其测量的有效性（具体位置请见修改稿第 10 页，最后一段绿色字体部分）。

意见 6:

讨论中“通过外显的反语等级评定任务，考察了说话人观念如何影响反语理解中的观点采择过程。”这里应该明确是对话篇章（短语对话）中说话人观念如何影响读者反语理解中的观点采择过程。否则，这个“说话人”好像变成主语了。文章的理解有点绕。

回应:

非常感谢审稿人的建议。我们将原句中“考察了说话人观念如何影响反语理解中的观点采择过程”修改为“考察了在对话篇章阅读中，说话人的观念如何影响读者在反语理解过程中的观点采择”（具体位置请见修改稿第 19 页，2.3 讨论第一段）。

同时，我们通读了全文，对类似可能引起主语歧义或缺少“读者”视角的表述逐一进行了修正，确保在讨论反语理解任务时，明确区分“说话人”（篇章内角色）和“读者”（实验被试），使研究问题中的观点采择过程更加清晰，增强文章的可读性。

意见 7:

“这一分化模式反映出高社会认知能力个体在反语加工中具备更高的意图整合效率，表现为更短的阅读时间；”突然出现“高社会认知能力个体”这个高社会认知能力个体是从何而来的？

回应:

非常感谢审稿人的建议。在上一版中，我们直接使用“社会认知能力”这一上位概念来指代本研究中测量的具体指标（如观点采择、共情关注），确实存在逻辑跳跃，给读者造成了困扰。

为了减少阅读障碍，在修改稿中，我们在深入讨论具体调节模式之前，增加了一段总括性文字（具体修改见 20 页第二段绿色字体部分“关于个体差异，……，均未表现出显著的调节作用”），梳理了各子变量与上位维度之间的逻辑关系。其中，社会认知能力维度只有观点采择和共情关注两个分维度表现出了调节作用，因此后文在讨论“高社会认知能力个体”时，读者可以清晰地对应到高观点采择或高共情关注能力的个体。

意见 8:

“字面表达条件下，内隐态度与反语等级评分呈显著正相关，而在反语表达条件下二者无显著关联；”这是否跟字面表达条件和反语表达条件下反语等级评分的得分范围有关，字面条件的反语等级得分范围相对更集中在低分，而反语可能全距更大？

回应:

非常感谢审稿人的宝贵意见。我们核查了两个条件下反语等级评分的全距，字面条件为 3.53，反语条件为 3.73，两者非常接近。因此，相关显著性的差异并非源于得分范围总宽度的不同。

进一步观察分布形态，字面条件均值为 1.51，标准差为 0.61，强烈正偏（偏度为 3.47），评分集中于低分区；反语条件均值为 3.88，标准差为 0.73，中等负偏（偏度为 -1.82），评分集中于高分区。根据一般统计学规律，得分范围受限（比如字面意思的“地板效应”）通常会削弱变量间的相关性，使显著性更难达成。字面条件在范围受限的不利条件下仍表现出内隐态度的显著效应，而反语条件变异范围更广、统计力理应更高，却未达显著。这一模式表明，字面条件下的显著关联并非由得分范围或分布形态导致的假象，而是真实效应。因此，我们判断观察到的显著性差异反映了内隐性别刻板印象在不同表达条件下的真实调节作用差异。

意见 9:

P23，讨论的第三段“实验 2 的另一个关键发现在于溢出区段。”上一段就在写溢出区段的结果，与这里的表述矛盾。

回应:

非常感谢审稿人的指正。在上一轮修改过程中，我们为了回应某条审稿意见，在讨论部分插入了一段关于“溢出区段”的说明。遗憾的是，这段插入未能充分顾及上下文的连贯性，产生了表述上的重复与逻辑矛盾。

在本轮修改中，我们已对讨论部分进行了重新梳理和调整：恢复原有的讨论顺序，即先集中讨论阅读过程中各个区段的结果，在适当位置插入的那段关于溢出区段的说明，作为可能没有进行反语加工猜测的讨论（具体见第 25 页第一段第四行“需要注意的是，此时的“无显著差异”并不意味着读者忽略了反语含义——如果仅按字面理解，反语条件的字面意思与已建立的说话人观念之间会产生冲突，反而会引发加工困难；实际并未出现困难，说明读者高效把握了说话人的真实意图，无需额外认知资源即可识别反语意图。”）。

此外，我们已通读全文，仔细排查了所有因修改可能导致的上下文断裂或不一致之处，并逐一进行了修正。

意见 10:

P24，4.1 中第 3-5 行，“这一结果支持假设 1，即语境强度与实验任务存在交互作用。”但是后续的解释中，“强预期线索(如经典刻板语境)在两类任务中均可提升语境权重。在无稳定社会预期条件(中性观念)下，两个实验均在不同时间节点观察到了明确的反语加工代价。”似乎说明没有交互作用，强和无社会预期条件，两种任务的效应是一样的？

