

《心理学报》审稿意见与作者回应

题目：汉语语义双关与谐音双关笑话加工的神经机制

作者：靳婷婷 钱昱夫 范昕茹 王倩 任桂琴 李卫君

第一轮

审稿人 1 意见：

本研究（稿号：xb25-714）采用 ERP 技术考察了汉语语义双关笑话和谐音双关笑话加工的时间动态特征，结果发现与非笑话相比，笑话诱发了更大的 P200、P600 和 LPP，并且语义双关条件下的 P600 和 LPP 效应大于谐音双关条件。作者认为三个不同的脑电成分反映了笑话加工的不同认知过程，并且相对于谐音双关，语义双关笑话的不协调解决需要更多的认知资源，并且会带来更强的情绪奖赏体验。本研究结果相对明确，有一些问题需要进一步厘清或解释。

意见 1：作者需要在引言部分明确本研究与 Chang 等(2019)的区别，并在讨论部分强调本研究的进一步贡献。比如 Chang 等(2019)的研究采用的是单因素三水平的实验设计，本研究是 2x2 实验设计，本质上有何区别？看上去语义双关非笑话和谐音双关非笑话二者并没有太大的差别。

回应：审稿老师说得非常有道理。我们原稿中对本研究与 Chang 等(2019)的区别交代得不够清楚。针对这一问题，我们首先在修改稿前言部分第 19 页通过举例表明我们和 Chang 等(2019)的研究使用了不同的实验材料：“Chang 等(2019)的研究比较了语义笑话(例如：房租最便宜的地方在哪里？监狱)与谐音双关笑话(例如：如果麒麟飞到北极，他会变成什么？冰淇淋)的脑电反应，发现双关笑话比语义笑话诱发更大的 N400 波幅，提示双关笑话可能带来更强的语义预期违背和整合困难。”

此外，我们在方法部分(第 20 页)进一步阐明了我们实验设计和材料的特点：“本研究采用 2(材料类型：语义双关/谐音双关)×2(幽默性：笑话/非笑话)被试内设计，因此实验包括四种条件：语义双关笑话、谐音双关笑话、非语义双关笑话和非谐音双关笑话。”

“语义双关笑话与非语义双关笑话、谐音双关笑话与非谐音双关笑话共享相同“笑点”，仅通过铺垫句改变其语境解释方向，从而最大程度控制了关键末词的物理差异与低层词汇差

异(示例见表 1)。所有材料总字数控制在 37~42 字之间,且句末关键词均为 2 个字(最常见、最“标准化”的词汇单位),而不是以往研究中句末关键词为 2~4 个字(如 Chang et al., 2019),避免整词识别完成时间不同影响 ERP 效应。”

此外,关于“语义双关非笑话”和“谐音双关非笑话二者并没有太大差别”这一点,我们非常认同审稿人的这一观察。实际上,这恰恰说明在排除了“笑话性”之后,两类材料在非幽默语境中的加工较为接近,因此更有利于将笑话条件下出现的差异归因于幽默语境中的双关加工机制,而不是一般性句子加工差异。我们也将这一点补充在了方法部分第 22 页。

最后,我们在讨论部分进一步强调了本研究的贡献:“与谐音双关笑话相比,语义双关笑话诱发了更大的 LPP 波幅,表明语义双关在意义整合和后期评价加工阶段可能需要更多认知资源,并伴随更强的持续性情绪评价或奖赏相关加工。本研究揭示了汉语双关笑话加工的阶段性神经特征,并为理解语义双关与谐音双关在认知和情绪加工机制上的差异提供了电生理证据。”并在接下来的讨论中进行了详细的阐述和修改(详见正文第 33 页)。

意见 2: 采用脑电技术考察笑话的认知加工阶段的研究中,是如何将不同阶段的 ERP 成分与认知加工阶段对应起来的。在引言中似乎有一点循环论证,比如“在不协调检测阶段常观察到 N400 效应,其在刺激后约 400 ms 出现,并呈中央—顶叶分布 (Kappenman & Luck, 2012)。”。如何定义的不协调检测阶段?是否是因为语义违反诱发了 N400 效应,所以认为是不协调检测?

回应: 感谢审稿专家指出这一重要的理论表述问题。我们根据该意见已对引言中有关不同阶段的 ERP 成分与认知加工阶段对应关系重写,因为该部分内容改动较大,请审稿人详见第 16 页。本研究的逻辑是:理论定义认知阶段 → 实验操纵关键变量 → 观察预期 ERP 效应 → 为理论提供证据。我们更倾向于认为这是一个演绎验证的过程,而非循环论证。我们的研究主要基于幽默加工的三阶段理论(Chan et al., 2013; 2016),该理论从认知功能层面将不协调检测定义为“个体识别笑话结尾与铺垫预期之间语义/逻辑冲突的过程”。N400 成分并非幽默研究独有的指标,而是语言加工研究中已被广泛验证的、反映语义整合困难和预期违背的通用电生理指标(Kutas & Federmeier, 2011)。在幽默研究中,研究者通过对比笑话与非笑话条件,发现笑话条件诱发更大的 N400 波幅,并将这一成分作为不协调检测阶段存在的证据。

意见 3: 本研究是如何区分 P600 和 LPP 的?即认为本研究中 P600 反映语义整合需求而 LPP

反映愉悦体验与奖赏效应的依据是什么？从波形图上看，P600 和 LPP 并没有明显的分界，从 500ms 之后，更像是一个持续的晚期正波效应。换句话说，在结果和讨论部分，作者将 500ms 之后的效应分成了两个时间窗口的依据是什么？

