《心理学报》审稿意见与作者回应

题目：置信区间宽度等高线图在线性混合效应模型样本量规划中的应用
作者：刘玥，徐雷，刘红云，韩雨婷，游晓锋，万志林

第一轮

审稿人 1 意见：

该研究介绍了适用于实验心理学研究数据的基于检验力以及效应量估计准确性的样本量规划分析方法。文章立足于实验心理学研究。实验心理学研究中，由于每位被试需要完成多个试次从而导致实验数据形成层级关系(试次嵌套于被试以及刺激)，因此适合用线性混合效应模型来进行数据分析。目前许多研究者对线性混合效应模型的分析如何进行有效的样本量规划分析较为陌生。因此，本研究具有较高的实际价值和指导意义。文章的论述条理清晰，研究方法也合理恰当。我对该研究有以下一些建议和疑问。

意见 1：

心理学的实验研究设置条件变化可能比较多，似乎文中绘制某些具体情境下的样本量表格，仍然难以让实际研究者使用起来。我会建议作者精简一下目前的内容，同时增加一个实例展示小节，专门展示如何使用 R 语言获得文中的图表，着重在对统计软件程序的介绍。

回应：

感谢审稿专家的建议。已经在原来版本基础上对全文进行了精简，删除了与主题关系不大的内容，并将一些与样本量规划无直接关联的结果放到附录中。同时，增加了实例展示小结“6 实例演示”，其中重点说明了如何使用 R 语言得到结果。具体请参见修改稿中的“6 实例演示”和附录中标红部分。

意见 2：

文中考虑的例子，Level-1 的协变量 X_{ji} 是二分的且两个水平出现的概率相等。由于心理学实验研究者可能出现某些设计，导致该出现的比例是不平衡的，比如实验条件的刺激出现频率可能比较少。这个比例的改变会如何影响你的研究结果？同时，目前有研究表明在多层模型中，即使 Level-1 的协变量是分类型变量，也需要进行中心化处理，以避免组内和和组间效应的混淆。这又会对你的研究产生什么影响？

回应：

感谢审稿专家的建议。不平衡设计的确会影响检验力，进而影响样本量规划结果。因此，我们在修改稿中两个模拟研究部分加入了不平衡设计下的模拟结果。具体请参见修改稿第 8 页最后一个自然段，第 14 页“5.2.2 变化参数设置”中标红文字，以及附录中标红的部分。同时，在研究编写的 R 函数中也加入了输入不同水平样本量占总体比例的参数部分，以灵活适应不平衡设计的情况。具体请参见附件 1 中标红文字。另外，本研究中水平 1 的自变量是经过偏差编码的分类变量，编码为-0.5 和 0.5，因此不再需要中心化处理(修改稿第 8 页脚注 5)。未进行中心化的确可能影响检验力和参数估计结果，进而影响样本量规划结果。但在我们研究中默认研究者已经进行了水平 1 变量的中心化，不再探讨未中心化的影响。

意见 3：

4.1.1，ICC 设置数值为 0.1，0.3 和 0.5。是否能提供该标准的依据？据我了解，不同类型的含层级结构的数据其 ICC 取值差异巨大。被试内数据(如文中的实验心理学研究数据或纵向追踪数据)其ICC往往高于诸如来自不同学校或不同地区而形成的层级结构数据的ICC。因此会对文中所选择的 ICC 有所疑问。

回应：

感谢审稿专家的意见。不同的嵌套结构确实会得到不同大小的 ICC。Arend 和 Schäfer(2019)对具有嵌套结构的心理学研究进行了总结(Bliese, 2000; Dziak et al., 2012; James, 1982; Maas & Hox, 2005; Mathieu et al., 2012; Snijders & Bosker, 1999)，发现 ICC 的平均值为 0.3，可以用 0.1，0.3，0.5 代表小，中，大水平。并且，一般来说，被试嵌套于组的研究设计所得到的 ICC 要小于测量嵌套于被试的 ICC。在他们总结的 151 篇嵌套结构数据的研究中，被试嵌套于组的研究所得到的平均 ICC 为 0.19，属于 ICC 小的程度；测量嵌套于被试的研究所得到的平均 ICC 为 0.42，属于 ICC 大的程度。但是，整体来说仍可以用 0.3 代表中等程度 ICC。在修改稿中加入了 ICC 小，中，大水平的参考文献，以及不同类型嵌套结构 ICC 大小差异的一般规律。具体请参见修改稿第 7-8 页脚注 2。

补充文献：

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Dziak, J. J., Nahum-Shani, I., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods, 17,* 153–175.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219–229.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 86–92.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97*, 951–966.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: SAGE.

