《心理学报》审稿意见与作者回应

题目：《基于合成平均刺激的平均表征机制--来自平均面孔吸引力的证据》
作者：田欣然；侯文霞；欧玉晓；易冰；陈文锋；尚俊辰

第一轮

审稿人 1 意见：

本研究考察了对多个面孔平均吸引力的加工，以验证平均表征的形成机制。作者认为结果说明面孔数较多时，平均吸引力的表征需形成平均面孔，而非简单对面孔吸引力进行平均。而面孔数量较少时，则更难形成平均面孔。该研究具有一定理论意义，但作者需要对其结果进行更进一步的解释，并加入新的分析方法，才能支持其结论。文章中较重要的问题如下：

意见 1：采用面孔吸引力考察平均表征的最主要的问题是面孔吸引力缺乏客观的量化指标，个体间的差别非常大，因此不适宜用群体的平均指标来代表个体的主观感受。例如，在文中作者计算集合面孔吸引力平均值时，采用的是所有人的平均值。在此处，建议作者使用每个被试单独对面孔的评价值来计算。即，作者应首先让参与实验的被试对每张原始面孔和合成的平均面孔一一进行评价，然后比较平均面孔的吸引力是否高于原始面孔吸引力的平均值。

回应：谢谢审稿专家的建议。研究中，实验 1 与实验 2 的主要目的是判断平均面孔是否形成，而专家提及的“平均面孔的吸引力是否高于原始面孔吸引力的平均值”在以往已经有一个较为稳定的结果，Francis Galton (1878)就已经通过摄影叠加技术发现，一组面孔合成的平均面孔的吸引力会高于各个面孔吸引力的平均值，并且在计算机生成和漫画脸等多种合成技术下得到验证（如 Langlois & Roggman，1990；Langlois et al.，1994；Rhodes & Tremewan，1996 等）。因此，原文中我们没有对此再进行评分验证分析。当然，吸引力的评价确实具有一定主观性，为了使结论更具说服力，我们在修改版采用参与实验被试的主观评分数据（实验 2 以及新增加的以主观评分为因变量的实验 3 和实验 4）再次验证了该结论。

具体来说，我们分别统计了实验 1 中核心实验条件下所有集合的成员吸引力的平均值 $M1 = 49.2$，又将生成的平均面孔吸引力也计算进集合的平均值得到 $M2 = 50.5$。t 检验结果表明，平均面孔提高了集合的吸引力平均值，$t(19) = 22.82$，$p < 0.001$，95%CI = [1.14，1.37]，Cohen’s $d = 10.47$，，即平均面孔的吸引力要高于原始集合成员面孔的吸引力平均值。实验 2 当时还让每个被试在主任务完成后对每张图片进行了吸引力评分。利用这些评分进行比较，

$t(33) = 2.35$，$p = 0.02$，95%CI = [0.51，7.05]，Cohen's $d = 0.82$，结果表明平均面孔的吸引力（55.2）确实高于原始集合成员面孔的吸引力平均值（51.7）。补充的实验 3、4 则采用直接评分的证据说明了平均面孔吸引力高于集合吸引力，也高于各个原始集合成员面孔的吸引力平均值（基于每个被试的单独评分）。

意见 2：作者对当前数据的解释还有不足之处。例如，按照形成平均面孔的假设，面孔集的吸引力应和平均面孔的吸引力相当，即判断探测刺激高于面孔集吸引力的比例应为 50%左右。作者对于不符该假设的解释是，平均表征形成之后，被误认为是面孔集的成员。这一解释还不够深入。首先，这部分解释缺乏支持证据，作者主观推断的成分较多。其次，作者还举出了反面证据，即面孔集中的表征会比单独出现的平均面孔更高。因此对于作者的解释，更加需要更多证据的支持。需要注意的是，面孔集和平均面孔吸引力的大小判断结果是比设置集合中是否包含平均面孔条件更直接的证据（部分原因见第 3 点），如果作者无法令人信服的解释这一现象（平均面孔吸引力高于面孔集吸引力的比例远高于 50%）出现的原因，则该文章无法建立牢固的结论。另外，如按照形成平均表征的假设，实验 1 的面孔集更容易形成平均面孔表征，实验 2 的面孔集较难形成平均面孔表征，则实验 1 的上述比例应该低于实验 2 才对，但是结果正好相反，如何解释这一现象？

