《心理学报》审稿意见与作者回应

题目：学步期至青年期社交焦虑的发展轨迹和稳定性：一项基于纵向研究的三水平元分析
作者：陈必忠、黄璇、牛更枫、孙晓军、蔡志慧

第一轮

编委初审意见：

意见 1：我看纳入元分析的文章的年份跨度有 30 年（1992-2022)，那不同年份发表的文章数是多少篇？

回应：感谢编委专家的意见。考虑到文章的年份跨度较大和期刊的版面限制，我们按区间给出了各时间段内发表的文章数，并补充在“3.1 描述性统计”部分：“*文章发表在 1992 年至 2022 年间，其中 1992 至 1999 年 3 篇，2000 至 2004 年 4 篇，2005 至 2009 年 17 篇，2010 至 2014 年 33 篇，2015 至 2019 年 60 篇，2020 至 2022 年 41 篇。*”对于各年份的详细篇数，专家可以下载补充材料 2 的 Excel 文档，并在“发表年份”一列中点击筛选即可识别(https://osf.io/3vw74)。

意见 2：被试的年龄划分了为 11 个年龄组，首先，每个年龄组的文章有多少篇？其次，有些文章可能同时涵盖好几个年龄组，但在给出数据的时候可能只给出了一个平均数，遇到这种情况如何处理？

回应：对于第一个问题，专家可以看到独立样本数(173 个)略多于文章数(158 篇)，这是因为有的文章提供了不止一个独立样本。而我们以独立样本为单位进行分析，因为在三水平元分析中，效应量是嵌套在独立样本中而非文章。因此，我们报告的是每个年龄组独立样本的个数(见表 3 和表 5 的第二列)，而不是文章的篇数。对于第二个问题，的确，有些研究没有报告样本的详细信息(比如，仅描述为“中学生”)，故我们确实不能保证每个年龄组的所有被试都落在相应的年龄区间内，我们只能根据研究所报告的年龄均值进行估计。同时要确保年龄的离散程度在可接受范围内(比如标准差在 5 以下)，使得该年龄均值能较好地代表样本所属的年龄段；一些来自顶级期刊的研究采取的都是类似的做法(Mund et al., 2020; Orth et al., 2021; Orth et al., 2018)。从实际数据来看，本元分析所纳入的样本在 T1 时均龄的标准差也较低(M = 1.015, 在 0.09~4.08 间)，与上述的以往研究相当。

参考文献

Mund, M., Freuding, M. M., Möbius, K., Horn, N., & Neyer, F. J. (2020). The stability and change of loneliness across the life span: A meta-analysis of longitudinal studies. *Personality and Social Psychology Review, 24*(1), 24–52.

Orth, U., Dapp, L. C., Erol, R. Y., Krauss, S., & Luciano, E. C. (2021). Development of domain-specific self-evaluations: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology, 120*(1), 145–172.

Orth, U., Erol, R. Y., & Luciano, E. C. (2018). Development of self-esteem from age 4 to 94 years: A meta-analysis of longitudinal studies. *Psychological Bulletin, 144*(10), 1045–1080.

……

审稿人 1 意见：

意见 1：一个最大的问题是，纵向研究的发展轨迹的变化，需要测量等值性，另外效应量的合并也需要测量内容的一致性。而从婴儿期到成年期的社交焦虑的测量是否是等值、以及不同方法测出的结果是否能合并，都是值得商榷的。而该元分析并未对此做出讨论，而是假设不同年段测的社交焦虑是一个概念，或者可以直接合并。因此，存在极大的缺陷。

回应：接下来的部分，我们依次从测量等值、测量内容和测量方法三个部分对专家的质疑进行回应。

首先，我们肯定专家的看法，测量等值是纵向研究的关键环节之一，对此我们有几点需要说明。第一，本研究是元分析，使用的是二手数据，的确无法进行纵向等值，只能依赖所纳入的原始研究都进行了纵向等值检验；不过，遗憾的是，测量等值（尤其是纵向等值）是近年来才普遍使用在追踪研究中的方法，因此本文纳入的大部分研究都没有报告这个统计指标。有些研究虽然未检验社交焦虑的纵向等值，但检验了一些自行修订的量表的纵向等值性(郭海英 等, 2017)，这可能是因为很多社交焦虑的测量工具较为经典，其纵向等值性已被很多量表类文献所证实。比如，常用的青少年社交焦虑量表(SAS-A; Nelemans et al., 2019)、Spence 儿童焦虑量表(SCAS; Gong et al., 2021)和儿童焦虑性情绪障碍筛查表(SCARED; Behrens, 2019; Robe et al., 2022)均被专门的量表类研究发现其至少在特定的年龄组内是纵向等值的；而我们的研究也是首先在既定的年龄组内合并效应量，进而才根据先后顺序将 11 个已合并的效应量串联起来，而非直接合并所有效应量。应专家意见，我们再次阅读了所纳入的文献，发现 195 个效应量中，仅 16 个(8.1%)进行了纵向等值检验。同时，我们在原有文献质量评价的 5 个指标上，再将是否进行纵向等值及其结果做为评价指标之一，参见“2.2.2 研究特征”部分。此外，我们单独以是否进行纵向等值检验(0=没做，1=做了)为调节变量，

发现均未显著影响平均水平变化(B = 0.104, p = 0.275)和稳定性(B = 0.011, p = 0.847)，这说明未进行纵向等值检验的研究的效应量并没有显著偏离做了纵向等值检验的研究。

第二，“从婴儿期到成年期的社交焦虑的测量是否是等值”这个问题即便在实证研究中也难于解答。因为不同年龄的社交焦虑本来就不完全一致，幼儿的社交焦虑涉及游戏中对其他孩子的行为抑制，而青春期的社交焦虑则可能是与异性交谈时的紧张感。而且，目前几乎没有一个测量工具的适用范围可以横跨婴幼儿至成年期。以 Tang 等人(2017)的研究为例，该研究探讨了儿童早期(8 岁)至青年中晚期(30-35 岁)羞怯的发展轨迹，且在儿童期、青少年期和成年期分别使用了三种不完全相同的量表。这并非不合理，因为三个时期羞怯的具体内容本就不可能完全相同，这种量表不一致的情况也就无法检验纵向等值，因此他们使用了标准分数来合成发展轨迹，这时不同量表的分数就是可比的。本研究，以及其他顶级期刊的元分析也是如此(e.g., Mund et al., 2020; Orth et al., 2021; Orth et al., 2018)，通过标准化均值差这一效应量，使得不同测量工具和不同年龄段所测的水平变化是可比的；