回应：非常感谢审稿老师提出此疑问。我们同意审稿人的说法，即从波形连续性上看，500 ms 之后的效应确实可被视为一个持续的晚期正波。根据审稿老师的建议，我们不再将 500–800 ms 和 800–1000 ms 的正波效应表述为两个完全独立的 ERP 成分，而是将其整体界定为持续 LPP，并进一步区分为早期 LPP 和晚期 LPP 两个时间阶段。早期 LPP 主要用于描述偏向意义重构和语境整合，晚期 LPP 则用于描述偏向持续评价和情绪加工的后续阶段。这一部分的内容修改较大，体现在了“数据分析”部分：“在本研究的波形中，笑话条件相较于非笑话条件诱发的晚期正波呈现出较为连续的时间进程，我们将其界定为晚期正波的早期阶段和后期阶段。早期和晚期阶段的时间窗口分别为 500–800 ms 和 800–1000 ms，均选取前中部区域的 9 个电极(F2、FC1、FCz、FC2、F1、Fz、C1、C2、Cz)作为 ROIs 进行统计分析，比较条件间该成分的波幅变化。”

讨论部分我们也进行了相应修改，详见第 35 页：“需要指出的是，早期 LPP 与晚期 LPP 并非两个完全独立的神经成分，而是持续晚期正波在不同时间阶段中的表现。前者可能更偏向意义重构和语境整合，后者可能进一步叠加了持续评价和情绪奖赏加工。换言之，双关笑话的晚期加工并不是“先完成认知整合，再进入情绪评价”的截然分离过程，而更可能是认知整合、语境更新、幽默评价和情绪体验逐步展开、相互重叠的连续加工过程。”

意见 4：在 N400 窗口发现了材料类型的主效应，谐音双关比语义双关诱发了更大的 N400。两种条件下的目标词不同，作者是否控制了相应的词汇特征，比如词频、笔画数、具体性、情绪性等。

回应：感谢审稿专家指出这一潜在混淆变量问题。现有材料在编制和筛选时主要控制了总句长、关键词的意外程度、可理解程度与好笑程度，但原稿未单独报告关键词的词频、笔画数、具体性、情绪性等词汇变量。

我们现在已经补充了一个行为实验，发现两种条件下的目标词在词频和具体性上没有差异，在笔画数和情绪性上二者存在显著差异。请详见第 21 页：“此外，我们招募 20 名大学生(男性 8 名，年龄：21.70±1.84)检验语义双关与谐音双关语目标词在具体性、词频、情绪效价及笔画数这些方面是否存在词汇特征上的差异。独立样本 t 检验结果显示，语义双关语与谐音双关语目标词在词汇具体性($t(150)=-0.401, p=0.690$)和词频上($t(150)=0.524, p=0.603$)均不存在显著差异。不过，二者在情绪性($t(150)=2.091, p=0.043$)和笔画数上($t(150)=-4.17$,

$p<0.001$)差异显著;具体表现为,语义双关语目标词的情绪性评分高于谐音双关语目标词,笔画数则少于谐音双关语目标词。语义双关语目标词在具体性、词频、情绪性和笔画数上的均值 $\pm$ 标准差分别为 4.61 ± 0.73 、 3.971 ± 0.876 、 3.898 ± 0.292 和 13.01 ± 3.99 ;谐音双关语目标词分别为 4.66 ± 0.83 、 3.809 ± 1.067 、 3.695 ± 0.321 和 15.89 ± 4.51 。上述结果表明,语义双关和谐音双关这两类材料在具体性和词频方面匹配较好,但在情绪性和笔画数方面未能完全匹配。”

同时,我们在讨论第 36 页增加了对 N400 更审慎的讨论:“不过,这一结果需要谨慎解释。本研究中语义双关与谐音双关条件下的关键末词并不完全相同,因此 N400 的材料类型效应可能部分受到词汇属性差异的影响。材料分析显示,谐音双关目标词与语义双关目标词在情绪性和笔画数等方面存在差异,而既有研究表明,词汇情绪性也可能影响 N400 波幅。例如,消极词汇相较中性词汇可能诱发更负的 N400 效应(王振宏,姚昭,2012)。因此,本研究中谐音双关诱发更负 N400 的结果,一方面可能反映其相对更高的语义预期违背或整合难度,另一方面也可能受到目标词词汇属性差异的影响。未来研究有必要在更严格匹配词频、笔画数、具体性、情绪效价和语义关联强度的基础上,进一步检验语义双关与谐音双关在 N400 阶段的差异。”

意见 5: 作者认为 P200 效应反映了笑话理解的早期感知和注意分配过程,一般认为前部分布的 P200 效应与注意有关,而图 3(A)中的地形图显示本研究中的 P200 效应主要分布后中后部。

回应: 非常感谢审稿老师提出此疑问,也感谢审稿老师的细致审阅。我们根据该意见,重新选取了 P200 的电极点进行统计分析: C1、Cz、C2、CP1、CPz、CP2, P1、P2、Pz,覆盖中后部头皮区域。

在讨论部分,我们在现有讨论基础上对 P200 做了进一步解释(详见第 33 页):“此外,既有研究表明, P200 与刺激新奇性、知觉显著性、注意分配和早期情绪加工密切相关 (Carretié et al., 2001; Huang et al., 2015)。因此,本研究中笑话句诱发更大的 P200,可能说明双关笑话中的语义突转或幽默线索能够在较早阶段引发大脑对预期违背信息的快速探测。值得注意的是,本研究的相关分析进一步显示, P200 波幅与个体意外程度评分显著相关。这一结果从行为—脑电关联层面支持了 P200 对主观意外感或早期不协调觉察的敏感性。也就是说,刺激越被参与者评定为意外,其诱发的 P200 波幅越大,提示 P200 可能是双关笑话早期预期违背加工的重要时间窗口。不过, P200 并不应被简单等同于“不协调检测”本身,它也可能同时反映注意定向、语义显著性觉察和早期情绪唤醒等过程。”

审稿人 2 意见：

本研究采用事件相关电位技术，探究语义双关笑话与谐音双关笑话的加工差异。结果显示，相较于非笑话，笑话能诱发更大的 P200、P600 和 LPP 成分，这反映出笑话加工涉及不协调检测、不协调解决及情绪奖赏三个阶段。此外，与谐音双关笑话相比，语义双关笑话诱发的 P600 和 LPP 更大，表明语义双关笑话的不协调解决需消耗更多认知资源，且能带来更强的情绪奖赏体验。本研究的问题具有理论价值，方法选择恰当，研究结果有助于揭示双关笑话的认知神经机制，对补充和完善笑话加工的三阶段理论也具有启发意义。