回应：谢谢审稿专家的意见。可能我们的表述不够清晰（包括平均面孔的吸引力、集合吸引力、集合成员吸引力的平均值等概念的混淆）产生了某种误导。接下来我们尝试澄清这部分混淆：当被试观看一个面孔集（例如四张面孔集合）时，面孔集合各个原始面孔吸引力的平均值为 30，而基于各成员合成的平均面孔的吸引力为 50；如果加上合成的平均面孔，则整个集合的成员面孔平均值为 34（即（30*4+50）/5），整个集合的整体表征（作为整体评判）可能为 40。因此，平均面孔的吸引力始终是高于面孔集合吸引力的。就面孔集和平均面孔吸引力的比较而言，我们也同意专家关于"面孔集和平均面孔吸引力的大小判断结果是比设置集合中是否包含平均面孔条件更直接的证据"。因此，修改版中补充了实验 3 和 4，采用评分任务重复实验 1 和实验 2。补充的实验 3 和 4 数据也支持这种结果模式，为实验 1 和 2 面孔集和平均面孔吸引力的大小判断结果提供更直接的结果支持。但平均面孔参与到集合表征的形成只是一种尚未验证的可能性，是本研究的主要假设。平均面孔如何参与集合吸引力表征并不明确，一种可能的机制是平均面孔作为成员之一参与平均值计算来表征集合吸引力。专家所提的"按照形成平均面孔的假设，面孔集合吸引力应和平均面孔的吸引力相当"是另一种可能机制之一，即集合吸引力是通过平均面孔的吸引力来表征的，提供一个很好的

思考角度帮助澄清平均面孔对集合吸引力的表征机制。我们在讨论部分“6.3 集合吸引力与平均面孔的关系”进行了讨论，数据结果表明面孔集合吸引力低于平均面孔的吸引力，这个机制没有得到支持。

其次，关于反面证据“面孔集中的表征会比单独出现的平均面孔更高”，也可能是我们表述得不够清楚引起的误导。实际上这部分指的是说单个面孔吸引力在集合中比原面孔更高，而不是集合成员吸引力比平均面孔高。这更接近于一种同化现象，认为背景面孔导致会将人脸的感知偏向群体的平均水平。因此，伴随同一目标面孔出现的背景面孔吸引力越高，目标面孔就会被评价得越高（DeBruine et al.，2007；Perrett et al.，1994；Walker & Vul，2014）。因此，平均面孔产生可能会对集合中的其他成员产生影响，由于平均面孔的吸引力高，导致集合其他成员的吸引力也得到了抬升。这是平均面孔对其他面孔产生的影响，也可能是集合吸引力高评现象产生的路径之一。

最后，关于实验 1 的比例应该低于实验 2 的推论与我们的结果相反的解释。我们在原实验 2 结果中涉及到这个问题，但没有详细说明，在此重新澄清一下这个问题。实验 1 相对于实验 2，平均面孔表征确实更容易形成。然而，实验 1 中使用的是 12 张原始面孔形成的平均面孔（记作 A1），实验 2 使用的是 4 张原始面孔形成的平均面孔（记作 A2）。合成平均面孔的原始面孔数量越多，平均面孔吸引力越高，即小容量集合面孔合成的平均面孔吸引力也低于大集合合成的平均面孔(Langlois & Roggman，1990)。对实验 2 和实验 1 中平均面孔(探测刺激)和集合平均值（分别记作 S1 和 S2）的差值(9.5 vs. 16.4)进行跨实验比较，也发现实验 2 小集合平均面孔和集合平均值的差异更小，校正 $t(53.8) = 112.13$，$p < 0.001$，95%CI = [6.70，6.94]，Cohen’s $d = 27.53$。尽管实验 1 平均面孔更容易形成而提高集合吸引力，但平均面孔本身的吸引力也更高，从而更可能被判断为比集合吸引力高。也就说，A1–S1=16.4 > A2–S2=9.5，那么 P(A1>S1) > P(A2>S2) 也就顺利成章了。这是实验 1 中选择探测刺激吸引力更高的比例高于实验 2 的原因，因此实验 1 的比例应该低于实验 2 的推论在本研究中平均面孔吸引力也变化的情况下并不成立。

意见 3：作者的主要结论是基于对比包含平均面孔的面孔集和不含平均面孔的面孔集的结果。我们看到，实验 2 中，两者确实出现了差异，即包含平均面孔的面孔集形成的平均表征吸引力更高。而在实验 1 中未发现两者的差异，作者据此认为实验 1 中平均面孔的存在并未提升平均表征的吸引力。但是存在另外一种解释，即实验 1 使用了 12 张面孔，每一张面孔对平均表征形成的贡献都很小，因此包含平均面孔只能略微提升平均表征的吸引力，以至于

在行为上不能表现出显著的差异。如何排除这一可能？另外，从结果上可以看出，实验 1 的两种条件的比例均达到了 80 以上，是否已达到天花板，导致集合不含平均面孔条件下的比例无法继续进一步提升？为了排除这一可能，建议作者挑出平均面孔吸引力或面孔集吸引力均值较低和较高的试次，考察不同情况下的比例有无差别。