意见 4：

4.1.1，SDpredictor 为什么是标准差为 1？按照-0.5，0.5 的编码方式且两个水平出现概率相等来看，标准差不是 0.5=sqrt(0.5^2*0.5+0.5^2*0.5)?

回应：

感谢审稿专家仔细的阅读。按照这种编码方式自变量的标准差应当是 0.5。此时，由于自变量是含有两个水平且进行偏差编码的分类变量，因变量已经标准化，因此无需再用原公式(16)进行标准化转换，回归系数即为标准化回归系数。设$Y = \gamma_0 + \gamma_1 X + r$，对组 1($X = -0.5$)，则有$Y_1 = \gamma_0 - 0.5\gamma_1 + r$，对组 2($X = 0.5$)，则有$Y_2 = \gamma_0 + 0.5\gamma_1 + r$，因此标准化的回归系数$\gamma_1 = (Y_2 - Y_1)/SD$，SD 为两组的联合标准差=1。因此标准化回归系数的含义为两个组的标准化均值差异(即 Cohen's d)。已经对原文中不准确的表述进行修改，具体请参见修改稿第 8 页脚注 5。

意见 5：

4.1.1, 可否用文字解释 gamma10 的含义？例如我可以理解为实验组和控制组在结果变量上的平均差异吗？

回应：

在一般情况下，γ_{10}表示被试随机斜率的均值，即为实验效应的固定部分，是研究重点考察的效应量指标(见修改稿第 4 页标红部分)。在本研究中，由于自变量有两个水平且采用偏差编码(-0.5 和 0.5)，因变量已经标准化($\sigma^2 = 1$)，因此，γ_{10}即为标准化的回归系数$\gamma_{10.std}$，其值代表了两个水平的标准化均值差异(Cohen's d)。可以理解为实验组和控制组在结果变量上的平均差异。见修改稿第 8 页脚注 5。

意见 6：

4.1.2，gamma10std 是对应 Cohen d 吗？Cohen d 似乎只标准化了结果变量，而没有标准化分组变量。所以不应该是偏标准化回归系数才对应的 Cohen d 吗？

回应：

在本研究中$\gamma_{10.std}$是对应的 Cohen's d。具体原因请参见上两条回复。

意见 7：

图 2，建议在图标题后注释阴影部分的含义。如阴影部分表示检验力高于 0.8 的样本量组合。

回应：

感谢审稿专家的建议。已经在图标题后注释了阴影部分的含义。具体请参见修改稿中图 2 标红的注释。

意见 8：

5.1.1，情境 1 考虑 level-2 的分类型预测变量的时候，考虑两个类别的被试量是相等的是否不太符合实际情况？例如，W 是性别，一般心理学实验中女性被试数量远高于男性被试数量。这会否进一步影响你的结论？

回应：

感谢审稿专家的建议。已经对情境 1 加入了两个类别样本量不等的模拟研究，具体请参见修改稿第 14 页“5.2.2 变化参数设置”中标红文字，以及附录中标红的部分。同时，在研究编写的 R 函数中也加入了输入不同水平样本量占总体比例的参数部分，以灵活适应不平衡设计的情况。具体请参见附件 1 中标红文字。

意见 9：

5.1.1，公式 16，请用文字解释清楚 gamma11 的含义。尤其是情境 1 的含义。

回应：

感谢审稿专家的建议。已经在首次引入γ_{11}处(公式(6))介绍了该参数的一般含义(见修改稿第 5 页公式(6)后标红文字)。然后在公式(15)(原公式(16))后解释了γ_{11}的意义，并分两种情境介绍了该参数的具体含义(见修改稿第 13 页公式(15)后标红文字)。

意见 10：

英文摘要太长了(编辑注：本刊英文摘要要求是 500 单词左右)。

回应：

已经对英文摘要进行了精简。请参见修改稿中的英文摘要。

……

审稿人 2 意见：

作者介绍了在混合线性模型中的进行样本规划的方法，为心理学研究者提供了检验力和效应大小分析的依据，有较强的实用性。希望作者考虑以下建议进一步提升稿件质量。

意见 1：

文章的引言部分过长，其中有多段本身也较长，所以给读者“目的”与“综述”混搭的感觉。例如，效应量一段文字过多，且包含两个“本研究”的内容，然而这些本研究并不应该是背景介绍，而是问题提出。建议作者修改引言，将其梳理成“引言”、 “文献综述”和“问题提出”三个部分。用一到两段切入主题作为“引言”（简要的背景介绍和本研究的最主要目的），其余的问题可以放到“文献综述”当中（例如，综述可以分为①检验力/样本量分析的研究及综述；②混合模型的定义和相关研究；③置信区间宽度的相关研究），以此进入“研究问题”模块（以往研究的缺陷和本研究的目的、亮点和特色，也包含评论 2 中的一些问题)。