其次，元分析的一个最大优势在于整合以往某一领域的相关研究，进而得出更一般性、综合性的结论(Lipsey & Wilson, 2001)。如果要求测量内容完全一致才能进行效应量合并，这几乎挑战了心理学领域大多数的元分析工作。比如，一项纵向元分析探讨了自尊和社会关系的相互预测关系(Harris & Orth, 2020)，而这其中社会关系整合了同伴关系、家庭关系和浪漫关系，很明显测量内容并不一致；但是，得出一般性结论的同时并不妨碍考虑具体某一类关系的特殊作用，因为这正是调节效应检验发挥作用的时候，这也是单个纵向研究难以做到的。再比如，有元分析探讨了孤独感的发展轨迹和稳定性，这其中既包括一般性的孤独感，也包括学校中的孤独感；既包括直接测量的孤独感，也包括间接测量的孤独感(Mund et al., 2020)。总而言之，元分析的价值就在于在合理的概念界定下（比如本研究在引言部分论证了情感、认知、行为和羞怯都是社交焦虑的一部分），尽可能地纳入更多的研究，从而得出更全面和准确的结论，但是又可以利用调节效应检验发现一般性结论外的特殊性结论。试想，如果本研究为了追求测量内容高度一致而只纳入社交焦虑的情感维度，不考虑认知维度和行为维度，这是否说明我们发现的年龄趋势并没有揭示社交焦虑的全貌？毕竟很多量表本身就综合了多个维度，比如社交回避及苦恼量表就包含了情感和行为两个维度。更为重要的是，本研究结果发现社交焦虑类型的调节效应均不显著，这也提示社交焦虑的平均水平变化和稳定性具有一定的跨维度不变性。

最后，根据我们的理解，专家提到的“不同方法”可能指的是问卷填答者的不同(自我报告 vs 他人报告)。考虑到社交焦虑既有内化体验(紧张不安)也有外化表现(回避退缩)，因

此基于他人报告的指标具有一定的客观性和合理性。特别是对于无法自主填写问卷的婴幼儿而言，父母报告和教师报告等手段是研究其心理与行为发展的常见做法。如果只纳入自我报告的效应量，本研究将损失较多婴幼儿期的数据，这难免降低了本研究的理论价值。同样地，研究结果也揭示了测量方式不显著的调节效应。

意见 2：在引言部分，作者没有给出足够的理由来选择“社交焦虑的等级排序稳定性”方法。因此，作者或许可以考虑说一下为什么选择“社交焦虑的等级排序稳定性”。

回应：关注社交焦虑的等级排序稳定性，有三点理由。**第一点是无法回避，在 1.2 部分第一段提到。**若想在群体层面系统地探讨某一构念的毕生发展，除了从绝对水平(即发展轨迹)上探讨构念的发展变化外，还需要关注相对水平(即稳定性)的年龄趋势。这是因为两种方法是相辅相成的，前者关注群体的平均水平趋势，而后者关注群体中的个体差异，只选择其一可能导致研究结论的片面性。**第二点是研究空白，在 1.2.2 部分提到**。几乎所有的研究者关注的都是社交焦虑的绝对水平变化(即发展轨迹)，即使他们在描述性统计中也会报告稳定性的指标。但是，从已有文献上看，社交焦虑的稳定性总体的程度究竟如何（或者说，其特质性程度如何）尚不清楚，有学者认为是相对稳定人格特质，有学者认为是频繁波动的心理状态。其次，社交焦虑的稳定性呈现出什么样的年龄趋势也不清晰。关注社交焦虑稳定性的发展有助于澄清上述问题，有较大的理论意义。**第三点是实践意义，我们将其补充在 1.2.2 部分。**发现社交焦虑稳定性何时较高、何时较低可能对其干预有较大启发。社交焦虑的稳定性较低的时期提示该阶段最适合对社交焦虑进行干预，因为这是最容易改变的“黄金时期”，且效果可能更好；而对于稳定性较高的时期，说明此时的社交焦虑较难改变，需要本人付出更多的努力和外界的更多关注。

意见 3：作者或许应该在引言部分，重点对“社交焦虑”的概念进行界定，并下操作性定义，什么属于“社交焦虑”，什么不属于“社交焦虑”。目前，该概念还很模糊，导致什么需要纳入元分析什么不需要纳入元分析不是很清楚。

回应：事实上，初稿也花了较大篇幅（1.1 部分）梳理了社交焦虑的结构和内涵，并指出了本研究所包含的社交焦虑有哪些内容。鉴于初稿可能说的不够详细，我们根据专家意见进行了稍微扩充。1.1 部分的逻辑是，第一段借鉴以往综述和元分析类研究指出了社交焦虑一般包括主观体验、认知、行为倾向和生理表现四个方面，随后对这四个方面进行简单介绍；第二段对社交焦虑和羞怯的概念进行了辨析，强调羞怯属于社交焦虑的亚类，即社交焦虑的类

特质成分；第三段，我们进行总结，并指出——“*综上所述，本研究所关注的社交焦虑是指个体在面对或想象中的社交场合时所表现出的情感状态(社交紧张、焦虑、苦恼)、认知(负面评价恐惧)、行为倾向(社交退缩、社交回避、社交抑制)和类人格特质(羞怯)四个方面。*”

同时，该部分也是对意见 1“而该元分析并未对此做出讨论，而是假设不同年段测的社交焦虑是一个概念，或者可以直接合并”的回应。正因为情感状态、行为倾向、认知特征和类特质成分四个都属于社交焦虑的心理与行为特征，即相当于社交焦虑这个潜变量下的维度或指标，故我们将这些有共同成分但又有其独立性的构念进行合并，但也探讨其调节效应。