意见 1：本研究旨在对比语义双关笑话与谐音双关笑话的加工差异，但前言中缺乏对这两种笑话类型的明确界定，建议在文章开头补充两者的明确定义与区别。

回应：谢谢审稿老师指出我们文稿中概念不清的地方。我们已在修改稿前言第 15 页补充了两类双关笑话的定义：

“双关语笑话是言语幽默的重要类型之一，指说话者有意利用词语或表达在语音、语义或语境层面的双重可解释性，使受众在同一语境中激活两个相关但不同的意义，并在意义错位、并置或转换中获得幽默效果。根据双关形成机制的不同，双关语笑话大体可分为语义双关笑话和谐音双关笑话。前者主要依赖同一语言形式的多义性、语义关联及语境驱动的意义重构，使同一表达在不同语义框架中获得双重解读；后者则主要依赖同音或近音词之间的语音联系，在呈现一个词项的同时激活另一个未直接呈现但与语境相关的意义(王夏菲, 佟和龙, 2024; Zheng & Wang, 2022)。”

意见 2：前言未提及 P200 相关成分的介绍，鉴于本研究将 P200 的结果作为核心发现之一，建议在前言部分补充相关内容。

回应：感谢审稿老师的宝贵建议。我们已在修订稿前言中补充 P200 相关研究，详见前言第 18 页：“除上述成分外，部分研究还发现笑话或非字面语言加工可在较早时间窗诱发 P200 效应。P200 通常出现在约 200—300 ms，可能与早期注意分配、语境驱动的语义预激活、显著性觉察或初步预期匹配有关。Shibata 等人(2017)发现，笑话相较于非笑话可诱发中央—顶叶分布的 P200 波幅，提示大脑可能在较早阶段就对笑点带来的预期违背作出反应。Regel 等人(2014)也发现，反语相较于字面语句可诱发 P200 效应，认为该成分可能反映早期语境驱动的语义关联加工。字面语言研究同样表明，P200 可受语义预期变化影响，说明大脑可能在 200 ms 左右即开始对输入信息与当前语境预激活之间的匹配程度进行快速评估(García-Sierra et al., 2021)。因此，P200 可能为笑话加工早期的显著性觉察或预期匹配提供时间窗口上的证据。”

意见 3: N400 与 LAN 的主要区别是什么? 研究假设中涉及 LAN 成分, 但数据分析部分未对该成分进行分析, 需说明原因。

回应: 感谢审稿老师的宝贵建议。我们已在修订稿中作两方面调整。首先, 在前言中补充说明 N400 与 LAN 的区别, 详见前言部分第 17 页: “在不协调检测阶段, N400 和 LAN 常被用来反映语言信息与语境预期之间的不匹配。N400 通常出现在刺激呈现后约 300—500 ms, 在中央—顶叶区域较为明显, 其波幅通常随着语义整合难度或语义预期违背程度的增加而增大(Kappenman & Luck, 2012; Kutas & Federmeier, 2011; Lau et al., 2008; Van Petten & Luka, 2012)。在笑话加工研究中, Coulson 与 Kutas(2001)发现, 与非笑话结尾相比, 笑话结尾可诱发更大的 N400 波幅, 表明笑点可能引发了更强的语义预期违背和语境整合需求。后续研究也在不同类型的幽默材料中观察到类似的 N400 效应(Benau et al., 2011; Du et al., 2013; Feng et al., 2014)。与此不同, LAN 通常表现为左前部区域的负波, 可能与局部语言冲突、结构性不一致或工作记忆负荷增加有关。部分研究发现, 在笑话或双关语加工中, LAN 可反映早期不一致检测或局部语义—结构冲突的处理(Canal et al., 2019; Coulson & Lovett, 2004)。例如, Li 等人(2025)在汉语谐音双关笑话研究中发现, 谐音双关可诱发明显的 LAN 效应, 提示该成分可能参与汉语双关语笑话中早期不一致信息的检测。需要指出的是, N400 与 LAN 在笑话加工研究中的出现并不稳定, 通常当冲突主要体现为词汇意义与整体语境之间的全局语义不匹配时, N400 更可能成为主要指标; 而当冲突更多涉及局部语言结构、语音—字形竞争、句法或工作记忆负荷时, LAN 或类似左前负波可能更容易出现(King & Kutas, 1995; Molinaro et al., 2011)。因此, N400 和 LAN 可以共同反映笑话理解早期的不协调检测过程, 但二者对应的认知机制可能并不完全相同。”

此外, 对于为何未对 LAN 单独分析, 我们已在修订稿中补充如下内容(详见修改稿第 24 页): “虽然基于 Li 等(2025)我们原先预期谐音双关可能诱发 LAN, 但在本研究总平均波形和地形图中, 并未观察到一个稳定、清晰、可独立界定的左前负波时窗与兴趣区, 因此正式统计分析未单独设立 LAN 窗口。”

意见 4: 第 5 页最后一行提到“Li 等人(2025)在谐音双关成语研究中发现, 相较于意义积极或中性的成语, 谐音双关笑话依次诱发了 LAN、LLAN 及 LPP 效应, 印证了笑话加工的三阶段理论”, 此处的 LLAN 具体指什么? 需明确。

回应: 感谢审稿老师的问题。我们在修订稿第 18 页增加了对 LLAN 成分的阐述: “LLAN(left late anterior negativity)通常出现在较晚时间窗, 分布于左前额区域, 可被视为与持续性语义

搜索、替代解释提取和框架转换有关的负向成分。Coulson 及其合作者的一系列研究表明，笑话理解常伴随框架转换(frame-shifting)过程，即个体需要从原先被激活的语义框架转向新的解释框架，以重新理解前文信息并形成连贯的幽默解释(Coulson, 2000; Coulson & Kutas, 2001; Coulson & Lovett, 2004; Coulson & Williams, 2005)。Mayerhofer 与 Schacht(2015)以及 Canal 等人(2019)的研究也进一步表明，笑话加工中的 LLAN 可能与替代意义搜索和语境重组有关。”

