《心理学报》审稿意见与作者回应

题目：一种基于进化算法的概化理论最佳样本量估计新方法：兼与三种传统方法比较
作者：黎光明，秦越

第一轮

审稿人 1 意见：

意见 1：研究意义阐述得不够好，最好重写这部分。
回应：感谢专家的意见。文章研究意义部分确有不清晰的地方，已重写。重新梳理了引言部分的逻辑，着重体现了改善概化系数的多种方法的递进关系，使主题“最佳样本量选择”的引出更加流畅。修改了部分语句的表达，使逻辑更清晰。

调整和修改了原文引言，重新撰写了原文引言某些段落，使得引言更具有逻辑性，如下（所修改过的内容全文都用蓝色进行了标识）：

修改说明：第一，为增加准确性，调整了原文第一段对概化系数定义的表述。将原文：概化系数指的是从一个测验或是测量的被试者得分到测验程序同等接受度的条件全域中被试均分的估计的准确性。修改为：概化系数是指从一个测验或是测量的被试者得分拓广到测验程序同等接受度的条件全域中，被试均分估计的准确性(Zhang & Lin, 2016)。

第二，在原文引言中增加了过渡段落第二段。新增加内容为：在概化理论中，随着侧面水平数量的增加，概化系数会随之提高。然而，如果在研究过程中存在限制条件，比如受人力、物力、财力等所限，那么就需要权衡是否需要对研究设计做出改变。在某些情况下，让概化系数提高所要增加的某一侧面的观察数量较大，这时需要的经费可能超出预算。当出现这种情况时，就需要权衡增加侧面的观察数量的必要性(Brennan, 2001)。由此看来，预算和成本是进行测量研究时不可忽略的问题，预算的高低在某种程度上会影响测量结果的正确性。在预算限制下，找到一个高可靠性的测量程序是研究者关注的主要问题之一。这个过渡段有助于说明在预算限制下，最佳样本量估计是有意义的，从而引出本文主题。

第三，调整了原文第二段和第三段的部分文字。例如，在第二段增加了一些文字：增加样本量是减少误差方差最简单也是最直观的方法。但是，样本量大小受到经费和测量内容的限制，不可能无限增大。 删除了原文内容：在基于概化理论设计的实际的研究中，一般存在多个测量侧面，也就是存在多个测量侧面样本量变量。而多个样本量孰高孰低，对测量设

计的概化系数的影响是不一致的。修改了原文内容： 概化系数是代表研究结果可以从一个观测样本推广到可接受的观测全域的可靠性程度的重要参数(Zhang & Lin, 2016)。

第四，修改了原文第四段部分文字。将原文：但前人的研究主要集中在传统算法领域，影响较大的有微分优化法、拉格朗日法、柯西-施瓦茨不等式法等等。这些传统算法基本都是从数学规划理论出发，通过适用性研究，得到较优解。但迄今为止，各种传统算法都存在计算复杂、适用性差、使用条件苛刻等问题。并且前人的研究都较为零散，没有比较系统地进行统一的比较和讨论。 修改为：但是，前人的研究主要集中在传统算法领域，影响较大的有微分优化法、拉格朗日法、柯西-施瓦茨不等式法等。这些传统算法基本都是从数学规划理论出发，通过适用性研究，得到较优解。但迄今为止，各种传统算法都存在计算复杂、适用性差、使用条件苛刻等问题，当情况较为复杂时，并不大适合在实际研究中指导最佳样本量的选择。

意见 2：模拟研究设计最好能联系考试或评价的实际场合，现实中不太可能出现的场景（如，试题数量 260 以上，评分者数量 34 人以上，测评场景 19 种以上）不值得讨论。

回应：感谢专家的意见。当初考虑到模拟研究可以验证一些数学上可行的极端情况，所以对参数进行了比较夸张的设计。为了体现模拟研究对现实的参考意义，我们重新设计了一组符合实际研究情况的参数，并进行了补充实验，所得的研究结论与原模拟研究的研究结论一致。

修改说明：p×i×r 设计模拟研究中，将原规定参数：n_i =（30, 90），n_r =（50, 100），修改为：n_i =（20, 40），n_r =（5, 10），形成 2×2 四种不同的模拟设计，每种设计模拟 500 次，共 2000 次。同时，也更新修改了表 3 的结果（见修改后的正文，所修改过的内容全文都用蓝色进行了标识）。

p×i×r×o 设计模拟研究中，将原规定参数：n_i = （15, 25），n_r =（5, 10），n_o = (10, 20)，修改为：n_i =（15, 25），n_r =（5, 10），n_o= (5, 10)，形成 2×2×2 八种不同的模拟设计，因数据规模较大，每种设计模拟 200 次，共 1600 次。同时，也更新修改了表 4 的结果（见修改后的正文，所修改过的内容全文都用蓝色进行了标识）。