回应：

感谢审稿专家的建议。已经按照要求对原文的引言部分进行了修改。将原引言划分为三大部分。第一部分简要介绍研究背景和目的。第二部分是文献综述，主要包括“1.1 线性混合效应模型的样本量规划问题”，“1.2 基于检验力分析规划样本量”和“1.3 基于效应量准确性分析规划样本量”三个部分。在每个部分先定义核心概念，再介绍相关研究(程序)，最后总结已有研究(程序)的缺陷和不足。第三部分是问题提出，通过对文献综述的总结，提出本研究主要解决的问题，突出亮点和特色。具体请参见修改稿引言部分标红文字。

意见 2：

本文以“样本规划”的范式对样本的大小提出建议，这一点和以往研究很不同。作者主要对比了 Arend 和 Schafer(2019)的研究；而 Murayama 等人(2022)开发了一个在线程序，对给定的先验 t 值和变量个数对样本量进行建议。基于多水平模型，de Jong 等人(2010)、Lane 等人(2018)也对样本量进行了研究。请作者进一步加强文献综述，论述本研究的优势和亮点，以及和已有文章之间的区别和联系。

回应：

感谢审稿专家的建议及提供的补充文献。在阅读这些补充文献的基础上，我们对引言部分进行了补充和修改。主要包括以下几点。(1)在“1.1 线性混合效应模型的样本量规划问题”的最后，总结线性混合效应模型推广应用受到限制，明确提出“研究者对基于 LMEMs 如何科学地规划实验设计，设置合理的被试量和试次数感到无所适从，急需方便易用的程序或图示，指导样本量规划”。(2)在“1.2 基于检验力分析规划样本量”部分，加入了审稿专家提供的补充文献作为检验力折线图的举例，并在最后总结了这些研究的缺陷，即“但是，嵌套结构的数据需要确定两个水平样本量，不同实验设计下增加不同水平样本量的成本不同。折线图仅能固定某个水平样本量，以另一个水平样本量为横坐标生成，无法同时呈现两个水平样

本量与检验力的关系”。并且，介绍了能够同时反映检验力随两个水平样本量变化的图示，明确提出了需求“综上，对于嵌套数据，研究者需要在同一个图内观察到两个水平样本量在检验力上的补偿关系，并在考虑实验成本的基础上综合权衡，得到合适的各水平样本量”(3)在“1.3 基于效应量准确性分析规划样本量”部分，介绍了包含两个水平样本量及其应的效应量准确性信息的图示，并说明其缺陷“但该图并未考虑检验力，并且色块仅表示综合得分，具有一定的主观性，研究者无法从图中清晰了解所关心的参数估计的准确性”。(4)在“1.4 问题提出”部分，进一步对比，总结了本研究与已有研究的区别和联系。具体请参见修改稿引言部分标红文字。

意见 3：

在讲具体研究的时候，尽量避免“夹叙夹议”。目前在参数设置等模块有一些介绍前人研究以及本研究和前人研究的区别的内容。例如 4.1.1 的第一段、4.1.2“本研究与 Arend 和 Schafer(2019)研究的区别主要有三点”这一部分、4.1.1 对虚无编码和偏差编码的论述(如果不是研究重点，这一段可以大幅度删减)、4.2.3 第二段前三句话等。这些内容应该都移动至文献综述，请作者对研究内容本身和以往研究的对比进行分别论述。建议作者对研究结果进行优化，做到简洁明了。

回应：

感谢审稿专家的建议。已经检查了相关章节，对出现“夹叙夹议”的部分进行了修改。为保证文章的完整性，体现模拟参数设置的依据，将阐述理由的部分放到了脚注中。同时，对于结果中出现置信区间宽度标准的推导过程这一问题，我们在文献综述部分举例说明了以往研究中对于“可接受的最宽置信区间宽度”的定义，在“置信区间宽度等高线图函数说明”部分阐述了本研究建议的计算方法。最后，在研究结果部分直接计算可接受的最宽置信区间宽度作为判断标准。具体请参见修改稿标红的脚注，以及第 3、6、10、14-15 页标红文字。

