《心理学报》审稿意见与作者回应

题目：简化的联合再认范式中情绪对错误记忆影响的年龄差异
作者：肖红蕊　黄一帆　龚先旻　王大华

第一轮

审稿人 1 意见：该论文采用 Stahl 和 Klauer（2008）提出的简化联合再认范式和多项式测量模型，考察了词语的情绪效价对年轻人和老年人关联性记忆错觉的影响。这个模型在分离字面记忆（verbatim memory）、要点记忆（gist memory）以及基于字面记忆的回想-拒绝过程（recollection–rejection processes based on verbatim memory，属于记忆源监控的重要部分）方面很有优势。该论文采用这一范式考察了情绪对关联性记忆错觉影响的年龄差异，选题具有科学意义和创新性。有如下问题或意见供作者参考：

意见 1：在多项式测量分析中，只选择了 Gr、Vr 和 a、b 几个指标，对于分析情绪如何影响错误记忆的机制，感觉不够完整和全面。尽管文章的重点是分析错误记忆，但对两种记忆痕迹（verbatim vs. gist）的全面分析是非常必要的，例如 Gt 反映了目标探测项引发的要点记忆痕迹的提取，Vt 反映了目标探测项引发的字面记忆痕迹的提取，Vr 反映了相关探测项目引发的回想-拒绝的源监控过程，而不只是字面痕迹的提取，等等，这些指标的测量和分析对于考察情绪如何影响年轻人和老年人关联性记忆错觉的内在机制都是非常必要的。

回应：感谢专家的宝贵意见。在简化后的联合再认范式中，通过之后的多项式建模能得到四个记忆参数：Vt(由目标探测引发的一致性判断过程)、Gt（由目标探测引发的相似性判断过程）、Vr（由相关探测引发的不一致性判断过程，也称为回想-拒绝过程）、Gr（由相关探测引发的相似性判断过程），以及两个猜测参数 a、b。我们的研究仅选取了 Gr、Vr 和 a、b 这四个指标来分析，确实没做到全面兼顾。我们之所以这样分析是基于以下原因：

1）如文章中所述，Brainerd 和 Reyna 在 2005 年的研究中指出可能造成错误记忆的主要原因，①由相关探测引发的正确项目字面痕迹（即 Vr）的缺失，字面痕迹可以通过回想拒斥（recollection rejection）来抑制错误记忆，比如说“我没有听到过“凳子”这个词，我听到的是“椅子”和“座位”这两个词”。②由相关探测引发的相似性(即 Gr) 。要点痕迹通过增强相关探测与目标探测之间的相似性判断（similarity judgment）来增加错误记忆。③反应偏向。本研究中也是从这三个方面对错误记忆的机制进行分析的。而目标探测相关的指标（Gt、Vt）与错误记忆本身没有直接关系。

2）已有的使用联合再认范式来研究情绪错误记忆的研究也是仅仅分析与相关探测 R 有关的指标（如 Gr、Vr）来考察错误记忆的加工机制(如 Brainerd, Stein, Silveira, Rohenkohl, & Reyna, 2008)；Gt、Vt 是作为正确记忆加工机制的指标来进行分析的。

3）为了突出情绪对错误记忆影响的研究主线，而没有对正确记忆的结果进行报告。与虚报率（作为错误记忆表现的指标）分析相对应，在多项式建模的结果中仅呈现了错误记忆的指标 Gr、Vr，并且结合反应偏向，对情绪错误记忆背后的机制进行综合分析。如果加入对正确记忆的分析，对于错误记忆的理解和解释似乎并无明显助益，反而使结果与分析显得庞杂，模糊了错误记忆这条研究主线。

4）正如专家指出，“Vr 反映了相关探测项目引发的回想-拒绝的源监控过程，而不只是字面痕迹的提取”，在原稿中我们没有对 Vr 的含义进行详细说明，本稿中我们在相应位置表述了 Vr 的这一含义，见文章 3.2.1 部分和 4.3 部分的蓝色字体。

意见 2：某些统计结果的描述不准确，如“年龄主效应边缘显著（F（1,59）=5.27, p<0.05,η2 p =0.08）”，p 值小于.05 不应该称作边缘显著。p=.08, .09 称作边缘显著是否合适？