意见 5：方法部分：第 7 页仅报告了实验材料前测的统计分析结果，需补充不同实验条件下的基本描述统计结果(包括平均数和标准差)。

回应：感谢审稿老师的意见。我们已经在修订稿中补充不同实验条件下的基本描述统计结果(包括平均数和标准差)，详见第 21 页：“对所有材料进行 2(材料类型：语义双关/谐音双关)×2(幽默性：笑话 / 非笑话)统计分析，结果发现，在意外程度评分上，笑话的意外程度评分显著高于非笑话($F(1, 19) = 68.95, p < 0.001, \eta^2p = 0.78$)；谐音双关的意外程度评分大于语义双关($F(1, 19) = 6.44, p = 0.020, \eta^2p = 0.25$)；二者交互作用不显著($F(1, 19) = 0.48, p = 0.475, \eta^2p = 0.020$)。(谐音双关笑话：5.745±0.687，非谐音双关笑话：3.161±1.16，语义双关笑话：5.636±0.707，非语义双关笑话：3.087±1.206)。

理解程度评分上，笑话与非笑话的理解程度评分无显著差异($F(1, 19) = 3.52, p = 0.076, \eta^2p = 0.160$)；语义双关与谐音双关的理解性评分也无显著差异($F(1, 19) = 0.215, p = 0.648, \eta^2p = 0.011$)；二者交互作用也不显著($F(1, 19) = 0.057, p = 0.813, \eta^2p = 0.003$)。(谐音双关笑话：6.048±0.652，非谐音双关笑话：5.786±0.896，语义双关笑话：6.040±0.706，非语义双关笑话：5.759±0.915)。

好笑程度评分上，笑话的好笑程度评分显著高于非笑话($F(1, 19) = 134.26, P < 0.001, \eta^2p = 0.88$)；谐音双关语的好笑程度评分显著高于语义双关语($F(1, 19) = 134.26, P < 0.001, \eta^2p = 0.88$)；交互作用不显著($F(1, 19) = 0.29, P = 0.596, \eta^2p = 0.02$)。(谐音双关笑话：5.232±1.268，非谐音双关笑话：2.225±1.040，语义双关笑话：5.113±1.226，非语义双关笑话：2.142±1.022)。”

意见 6：第 8 页需补充理解程度中两个因素的交互作用情况，以及好笑程度统计结果的自由度。

回应：感谢审稿老师的意见。如上一条的意见回复，我们已经在正文补充审稿人提到的信息。

意见 7：材料前测采用 7 点李克特量表，而正式实验采用 4 点李克特量表，原因是什么？结果部分同样缺乏正式实验中行为结果的描述统计，需补充。

回应：感谢审稿老师的意见。我们已在修订稿方法部分(第 22 页)补充说明设置了不同量表的原因：“材料前测采用 7 点评分，是为了在材料筛选阶段获得更细致的分辨率，便于筛除边界材料。正式实验采用 4 点评分，主要考虑到 ERP 实验中每个试次后均需进行主观评分，若继续采用 7 点评分，会增加任务负担、延长实验时长，并可能引入额外的疲劳与注意波动；4 点评分更有利于被试快速完成在线判断，减少脑电实验中的非目标负荷。”

正式实验中行为结果部分的描述统计也已经补充在修订稿第 26 页：

表 2.四种条件下主观评分的描述性统计(M±SD)

	语义双关语笑话	非语义双关笑话	谐音双关笑话	非谐音双关笑话
意外程度	2.741±0.338	1.818±0.329	2.809±0.314	1.906±0.314
理解程度	3.390±0.353	3.363±0.406	3.340±0.331	3.318±0.400
好笑程度	3.031±0.357	1.717±0.328	3.092±0.353	1.725±0.313

意见 8：P600 和 LPP 成分的分析电极相同，结果模式也一致，二者是否为同一成分？如何通过这两个成分区分不同的加工阶段？需进一步论证。

回应：非常感谢审稿老师的细致审阅与宝贵建议。根据审稿人的建议，我们不再将 500–800 ms 和 800–1000 ms 的正波效应表述为两个完全独立的 ERP 成分，而是将其整体界定为持续 LPP，并进一步区分为早期 LPP 和晚期 LPP 两个时间阶段。早期 LPP 主要用于描述偏向意义重构和语境整合，晚期 LPP 则用于描述偏向持续评价和情绪加工的后续阶段。

这一部分的内容修改较大，体现在了“数据分析”部分：“在本研究的波形中，笑话条件相较于非笑话条件诱发的晚期正波呈现出较为连续的时间进程，我们将其界定为晚期正波的早期阶段和晚期阶段。早期和晚期阶段的时间窗口分别为 500–800 ms 和 800–1000 ms，均选取前中部区域的 9 个电极(F2、FC1、FCz、FC2、F1、Fz、C1、C2、Cz)作为 ROIs 进行统计分析，比较条件间该成分的波幅变化。”

讨论部分我们也进行了相应修改，详见第 35 页：“需要指出的是，早期 LPP 与晚期 LPP 并非两个完全独立的神经成分，而是持续晚期正波在不同时间阶段中的表现。前者可能更偏向意义重构和语境整合，后者可能进一步叠加了持续评价和情绪奖赏加工。换言之，双关笑话的晚期加工并不是“先完成认知整合，再进入情绪评价”的截然分离过程，而更可能是认知整合、语境更新、幽默评价和情绪体验逐步展开、相互重叠的连续加工过程。”

意见 9：图 2 的标题显示最右侧小提琴图的纵坐标应为“好笑程度”，但图中标注为“意外程度”，是否存在错误？请核查。

回应：谢谢审稿老师这么认真地阅读文稿，惭愧文中存在这样的问题。我们已核查图 2，确认原图中最右侧纵坐标标注存在错误，现已更正为“好笑程度”。

意见 10：图 3、4、5、6 的波形图是标注电极的平均值结果，还是某一个电极的结果？建议进一步澄清。此外，建议将这几个图整合，以方便读者阅读。

回应：谢谢审稿老师认真阅读本文稿。我们已在各图图注中统一明确：波形图均为兴趣区电极的平均波形，而非单一电极结果。

同时，我们同意图形较多会影响阅读流畅性，因此已在修订稿中做进一步的修改。目前 P200 和 N400 分析不同电极点，所以分别呈现，早期和晚期 LPP 分析了相同的电极点，所以在波形图中将二者合并一起呈现。