意见 4：

4.1.1 参数设置中，ICC 值的大中小有何文献依据？一般在做多水平模型的时候，组间 ICC 大于 0.08 即可用多水平分析。实验设计的混合模型与调查研究的混合模型是否有不同的地方？

回应：

感谢审稿专家的意见。实验设计与调查研究得到的嵌套数据的 ICC 分布的确不同。参考对审稿专家 1 意见 3 的回复。Arend 和 Schäfer(2019)对具有嵌套结构的心理学研究进行了

总结(Bliese, 2000; Dziak et al., 2012; James, 1982; Maas & Hox, 2005; Mathieu et al., 2012; Snijders & Bosker, 1999)，发现 ICC 的平均值为 0.3，可以用 0.1，0.3，0.5 代表小，中，大水平。并且，一般来说，被试嵌套于组的研究设计所得到的 ICC 要小于测量嵌套于被试的 ICC。在他们总结的 151 篇嵌套结构数据的研究中，被试嵌套于组的研究所得到的平均 ICC 为 0.19，属于 ICC 小的程度；测量嵌套于被试的研究所得到的平均 ICC 为 0.42，属于 ICC 大的程度。但是，整体来说仍可以用 0.3 代表中等程度 ICC。在修改稿中加入了 ICC 小，中，大水平的补充文献，以及不同类型嵌套结构 ICC 大小差异的一般规律。具体请参见修改稿第 7-8 页脚注 2。

意见 5：

4.2.1 对“区间过宽”的说明，何为过宽？标准是否有参考依据？

回应：

感谢审稿专家的意见。对于置信区间宽度标准，学术界目前还没有统一公认的方法，仍是需要解决的问题。本研究中的标准参考了 Usami(2020)的做法，根据期望的置信区间上下限，倒推可接受的最宽置信区间宽度。具体来说，“例如，在效应量的点估计值为 0.5 的情况下，计算得到其 95%置信区间(以下简称“95% CI”)宽度为 0.6，则 95%CI 约为[0.2,0.8]。根据 Cohen 的标准，该区间涵盖了效应量小、中、大的条件(0.2,0.5,0.8)，为不准确的估计结果(Maxwell et al., 2008; Usami, 2020)。因此，在此情况下可接受的最宽 95%CI 宽度为 0.6(0.8-0.2)。”(修改稿引言部分，第 3 页标红文字)。在函数说明部分进一步定义，“本研究建议采用效应量标准的最高水平减去最低水平作为可接受的最大 CI 宽度。”(修改稿第 6 页标红文字)。最后，在结果部分，对每种情况下可接受的最宽置信区间进行了计算(修改稿第 10、14-15 页标红文字)。

意见 6：

图 2，(1)线条的含义不太明确。因为颜色梯度对应到线条上比较困难。(2)阴影部分的含义没有明确，正文中似乎也没有相关介绍，是否可以不用阴影？(3)图 2b 似乎有误。对应的置信区间宽度为 0.6，水平 1 为 50 对应的水平 2 样本量应该在 720 左右(而并非 400)？

回应：

感谢审稿专家的意见。(1)已经在第一次出现等高线图时，对等高线进行了说明。读者可以根据图例中显示的等高线条数，按顺序依次对应到图中。即，“如图例所示从 0.3 到 1.0

间隔 0.1，在图中共有 8 条依次排列的等高线。例如，0.3 对应的等高线表示线条以上的区域 95%CI 宽度在 0.3 及其以下。后同。”。具体请参见修改稿图 2 标红的图注。(2)在图注中对阴影部分的含义进行了说明“图(a)中阴影区域表示符合检验力大于等于 0.8 标准的条件，图(b)中阴影区域表示符合检验力大于等于 0.8 且所有随机效应估计值 rbias 小于 0.1 的条件。”。阴影部分表示了检验力或检验力+随机效应估计准确性基本要求的样本量范围，读者应当在阴影区域内结合置信区间宽度等高线设定合理的样本量。这部分内容已经在置信区间等高线图的应用部分进行了说明。具体请参见修改稿敏感性分析研究中“样本量规划建议”和实例演示部分标红文字。(3)图 2(b)中，阴影区域中的置信区间宽度均小于 0.6。当水平 1 样本量=50，水平 2 样本量=400 时，对应的置信区间宽度等高线为从上往下数第 3 条，对照图例可知表示置信区间宽度在 0.5 及其以下，因此，符合置信区间宽度小于 0.6 的标准。为帮助读者理解，已经在图注中对置信区间等高线进行了详细说明。