回应：感谢专家的提醒。“年龄主效应边缘显著（F（1,59）=5.27, p<0.05,η2 p =0.08）”是笔误，已在原文改正。

边缘显著本是一个比较模糊的概念，很多研究将 0.05＜p＜0.1 的差异检验结果视为边缘显著，虽然这样只是一种经验范围，而非理论范围，但是考虑到本研究为被试间设计，各年龄组的被试样本并不算大，如果增加样本数有可能使差异趋于显著。

使用边缘显著的结果，能更好的将情绪对不同年龄错误记忆影响的趋势呈现出来，并且这种趋势是比较符合理论解释。

为了使结果呈现的更加谨慎、科学，在边缘显著的结果处特添加了“有……的倾向”，并用蓝色字体标示，见文章 3.2 节。

意见 3：整体感觉 η2 比较小，效应不是很强。

回应：感谢专家的意见。

本研究中关于虚报率的效应量不高确实可能是本研究的一个不足之处。根据 Cohen 在 1988 年提出的效应量 η2 大小范围的划分，~0.02 为小，~0.13 为中等，~0.26 为大，本研究虚报率相关的效应量介于小到中等之间。

关于记忆痕迹和反应偏向的效应量，虽然多项式加工树建模（采用模型拟合指标 ΔG2 来衡量差异显著性，ΔG2 与 Δχ2 的意义相似）无法给出效应量，不过从描述值（均值及置信区间）来看，指标之间的差异还是比较明显的；统计推断值 ΔG2 也较大，推测其效应值还是能接受的（对于同一组数据，当自由度一定的时候，统计值如 F，t 等的相对大小能在一定程度上反映效应值的相对大小）。

参考文献：

Brainerd, C. J., Stein, L. M., Silveira, R. A., Rohenkohl, G., & Reyna, V. F. (2008). How does negative emotion cause false memories? Psychological Science, 19(9), 919-925.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

审稿人 2 意见：论文还存在如下问题，请作者予以斟酌和思考：

意见 1：DRM 范式错误记忆中，错误记忆的出现概率与词表本身的特殊性有密切的关系，可以说词表的负向关联强度越高，错误记忆越容易被诱发出来。作者是如何选择相关探测的呢？作者在选择相关探测项目的时候，有没有考虑到各个词表之间的同质性？因为有些相关探测如果恰好是 DRM 范式中的关键诱词的话，被错误记忆的概率就会提高；而相关探测的负向关联强度不是特别高的话，发生错误记忆的概率就会下降；在作者呈现的附录中，可以看到前 3 个中性词表中的相关探测都是关键诱词，而情绪词表则不得而知。请作者提供完整的词表，并说明项目的负向关联强度可能产生的影响；

回应：感谢专家的宝贵意见。

完整的词表已在附录中的实验材料中呈现。

由于原文中给出的相关探测选取方式的描述并不清楚，以至于使专家提出相关探测的不同关联强度可能是虚报率的混淆变量。相关探测的具体选取方法是这样的：选择每列学习词表的关键诱饵，以及每列学习词表中那个没有被学习过的词语，作为测验阶段的两个相关探测呈现。这种相关探测的具体选取方法已在文中 2.2 节修改以表述更清楚，用蓝字标示。

的确，与从词表内部选取的相关探测相比，关键诱饵诱发的错误率也许会更高，但是由于每个词表的相关探测都包括其各自关键诱饵，可以平衡这种干扰。但是另一个相关探测是从词表内部随机选取的，因此这种类型的相关探测的负向关联强度的大小不能保证一致，这的确是本研究在实验设计上存在的缺陷，这个问题在文章中的讨论部分的倒数第二段作为研究缺陷进行讨论，用蓝色字体标示。

本研究使用的 DRM 词表在以往研究中被多次使用，具有较高的适用性。但是由于实验材料来源国外的经典研究，在中国并没有本土化的词库版本，词表很多的重要的特征参数，均没有标准的评定值。国内相关 DRM 范式下错误记忆的研究，一部分使用国外翻译词表，正如本研究；一部分采取自编的 DRM 材料。这一方面增加了研究过程的繁琐和产生误差的可能，另一方面这种在材料上的不一致性，使得各研究所得到的结果的外部效度有所折扣。对于 DRM 实验材料的反思与展望，在讨论部分的最后一段具体说明，用蓝色字体标示。