回应：首先，专家所说的“平均面孔的存在并未提升平均表征的吸引力”问题可能是由于本文关于几个概念（包括平均面孔的吸引力、集合吸引力、集合成员吸引力的平均值的表述还不够清晰产生了干扰。我们实际上认为平均面孔的存在是提升集合吸引力的原因（主要研究假设）。实验 1 中，未发现包含/不含平均面孔集合两者的差异，并不是由于平均面孔的存在并未提升平均表征的吸引力，而是由于在不包含平均面孔的集合里，被试也自动合成了平均面孔，因此两种集合的差异不显著。而平均面孔的出现实际上是提升了集合表征的吸引力的（见各实验结果分析的 *M*1，*M*2 部分和意见 2 的回应）。

其次，专家提出一种可能的解释：实验 1 集合容量大导致个体贡献小，从而“包含平均面孔只能略微提升平均表征（集合表征）的吸引力”，我们补充了实验 3 和 4（评分任务）来进行检验，结果并不支持这种解释。实验 3 中 12 个成员面孔的吸引力平均值 *M*31 为 45.8，包含平均面孔的平均值 *M*32 为 48.8，包含平均面孔的集合吸引力 *S*32 为 53.7；实验 4 中 4 个成员面孔的吸引力平均值 *M*41 为 46.5，包含平均面孔的平均值 *M*42 为 47.6， 包含平均面孔的集合吸引力 *S*42 为 50.2。从这些数据看，集合容量大，平均面孔对集合个体成员平均值提升也更大（$M32 - M31 = 3.0 > M42 - M41 = 1.1$），并且对集合吸引力提升也更大（$S32 - M31 = 7.9 > S42 - M41 = 3.7$）。因此，集合容量大导致平均面孔贡献更小的结论并成立。此外，根据实验 1 和 2 的数据分析，也同样不支持这个结论。对实验 1 和实验 2 中平均面孔(探测刺激)和集合平均值的差值(9.5 vs. 16.4)进行跨实验比较，也发现实验 2 小集合平均面孔和集合平均值的差异更小，校正 $t(53.8) = 112.13$，$p < 0.001$，95%CI = [6.70，6.94]，Cohen’s $d = 27.53$，即大集合中平均面孔对集合吸引力的提升也更大。

最后，为了排除天花板效应，根据专家建议，分别挑出集合包含/不含平均面孔条件下，平均面孔吸引力最高和最低的 5 个试次，共 20 个试次。结果发现，在吸引力最高的试次下，集合包含平均面孔与不包含平均面孔的条件无显著差异，$t(30) = 0.55$，$p = 0.586$，在吸引力最低的试次下，两条件也无显著差异，$t(30) = 0.99$，$p = 0.331$。由此可以排除天花板效应的影响。

意见 4：进行模型拟合时，需要报告模型的拟合参数和拟合度。

回应：扩散模型是对每个被试进行单独拟合，一般认为，如果模型拟合优度参数 *R*hat（Gelman & Rubin，1992）小于 1.05（也有认为小于 1.1），则拟合度较优，我们对所有的结果进行单样本 *t* 检验发现，各拟合参数均显著小于 1.05。

实验一 a（*M* = 1.0，*SD* = 0.00017）：*t*（31）= 1625.12，*p* < 0.001，95%CI = [0.0498，0.0499]，Cohen's *d* = 583.76；实验一 v（*M* = 1.0，*SD* = 0.00015）：*t*（31）= 1764.96，*p* < 0.001，95%CI = [0.0498，0.0499]，Cohen's *d* = 634.00；实验一 t（*M* = 1.0，*SD* = 0.00072）：*t*（31）= 385.93，*p* < 0.001，95%CI = [0.049，0.050]，Cohen's *d* =138.63；

实验二 a（*M* = 1.0，*SD* = 0.00013）：*t*（31）= 2043.97，*p* < 0.001，95%CI = [0.0498，0.0499]，Cohen's *d* = 734.22；实验二 v（*M* = 1.0，*SD* = 0.00015）：*t*（31）= 1828.95，*p* < 0.001，95%CI = [0.0498，0.0499]，Cohen's *d* =656.98；实验二 t（*M* = 1.0，*SD* = 0.00065）：*t*（31）= 430.91，*p* < 0.001，95%CI = [0.0494，0.0499]，Cohen's *d* = 154.79。

意见 5：正文中标注的 2.2.1 中 *M*1,*M*2“这一结果是如何计算出来的？什么叫假设合成平均面孔时集合成员的吸引力？为何不直接评价生成的平均面孔的吸引力？”