意见 7：

跨水平交互作用大小是否有依据？交互作用效应量应小于主效应，一般研究认为 R^2 在 0.02 即有实质的效应(温忠麟, 刘红云, 2020)，对应到科恩 d 值一般在 0.1 左右。本研究的交互效应量较大(0.3 和 0.5)，在实证研究中情况并不多见，一般情况下交互项的效应量都比较小(Liu & Yuan, 2021)。

回应：

感谢审稿专家的意见。Liu 和 Yuan(2021)的研究主要关注了相关关系研究的效应量，而实际中实验研究跨水平交互作用的效应量也可能较大。例如，蒋元萍等人(2022)研究发现，在情绪自评量表测试中，测量时间(水平 1 自变量)和情绪组别(水平 2 自变量)对积极情绪和消极情绪评分的交互作用均显著，$\eta_p^2 = 0.36$和 0.55；在情绪诱发效果测试中，测量时间(水平 1 自变量)和情绪组别(水平 2 自变量)对积极情绪和消极情绪评分的交互作用均显著，$\eta_p^2 = 0.59$和 0.56；延迟金额(水平 1 自变量)与情绪组别(水平 2 自变量)对时间折扣率影响的交互作用显著，$\eta_p^2 = 0.14$。张环等人(2021)研究发现，年龄(水平 2 自变量)、关系类型(水平 2 自变量)和提取方式(水平 1 自变量)对提取正确率的三因素交互效应显著，$\eta_p^2 = 0.12$。根据 Cohen 对η_p^2的标准(0.02,0.06 和 0.14 分别对应小、中和大效应)，这些跨水平交互均属于大效应(Cohen, 1988, p. 24)。本研究中的交互作用效应量的取值参考了 Arend 和 Schäfer(2019)的研究(参见原文表 4)，采用$\gamma_{11.std} = 0.1,0.3,0.5$代表小，中，大水平。此外，本研究只是进行模拟演示，旨在说明样本量规划的方法，不一定代表具体研究的情况。此局限性已经在讨论

部分进行说明，“特别说明的是，本研究主要目的是说明样本量规划的方法及 CI 等高线图的使用，参数设置不一定代表实际中的大多数情况。”(修改稿第 21 页标红文字)。

补充文献：

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

蒋元萍, 江程铭, 胡天翊, 孙红月. (2022). 情绪对跨期决策的影响: 来自单维占优模型的解释. *心理学报*, *54*(2), 122−140.

张环, 王欣, 刘一贝, 曹贤才, 吴捷. (2021). 成员关系对协作提取成绩的影响. *心理学报*, *53*(5), 481−493.

意见 8：

表 6 上面一段，第二句话又在介绍效应量大小，这一部分应该在前面整体对效应量进行统一论述，而不是分散到文章的各个地方。

回应：

感谢审稿专家的意见。已经整体进行了修改。首先，在“3 置信区间宽度等高线图函数说明”中说明了本研究采用的可接受的最宽置信区间宽度计算方法(修改稿第 6 页标红文字)。然后，使用公式(15)说明了产生模型参数与标准化参数的转换，得到效应量不同水平转换后的数值。最后，在结果部分的每种情况下，根据定义计算了可接受的最宽置信区间宽度作为比较标准。具体请参见修改稿第 14-15 页标红文字。

意见 9：

建议作者加上页码，以便审稿修改。

回应：

感谢审稿专家的提醒，已经加上页码。

补充文献：

温忠麟，刘红云. (2020). *中介效应和调节效应:方法及应用*. 北京教育科学出版社.

de Jong K., Moerbeek, M., & van der Leeden, R. (2010) A priori power analysis in longitudinal three-level multilevel models: An example with therapist effects, *Psychotherapy Research*,20(3), 273–284.

Lane, S. P. & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevelrelationships research. Journal of Social and Personal Relationships. 35(1) 7–31.

Liu, H., & Yuan, K-H. (2021). New measures of effect size in moderation analysis. *PsychologicalMethods, 26*(6), 680–700.

Murayama, K., Usami, S., & Sakaki, M. (2022). Summary-statistics-based power analysis: a newand practical method to determine sample size for mixed-effects modeling. *Psychological Methods*. 27(6), 1014–1038.