意见 4：调节变量的选择上，作者或许应该考虑在 1.3 调节变量下面加一段，为什么要做调节变量分析，以及综合文献，以下这些变量最可能是调节变量。

回应：感谢专家的建议，添加该段文字后的确使得文章更为流畅且有逻辑性。我们在“1.3 调节变量”加了一段文字，如下：

“由前所述，社交焦虑的平均水平和稳定性的年龄趋势在以往研究中存在较大的异质性，这提示需要进一步探索潜在的调节因子。综合已有文献，以下变量可能会影响社交焦虑的平均水平变化和稳定性。”

意见 5：文献的检索和编码描述部分信息存在缺失。例如，作者可以考虑请添加检索关键词，检索的策略（包括布尔逻辑词），如何编码的， 多少人参与，不一致怎么办，等等。

回应：总的来说，以上问题在初稿中均有交代。首先，我们检索了 5 个数据库，而每个数据库的检索策略都有微小的差异，考虑到版面的限制和文章的可读性，我们没有把检索词和检索策略放在正文中，但是已经指出可以在开放数据库中找到(https://osf.io/7rj8e)，见“2.1 文献检索”部分的第一段。其次，在“2.2 文献编码”部分的第一段，我们也提到“*编码前参考以往研究和元分析专家的建议制作编码书，编码工作由两位拥有心理学硕士学位的学生独立进行。对于编码不一致的内容，通过协商讨论或再次阅读原文后达成一致。采用 Cohen Kappa 值衡量编码者间的一致性，0.8 及以上表示编码者一致性高(McHugh, 2012)。*”同时，在随后的 2.2.1 至 2.2.4 部分，我们也详细介绍了效应量、样本特征、变量特征和研究特征的编码，并报告了每个编码变量的 Kappa 值，而且详细的编码书仍然可以在开放数据库中看到(https://osf.io/3vw74)。

意见 6：作者可以考虑添加，文献质量的评估的环节，并做敏感性分析。

回应：初稿就已经做了文献质量评估，并根据意见 1 纳入了是否进行纵向等值检验做为新增的评价标准，见“2.2.2 研究特征”部分。同时，我们也在调节效应部分检验了其统计显著性。结果显示(见表 4 和表 6)，在平均水平变化和稳定性上，文章质量的调节效应均不显著，这提示研究结果对文献质量的高低不敏感。

参考文献

Behrens, B., Swetlitz, C., Pine, D. S., & Pagliaccio, D. (2019). The screen for child anxiety related emotional disorders (SCARED): Informant discrepancy, measurement invariance, and test–retest reliability. *Child Psychiatry & Human Development, 50*(3), 473–482.

Dirks, M. A., Weersing, V. R., Warnick, E., Gonzalez, A., Alton, M., Dauser, C., ... & Woolston, J. (2014). Parent and youth report of youth anxiety: Evidence for measurement invariance. *Journal of Child Psychology and Psychiatry, 55*(3), 284–291.

Gong, J., Wang, M. C., Zhang, X., & Yang, W. (2021). Measurement invariance and psychometric properties of the Spence Children's anxiety scale-short version (SCAS-S) in Chinese students. *Current Psychology*. Advance online publication.

Guo, H., Chen, L., Ye, Z., Pan, J., & Lin D. (2017). Characteristics of peer victimization and the bidirectional relationship between peer victimization and internalizing problems among rural-to-urban migrant children in China: A longitudinal study. *Acta Psychologica Sinica, 49*(3), 336–348.

[郭海英, 陈丽华, 叶枝, 潘瑾, 林丹华. (2017). 流动儿童同伴侵害的特点及与内化问题的循环作用关系：一项追踪研究. *心理学报, 49*(3), 336–348.]

Harris, M. A., & Orth, U. (2020). The link between self-esteem and social relationships: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology, 119*(6), 1459–1477.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282.

Mund, M., Freuding, M. M., Möbius, K., Horn, N., & Neyer, F. J. (2020). The stability and change of loneliness across the life span: A meta-analysis of longitudinal studies. *Personality and Social Psychology Review, 24*(1), 24–52.

Nelemans, S. A., Meeus, W. H., Branje, S. J., Van Leeuwen, K., Colpin, H., Verschueren, K., & Goossens, L. (2019). Social Anxiety Scale for Adolescents (SAS-A) Short Form: Longitudinal measurement invariance in two community samples of youth. *Assessment, 26*(2), 235–248.

Orth, U., Dapp, L. C., Erol, R. Y., Krauss, S., & Luciano, E. C. (2021). Development of domain-specific self-evaluations: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology, 120*(1), 145–172.

Orth, U., Erol, R. Y., & Luciano, E. C. (2018). Development of self-esteem from age 4 to 94 years: A meta-analysis of longitudinal studies. *Psychological Bulletin, 144*(10), 1045–1080.

Robe, A., Dobrean, A., Balazsi, R., Georgescu, R. D., Păsărelu, C. R., & Predescu, E. (2022). Factor structure and measurement invariance across age, gender, and clinical status of the screen for children anxiety related emotional disorders. *European Journal of Psychological Assessment*. Advance online publication.

Tang, A., Van Lieshout, R. J., Lahat, A., Duku, E., Boyle, M. H., Saigal, S., & Schmidt, L. A. (2017). Shyness trajectories across the first four decades predict mental health outcomes. *Journal of Abnormal Child Psychology, 45*(8), 1621–1633.

.....