意见 11：第 16 页提到“本研究中笑话句的‘意外程度’评分显著高于非笑话句，这一行为结果进一步支持了 P200 与笑话加工早期‘惊讶感’或‘意外感’体验的关联”，建议补充意外程度评分与 P200 波幅的相关分析，以佐证这一观点。

回应：感谢审稿老师这一建设性建议。我们认同若能在行为评分与 ERP 成分之间建立直接关联，将显著提升结果解释的严谨性与说服力。

为了明确 P200 与意外程度的相关关系。我们采用线性混合模型，考察意外程度评分对 P200 波幅的影响，发现了意外程度评分与 P200 波幅的显著相关关系。具体结果见修改稿第 31 页：“为进一步考察主观评定指标与 ERP 波幅之间的关系，本研究采用线性混合效应模型进行分析。各模型均纳入被试随机截距，以控制重复测量数据中的个体差异。首先，以意外程度评分作为预测变量，分别考察其对 P200 和 N400 波幅的影响。P200 模型结果显示，意外程度评分能够显著正向预测 P200 波幅， $b = 0.73$ ， $SE = 0.24$ ， $t(95.63) = 3.00$ ， $p = .003$ ，表明刺激被评定为越意外，其诱发的 P200 波幅越大。模型截距显著， $b = 4.02$ ， $SE = 0.86$ ， $t(82.73) = 4.70$ ， $p < .001$ 。”。

意见 12：第 17 页提到“行为结果显示，谐音双关材料的意外程度和好笑程度评分均高于语义双关材料，由此推测谐音双关可能比语义双关更意外、语义不协调程度更高，进而导致其诱发更大的 N400 波幅”，建议补充行为评分与 N400 波幅的相关分析，以佐证这一观点。

回应：感谢审稿老师的宝贵建议。我们认同该分析对验证“谐音双关更意外，因而诱发更大N400”这一解释具有重要意义。

在修订稿中，结果补充在修订稿第31页：“N400模型结果显示，意外程度评分对N400波幅的正向预测作用未达到传统显著性水平，但呈边缘显著趋势， $b = 0.43$ ， $SE = 0.24$ ， $t(95.35) = 1.74$ ， $p = .084$ 。模型截距显著， $b = 2.10$ ， $SE = 0.91$ ， $t(75.47) = 2.31$ ， $p = .024$ 。该关系见图5B。

其次，以好笑程度评分作为预测变量，分别考察其对N400和早期及晚期LPP波幅的影响。N400模型结果显示，好笑程度评分对N400波幅的正向预测作用同样未达到传统显著性水平，但呈边缘显著趋势， $b = 0.31$ ， $SE = 0.18$ ， $t(95.02) = 1.78$ ， $p = .079$ 。模型截距显著， $b = 2.35$ ， $SE = 0.83$ ， $t(55.73) = 2.84$ ， $p = .006$ 。该关系见图5C。”

意外程度和好笑程度与N400均呈边缘显著关系。因此在讨论部分对原有解释进行了修正，不再将N400效应归因于意外程度差异，转而从表面语义关联、汉语形音分离特性、双关类型的表层特征差异以及词汇特征等角度对N400结果做出更贴合数据的阐释，使结论更加严谨。具体内容详见修订稿第35页：

“在本研究中，笑话性并未在经典N400时间窗内产生显著主效应；相反，结果显示谐音双关条件比语义双关条件诱发了更负的N400波幅，且该效应并不完全依赖于材料是否具有幽默性。这一结果提示，在汉语双关语加工中，N400可能更多反映双关类型所带来的语言表征差异，而不只是笑话性或幽默性本身。Zheng和Wang(2023)认为，双关语理解中未必存在一个典型的“不一致探测”阶段，因为双关词的两个意义在一定程度上均可与语境建立联系，第二意义可能通过语音或语义关联被快速激活，而不一定需要以明显冲突为触发条件。结合本研究行为结果，谐音双关材料的意外程度和好笑程度评分均高于语义双关材料，说明谐音双关可能给参与者带来更强的主观突转感和语义不协调体验，这可能是其诱发更大N400波幅的重要原因。

这一发现也可以从“表面语义关联”或形式关联角度加以解释。Marinkovic等(2011)的脑磁图研究表明，双关类幽默若与前文语境存在语音、字形等表面关联，可能会影响早期语义加工过程。对于汉语而言，字形、语音和语义之间既相互联系，又可相对分离(Zhang & Yang, 2004)。本研究中的语义双关和谐音双关在表层形式上存在差异：语义双关通常依赖同一词形或同一语言形式的多义性，其语音与字形线索较为一致；谐音双关则主要依赖语音相似性，目标词与被激活的另一意义之间往往存在字形差异。由此，语义双关可能因形式一致性更高

而较容易与语境建立关联，而谐音双关则需要在语音相似、字形差异和语义转换之间进行协调，因而可能带来更大的语义整合负荷，表现为更负的 N400 波幅。

不过，这一结果需要谨慎解释。本研究中语义双关与谐音双关条件下的关键末词并不完全相同，因此 N400 的材料类型效应可能部分受到词汇属性差异的影响。材料分析显示，谐音双关目标词与语义双关目标词在情绪性和笔画数等方面存在差异，而既有研究表明，词汇情绪性也可能影响 N400 波幅。例如，消极词汇相较中性词汇可能诱发更负的 N400 效应(王振宏，姚昭，2012)。因此，本研究中谐音双关诱发更负 N400 的结果，一方面可能反映其相对更高的语义预期违背或整合难度，另一方面也可能受到目标词词汇属性差异的影响。未来研究有必要在更严格匹配词频、笔画数、具体性、情绪效价和语义关联强度的基础上，进一步检验语义双关与谐音双关在 N400 阶段的差异。”