回应：在全部由原始面孔组成的集合中，在物理上并不存在这些原始面孔的平均面孔，而根据我们的实验假设，应当会生成平均面孔。我们首先计算所有原始面孔的吸引力算数平均值，得到 *M*1，而根据假设，在直接评价了生成的平均面孔吸引力之后，我们再次计算集合的算数平均值，这一次让平均面孔的吸引力也参加到平均过程中，进而得到 *M*2。*M*2 显著大于 *M*1，实际上是说明平均面孔的吸引力显著高于面孔集中的原始面孔吸引力平均值，进而说明合成平均面孔提高了集合吸引力平均值。补充的实验 3 和实验 4 给出了更直观的评价数据。

意见 6：“模型的介绍、拟合方法、参数设置、使用软件等信息均应该在方法部分介绍，结果部分仅介绍拟合结果即可。”

回应：已按照专家建议调整了模型介绍的位置。

意见 7：“在探测刺激为平均面孔条件下，什么反应是正确反应，什么反应是错误反应”。

回应：在模型拟合中，默认平均面孔吸引力高于集合吸引力，因此在探测刺激为平均面孔条件下，将判断平均面孔吸引力更高设定为正确反应，反之为错误反应。

意见 8：“实验 2 尺寸为何不与实验 1 保持一致？”

回应：实验 1 与实验 2 不是在同一时间段进行，由于实验 2 相对更简单，先于实验 1 完成。实验 1 根据屏幕显示效果调整了材料尺寸。补充的实验 3 和实验 4 保持了一致。

意见 9：实验 2 被试信息“是否与实验 1 一致，一致则无需报告”

回应：实验 1、2 被试不一致，采用的是被试间设计。

意见 10：正文中的其他细节问题（如首字母大小写、词语搭配，伦理规范报告，软件引用信息、图表等问题）。

回应：均已按照专家建议修改。

……

审稿人 2 意见：

平均表征的形成机制作为集合刺激知觉加工的一个关键性问题，是当前研究中的热点话题。本文另辟蹊径，拟采用集合面孔吸引力的平均表征和集合中所有面孔的吸引力平均值的差异来考察知觉平均过程中是否形成平均刺激的表征，试图分离集合平均刺激的特征值和集合各成员特征值的平均。该选题具有一定的新颖性和理论价值，但还存在如下问题：

意见 1：实验所采用的平均辨别任务用知觉比较后对集合平均表征的反应作为因变量来推断平均表征是否存在，主要依据行为反应数据来判断（主观判断），而且平均面孔是否包含在集合中，也只是操控了集合的个体信息特征。这种做法可以直接考察被试对个体信息的加工程度，但是用来考察平均面孔表征的形成机制，证据并不充分。

回应：为了提供更充分的证据，我们补充了实验 3 和实验 4，采用主观评分任务重复了实验 1 和实验 2 的结果。

意见 2：作者对面孔材料的选取标准并不严格，而且样本量没有达到 GPower 给出的最小样本量，这将降低实验结果的可靠性和可重复性。

回应：已补充被试数量达到 Gpower 要求的最小样本量，重新分析的结果已更新在正文实验 1 结果中。

意见 3：基于低级皮层表征的内隐知觉过程，究竟是先低级后高级 ？

回应：传统上，视觉系统被视为皮层区域和细胞类型的层次结构。低水平区域（V1，V2）的神经元接收视觉输入，并表现出简单的特征，如特定方向和位置的线条。它们的输出被高水平区域（V3，V4）整合和处理，这些水平逐渐概括空间参数，并表示全局特征。最后，更高水平的区域整合它们的输出来表示抽象的形式、对象和类别。

而逆层级理论（Hochstein & Ahissar, 2002）认为，上述的自低级向高级的过程是内隐的，而外显的感知则从高级皮层开始，外显的感知是基于低级皮层输入的近似信息的整合。因此，如果讨论“基于低级皮层表征的内隐知觉过程”，可以认为内隐知觉是先低级后高级的，而清晰的有意识的知觉是先高级后低级的。为了避免混淆，在修改版中我们删除“内隐”，改为“基于低级皮层表征的内隐知觉过程”。

意见 4：从互联网和中国化面孔情绪图片库中选取的图片数量应详细报告。

回应：评定后选择的完全不经过合成的原始面孔，其中包含 6 张互联网材料，高吸引力组有 4 张，低吸引力组有 2 张，其余所有图片都是从中国化面孔情绪图片库中选取的。

意见 5：面孔材料的吸引力水平（101 点量表评分）和效价评分究竟是几点量表？

回应：吸引力水平和效价评分都使用了 101 点量表，即 0~100 分评分。

意见 6：对高中低吸引力面孔的分类只是通过平均分，选择这三个分数的依据是什么？三组水平并没有进行差异显著性分析，这样的分类并不可靠。

回应：选择高中低吸引力面孔时，参照了吸引力水平评分和效价评分。效价评分中以指导语告知被试 50 分即为中性效价面孔，评分越高则情绪越积极，越低则情绪越消极。选择时首先排除了效价评分与 50 分存在显著差异的面孔，在剩余的面孔中，选择了吸引力评分最低、最高和最居中的面孔作为高中低三组。