审稿人 2 意见：

意见 1：该研究基于纵向研究的三水平元分析，刻画了社交焦虑平均水平变化和稳定性的年龄趋势，探讨了人格成熟假说和非适应困难假说，研究视角比较新颖，具有一定的理论意义和应用价值。此外，筛选流程规范、尽可能考虑了潜在调节变量对结果的影响，结果具有一定可信性。但由于研究年龄跨度较大，所纳样本在取样年代、取样工具等方面存在较大差异，分析较为复杂，分析结果的可重复性、以及能否科学回答“社交焦虑是相对稳定的人格特征还是频繁波动的心理状态？”这一问题上还需慎重。

回应：衷心感谢审稿专家对本研究方法和价值的肯定。接下来的部分，我们对结果的可重复性和能否科学回答“社交焦虑是心理特质还是心理状态”两个问题做出回应。

首先，对于研究结果的可重复性，我们的确不能保证该结果能在每一个未来的实证研究中被复制，因为无论是追踪的年份(婴儿中后期到青年晚期)、还是测量工具的多样性、抑或是样本来源的广泛性，这些都是单个追踪研究难以实现的。但是，这也是本元分析的优势所在，正是因为综合了来自不同地区样本、采用不同工具、测量不同方面的研究，我们才更有底气去揭示社交焦虑发展轨迹和稳定性的总体的年龄趋势，并可以利用调节效应检验去探讨这些潜在的影响。

其次，对于“社交焦虑是心理特质还是心理状态”的回答，我们之前的做法与表述的确欠妥。我们最初的思考是，将社交焦虑的稳定性程度和年龄趋势与 Roberts 和 DelVecchio(2000)所发现的大五人格的稳定性做对比，如果程度和年龄趋势与之相近，即可认为该构念是特质性构念。但是我们继续阅读了相关文献后，发现检验某个心理构念的稳定性，除了参考等级排序相关系数的点估计值外，还应考察该相关系数随时间间隔的下降趋势(Fraley & Roberts, 2005)。相关系数一般与时间间隔呈负相关，如果相关系数随着时间间隔的增大呈线性下降并趋近于 0，说明该构念的特质属性较弱；如果相关系数随着时间间隔的增大出现先快后慢的趋势，在较长的时间间隔后不再明显下降，则该构念的特质属性较强(Fraley & Roberts, 2005)。为此，我们使用 SPSS 24.0，以时间间隔为自变量，以相关系数为因变量，通过曲线估计，判断稳定性系数随着时间间隔的变化趋势(**详细的方法部分我们已添加在“2.4 统计分析”部分第三段，结果部分添加在“3.4 社交焦虑的稳定性”部分**)。从稳定性程度上看，控制时间间隔为一年的社交焦虑的相关系数在 0.467 到 0.686 之间，与传统人格特征的稳定性范围(0.31 到 0.71 之间)基本相当(Roberts & DelVecchio, 2000)，说明在较短的间隔内，社交焦虑是高度稳定的。此外，社交焦虑的稳定性与时间间隔呈对数函数关系，随着时间间隔的增加，稳定性的降幅逐渐放缓并进入稳定期。如图 4 所示的对数曲线，16 年时间间隔下的

稳定性系数仍在 0.3 左右，而根据最新的效应强度判断标准，相关系数在 0.3 及以上即为相对较大的效应量(Gignac & Szodorai, 2016)，这提示初始水平的社交焦虑对多年后的社交焦虑仍有一定的预测力度。根据 Fraley 和 Roberts(2005)的观点，可推测社交焦虑具有较强的特质属性，倾向于特质性构念。为了研究的谨慎性，我们将相关部分改为更 soft 的表述，如“社交焦虑的特质性属性更强”、“更近似特质性构念”，而不直接断言社交焦虑是一种特质。

意见 2：引言部分，第一段提到“来自美国的调查显示” 和“基于我国民众的元分析也表明”，建议补充文章的针对人群、年龄段。

回应：我们已经补充了两个研究所对应的年龄群体，分别是“18 岁以上人群”和“15 岁以上人群”。

意见 3：文中有几处出现社交焦虑的“高发生率”字样，但文中以往研究更多的为患病率，需确定是“高患病率”还是“高发生率”。

回应：感谢审稿人的细心审阅。我们查阅原文后发现所使用的单词是“prevalence”，我们将相关表述更改为“患病率”。

意见 4：方法部分，文献纳入部分中提到“波段的间隔至少在 6 个月及以上”，但未界定最长波段间隔，建议补充。此外，“干预研究中未经历任何干预且满足其他标准的对照组可以纳入”，但干预研究中的样本多为筛选后符合一定标准的，是否与其他文章中样本有差？

回应：感谢专家的建议，我们补充了最长间隔的纳入标准“*最长间隔至少与计划年龄组范围有 50%的重叠(Hoff et al., 2018)*”。的确，干预研究中的对照组既可能是社交焦虑最低的人群，也可能同样是高度的社交焦虑者，因此我们删去了此条纳入标准。上述标准的变动没有使得所纳入文献同步变动，故结果部分没有改动。

意见 5：结果部分，虽然汇报了出生年代，但对于不同出生年代下接受调查的年龄未见报道，建议补充。正如文中所提到，出生组效应在其中起重要调节作用，那么本文中不同年龄段中样本群体的出生年代分布是否存在差异？

回应：根据专家的意见，我们在表 3 和表 5 的第 4 列和第 5 列分别补充了对应年龄组的出生年代范围和加权出生年代。从数据上看，婴儿中晚期(1~3 岁)的被试平均出生于新世纪以后，而青年晚期(25~35 岁)被试平均出生于 70~80 年代，其他年龄组则主要出生在 90 年代。

意见 6：讨论部分，有关社交焦虑的稳定性、调节效应方面的讨论不足，有待进一步解释。

回应：应专家意见，我们进一步扩充了对社交焦虑稳定性和调节效应的讨论。相应内容补充在 4.2 和 4.4 部分，由于内容较多，我们没有粘贴在此处。

意见 7：结果中未见有关自尊、孤独感等变量，那么“与其他相近构念的比较”该部分的讨论目的是什么？

回应：“4.3 与相近构念的比较”这一部分主要是基于两点考虑。一方面，由于研究总是站在巨人的肩膀上前进，那么讨论部分我们常常需要指出研究结果验证或反驳了哪些理论，与哪些类似研究相符或相悖；而这一部分内容偏多，如果分散在 4.1 和 4.2 部分可能会造成可读性变差。另一方面，这一部分我们提到的变量都与社交焦虑有较大联系，我们希望与相近构念的元分析进行承接，可以引导读者阅读相关文献并将这些与心理健康密切相关的构念的发展趋势和稳定性进行对比，进而发现心理与行为毕生发展的潜在规律；比如，与人格的比较能够帮助推断社交焦虑的特质属性，与孤独感的比较能够给两者长期存在的异同辩论提供证据。