意见 3：概化系数不能设置得近乎完美（0.9857 以上），若能设置在 0.70～0.90 之间，则可以更好地看出不同算法在保障测评精度的前提下谁的成本最低。

回应：感谢专家的意见。当初考虑到模拟研究可以验证一些数学上可行的极端情况，所以数值设计上比较夸张。为了使模拟研究与实际情况相吻合，我们重新设计了新的预算参数，使概化系数尽量在 0.80～0.94 之间分布，所得的研究结论与原模拟研究的研究结论一致。

修改说明：p×i×r 设计模拟研究中，将原规定四种设计下预算参数：B =（2000，5000,10000,18000），修改为：B =（100，150，200，250）。同时，也更新修改了表 3 的结果（见修改后的正文，所修改过的内容全文都用蓝色进行了标识）。

p×i×r×o 设计模拟研究中，将原规定八种设计下预算参数：B =（1000,1500,2000,2500,3000,3500,4000,4500），修改为：B =（800,900,1000,1100,1200,1300,1400,1500）。同时，也更新修改了表 4 的结果（见修改后的正文，所修改过的内容全文都用蓝色进行了标识）。

审稿人 2 意见：

该研究比较了不同设计情景条件下，基于预算限制的不同侧面最佳样本容量估计算法，还是具有一定的实际意义。建议作者编制适用于实际使用的简单易用的软件代码作为附件，因为许多实际工作者可能并不清楚如何估计最佳样本容量，这样可能更有利于该研究的实际价值。

回应：感谢专家的意见。本研究所用主要编程语言为 R 语言和 python 语言。整项研究代码较为复杂，经过整理，在附件中添加了易于实际工作者使用的进化算法程序代码。作为比较和补充，也附上了传统算法编写的最佳样本量估计程序代码。

第二轮

审稿人 1 意见：

文章修改后较好，附上程序更具有应用价值，文献综述中建议添加国内的部分研究文章或著作，方便读者把握概化理论的全貌。

意见 1：文献综述中建议添加国内的部分研究文章或著作，方便读者把握概化理论的全貌。

回应：感谢专家的意见。已在文献综述中添加了国内的部分研究文章或著作，以方便读者能够把握概化理论的全貌。第二次修改之处已用红色标识出来，以区别于第一次修改的蓝色标识。

修改说明：已在文献综述中添加了国内有关概化理论的三部著作和四篇期刊文献，如下：

Qi, S., Dai, H., & Ding, S. (2002). *Modern educational and psychological measurement*. Beijing, China : Higher Education Press.

[漆书青，戴海崎，丁树良. (2002). *现代教育与心理测量学原理*. 北京:高等教育出版社.]

Yang, Z., & Chang, L. (2003). *Generalizability theory and its applications*. Beijing, China: Educational Science Publishing House.

[杨志明，张雷. (2003). *测评的概化理论及其应用*. 北京：教育科学出版社.]

Li, G. (2019). *Psychological measurement*. Beijing, China: Tsinghua Universiy Publishing House.

[黎光明. (2019). *心理测量*. 北京：清华大学出版社.]

Zhu, Y., Fung, S., & Xin, T. (2013).Improving dependability of new HSK writing test Scores: A generalizability theroy based approach. *Journal of Psychological Science*, *36*(2), 479–488.

[朱宇，冯瑞龙，辛涛. (2013). 新 HSK 书写成绩可靠性影响因素的概化理论分析. *心理科学*, *36*(2), 479–488.]

Luo, Z., & Guo, X. (2014). The optimal size of material in psychological experiment: The applications of multivariate generalizability theory. *Acta Psychologica Sinica*, *46*(6), 876−884.

[罗照盛，郭小军. (2014). 认知行为实验研究中最佳素材容量的选择与确定：多元概化理论应用. *心理学报*, *46*(6), 876–884.]

Li, G., Chen, Z., & Zhang, M. (2020). Estimating the best sample size of teaching level evaluation for college teachers under budget constraints in generalizability theory. *Psychological Development and Education*, *36*(3), 378–384.