意见 2：与上一个问题相关，老年人在记忆痕迹上表现出的差别，即积极情绪和消极情绪的要点提取痕迹概率都低于中性词表。除了作者解释的原因外，这一结果还可能与情绪词表相对于中性词表而言，更难产生语义激活，从而更难诱发错误记忆的出现有关，作者如何排除这一可能性呢？

回应：感谢专家的宝贵意见。

在文中我们基于社会情绪选择理论，推测可能是因为老年人对情绪词语编码加工得更好。专家指出“还可能与情绪词表相对于中性词表而言，更难产生语义激活”，这点意见非常好，从逻辑上讲这确实是一种可能性。

关于情绪与词列语义之间的关系，目前有两种观点：一种观点则认为情绪增加了词语间的语义关联；另一种观点认为情绪能增加词语间的区分度（distinctiveness）（见 Talmi et al., 2007 综述部分）。按照前一种观点，情绪是能强化语义激活的，而后一种观点强调的是感知觉上的区分度，未说明情绪如何影响语义激活。而有神经影像学的研究也指出情绪能促进词汇加工的各个过程（如 Kissler et al., 2006）。因此将老年人情绪与中性词语之间记忆痕迹的差异归因于“情绪更难产生语义激活”可能缺乏理论支持。

另外，如果说情绪阻碍了语义激活，那应该在年轻人中也观察到类似的现象（因为没有理论和研究指出这个在老年人与年轻人中存在差异），但事实上在本研究年轻人中情绪与中性词语的语义痕迹并没有差别。因此，“情绪词表阻碍语义激活”这种可能性很难用来解释本研究中老年人和年轻人的结果。

意见 3：从作者的描述统计结果中，如

T-u　　0.11±0.14　　0.44±0.10　　0.02±0.04　　0.21±0.15　　0.20±0.24　　0.14±0.13

这一栏里，可以看出，作为目标项目结果被判断为“未出现且无关”的概率在中性词表中是最低的，而在情绪词表中则比较高，这说明情绪词表的构建跟中性词表可能还是有一定的出入，而作者恰恰是以之作为比较的起点，所以需要对项目的确定情况做一些解释；

回应：感谢专家指出的错误。

T-u 这样的结果的确不正常，在仔细核对各判断类型的反应率之后，发现原文中描述统计结果有些数据是错误的，这的确是我们的粗心造成的错误，对此我们非常抱歉。我们在表上将改正的结果用蓝色字体标示，并且再次认真核查了文中所有的数据结果。

改正后的 T-u 结果为：

T-u　0.11±0.14　　0.15±0.14　　0.14±0.17　　0.21±0.15　　0.20±0.24　　0.14±0.13

对此结果进行方差分析，发现各效价间的 T-u 反应概率没有差异。

意见 4：是否可以对老年人和年轻人之间的差异进行直接比较呢？而不仅仅是从各种的结果中进行推论。

回应：感谢专家的宝贵意见。

年龄间直接比较的方法我们在结果分析时也曾考虑到，但是鉴于老年人和年轻人的记忆基线本身就有差异，考察情绪对不同年龄组记忆的影响时不能排除记忆基线的年龄差异，因此直接比较老年人与年轻人在某种特定刺激（比如积极情绪刺激）上的差异是没有意义的。在结果分析时，我们也尝试直接将 2（年龄）×3（效价）的两因素混合设计下各条件的记忆参数进行类似交互作用的分析，但是由于 MPT 软件和方法本身的限制，交互作用分析结果是不稳定的。已有的使用 SCR 范式的研究还没有人使用过多因素混合设计，因此无法提供更好的分析方法的参考。因而我们没有对老年人与年轻人在各个指标上的直接比较。

本研究采用了不同年龄组内分别进行比较的方法，得到不同年龄组的情绪-记忆作用模式，结果发现老年组与年轻组的情绪-记忆作用模式截然不同，这在一定程度上可以作为存在年龄差异的支持性证据。

在此稿中，我们做出推测的地方在于“老年人与年轻人在记忆痕迹和反应偏向为什么会出现差异？”，其背后的原因无法通过本研究的数据分析得到，因而只能根据已有的研究和理论进行推测。

参考文献

Talmi, D., Luk, B. T. C., McGarry, L. M., & Moscovitch, M. (2007). The contribution of relatedness and distinctiveness to emotionally-enhanced memory. *Journal of Memory and Language, 56*(4), 555-574.