使用检验三组评分之间的差异发现，组别两两直接差异显著，低吸引力组与中吸引力组：$t(19) = 7.91$，$p < 0.001$，Cohen’s $d = 3.63$；中吸引力与高吸引力组：$t(19) = 9.77$，$p < 0.001$，Cohen’s $d = 4.48$；低吸引力组与高吸引力组：$t(19) = 15.92$，$p < 0.001$，Cohen’s $d = 7.30$。

意见 7：不同吸引力分组的面孔集合分组的依据或标准是什么？

回应：面孔集合并未进行吸引力方面的分组，主要的分组标准就是是否包含平均面孔在内，其余面孔的选择分为若干种，如实验 1 不包含平均面孔条件下，集合刺激包含两高一中一低

（吸引力分组），或一高两中一低，主要目的是使实验中所有面孔集合使用的高中低三组面孔数量相同，来避免面孔集合吸引力过高或过低产生的潜在影响。

意见 8：本文的一个重要问题就是考察面孔的吸引力，可是实验设计部分的自变量只有集合类型，并不包含面孔吸引力因素。

回应：本研究的因变量是吸引力相关的变量，主要关注的是平均面孔的吸引力和面孔集合吸引力之间的关系，通过被试判断哪一方吸引力更高的比例来反映平均化的加工过程。因此，集合面孔的吸引力水平可以是自变量，但对于本研究问题并不是核心变量。

在原始材料中，我们选定了高中低三组原始面孔，是为了平衡可能存在的单张面孔吸引力对实验结果的潜在影响，在组织面孔集合时，也确保了高中低三组面孔被使用的次数一致，以此确保被试所评定的面孔集合整体上不是吸引力偏高或偏低的。

意见 9：集合平均面孔是怎样制作的，依据是什么？被试是否知晓平均面孔的定义和具体图片？

回应：集合平均面孔是使用计算图形合成软件 Abrosoft FantaMorph 制作的，该软件可以将两张面孔按照一定的比例融合。例如，当我们希望制作 4 张原始面孔的平均面孔，就将原始面孔两两一组，再按照 50:50 的比例进行合成取中，将合成的两张图片再次按照 50:50 比例合成，就相当于每张原始面孔在合成面孔的贡献比例为 25%，得到了 4 张原始平均面孔的平均面孔。如果要制作 3 张原始面孔的平均面孔，则控制每张面孔的贡献比例为 33.3%即可。

这类计算机 Morphing 程序是将面部各特征以众多关键点来标注，如嘴角的位置，大小，弧度，随后取关键点的平均值来合成图像。这种方法得到的合成图像质量很高，没有因污点和柔焦而产生的重影, 而且质量相当高。因此在面孔研究的应用上也较多（如 Rubenstein，Langlois，Kalakanis，1999）

被试在实验之前并未被告知研究目的，也并不清楚实验中存在原始面孔和平均面孔两种类型，两类面孔乱序、混合呈现。在实验结束后告知实验目的时向被试告知了平均面孔的存在但并未告知哪些面孔是平均面孔。

意见 10：指导语对结果的影响很大，请提供实验时的具体指导语。

回应：指导语如下：

接下来，您需要完成一个面孔吸引力判断任务。屏幕上将首先呈现一组多张面孔，随后

消失，消失后呈现单张面孔。您需要判断：先呈现的一组面孔的平均吸引力水平与后出现的单张面孔的吸引力水平相比如何。

如果您认为一组多张面孔的吸引力水平高于单张面孔，请按 F 键，否则请按 J 键。

意见 11：作者说所选图片均为中性图片，可是流程图上很明显有高兴的面孔，这是什么意思？

回应：实验选定使用的面孔的效价评分，与中性表情没有显著差异，$t(19) = 0.35$，$p = 0.732$。由于涉及到版权的问题，流程图的面孔并不是实验真实刺激材料，而是作为示意图，已采用无版权的图片重新制作流程图。