第二轮

审稿人 1 意见：

作者回答了我上一轮绝大多数问题。我仅对审稿意见的第4问仍有一点疑虑。对于作者的回答“$Y=\gamma0+\gamma1X+r$，对组1($X=-0.5$)，则有$Y1=\gamma0-0.5\gamma1+r$，对组2($X=0.5$)，则有$Y2=\gamma0+0.5\gamma1+r$，因此标准化的回归系数$\gamma1=(Y2-Y1)/SD$，SD为两组的联合标准差=1。”没有充分理解。显然第二个式子求条件期望减去第一个式子求条件期望是gamma1 = E(Y2|X)-E(Y1|X)，并没有SD的出现。其次SD是什么变量的联合标准差呢？如果是X，那么当X采用偏差编码后，两组的联合标准差就是0.5啊，并不是1。再次，作者是否对于含分类预测变量的标准化回归系数有特殊的定义，即只需标准化结果变量Y而不需要标准化预测变量呢？要知道传统定义，标准化回归系数是指X变化一个X的SD，Y平均变化了多少个Y的SD。操作上相当于回归分析中所有变量，即结果变量和预测变量均先标准化再进行回归分析所得的回归系数。显然偏差编码后的二分预测变量并非标准分数(z分数；因其标准差不等于1)。这就是为什么我之前有疑问为何作者声称采用偏差编码后的回归系数即标准化回归系数。最后，作者新的脚注5声称采用了偏差编码后，无需中心化或标准化。且不论标准化的问题，中心化似乎也只有在平衡设计(实验组和控制组对半)情况下不需要中心化吧。

回应：

感谢审稿专家的意见。首先，非常抱歉在上一轮没有解释清楚效应量的公式。根据定义，Cohen’d值(Cohen, 1988)是用于表示两组连续数据差异的效应量，公式中的联合标准差是因变量的标准差。特别的，在多水平模型中，如果自变量是分类变量，则Cohen’d值的公式为$d=\frac{\gamma_{10.std}}{\sigma_{pooled}}$(Elliot, 2004)，其中，$\gamma_{10.std}$表示部分标准化的回归系数，$\sigma_{pooled}$表示分类自变量分成的不同组其水平1残差($\sigma$)的联合标准差。在本文的例子中，由于两组残差都随机取自同一总体，因此不管两组样本量是否平衡，加权后的σ_{pooled}都等于水平1残差的标准差，为1。所以，Cohen’d值即为$\gamma_{10.std}$。

其次，在多水平模型中，如果预测变量为分类变量，为了系数具有良好的解释性，采用部分标准化（partially standardized, MacKinnon, 2008），即，只对因变量标准化。本研究中，分类自变量仅含两个类别，采用偏差编码(-0.5和0.5)后，部分标准化回归系数表示自变量两个类别(如实验组，控制组)在因变量上的标准化均值差异(Cohen’s d)。如果预测变量为连续

变量，则采用完全标准化（completely standardized, Snijders & Bosker, 2012），即自变量和因变量都需标准化。此外，根据Arend和Schäfer(2019)，在多水平模型中固定效应的标准化回归系数需要根据不同系数在不同水平对应的随机效应来对结果变量标准化。例如，当连续自变量已经标准化后，水平1固定斜率的完全标准化形式为$\gamma_{10.std} = \gamma_{10}/\sigma$，水平2固定斜率的完全标准化形式为$\gamma_{01.std} = \gamma_{10}/\tau_{00}$，跨级交互固定效应的完全标准化形式为$\gamma_{11.std} = \gamma_{11}/\tau_{11}$。

反思我们之前的脚注5，以及全文中对于标准化系数的描述的确有一些表述不准确的地方。在修改稿中已经删除原来的脚注5，改为在模拟研究一实验效应的大小($\gamma_{10.std}$)处增添脚注，“在多水平模型中，$\gamma_{10.std} = \gamma_{10} * SD_{predictor}/SD_{outcome}$。当自变量为分类变量时，$\gamma_{10.std}$为部分标准化的回归系数，即只对因变量标准化($SD_{outcome}= \sigma$，$\gamma_{10.std} = \gamma_{10}/\sigma$)。该系数代表了自变量两个类别在因变量上的标准化均值差异（Cohen’s d）。”。具体请参见修改稿第7页的脚注5。并且，对于自变量是连续变量的情况，在模拟研究二中增添了脚注7，“在多水平模型中，$\gamma_{11.std} = \gamma_{11} * SD_{predictor}/SD_{outcome}$。当$W_i$为分类变量时，$\gamma_{11.std}$为部分标准化的回归系数，即只对因变量标准化($SD_{outcome}= \tau_{11}$，$\gamma_{11.std} = \gamma_{11}/\tau_{11}$；当$W_i$为连续变量时，由于自变量已经标准化（$SD_{predictor} = 1$），则$\gamma_{11.std} = \gamma_{11}/\tau_{11}$为完全标准化的回归系数。”。具体请参见修改稿第13页的脚注7。