Kissler, J., Assadollahi, R., & Herbert, C. (2006). Emotional and semantic networks in visual word processing: insights from ERP studies. *Progress in Brain Research, 156*, 147-183.

最后，文章中其它地方有些细节进行了修改，用红色字体进行了标注。

第二轮

审稿人 1 意见：作者针对审稿意见做出了认真的回复和相应的修改，对于文章的不足也在讨论中进行了阐述，修改后的文章有了较大的改进。但出现数据错误是令人遗憾的，建议进一步仔细核查所有的数据，保证数据的可靠性。

此外，既然作者也提到“考虑到本研究为被试间设计，各年龄组的被试样本并不算大，如果增加样本数有可能使差异趋于显著”，那为什么不补充一些被试数据让结果更坚实一些呢？

回应：非常感谢专家的意见。对前次的数据分析结果呈现的低级错误，再次表示抱歉。引以为戒，我们怀着谨慎的态度多次核查了数据，确保文章中呈现的数据都是真实可靠的。

对于被试样本问题，按照推论统计对样本量要求，我们在最初实验设计时选取老年人 30 人（删除 2 名无效数据），年轻人 34 人的样本量是适当的。但是，如一审专家所述，数据分析的结果中有几处是边缘显著，没有达到我们理想中显著的水平。根据经验，我们猜测样本量增加可能提升显著性。因此，我们在一审的“修改说明”中提到“如果增加样本数有可能使差异趋于显著” 这种可能性。很遗憾我们未能实现补充被试的弥补措施，主要原因是项目管理的客观局限和课题组下一阶段的项目实施计划。该实验数据于 3 年前由一位硕士生（即现在本文第二作者，已毕业）负责收集和管理，现已结项，遗憾的是我们与当时取样的社区负责人失去了联系，补充数据过程中可能产生的主试及取样误差会较大。并且我们课题组下一阶段的研究任务是拟对老年人错误记忆的情绪效应的产生机制（行为与神经两个层面）进行进一步的研究，其中就包括对前一阶段研究结论的部分验证。届时将另外取样提供基于充分的样本证据对现有结论进行验证。除此之外，考虑到本研究的主要结论是基于显著的结果之上。

因此，本研究没有选择追加少量被试的做法。谨慎起见，在几处差异边缘显著的结果解释上，均标明“有……的倾向”。

审稿人 2 意见：经修改后，论文达到了发表水平，建议发表。同时建议作者将后面修改时增加的关于词表问题所做说明的一段文字删除。即删除“有本土化的词库版本，词表很多的重要的特征参数，如负向关联强度（Backward Associative Strength）、正向关联强度(Forward Associative Strength)、内部关联强度（Interitem Associative Strength）、熟悉度（familiarity）、具体度（concreteness）等参数均没有标准的评定值。国内关于 DRM 范式下错误记忆的研究，一部分使用国外翻译词表(郭秀艳, 周楚, &周梅花, 2004;李林, 张金璐, &高旭辰, 2010)；一部分采用自编的 DRM 材料(张蔚蔚, 高飞, 蒋军, 张继元, &张庆林, 2012;白学军, 巩彦斌, &刘湍丽, 2014)。这一方面增加了研究过程的繁琐和产生误差的可能，另一方面所用材料的不一致性，使得各研究结果的外部效度有所折扣。在未来的研究中有必要对 DRM 词库进行本土化的评定和修正，以提供一套标准化的 DRM 范式的实验材料。”
回应：非常感谢专家的修改意见，已在文中将此段删除。

第三轮

主编意见：为了让结果更精确，表 2 和表 3 中的数值结果最好保留到小数点后三位数字。
回应：非常感谢主编的建议，已经在原文中将表 2 和表 3 中的结果增加到小数点后三位。