意见 8：局限性中“正性价恐惧”的讨论有点突兀，建议进一步阐述此不足给研究带来的影响。

回应：原文为笔误，应为“正性评价恐惧”。的确，我们在阅读时发现该部分没有较好第承接原文。对此，我们增加了表述“*Weeks 等人(2009)强调，评价恐惧是社交焦虑的核心特征，包括正性和负性两个方面*”。我们的考虑是，社交焦虑的认知特征应该是“评价恐惧”而非仅仅是“负面评价恐惧”，很多综述论文都强调了两个构念是相互独立的，而且正性评价恐惧对社交焦虑的特异性更强(刘洋，张大均, 2010; 叶友才 等, 2021)。但是，以往的追踪研究基本缺乏对正性评价恐惧的关注，这正是局限之一。

参考文献

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112*(1), 60–74.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102,* 74–78.

Hoff, K. A., Briley, D. A., Wee, C. J., & Rounds, J. (2018). Normative changes in interests from adolescence to adulthood: A meta-analysis of longitudinal studies. *Psychological Bulletin, 144*(4), 426–451.

Liu, Y., & Zhang, D. J. (2010). On theory of fear of evaluation and its relative research. *Advances in Psychological Science, 18*(1), 106–113.

[刘洋，张大均. (2010). 评价恐惧理论及相关研究述评. *心理科学进展, 18*(1), 106–113.]

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*(1), 3–25.

Weeks, J. W., Rodebaugh, T. L., Heimberg, R. G., Norton, P. J., & Jakatdar, T. A. (2009). "To avoid evaluation, withdraw": Fears of evaluation and depressive cognitions lead to social anxiety and submissive withdrawal. *Cognitive Therapy and Research, 33*(4), 375–389.

Ye, Y., Lin, R., & Yan, Y. (2021). Tall trees catch much wind: Fear of positive evaluation in social anxiety. *Advances in Psychological Science, 29*(6), 1056–1066.

[叶友才，林荣茂，严由伟. (2021). 树大招风：社交焦虑者的正性评价恐惧. *心理科学进展, 29*(6), 1056–1066.]

第二轮

审稿人 1 意见：

意见 1：作者很好地回应了我的评价。

回应：感谢审稿人对本研究的认可。

审稿人 2 意见：

论文对 158 项研究进行了三水平元分析，探究了婴儿中后期至青年晚期社交焦虑平均水平变化和稳定性的年龄趋势，研究视角比较新颖，具有一定的理论意义和应用价值。逻辑思路清晰，语言流畅。从读者的角度来看，希望补充以下信息：

意见 1：文献综述部分，建议作者增加对三水平元分析方法的介绍。

回应：感谢专家的建议，但我们认为这方面内容不宜放在引言部分。一方面，三水平模型虽然相比传统元分析较新，在医学领域使用的也不多，但在心理学领域较常见，知网检索后也发现了三篇采用该方法的文章(陈静 等, 2022; 雷丽丽 等, 2020; 魏星 等, 2022)，相关的方法介绍类文章早在 2014 年就有详细的介绍(Cheung, 2014)。另一方面，该方法的原理并不复杂，只是在传统元分析两水平(抽样误差、研究间误差)的结构上扩展为三水平结构(抽样误差、研究内误差、研究间误差)，此时允许从一个研究中提取多个效应量，单个研究内效应量间的误差即为研究内误差，解决了传统元分析中无法处理效应量依赖的问题。此外，本研究的主要贡献是刻画社交焦虑的平均水平和稳定性的发展趋势，虽然三水平元分析有其优势，但并非本研究的"主要卖点"。综上，从方法的新颖度、原理的复杂性、以及对本研究的重要性上都不太有必要在引言中开辟某一小节对该方法做专门介绍。而且，心理学报对引言的字

数要求是 3500 字内，而目前的引言部分已近 6000 字，如果再增加这部分内容可能不合适。综合考虑，如初稿所示，我们选择在“2.5 统计分析”部分第二段(P29)对三水平元分析进行简单介绍，如果读者希望进一步了解该方法，可以阅读该部分提到的 Cheung 及其同事(2014, 2019)的文章。如有不合理之处，还请专家批评指正。

意见 2：文献检索部分，作者仅选取两个中文数据库(中国知网、万方数据)和三个外文数据库(Web of Science, PsycINFO, PubMed)的依据是什么？为什么没有纳入其他常见的元分析所用的数据库，如，维普、MEDLINE 和 EMBASE 数据库？另外，作者针对灰色文献进行了哪些检索？文献的纳入时间范围是什么？另外，建议作者在正文中补充关键检索词。

回应：我们选择这五个数据库主要是基于全面性和工作量的权衡。比如，当前的知网其实已经涵盖了心理学相关的所有期刊和学位论文，而万方能检索到部分知网没有收录的医学期刊和学位论文；但是维普则没有特殊性，其收录范围和时效性都不如前两个数据库，故我们没有检索维普。另外需要表达歉意的是，我们之前检索的是 ProQuest 库而非 PsycINFO，仔细检查后发现作者学校的 ProQuest 库中不含 PsycINFO 库。因此，应专家建议，除在上述五个数据库的基础上进行更新外，我们还新增检索 EBOSCO(含 MEDLINE、PsycINFO 和 PsycArticles)数据库，中英文检索词已添加在“2.1 文献检索”部分。最终新增了 15 项研究，19 个独立样本，26 个效应量，所有相关数据同步更新，更新后的结果与更新前基本一致。

灰色文献主要是指没有正式发表的文献，本研究的效应量中有 15.8%来自中英文学位论文，也属于灰色文献的范畴。另外，在一篇论述发表偏倚的经典文献中，研究者指出，通过电子邮件索取的所谓“灰色文献”几乎不可能获得随机且未选择性的样本数据，甚至会导致更多的偏倚(Ferguson & Brannick, 2012)。