意见 12：屏幕上的集合刺激究竟有多少张图片，似乎与流程图不符。

回应：为了让局部信息清晰可见，流程图为实验 2 的流程图，集合刺激包含 4 张图片，实验 1 的流程除了集合刺激数量较多，以 4×3 矩阵排列以外没有差别。

意见 13：将两批不同被试完成的实验结果直接进行比较，这样不太恰当。

回应：实验中所用的被试都来自于同一高校的在校生，年龄、性别比例等人口学信息相近，可以看作被试来自同质群体随机抽样的被试间设计。

意见 14：数据的分析方法和统计检验的标准应该报告。

回应：主要的分析方法在实验结果部分进行了报告，主要的条件之间采用的是 *t* 检验，α 水平为 0.05。

意见 15：实验 2 中，“根据事先评定的得分，在探测刺激类型为新面孔和集合成员两种条件下的反应正确率为 84.17%，评定人员的详细信息以及面孔刺激的具体信息也需要报告。”

回应：评定人员的信息为：20 名北京某高校在校生，其中 10 名女性，平均年龄为 20.35 岁，标准差 2.03。所评定的面孔刺激即 2.1.2 中提及的各个集合的平均面孔以及原始面孔。

意见 16：讨论部分过度推论的现象严重，不够客观严谨。

回应：为了避免平均辨别任务间接证据的过度推论问题，我们采用直接的评分任务，补充了实验 3、4，希望能用较为直接的证据说清楚实验逻辑。

第二轮

审稿人 1 意见：

作者已对文章作出实质性修改并增加了控制实验，主要问题基本已得到解决。在此次修改中，最重要的问题是实验 3 和实验 4 的统计分析方法较为混乱，采用的统计方法和实验设计不对应，较为随意和不符合惯例，很多统计分析的目的不明。建议作者加强统计分析的逻辑性和较详细的说明，以更好的支持作者的结论。其他细节问题已在文章中标出。

回应：感谢专家的肯定和指正。在设计中集合平均值是计算值，主要是为了作为比较标准使用，与以往研究的分析设计不同，并非典型的单因素四水平设计。本研究的分析重点是检验不包括平均面孔的集合的吸引力与哪种条件或假设的结果值更接近，以此来推测没有物理平均面孔的集合是否与形成了平均面孔的结果更接近。为了更突显分析的目的，修改版中区分了包括/不包括平均面孔的集合成员平均值，按照单因素 5 水平设计进行分析。我们调整了统计分析部分并且重新斟酌了一些可能造成混淆的表述。其他细节问题也已经修改。

意见 1： 没有显著差异，p 却小于 0.05，相互矛盾。

回应：此处为笔误，已经更正统计分析，应为“选定的材料效价评分（M=49.94，SD=0.77）与中性（评分 50）没有显著差异，$t(19) = 0.35$，$p = 0.732$”。选择的图片从数据库中性图片，但我们另外评定了一遍，还是有个别图片被评为非中性而被排除在选择范围之外。

意见 2：Cohen’s d 属于人名，首字母应当大写

回应：感谢专家的指正，已经对全文的 Cohen’s d 进行修改

意见 3：M 是均值吗？st 代表什么统计量？需要事先说明。这里的 CI 是哪个统计量的置信区间？

回应：M 代表均值，st 应为 sd 即标准差，已在正文更正。这里的 CI 指的是拟合优度参数 Rhat 与拟合优度标准 1.05 的差值的置信区间。根据学报自检报告要求，统计分析比较时提供差异量的 95% CI，如果 CI 区间包含 0，代表差异和 0 没有显著不同，即差异不显著。考虑这部分内容并非主要部分，修改版删去了详细的细节分析，只保留了整体的拟合优度表述，以减少可能的误解。

意见 4 ：哪一个是平均面孔，哪一个是原始面孔平均值呢？建议直接将 *M* 和 *SD* 标注在相应的名称后面。另外，原始面孔吸引力的平均值和前面的集合中所有成员吸引力的平均值是一个概念吗？为何两者差异这么大（47.8 和>51）？

回应：谢谢审稿人的建议，已经将 *M* 与 *SD* 加入正文的相应名称后。平均面孔的吸引力为 55.2，“原始面孔平均值”为 51.7，该表述更正为“集合成员面孔平均值”以避免混淆。而前面所说的 *M*1=47.8，其实是不含平均面孔的集合中所有成员吸引力的平均值，也就是说 *M*1 当中是不包含“集合中存在物理合成的平均面孔”这一条件的，而 51.7 为成员面孔平均值，是将集合包括/不包括平均面孔两个条件合并计算的，当中包含了吸引力较高的平均面孔，因此两者差异较大。本研究的分析重点是检验不包括平均面孔的集合的吸引力与哪种条件或假设的结果值更接近。为了更突显分析的目的，修改版中区分了包括/不包括平均面孔的集合成员平均值。

意见 5 ：实验 3 既然在实验设计中已经说了实验是单因素被试内设计，应该先做方差分析，再做配对检验。

回应：感谢专家的建议，已经补充方差分析结果

意见 6：平均面孔吸引力显著高于整个集合吸引力，这里的集合是包括平均面孔的集合还是不包括的集合

回应：原本这里的“整个集合吸引力”是将集合包括/不包括平均面孔两个条件合并计算的。为了更突显分析的目的，修改版中区分了两个条件的集合吸引力，按照单因素 5 水平设计进行分析。