补充文献：

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

Elliot, K. (2004). *But what does it mean? The use of effect sizes in educational research*. NFER.

MacKinnon, D. P. (2008). Introduction to statistical mediation analysis. Mahwah, NJ: Erlbaum.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage Publishing.

审稿人 2 意见：

本文经过一轮修改，质量已经有大幅度提升，对审稿人的问题也有了全面且恰当的回应。修改稿还有一下小问题建议作者再整体把控和调整。

意见 1：

第5页，文章第3部分的标题和本节内容不符合。(1)标题为函数的说明，而文本内容并没有对“函数”本身的说明，本节没有介绍有关函数的内容；(2)本节更像是进行样本规划的

方法和步骤，是否考虑修改章节标题。如果是写“步骤”，流程图和过程就很合适。(3)另外，本节内容有一部分更像是“研究设计”，包含了模型介绍、评价指标的计算方法，应该是放在“研究设计”中，比如在第4节，4.2部分增加评价指标，4.3再介绍研究结果。建议作者再对本节内容进行调整。

回应：

感谢审稿专家的建议。已将本节标题改为“置信区间宽度等高线图生成步骤”。并且，把模型评价指标放到了“4.2 评价指标”中，将原来的研究结果放到了“4.3 研究结果”。这种结构更符合一般模拟研究的行文框架。具体请参见修改稿中标红文字。

意见 2：

为什么4和5的标题要称之为“敏感性”分析？文章在第7页第一段只是提出来通过两个敏感性分析来研究，但是没有对“敏感性”的含义进行说明，建议在研究设计部分进行简单介绍；不要“敏感性”是否可行？

回应：

感谢审稿专家指出的问题。最初版本称为敏感性分析是参考了Lafit等人(2022)的研究，他们开发了一个检验力分析的app，在文中说明其使用，然后通过敏感性分析探讨模型参数和其他变量对检验力的影响(To assess how the hypothesized values of the model parameters and of other characteristics of the variables, or the included number of measurement occasions influence statistical power, one can vary the specified values and/or the number of measurement occasions)。本研究使用敏感性分析的确没有说明原因，考虑到文中所指的敏感性分析其实也是常见的模拟研究。为便于读者理解，删除了“敏感性”，改为模拟研究。具体请参见修改稿中标红文字。

补充文献：

Lafit, G., Sels, L., Adolf, J. K., Loeys, T., & Ceulemans, E. (2022). PowerLAPIM: An application to conduct power analysis for linear and quadratic longitudinal actor–partner interdependence models in intensive longitudinal dyadic designs. *Journal of social and Personal Relationships*, *39*(10), 3085-3115.

意见 3：

文章还有一些需要小修，比如第2页1.2第一段、第3页1.3第二段、1.4最后一段以及第13页倒数第二段，举例的“e.g”在中文文章中应该用“例如”；中英文括号混用；中英文字体不统一(包括阿拉伯数字字体也不统一)。

回应：

感谢审稿专家仔细的阅读。已经对文中的这些问题进行了修改。具体请参见修改稿中的标红文字。

第三轮

审稿人 1 意见：

作者已经回答了我所有问题，我没有更多的问题了。

编委意见：

该文还有一些问题需要修改或者回应，详见审改稿上的评论和修改建议。【该文是为实验研究者写的方法论文，只是方法研究者能看懂是不够的。所以请一位实验方面的副主编看看是否能看懂和使用里面的方法（也包括有无实用价值）】

意见 1：

p.3 “根据Cohen的标准，该区间涵盖了效应量小、中、大的条件(0.2,0.5,0.8)， 估计精确性差(Maxwell et al., 2008; Usami, 2020)。” 原本这句话有点矛盾。此外，这里有两个问题，一个是这里所说的效应，与Cohon的标准中的效应是一样的吗？另一个问题是，如果CI涵盖了大中小效应量，这样的区间估计还有意义吗？