此外，本研究未限制检索的时间范围，初次检索时间在 6 月下旬，更新和补充检索时间在 12 月上旬。最后，中英文检索策略已添加在“2.1 文献检索”部分。

意见 3：文献纳入和排除标准部分：第一，建议作者按照 PICOS 的顺序梳理归纳文献的纳排标准。第二，作者的研究对象是婴儿中期至青年晚期，但是在纳排标准部分没有对研究对象的年龄进行限定。第三，纳入标准 3 中要求“每个波段所使用的社交焦虑量表在内容、题量、计分方式等方面必须完全一致”，这个标准的设置依据是什么？由于社交焦虑有很多种测量方法，作者在选取“完全一致的量表”时的选取依据（如，量表的测量学属性或量表的使用数量）是什么？对于其他使用了不一致量表的研究，作者的处理办法是什么？第四，标准

6 中“样本未经历特殊事件”中的“特殊事件”具体指什么？第五，建议作者将纳入标准和排除标准分开叙述。

回应：我们依次回应这 5 个问题。

第一，我们了解到 PICOS 原则的顺序依次为对象、干预措施、对照措施、结局和研究类型。该原则主要在循证医学或者是 RCT 元分析中使用，从本研究来看不是很契合。我们尽量按照样本、研究设计和结果报告的顺序重新梳理了纳排标准。

第二，本研究在文献检索前进行了预注册，注册时的文章标题是“The Stability and Change of Social Anxiety Across the Life Span: A Three-Level Meta-Analysis of Longitudinal Studies”，可见该研究的预期目标是探讨生命全程社交焦虑的发展，只是最终符合标准文献没有获取到中老年期的数据，最终才改为婴儿中后期到青年晚期，对此我们在纳入标准部分下利用脚注进行说明。

第三，由于我们编码的是每个波段社交焦虑的平均分和标准差，如果量表没有完全一致，则无法准确计算标准化均值差。比如，题量不一致会导致均分不可比、计分方式不一致会导致标准差不可比。举个例子，某项研究同时使用 SAS-A 和 SIAS 分两个时间点测量了青少年的社交焦虑，这时我们会分别提取两个效应量；但如果这项研究在时间点 1 只用了 SAS-A，时间点 2 只测了 SIAS，则被排除。如果原文没有特别说明，我们默认不同时间点所使用的测量工具是完全一致的。根据这条标准，仅 3 篇文献被排除在外。

第四，特殊事件主要是指重大灾害，比如对震后儿童的心理健康追踪调查。

第五，已经将纳入和排除标准分两段叙述，见“2.2 文献筛选”部分。

意见 4：PRISMA 文献筛选流程图中提到有 33 篇文献无法获取全文，作者采用了什么办法处理这 33 篇文献？

回应：无法获取全文的文献主要包括无法下载的学位论文、会议摘要和最新发布的期刊论文。对于前两者我们选择忽略，因为确实没有办法；而最后者是指我们利用 SCI-HUB、访问不同学校的图书馆资源、以及在 ResearchGate 上和作者索要后仍无法下载的期刊论文，这些论文主要是 2022 年最新发布的。

意见 5：研究在筛选标题摘要和全文时的方法与流程是什么？如果是两位研究者同时筛选，筛选的一致性是多少？

回应：感谢专家的建议，我们之前在筛选部分没有考虑到这个问题，对此我们开展了相应工作，并将下列内容补充在“2.2 文献筛选”部分：

筛选工作首先由第一作者独立进行，随后再随机选取 60 篇文献(30 篇已排除文献和 30 篇已纳入文献)对半分配给另外两名心理学研究生，要求他们根据上述标准判断文献是否纳入元分析。结果表明筛选的一致性高(一致性分别为 90%和 87%)，不一致的主要原因是没有发现正文未提供的补充材料，以及误将仅报告患病率的研究纳入。文献筛选流程图如图 1 所示。

意见 6：论文采用标准化均值差和相关系数作为效应量，对于分类变量，作者的处理办法是什么？

回应：与探讨变量间关系的元分析不同，本研究的主要目的是精确刻画社交焦虑平均水平和稳定性的年龄趋势，但是通过某个数值将社交焦虑得分进行分类的研究则掩盖了很多重要信息，不适合刻画发展轨迹。而且，大多数心理学研究通常报告连续型数据，已经可以获取充分的效应量进行分析。因此，与其他元分析一样(Orth et al., 2021; Orth et al., 2018)，我们不纳入报告分类数据（主要是报告 OR、RR 值等）的研究，这一点我们在排除标准部分进行了补充。

意见 7：发表偏倚部分，“由于本研究纳入的效应量为两个波段间的标准化均值差和相关系数，但大多数原始研究并不关注该系数，通常只附带在描述性统计中被报告，因此可以认为本元分析不受发表偏倚的影响”。作者的这个推论并不成立。

回应：这句话我们的思考是，发表偏倚主要来自研究者有目的地呈现变量间关系显著的结果，因为显著的结果更可能发表。但是本研究提取的效应量并非变量间的关系或某种干预的有效性，而是单一变量的均分、标准差和重测相关系数，而这些数值的大小或是否显著与发表的可能性通常无关。比如，横断历史元分析(探讨某一构念的平均分随年代的变化趋势)一般不检验发表偏倚，这是因为平均分的数值大小通常不是研究者和审稿人关心的结果，自然也不存在“平均分越大的研究越可能被发表”的问题。为了研究的谨慎性考虑，我们删去了这个表述。

意见 8：作者仅报告了文献质量评价的量表，具体进行文献质量评价的流程是什么？另外，有必要上传研究的具体文献评价质量结果。

回应：由于初次评价时两位编码者都只记录了总分，原始分数已丢失，因此我们只能重新对文献进行评价。所有文献由第一作者根据文献质量评估量表进行评价，再请一名心理学研究

生随机评价其中的 30 篇，结果显示评价的一致性高(Kappa = 0.808~1.000)。具体的文献质量评估结果已上传至开放数据库(https://osf.io/2sp6m)。