意见 13:行为结果显示谐音双关的有趣度高于语义双关，但 ERP 结果中语义双关诱发的 LPP 显著大于谐音双关。若 LPP 反映情绪体验强度，这两个结果是否存在矛盾？建议补充讨论。

回应:感谢审稿老师提出这一非常重要的问题。我们认为二者并不必然矛盾，原因已在修订稿第 37 页补充：

“因此，语义双关虽然不一定在显式评分上被评定为更“好笑”，但其理解过程可能需要更深层的语义加工，并在理解成功之后带来更强的认知完成感和持续评价投入(Zheng & Wang, 2023)。这一点也有助于解释本研究中行为结果与 ERP 结果之间看似不一致的现象。行为结果显示，谐音双关材料的好笑程度评分高于语义双关材料；而 ERP 结果显示，语义双关笑话在晚期 LPP 时间窗内诱发了更大的正波。二者并不必然矛盾。显式好笑评分更多反映被试对材料整体趣味性、通俗性或即时幽默效果的主观判断，而晚期 LPP 并不等同于“觉得更好笑”。它还可能反映理解成功后的持续性评价、情绪加工、动机性注意和奖赏相关资源投入。换言之，谐音双关可能更容易让被试迅速感到巧妙、有趣或意外，因此获得更高的显式好笑评分；而语义双关虽然显式趣味性评分较低，但由于其意义整合和框架重构要求更高，理解成功后可能带来更强的认知完成感和后期评价投入，因而表现为更大的晚期 LPP 波幅。”

第二轮

审稿人 1 意见:

感谢作者在第一轮修改过程中的工作，很好地回答了大部分问题。还有一些建议如下：

意见 1: 建议将理解程度作为预测指标考察其对 N400、早期 LPP 和晚期 LPP 的影响, 从而进一步明确各个脑电成分可能的认知意义。

回应: 感谢审稿专家提出的重要建议。我们认同, 将理解程度作为连续变量, 有助于进一步检验不同 ERP 成分所反映的认知加工过程。因此, 根据审稿意见, 我们补充采用线性混合效应模型, 以理解程度评分作为预测变量, 分别预测 N400、早期 LPP 和晚期 LPP 波幅。

结果显示, 理解程度对 N400、早期 LPP 及晚期 LPP 波幅均未达到显著预测作用, 相关统计结果见正文第 34 页: “其次, 以理解程度评分作为预测变量, 分别考察其对 N400 和早、晚期 LPP 波幅的影响。N400 模型结果显示, 理解程度评分对 N400 波幅的正向预测作用未达到显著性水平, $b = 1.120$, $SE = 0.795$, $t(118.69) = 1.409$, $p = 0.161$, 模型截距不显著, $b = -0.674$, $SE = 2.768$, $t(124.35) = -0.244$, $p = 0.808$ (该关系见图 5D)。早期 LPP 的模型结果显示, 理解程度评分对早期 LPP 波幅的正向预测作用未达到显著性水平, $b = 0.790$, $SE = 1.218$, $t(110.93) = 0.648$, $p = 0.518$, 模型截距不显著, $b = 3.319$, $SE = 4.137$, $t = 0.802$, $p = 0.424$ (该关系见图 5F)。晚期 LPP 模型结果显示, 理解程度评分对晚期 LPP 波幅的正向预测作用未达到显著性水平, $b = 1.242$, $SE = 1.319$, $t(102.96) = 0.942$, $p = 0.348$, 模型截距不显著, $b = 1.348$, $SE = 4.483$, $t(100.35) = 0.301$, $p = 0.764$ (该关系见图 5H)。”

我们认为这一结果具有两方面意义。首先, 它说明本研究中 ERP 效应并不能简单归因于句子的理解难易程度。由于四类材料在正式实验中的理解程度评分本身不存在显著差异, 且整体理解率较高, 说明被试能够较稳定地完成对所有材料的理解, 因此 ERP 差异更可能反映不同类型双关语所涉及的加工机制, 而非理解成功与否。其次, 这一结果进一步支持了不同 ERP 成分功能的相对独立性。N400 通常被认为反映在线语义整合及语义预期违背 (Kutas & Federmeier, 2011), 其变化并不必然对应最终理解程度; LPP 则更多反映持续评价、情绪加工及动机意义 (Hajcak et al., 2010; Schupp et al., 2006), 因此也未必随着理解程度线性变化。换言之, 个体即使最终获得相近的理解水平, 也可能经历不同程度的语义冲突、意义重构以及后续情绪评价。

考虑到本研究材料整体理解率较高, 理解程度评分分布范围相对有限, 确实可能降低了连续变量分析的统计敏感性。因此, 我们已在讨论部分增加这一局限性的说明, 并建议未来研究进一步扩大材料理解难度范围, 以更系统地检验理解程度与 ERP 成分之间的关系。

此外, 我们还在讨论部分 40 页补充说明: “值得注意的是, 我们发现理解程度评分不能显著预测 N400、早期 LPP 或晚期 LPP 波幅。该结果提示, 本研究中观察到的 ERP 效应并不能简单归因于材料理解难度的差异, 而更可能反映双关笑话加工过程中不同阶段的在线认知

加工。已有研究认为, N400 主要反映语义预期违背和在线语义整合过程(Kutas & Federmeier, 2011; Lau et al., 2008), 而 LPP 更多与意义重构后的持续评价、情绪加工和幽默欣赏等后期加工有关(Du et al., 2013; Feng et al., 2014; Canal et al., 2019)。因此, 即使不同材料最终具有相近的理解水平, 其加工过程中仍可能经历不同程度的语义整合和后期评价。另一方面, 本研究材料整体理解程度较高且评分变异较小, 也可能降低了连续变量分析的统计敏感性。综上所述, 理解程度作为最终行为结果, 并不能完全反映双关笑话理解过程中不同阶段的神经加工, 因此结合 ERP 时间进程分析对于揭示幽默加工机制具有必要性。”

意见 2: “不过, 这一结果需要谨慎解释。首先……。其次, ……”, 这两点实质是一个意思。另外, 除了语义和谐音条件下关键词的词汇特征(情绪性和笔画数)存在显著差异以外, 两种条件下的意外程度也存在显著差异: 谐音条件下意外程度更高, 也可能是该条件诱发了更大 N400 的原因。