意见 7：这部分的检验不清楚是什么意思。例如，什么叫做平均面孔的吸引力在包含或不包含平均面孔的集合间没有显著差异？做这些比较的意义何在呢？

回应：这部分的统计分析原本是尝试了将平均面孔与不同的集合做了匹配，尝试了比较不同集合类型的平均面孔吸引力是否有差异，考虑到这部分并未主要内容，已经在修改版中删除，保留了较为重要的平均面孔和集合吸引力的差值在包含或不包含平均面孔的集合间的差异，（实验 3：差异不显著，$t(28) = 0.19$，$p = 0.852$；实验 4：差异显著，$t(29) = 6.40$，$p < 0.001$，95%CI = [4.26，8.26]，Cohen’s $d = 2.38$），这部分结果是为了从另一角度验证实验 1 或 2 中不同条件下被试选择探测刺激吸引力更高的比例变化的结果。

意见 8：照此图所示，实验设计应该是 2*3 被试内设计，前面又说是单因素设计，到底是哪种设计呢？另外，平均面孔不是单独评价的吗，怎么在集合面孔里也有平均面孔的吸引力呢？

回应：原图中的 2*3 设计的 6 个水平中，其中包含平均面孔吸引力、集合成员吸引力平均值依据集合类型进行再匹配，各自分为两部分而产生的新水平，因此其中包含了填充条件。目前已经对图表进行了调整。另外，平均面孔既进行了单独评价，同时在集合评定中也包括“包含平均面孔的集合”这一条件，就是将平均面孔作为集合成员之一与原始面孔一起进行评定，实际上与实验 1、2 采用的“包含平均面孔的集合”是相同的，只是进行的是评分任务。本研究的分析重点是检验不包括平均面孔的集合的吸引力与哪种条件或假设的结果值更接近，以此来推测没有物理平均面孔的集合是否与形成了平均面孔的结果更接近。为了更突显分析的目的，修改版中区分了包括/不包括平均面孔的集合成员平均值，按照单因素 5 水平设计进行分析。

意见 9：此处不宜做肯定结论，仅为边缘显著且效应值较小

回应：感谢专家的指正，此处已经修改表述

……

审稿人 2 意见：

在修改稿中，作者增加了两个直接的吸引力评价任务来深入地考察平均表征的形成机制，但依然存在如下问题：

意见 1：“吸引力评价任务要求被试对集合表征和平均刺激进行评价，直接反映平均表征的知觉。”表述不当，平均刺激可以进行外显评价，而集合表征是抽象概念，无法进行外显评价。文中还有许多类似的问题，在此不一一指出。建议作者通读全文，更改相关表述，保证论文的可读性。

回应：谢谢专家的建议，我们对全文的表述进行了一定调整，希望能增加论文可读性。

意见 2：经过图片评定和筛选后应用于正式试验的图片数量（互联网图片、中国情绪面孔表情库系统）应报告出来

回应：评定后选择的完全不经过合成的原始面孔共 30 张，其中包含 6 张互联网材料，包括

高吸引力组 4 张，低吸引力组 2 张，其余图片都是从中国化面孔情绪图片库中选取的。这部分是原始图片的数量组成，而用于正式实验的其余 335 张图片都是在原始图片的基础上合成的非真实人脸，不涉及到来源问题。

意见 3：中国化面孔情绪图片系统后面应注明相关文献。

回应：谢谢专家的建议，已经将相关文献注明在正文以及参考文献中。

意见 4：材料评定选用的分析方法是什么，“选定的材料效价评分与中性表情没有显著差异，$t(19) = 9.72$，$p < 0.001$。”难道选择的面孔表情不都是中性表情吗？效价评分与中性表情没有显著差异是什么意思？后面的 p 值又是小于 0.001，这不是代表着差异极其显著吗？这里这样表达的逻辑是什么？另外，对原始材料吸引力高中低三组的评定不是应该使用重复测量方差分析吗，为什么此处用的是 t 检验？

回应：此处统计结果引用错误，已更正相关表述：“…其中被评价为非中性（与评分 50 有显著差异）的材料被剔除。选定的材料效价评分（M=49.94，SD=0.77）与中性（评分 50）没有显著差异，$t(19) = 0.35$，$p = 0.732$”。选择的图片从数据库中性图片，但我们另外评定了一遍，还是有个别图片被评为非中性而被排除在选择范围之外。另外，已经补充原始材料吸引力高中低三组评定的方差分析结果，$F(2,38) = 148.64$，$p < 0.001$，$\eta_p^2 = 0.89$。