意见 9：研究除了对文献质量进行敏感性分析外，是否考虑了其他重要变量的影响？如，缺失数据、其他合并精神障碍等。

回应：样本流失率合并在了文献质量评价中。如文中所示，文献质量评价包含了样本选取、T1 数据有效率、流失率、测验信度、纵向等值和出版物等级六个部分。如果拆开依次做调节效应分析，可能会使结果部分更繁杂。对于合并精神障碍，我们考察的是一般人群社交焦虑的发展，排除标准部分也说明不纳入临床或干预样本，故不存在合并精神障碍的问题。

参考文献

Chen, J., Ran, G., Zhang, Q., & Niu, X. (2022). The association between peer victimization and aggressive behavior in children and adolescents: A three-level meta-analysis. *Advances in Psychological Science, 30*(2), 275–290.

[陈静, 冉光明, 张琪, 牛湘. (2022). 儿童和青少年同伴侵害与攻击行为关系的三水平元分析. *心理科学进展, 30*(2), 275–290.]

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*(2), 211–229.

Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review, 29*(4), 387–396.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120–128.

Lei, L., Ran, G., Zhao, Q, Mi, Q., & Chen, X. (2020). The associations between parenting styles and anxiety in preschool-age children: A three-level meta-analysis. *Psychological Development and Education, 36*(3), 329–340.

[雷丽丽, 冉光明, 张琪, 米倩文, 陈旭. (2020). 父母教养方式与幼儿焦虑关系的三水平元分析. *心理发展与教育, 36*(3), 329–340.]

Orth, U., Dapp, L. C., Erol, R. Y., Krauss, S., & Luciano, E. C. (2021). Development of domain-specific self-evaluations: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology, 120*(1), 145–172.

Orth, U., Erol, R. Y., & Luciano, E. C. (2018). Development of self-esteem from age 4 to 94 years: A meta-analysis of longitudinal studies. *Psychological Bulletin, 144*(10), 1045–1080.

Wei, X., Gao, S., Xu, J., Wang, W., & Zhao, Y. (2022). Three level meta-analysis of the relationship between parental control and depression in children and adolescents. *Chinese Journal of Child Health Care, 30*(4), 417–421+430.

[魏星, 高双, 许江, 王文娟, 赵源. (2022). 父母控制与儿童青少年抑郁的三水平 Meta 分析. *中国儿童保健杂志, 30*(4), 417–421+430.]

第三轮

审稿人 2 意见：

论文采用三水平元分析的方法探究了婴儿中后期至青年晚期社交焦虑平均水平变化和稳定性的年龄趋势，研究具有一定的理论意义和应用价值。作者较好地回应了审稿人的意见，对论文进行了修订，研究过程更加清晰，结果更有说服力。从读者的角度来看，希望补充以下信息：

意见 1：论文进行元分析依据了哪些标准规范？建议在方法部分第一段补充说明。

回应：感谢专家的建议，我们在 PRISMA 2020 (Page et al., 2021)的规范下开展了元分析，已补充在方法部分第一段。

意见 2：文献检索部分，建议作者在正文中仅呈现出检索词。具体的检索策略以补充材料形式上传。

回应：已修改为仅呈现检索词，具体的检索策略见 OSF 中的补充材料。

意见 3：文献筛选部分，论文选取了 60 篇论文进行双人独立筛选，约占总文献的 0.94%（60/（6094+252））。这个筛选比例的制定依据是什么？如果论文采用的是快速审查的方法，建议在正文方法部分说明。另外，论文所报告的一致性 90%和 87%是标题摘要的筛选一致性，还是全文筛选的一致性？

回应：很抱歉我们在正文中没有论述清楚上述问题。由于文献初筛的数量较多，且排除较易(有很多文章的标题和来源期刊一看就和心理学无关)，故初筛(阅读标题和摘要)的工作仅由第一作者进行，即快速筛选的方法。复筛(阅读全文)则先由第一作者独立进行，随后选择其中的 60 篇给第二和第三个研究者，并计算后两人分别与第一作者的编码者一致性，即全文筛选的一致性。因此，我们筛选的比例约为 8% (60/(606 + 155)), 但这个比例当前没有固定标准，我们主要是参考了一些权威期刊近年的元分析论文，如 Orth 及其同事(2018, 2021)的两篇文章的也只在全文筛选时才采用“背靠背”的方法，其选取的比例分别为 3.7% (60/1853) 和 5.4%(120/2236)。

意见 4：PRISMA 文献筛选流程图中列出的文献筛除标准似乎与正文的纳排标准不完全一致。其中，格式不符指什么？因其他原因排除的有 59 篇，占比较大，具体指哪些原因？

回应：格式不符指文章所报告的数据为不可比的非连续型数据，如 OR 和 RR 等分类数据、同伴提名数据。其他原因中占比最大的是访谈类的质性研究、综述和元分析类研究(40 篇)，应专家建议，我们将其纳入到流程图中。余下的其他原因(n = 19)包括非中英文论文、有效样本量不足 30、样本均龄标准差大于 5 和量表前后不一致等。上述内容我们已在图 1 下方利用图注进行说明。

意见 5：论文编码部分，作者在 2.3 部分提出由两名研究生独立进行编码，但在 2.3.2 部分作者提出由一名研究生完成全部编码，另一名编码其中的 30 篇。请明确两人独立编码的比例，是全部由两人独立编码，还是仅选取了论文的一部分？选取这一部分比例的依据是什么？

回应：与筛选部分不同，编码最初是由两名研究生独立进行的，但上一轮审稿意见指出应上传具体的文献质量评价结果，由于最初编码时只记录了总分，故这一部分只能重新编码。因此，第二轮修改中，除文献质量部分为部分独立编码外，其余内容全部为完全独立编码。为了保持所有编码项一致，我们将其余的文章再次请另一名研究生独立评价，相关数据同步更新，结果变化微弱。

参考文献

Orth, U., Dapp, L. C., Erol, R. Y., Krauss, S., & Luciano, E. C. (2021). Development of domain-specific self-evaluations: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology, 120*(1), 145–172.