回应：

感谢编委的仔细阅读。已经按照您的建议修改了这段话。对于您提出的两个问题，解释如下：首先，这里所说效应量的大、中、小与Cohen定义的标准是对应的。其次，对于效应量考虑置信区间是有意义的，它提供了效应量估计可靠性的信息。如果CI涵盖了大中小效应量，此时，即使根据效应量的点估计值能够大概判断效应量大小，但是置信区间说明真实的效应量可能是小、中、大任一种情况，因此，称为没有得到高准确度的效应量估计值（does not provide a highly accurate estimate of the population Cohen’s d, Maxwell et al., 2008），这时仅关注点估计可能意义不大。在实际中，样本量小的情况下得到这样的置信区间并不少见（见模拟研究结果）。因此，这种区间估计是有意义的，正是这种区间估计才能说明效应量点估计值的不可靠，进而提醒研究者基于结果谨慎下结论。因此，研究报告应同时报告效应量(Wasserstein & Lazar, 2016)及其区间估计的结果。

意见 2：

p.5，脚注1，这个脚注的解释的并没有比原来的说法更清晰。

回应：

感谢编委的仔细阅读。已经将正文删减为“设置水平1、水平2样本量”，将脚注改为“当水平1、水平2自变量为分类变量时，可设定不同类别的样本量。”具体请参见第5页脚注。

意见 3：

p.5，正文中之前没有出现附件1 ，突然就来个附件2了。请调整附件，否则应当在前面合适位置引用附件1.如果是在文章后面附上的内容，建议改为附录。

回应：

感谢编委的仔细阅读。考虑到文章篇幅问题，已经把所有附录上传到科学数据银行（待审核后补充下载地址）。文章的在线补充材料包括：补充材料1：模拟研究结果的附表及附图；补充材料2：线性混合效应模型样本量规划R函数；补充材料3：调用线性混合效应模型样本量规划R函数语句及说明；补充材料4：实例数据分析R语句及结果。已在修改稿正文中重新明确引用了在线补充材料，具体请参见修改稿中标红文字。

意见 4：

p.6，intraclass correlation应当翻译为组内相关，当每个组内的个体完全相同、没有变异时，组内相关系数等于1。建议将需要的脚注编入正文，脚注影响阅读。略过也不影响阅读的才放脚注。

回应：

感谢编委。已经修改为组内相关。将脚注中阐述本研究固定ICC理由的部分移入了正文。在多水平模型中，ICC是一个常见概念，因此为保持行文简洁，默认在正文中不需要对其进行解释，在脚注中给出ICC的定义及其特点。具体请参见修改稿第6-7页标红文字。

意见 5：

p.7，公式（7），ICC涉及因变量组间方差和组内方差，为何根据ICC就能确定tau-00?

回应：

感谢编委。根据定义，在零模型中，$ICC = \tau_{00}/(\tau_{00} + \sigma^2)$。本研究中设定残差方差$\sigma^2 = 1$，因此已知ICC，可以计算$\tau_{00}$。为清楚说明，在修改稿中将原来公式（7）中的1改为σ^2，并在公式上加入文字解释“已知残差方差$\sigma^2 = 1$”。具体请参见修改稿第7页公式（7）及上方标红文字。

意见 6：

p.12， 水平2是试验次数对吗？这是每个被试需要的实验次数？如果是，那几乎没有什么实用意义了吧？

回应：

感谢编委。在本文中，水平1是试次，水平2是被试，具体请参见模型介绍部分（p.4公式前的介绍）。为避免读者在后面样本量规划建议部分忘记两个水平的含义，在第一次说明样本量规划建议时，又重新标注了两个水平。见“当水平1（试次）的样本量过小时(例如，小于30)，无论怎样增加水平2（被试）样本量，也无法使得检验力或检验力+随机效应估计准确性达到要求。”（修改稿第11页标红文字）。

意见 7：

p.22，这样的结论不如没有。

回应：

感谢编委的建议。原来的结论的确与正文的内容有很多重合。作为一篇为实验研究者写的方法论文，本文最重要的结论就是提供了一种简便可行的方法工具。因此，在修改稿中删除了原来的结论部分。

第四轮

审稿人 3（副主编）意见：

该文对心理学的实验设计之样本量规划有非常重要的实用价值。文章易懂，适合非统计专业的心理学研究者阅读和使用。

编委意见：我看了作者的修改和回应，建议发表。

主编意见：该论文对心理学研究的统计分析具有积极意义。同意发表。