意见 5：在正文最好把平均面孔的制作方法写明，便于读者理解。

回应：谢谢专家的建议，已经增添相关说明。

意见 6：预先的吸引力评定中是只评了单张面孔的吸引力还是说也对不同面孔集合的平均吸引力进行了评定？

回应：预先评定只评定了单张面孔的吸引力。最后的统计计算基于实际实验中每个被试自己的事后评分。

意见 7：实验材料和实验程序部分的语言表述逻辑混乱，生涩难懂。

回应：谢谢专家的建议，我们语言表述进行了一定调整，希望能增加论文可读性。

意见 8：实验程序里面部分内容属于实验材料方面的内容，建议把相关部分放入恰当的位置，

表述也应当通俗易懂。

回应：谢谢专家的建议，我们对全文的表述进行了一定调整。

意见 9：这 180 个试次分几个 block 呈现？

回应：所有试次都在同一个 block 中呈现，不同条件的试次按照混合随机顺序进行，每 60 个试次休息一次。

意见 10：图 1 是流程图，并非比例图。另外，这句话放在实验程序里面的目的是什么？

回应：想要表示实际实验中的面孔图片与显示屏大小比例并不如流程图所示。似乎多此一举，可能引起误解，因此本轮修改删除了这句话。

意见 11：被试与显示屏间的距离是多少？

回应：被试直坐时双眼与显示屏距离约为 70cm，已补充这个信息。

意见 12：前文提到在探测面孔为新面孔和集合成员之一两种条件下，探测刺激在预评中的吸引力高于集合刺激成员的平均吸引力的比例为 50%，基本处于机会水平。而此处探测刺激类型为新面孔和集合成员之一两种条件下的总正确率达到 84.7%，远高于机会水平。在预评中是 50%，而平均辨别任务中是 84.7%，这二者差距还是蛮大的，怎么能说被试的吸引力判断和事先评定基本一致呢？更重要的是，预评中是直接对图片进行吸引力评定，而平均辨别任务中涉及集合面孔刺激和探测刺激的评定和比较，所包含的认知过程都是不同的，这样比较是否合理和可靠？科学依据是什么？

回应：此处原本的表述错误，可能对专家产生了误导。我们修改了表述：“一半探测刺激在预评中的吸引力高于集合刺激成员吸引力平均值，一半低于平均值”。正文中说“根据预评分数计算集合成员的吸引力平均值，再和预评的平均面孔吸引力（表述有误，更正：应为探测面孔吸引力）比较来确定正确反应，结果表明探测刺激类型为新面孔和集合成员之一两种条件下的总正确率达到 84.7%。”这时 84.7%的正确率实际上并不是判断探测刺激在预评中的吸引力更高的比例，而是在正式测量中结果与正确答案一致的比例。而正确答案的确定是按照单张面孔的评定进行数值的比较来确定的，设置了探测刺激的吸引力高于/低于吸引力平均值各一半，正确答案也就是 j/f 各一半，因此正式测量中正确率较高一方面验证了被试的吸引力判断和事先评定基本一致，也就是与平均值比较结果基本一致。

意见 13：“统计了假设该集合合成平均面孔时集合成员的吸引力平均值，也就是将假设生成的平均面孔吸引力计算进集合的平均值 *M*2 = 50.5。”作者将假设生成的平均面孔吸引力计算进集合的平均值，这样的计算方法似乎并非完全客观的，有什么前期研究依据？在此基础上得出的合成平均面孔提高了集合吸引力平均值结果又是否真的能反映被试主观上对面孔集合吸引力平均值的评定呢？

回应：首先，这些参与计算的值都是被试的主观评分。其次，之所以计算 *M*1 和 *M*2 并计算其差值，主要是为了说明如果合成了平均面孔，的确会提高了集合成员吸引力的平均值；但这并非直接作为合成了平均面孔的单一证据，而是和其他数据一起支持我们的结论。在实验 1~4 里，当我们假设平均面孔存在，将其计算进平均值，都导致了平均值的提高，也就是从理论上来说，平均面孔的存在提升了集合平均值，应该会导致原本就包含平均面孔的集合与平均面孔的差值相对于不包含平均面孔的集合与平均面孔的差值更小，因此产生选择比例上的差异，进而论证原本不包含平均面孔的集合可能形成了平均面孔。*M*1、*M*2 和集合评分比较，可以根据判断/评分结果推断 *M*2 更接近集合评分。

意见 14：“无论集合中包含和不包含平均面孔，被试判断平均面孔吸引力更高的比例显著高于随机概率（50%）。”那么，每种条件下的比例具体值应该详细报告。。

回应：具体比例值为包含平均面孔：83.55%，不包含平均面孔：84.03%，已经在正文中报告