[黎光明，陈子豪，张敏强. (2020). 高校教师教学水平评价概化理论预算限制下最佳样本量估计. *心理发展与教育*, *36*(3), 378–384]

Liu, Y., Zhang, M., & Zhen, F. (2020). Estimating the best sample size for students' evolution of teaching——Based on the application of LaGrange multiplier method. *Journal of Psychological Science*, *43*(4), 857–863.

[刘颖，张敏强，甄锋泉. (2020). 学生评教研究的最佳样本量估算——基于拉格朗日乘数法的应用. *心理科学*, *43*(4), 857–863.]

因为增加了这些国内的著作和论文文献，因此我们对一些语句进行了梳理和调整，如下：

第一，为节省篇幅，取消了三篇原外文的引用，分别是 Gage, Prykanowski, & Hirn,

2014; Maulana, Helms Lorenz, & Grift, 2015; Wolbing & Riordan, 2016，同时考虑到应适应概化理论研究的前沿性和时空性，又增加了 Truong et al., 2021 这篇文献的引用，该文献如下：

Truong, Q. C., Choo, C., Numbers, K., Merkin, A. G., Brodaty, H., Kochan, N. A., Sachdev, P. S., Feigin,V. L., & Medvedev, O. N. (2021). Applying generalizability theory to examine assessments of subjective cognitive complaints: whose reports should we rely on – participant versus informant?. *International Psychogeriatrics*, *4*, 1-11.

第二，修改了一些语句。例如，将原文：概化理论(Generalizability Theory)属于现代测验理论，广泛用于心理与教育测量领域(Gage, Prykanowski, & Hirn, 2014; Maulana, Helms Lorenz, & Grift, 2015; Wolbing & Riordan, 2016)，修改为：概化理论(Generalizability Theory)属于现代心理测验理论，广泛用于心理与教育测量领域(杨志明, 张雷, 2003; 朱宇, 冯瑞龙, 辛涛, 2013; 罗照盛 , 郭小军, 2014; Truong et al., 2021)。将原文：决策研究是基于概化研究的进一步研究和讨论，研究者利用概化研究得到的各种方差分量，在考查测量设计结构的基础上调整测量侧面的数量以及各侧面之间的关系，通过得出的决策研究的各方差分量及其他结果，衡量研究的有效性和改进的方向，减少误差。修改为：决策研究的目的是为了某种特殊的决策需要，以概化研究所得到的这些方差分量估计值为基础，通过调整测量过程中各方面的关系（如调整各个侧面样本水平数、调整各个侧面之间关系、改变不同变量权重等），来探索如何控制和调节测量误差(黎光明, 2019)。将原文：但迄今为止，各种传统算法都存在计算复杂、适用性差、使用条件苛刻等问题，并不适合在实际研究中指导最佳样本量的选择。修改为：但迄今为止，各种传统算法都存在计算复杂、适用性差、使用条件苛刻等问题，在情况较为复杂时，并不大适合在实际研究中指导最佳样本量的选择。

第三，将增加的相应的文献引入正文中。如将漆书青, 戴海崎, 丁树良, 2002；黎光明, 2020；黎光明, 陈子豪, 张敏强, 2020; 刘颖, 张敏强, 甄锋泉, 2020 等分别加入相应的正文内容中，以方便国内外读者能够把握本文有关概化理论相关研究发展的全貌。

第四，对参考文献进行了调整。删除了原有的 3 篇英文文献，分别是：

Gage, N. A., Prykanowski, D., & Hirn, R. (2014). Increasing reliability of direct observation measurement approaches. *Behavioral Disorders*, *39*(4), 224–228.

Maulana, R., Helms Lorenz, M., & Grift, W. V. D. (2015). A longitudinal study of induction on the acceleration of growth in teaching quality of beginning teachers through the eyes of their students. *American*

Biology Teacher，*51*(1), 225–245.

Wolbing, T., & Riordan, P. (2016). How beauty works? Theoretical mechanisms and two empirical applications on students'evaluation of teaching. *Social Science Research*, *57*(5), 253–272.

参考文献增加了上述提及的新加入的 7 篇中文文献和 1 篇英文文献，新增加的中文文献和英文文献都已用红色标识出来了。

编委意见：

改论文选题有一定新意，经过审稿专家的挑剔性审读以及作者的修改，论文达到发表要求，建议发表

主编意见：

建议提交代码 py 文件和代码中调用的模拟数据 csv 文件，作为网络版的附件供读者参考。