回应: 感谢审稿专家的细致意见。我们重新审视了原讨论部分的相关表述, 发现原文中关于 N400 结果需谨慎解释的两点确有部分重复, 因此已对相关内容进行了合并和精简。同时, 我们采纳审稿专家的建议, 在讨论部分进一步补充了意外程度差异对 N400 效应的可能影响。具体而言, 本研究行为结果显示, 谐音双关条件的意外程度显著高于语义双关条件, 而更高的意外程度可能意味着更强的语义预期违背和语境整合需求。因此, 谐音双关诱发更负 N400 波幅的结果, 可能不仅受到目标词词汇属性差异, 如情绪性和笔画数差异的影响, 也可能与谐音双关条件下更高的主观意外程度有关。

据此, 我们采纳审稿人的观点, 在讨论部分第 38 页新增了关于意外程度差异也可能是 N400 效应的来源: “其次, 本研究行为结果发现, 谐音双关条件的意外程度显著高于语义双关条件, 而更高的意外程度通常意味着更强的语义预期违背与整合困难, 这也可能是谐音双关诱发更大 N400 波幅的重要原因。因此, 本研究中谐音双关诱发更负 N400 的结果, 一方面可能受到词汇属性差异的影响, 另一方面可能反映其相对更高的语义预期违背或整合难度。未来研究有必要进一步匹配不同双关类型的词汇属性, 并控制材料的主观意外程度, 以更准确地揭示语义双关与谐音双关在语义整合阶段的加工差异。”

意见 3: 作者认为语义双关在不协调解决和意义重构阶段可能需要更多的认知资源, 行为结果表明语义条件的理解程度高于谐音条件。能否进一步说明“需要更多的认知资源”和“理解程度”之间的内在联系?

回应: 感谢审稿专家提出这一重要问题。我们认同, 原文中对“认知资源投入”与“理解程度”之间关系的说明不够充分, 容易造成误解。在本研究中, 语义双关条件的理解程度高于谐音条件, 说明语义双关虽然可能需要更深入的意义重构, 但这种重构最终较容易形成连贯解释。相比之下, 谐音双关更多依赖语音相似性及音、形、义之间的转换, 可能具有更强的突转性

和意外性，但其最终语义解释未必同样稳定。因此，语义双关条件下较大的 LPP 和较高的理解程度可以被理解为：被试投入更多后期整合和评价资源，并最终形成了更充分、更连贯的理解。我们已据此在修改稿第 39 页进一步修改如下：“需要说明的是，较大的 LPP 所反映的“认知资源投入”并不等同于较低的最终理解程度。前者更多指双关笑话在线加工过程中语义整合、框架转换和意义重构所需的加工资源，后者则是加工完成后的主观理解结果。语义双关虽然可能需要更深入的意义重构，但其依赖语义关联和语境重释，最终可能形成更连贯、稳定的解释，因此表现出较高的理解程度；同时，其更大的 LPP 可能反映了这种成功理解过程中持续的整合、评价和意义重构资源投入。”

意见 4：建议进一步明确本研究与 Chang 等（2019）研究的差异。

回答：感谢审稿专家提出这一宝贵意见。我们重新阅读了 Chang 等（2019）的研究，并进一步梳理了两项研究之间的联系与区别。

回应：Chang 等（2019）主要关注不同类型笑话（语义笑话与双关笑话）的时间进程差异，重点讨论 N400 和 P600 所反映的不协调解决及框架转换过程。而本研究则聚焦于汉语两类双关（语义双关与谐音双关）的认知加工机制差异，不仅比较两类双关在不同加工阶段的 ERP 特征，还结合理解程度、意外程度等行为指标，对各 ERP 成分所反映的认知意义进行了进一步分析。

据此，我们已在引言部分第 22 页进一步补充与 Chang 等（2019）研究的差异：“Chang 等人（2019）比较了语义笑话(笑点主要依赖语义关系，需要通过语义推理解决不协调。例如：房租最便宜的地方在哪里？监狱)与谐音双关笑话(笑点主要依赖语音相似性/谐音关系。例如：如果麒麟飞到北极，他会变成什么？冰淇淋)的脑电反应。他们发现双关笑话比语义笑话诱发更大的 N400 波幅，提示双关笑话可能带来更强的语义预期违背和整合困难；同时，相较双关笑话，语义笑话在后部区域诱发了更大的 P600-like 波幅，提示语义笑话理解可能需要更多工作记忆资源，以建立新的语义关联并完成框架转换。”

意见 5：Punchline 在前言中译为“笑点”，而在实验材料部分被称为“妙语”，建议统一。

回应：感谢审稿专家指出该问题。我们已统一全文术语，将 Punchline 统一译为“笑点”。

意见 6：“我们招募 20 名大学生（男性 8 名，年龄：21.70±1.84)检验语义双关与谐音双关语目标词在具体性、词频、情绪效价及笔画数这些方面是否存在词汇特征上的差异”表述不完整。建议表述为：为了检验/控制语义条件和谐音条件下关键词的特征，我们招募了 20 名大学生，对关键词的具体性和情绪效价进行评分，同时查询了关键词的词频和笔画数……。另外，请明确词频是按照哪个数据库进行统计的。

回应：感谢审稿专家的建议。我们已按照建议在方法部分 25 页修改了相关表述：“为控制语义双关与谐音双关条件下关键词的词汇特征差异，我们招募了 20 名大学生（男 8 名，年龄 21.70±1.84 岁），采用 7 点量表对关键词的具体性、情绪效价进行评分，同时统计了词语

笔画数，并借助 SUBTLEX-CH 词频数据库查询了词频。结果发现，语义双关组 76 个关键词中仅 43 个可检索到词频数据，谐音双关组 76 个关键词中仅 39 个可检索到词频数据；且两组有效词频数据均呈偏态分布，不满足正态分布要求，因此无法采用独立样本 t 检验进行组间比较。鉴于此，本研究以被试主观评定的词汇熟悉性替代词频指标，完成组间匹配检验。”。