Orth, U., Erol, R. Y., & Luciano, E. C. (2018). Development of self-esteem from age 4 to 94 years: A meta-analysis of longitudinal studies. *Psychological Bulletin, 144*(10), 1045–1080.

第四轮

审稿人 2 意见：建议接受。

回应：感谢审稿人对本研究的认可。

编委复审意见：该文现在正文字数有 1.8 万之多，请作者压缩到 1.5 万字左右，压缩后可以发表。

回应：根据编委的意见，正文字数(不含参考文献和英文摘要)**从 18167 字压缩至本轮的 16744 字。**总的来说，本元分纳入效应量多，年龄跨度广，编码和分析工作繁杂，涉及到发展轨迹和稳定性两大块内容，故字数的确偏多。为保持文章结构的完整性和清晰性，我们很难再继

续压缩字数，如果编委或主编有更好的删减建议，不吝赐教！下面是对文章各部分的具体修改说明：

(1)摘要部分精简表述，**字数从 465 字压缩至 405 字**。

(2)引言部分，我们精简了一些冗余的表达，适当减少了参考文献的引用，**字数从上一轮的 5969 字缩减至 5529 字**。需要说明的是，当前引言的字数超过了投稿指南上引言 3500 字以下的要求，但考虑到元分析的特殊性和本研究的工作量，我们不得不暂时保留这样的篇幅。具体而言，前言首两段需要引入主题，强调研究社交焦虑的现实意义和以往研究的争议；1.1 部分需要论述清楚本研究所关注的社交焦虑到底包括什么；1.2 部分需要论述各年龄段社交焦虑的发展，由于本研究年龄跨度较大，因此字数难免偏多；1.3 部分以尽量简洁的方式阐述了 5 个关键的潜在调节变量。

(3)方法部分，精简表述的同时将文章质量评价量表移至附录，**字数从上一轮的 4616 字压缩至 4309 字**。由于该元分析的编码和统计分析过程相当繁杂，需要进行 22 个独立元分析，而且国内尚未见该类型的元分析，故文中我们举了一些例子来明晰部分编码的细节，以提高文章的可读性和研究的可重复性，因此字数也多了一些。

(4)结果部分主要有三处改动。第一处是删去了 3.1 中介绍每个年份发表了几篇文献的部分，因为表 1 中已有了初步的介绍；第二处是将表 2 中 Egger 线性回归法的 *t* 值替换为更常报告的截距值；第三处是将原表 4 和表 6 合并在同一个表格，即表 4。**字数从上一轮的 2641 字压缩至 2419 字**。

(5)讨论和结论部分。讨论部分删去了 4.1 前对研究结果再重复的表述，精简了“4.3 与其他相近构念的比较”部分。结论部分进行重写，避免与结果相同。**字数从上一轮的 4434 字压缩至 4044 字。**

(6)其他修改部分：压缩参考文献(**字数从上一轮的 3452 字压缩至 3225 字**)、优化英文摘要。

第五轮

主编终审意见：该文进行了四轮审稿，作者已就审稿人详尽的意见进行了充分的回复。但目前稿件字数确实还是比较多，我仔细读了一下文章，觉得至少在前言部分，如 1.2 各年龄段社交焦虑的发展，以及 1.3 关键的潜在调节变量方面，虽然我也同意有必要进行相关的回顾和论述，但建议是作者不一定要分点详细讲，因为每一点论述的方式是比较相似的，分别铺

开来写就会使字数变多，也许作者可以考虑进行综合介绍，如介绍几个年龄段的总体趋势，然后再综合说明一下在哪些年龄段中有比较大的差异，进而说明在元分析中重点会看哪些差别。同样，影响因素也可以做一下综合，毕竟这些因素都比较容易理解，不需要解释太多，只需要把有共识的进行综合介绍，而有分歧的可以重点提及，因为分歧点是元分析想要重点观察的点。这样进行整理可以突出重点，也应该可以同时精简字数。

建议作者再次进行删改，使文章结构更为紧凑，重点的研究问题更为突出。

回应：参考主编的意见，我们再次请一名发展心理学方向的讲师和一名副教授对本研究的引言部分进行了挑剔阅读，他们一致认为不同年龄段的论述无法整合在一起，因为即使存在相同的趋势(如小学儿童期和青年期)，但驱动因素和理论基础也不太一致。而且，按年龄段的顺序进行论述是发展心理文章的常见写作方式。我们同意该看法，仍采用基于五个年龄段的分段叙述，但对内容做了大幅简化，并在 1.2.1 部分的最后一段进行小结(以紫色字体呈现)，突出了本研究重点关注的分歧之处。对于调节变量部分，我们同意主编的看法，将五个部分进行压缩整合。**引言部分最终删减近 1200 字，字数降至 4339 字，正文字数(不含参考文献和英文摘要)来到 15291 字，参考文献压缩至 2931 字。**不得不承认，当前引言的字数还是超过了 3500 字的限制，但是考虑到元分析不同于其他实证研究，带有系统综述的性质，故引言的字数和引文数量往往偏多。当然，如果专家认为还可以继续删减，我们也愿意执行。

除此之外，经再次核查数据和请其他同行评阅论文后，我们还做出了以下修改：(1)由于当前的婴儿多指的是 1 个月到 12 个月的孩子，而 1~3 岁在英文中多称作“toddlerhood”，即学步期的幼儿。因此，我们将“婴儿中后期(1~3 岁)”修改为“学步期(1~3 岁)”，而原“幼儿期(3~6 岁)”修改为“学前儿童期(3~6 岁)”，其他相关内容同步变动。(2)原稿在绘制累计 d_{year} 值图(即图 2)时，忽略了每个年龄组跨度不等的情况，而直接使用了每个年龄组的估计值。在修改稿中，我们以 1 岁为 d 值的 0 点，根据每个年龄组的估计值，持续累加 d_{year} 值 34 次直到 35 岁，即为正确的发展轨迹图。