回应:

非常感谢您指出的问题。我们完全同意您的判断。原稿中“这一结果支持假设 1”的表述过于笼统且不准确，因为后续对强预期与中性条件的解释实际上并未清晰体现交互作用，甚至可能让读者误以为两种条件下效应模式相同。基于您的意见，我们已对第 24 页 4.1 节相关表述进行了修改（详见修改稿中第 26 页 4.1 部分第一段的绿色字体部分）。

我们明确区分了不同社会预期条件下的结果模式与假设之间的关系：在高社会预期（经典刻板）条件下，两个实验均未在两类任务中观察到明显的反语效应，这支持假设 1 中关于“强预期线索提升语境权重、促进观点采择”的预测；而冲突性线索（反向刻板）在实验 2（无外显监控任务）中并未放大自我中心倾向，反而实现了与经典刻板语境同样高效的信息整合，这与假设 1 中关于冲突预期线索的假设不符。

.....

审稿人 3 意见：

作者对审稿意见的回应总体认真，核心问题（反应时指标删除、理论框架重构、替代解释排除）处理得当。但边缘显著结果的处理、PT 量表信度、以及复杂交互作用的精简仍需进一步修改。

意见 1：

边缘显著的使用仍需收紧。虽然作者声称不以边缘显著作为推论依据，但正文多处仍在解读边缘显著结果。例如，实验 1 事件区段的交互作用仅为边缘显著（e.g., $p=0.06$ ），随后仍进行了简单效应分析并在讨论中提及；目标区段个体差异分析中，中性观念下 D 分数与阅读时间的关系仅为边缘显著（ $p=0.07$ ），却在讨论中被作为核心论据之一。建议作者明确标注哪些发现属于探索性，并在讨论中降低对这些效应的解读权重。

回应：

非常感谢审稿人的指正。在本轮修改中，我们重新梳理了所有结果，重点修改了对于边缘显著的讨论及结论的问题。

首先，我们已从讨论中删除了实验 1 事件区段边缘显著交互作用及其简单效应分析的任何解读，仅在结果部分客观报告数据。

其次，在个体差异的考察方面，除内隐和外显刻板印象之外，其他的个体差异分析，我们在正文中已明确标注其为探索性研究。

最后，在修改版中，我们重新梳理了讨论部分，先总体描述以上个体差异中有哪些存在调节作用，然后逐一展开。在每一个实验的讨论中，我们都先着重讨论显著的效应，再提及边缘显著，并且强调边缘显著只能作为可能的趋势；讨论信度低的维度（PT）时，我们强调出要谨慎看待相关结果。在总讨论中，我们不在关注边缘显著的结果，仅在显著结果的基础上得出结论（具体请见实验 1 讨论部分、实验 2 讨论部分和总讨论，修改处均用绿色字体标注）。

意见 2：

IRI 观点采择分量表信度问题。实验 1 中 PT 分量表的 Cronbach α 仅为 0.30，这一信度极低，严重影响该变量作为调节因子的解释效力。然而，观点采择恰恰是作者跨两个实验一致强调的核心调节变量之一。作者需要正面讨论这一信度问题对结果解释的影响，并说明实验 2 中该量表的信度如何。

回应：

非常感谢审稿人富有建设性的意见。我们完全认同，需要认真讨论 IRI 观点采择（PT）分量表较低的内部一致性信度对结果与结论的潜在影响。因此，我们在每个实验的讨论部分和总讨论部分，都说明低信度可能表明测量可靠性较低，应谨慎看待文章中得出的与 PT 相关结果（见第 20 页第三段最后两句、第 25 页第三段最后一句和第 28 页第一句，已用绿色字体标明）。此外，我们删除了有关 PT 的确定性结论。

需要说明的是，本研究报告的量表信度是基于全部被试数据，而非实验 1 的被试数据。因为，两个实验中 IRI 四个分维度的信度模式与全部被试的信度模式基本一致：全部被试中，FS、EC、PT 和 PD 的 Cronbach's α 系数分别为 0.76、0.62、0.30、0.60；实验 1 中分别为 0.64、0.51、0.33 和 0.61；实验 2 中分别为 0.74、0.71、0.26 和 0.60。可见，在两个实验中 IRI 四个分维度的信度模式基本相同，具有跨实验的稳定性。基于这一稳定的信度模式，我们认为采用全部被试数据计算量表信度既能提高估计的精确性，也不会掩盖各实验内部的差异特征。因此，后续结果部分所报告的信度指标均基于总样本（ $N=144$ ）。为避免读者对信度计算的样本来源产生误解，我们在报告中已明确注明信度系数是基于全部 144 名被试计算