意见 7：建议在地形图图例中补充单位。

回应：感谢审稿专家的建议。我们已在所有 ERP 地形图图例中补充单位(μV)。

审稿人 2 意见：

作者较好地回应了前轮审稿意见，修订稿在研究背景梳理、双关类型界定、ERP 成分解释、材料描述统计以及行为—ERP 关系分析等方面均有明显改进。本文选题具有一定理论价值，将汉语语义双关笑话与谐音双关笑话置于同一实验范式中进行比较，有助于深化对汉语言幽默加工时间进程及其神经机制的理解。总体来看，稿件已经较为完整，但在方法表述、统计报告和文字规范方面仍有若干细节需要进一步修改。

意见 1：方法部分 P200 的 ROI 列出的是 FC1、FCz、FC2、C1、Cz、C2、CP1、CPz、CP2，但图 2A 中又写 P200 波形来自 C1、Cz、C2、CP1、CPz、CP2、P1、P2、Pz，两处电极集合不一致，请核查并统一。

回应：非常惭愧出现这样的问题。我们已将图 2A 与方法部分的 ROI 描述统一，均选取额中央区的 9 个电极(FC1、FCz、FC2、C1、Cz、C2、CP1、CPz、CP2)作为分析 P200 的 ROIs。

意见 2：部分章节编号不够规范，例如“2.3.1 实验程序”前似乎缺少“2.3”标题，随后又直接进入“2.5 数据采集与处理”，建议统一全文编号。

回应：感谢审稿专家的细致审阅。我们已全面检查全文标题编号，并按照《心理学报》格式要求进行了统一调整，确保各级标题层级清晰、编号连续。

意见 3：统计报告中 p 值格式不统一，如“ $P < 0.001$ ”“ $p < .001$ ”“ $p = 0.05$ ”混用，建议全文统一格式；其中 N400 相关结果中的“ $p = 0.05$ ”建议报告精确 p 值。

回应：感谢审稿专家的建议。我们已按照 APA 格式统一全文 p 值书写形式，统一采用斜体小写 p 表示统计显著性。同时，对原文中“ $p = 0.05$ ”的结果补充报告了精确 p 值 $p = 0.050$ 。

意见 4：效应量报告仍不充分，行为结果报告了 η^2p ，但 ERP 混合模型结果主要报告 F 值或 β 值，建议补充相应效应量。

回应：感谢审稿专家的建议。我们已在 ERP 混合模型结果部分补充了相应效应量。

意见 5：材料前测中“剔除任一维度评分低于 4 分的材料”表述不够清楚。由于非笑话材料的好笑程度本应较低，若同样按该标准剔除可能不合理，请说明该筛选标准是否仅适用于笑话材料，或不同条件是否采用了不同筛选标准。

回应：感谢审稿专家指出该表述不够清晰的问题。我们已在材料筛选部分进一步补充说明。原筛选标准仅适用于笑话材料，非笑话材料并未依据好笑程度进行筛选，而是依据其他匹配

指标进行控制。相关说明已补充至方法部分第 24 页：“针对笑话刺激剔除任一维度评分低于 4 分的材料，以及非笑话刺激剔除理解性低于 4 分的材料后，最终保留 304 条实验材料作为正式实验刺激，每个条件下材料为 76 条。”

意见 6：材料前测中补充了具体性、词频、情绪效价及笔画数的结果，但目前尚不清楚词频和笔画数是否也来自被试评定，具体性和情绪效价的测量方式也需进一步说明。

回应：感谢审稿专家指出文章中不清楚的地方。我们已在修改稿中进一步补充了各指标的来源与测量方式。其中，词频和笔画数来源于相关语言数据库；具体性和情绪效价则通过独立被试评分获得。相应结果已补充至材料前测部分第 25 页：“为控制语义双关与谐音双关条件下关键词的词汇特征差异，我们招募了 20 名大学生（男 8 名，年龄 21.70 ± 1.84 岁）采用 7 点量表对关键词的具体性、情绪效价进行评分，同时统计词语笔画数，并借助 SUBTLEX-CH 词频数据库检索词汇词频。结果发现，语义双关组 76 个关键词中仅 43 个可检索到词频数据，谐音双关组 76 个关键词中仅 39 个可检索到词频数据；且两组有效词频数据均呈偏态分布，不满足正态分布要求，因此无法采用独立样本 t 检验进行组间比较。鉴于此，本研究以被试主观评定的词汇熟悉性替代客观词频指标，完成组间匹配检验。独立样本 t 检验结果显示，语义双关与谐音双关目标词在具体性 ($t(150)=-0.401$, $p=0.690$) 和熟悉性 ($t(150)=0.524$, $p=0.603$) 上均无显著差异；但在情绪效价 ($t(150)=2.091$, $p=0.043$) 和笔画数 ($t(150)=-4.170$, $p<0.001$) 上差异显著，表现为语义双关目标词的情绪性评分更高、笔画数更少。语义双关目标词各指标均值 $\pm$ 标准差为：具体性 4.610 ± 0.730 、熟悉性 3.971 ± 0.876 、情绪效价 3.898 ± 0.292 、笔画数 13.010 ± 3.990 ；谐音双关目标词为：具体性 4.660 ± 0.830 、熟悉性 3.809 ± 1.067 、情绪效价 3.695 ± 0.321 、笔画数 15.890 ± 4.510 。”

意见 7：讨论部分“本研究的揭示了……”存在语病，应改为“本研究揭示了”。

回应：感谢审稿专家指出该问题。我们已修改为“本研究揭示了。。。”。

意见 8：统计结果中小数点后保留位数不统一，有些保留两位，有些保留三位，建议全文统一。

回应：感谢审稿专家的建议。我们已对全文统计结果进行了统一检查，并按照期刊规范统一小数点后保留三位数格式，保证报告的一致性。

第三轮

审稿人 1 意见：

作者在前两轮修改中已经很好地回答了相关问题，无进一步修改意见。

编委意见：该稿件已达到发表水平，同意接收。

主编意见：同意发表。