得出，而非单个实验样本（见第 12 页最后一段和第 13 页前两段的最后一句，已用绿色字体标明）。当然，在进行个体变量的调节效应时，我们使用实验 1 和实验 2 被试各自的个体差异数据。

意见 3:

两个实验样本量相同但个体差异模式差异较大。实验 1 中目标区段的个体差异显著，实验 2 中目标区段完全不显著；而溢出区段则相反。作者将此归因于任务需求的差异，但也应考虑统计效力的问题——对于三重交互作用，72 人的样本量检验力可能不足。建议在局限性中明确讨论。

回应:

感谢审稿人指出这一重要的统计问题。我们已经在修改稿中增加了有关个体差异结果可能存在样本量不足导致的统计检验力问题的详细讨论（详见第 28 页倒数第二段绿色字体部分）。

意见 4:

表 1 中的语料举例，由于列和行的排列方式和对齐方式不同，阅读起来较为费劲。

回应:

感谢审稿人的宝贵意见。我们对表 1 进行了调整：修改前的表格中，对于各条件下内容完全一致的区段，我们采用了省略处理；在修改后的版本中，我们补充了所有区段的完整表述，仅在事件区段通过括号与注释说明字面表达（符合说话人观念的事件）和反语表达（不符合说话人观念的事件）两种情况的差异。这样既保证了表格结构的完整性与可读性，也避免了不必要的内容重复。

第三轮

审稿人 1 意见:

经过本轮修改，本文质量获得了进一步提升。不过本文所用量表的效度较低，如信度低于 0.6（或其他标准）的相关结果是否考虑淡化或取消？另外，观点采择是一个容易被误解的术语，行为实验中的观点采择和问卷中的观点采择测量的并非一个心理概念。本文是采用问卷评定观点采择，建议在观点采择前增加一个定语或其他限制性表达，以便读者可以快速理解观点采择是如何测量的。

回应:

感谢审稿人的宝贵意见。AQ 沟通维度和 IRI-PT 的 Cronbach's α 未达到可接受水平（分别为 0.54 和 0.30），测量误差占比较大，若作为固定效应纳入混合效应模型，可能低估参数估计精度并削弱统计检验力。因此，本次修改完全取消了相关结果。具体地说，我们重新设计个体差异模型，不再纳入这两个变量（详见 14 页第二段黄色高亮部分“需要说明的是，……”）。结果和讨论部分我们也进行了相应的更新，详见实验 1 和实验 2 目标区段（第 17 页前两段黄色高亮处）、溢出区段个体差异结果部分（第 17 页最后一段黄色高亮处、第 22 页倒数第一段黄色高亮部分）、实验 1 反语等级判断的个体差异结果部分（第 18 页最后一段黄色高亮处）、附录中相应的表格，以上内容修改之处均已用黄色高亮标明）。需要强调的是，无论是否纳入这两个变量，其他个体差异变量的结果(方向及显著性)均无实质性改变。

此外，我们重新绘制了个体差异相关的结果图。具体地说，去掉所有 IRI-PT 相关的结果，根据新的混合效应模型，重新绘制 IRI-EC 和 IAT 相关的简单斜率图（与之前的模型相比，只存在数值上的差别，结果模式、显著性等并无变化）。讨论部分详细介绍了本研究中

信度较低的量表的处理策略和稳健性检验的结果（详见第 27 页第 2-4 段黄色高亮部分）。

我们完全同意审稿人有关两个‘观点采择’易被误解的看法。行为实验中所涉及的“观点采择”（即读者根据说话人观念推测其意图的在线过程）属于一种状态性观点采择，而 IRI 量表所测量的“观点采择”则是一种人格特质，反映个体在日常生活中自发采用他人视角的稳定倾向。为避免混淆，在本文中，我们将 IRI 量表对应的维度明确称为“特质观点采择”，并在首次出现时给出完整表述：“特质观点采择（Perspective Taking, PT）反映个体推断他人心理状态的倾向”。在实验流程——人际反应指数量表部分，我们明确写出以后简称“IRI-PT”（见 13 页第二段黄色高亮部分）。

本次修改发现一处笔误，错将大学生现代性别偏见问卷表述为 Bem 性别角色量表。本研究实际使用的量表即为“大学生现代性别偏见问卷”，文中该量表的参考文献条目、具体题项内容、维度分类标准以及计分公式均准确无误。现已在修改稿中将该处表述更正为“大学生现代性别偏见问卷”或者其简写 CMSS（见 13 页第三段黄色高亮部分）。该笔误仅影响量表名称的表述，不涉及数据采集、统计分析及研究结论的有效性。

编委意见：建议接受

主编意见：同意发表