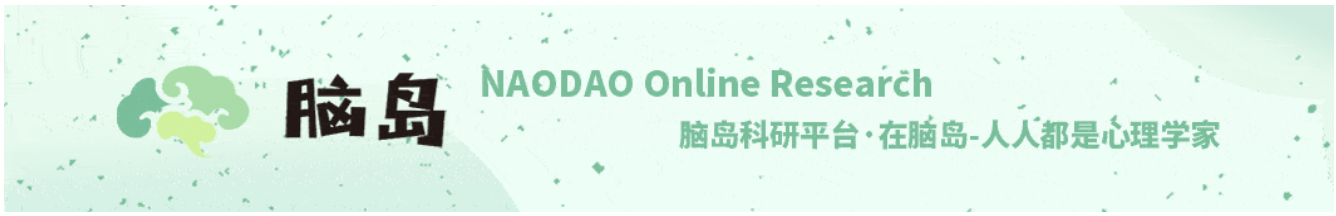
# 我的实验能在线上进行吗？到底需要多少样本量？ ——在线实验可行性&样本量

Original 刘雨晨 脑岛科研平台 2022-06-25 18:00 Posted on 四川




## 1.我们能够在线上平台进行哪些实验呢？

提到在线实验，你会想到什么？

“省时省力，不仅能收集更大的样本量，生态效度还高！很香很奈斯！”

“但是在线实验的结果会不会与传统的线下实验结果有显著差异啊？目前的技术手段可以支持哪些实验在线上进行呢？”

在线实验因其简便、快速收集大范围、大数量样本的能力而越来越被研究人员青睐。然而，也有人会担心这种方式下所收集的数据会受到被试状态、实验环境和实验设备软硬件的影响，导致其与传统实验室环境下的结果存在显著差异；这一担心也劝退了很多对在线实验感兴趣的研究人员。对此，越来越多的研究采用广泛的实验方法进行探索，表明在线实验产生的结果可以与传统实验室环境下获得的结果相媲美。

Psychonomic Bulletin & Review上发表的一篇文章《Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments》基于Mean performance、Performance variance、Internal reliability（内部一致性信度）三个关键指标探索了在线和传统实验室的认知/知觉实验结果有无差异，结果表明就算是对于涉及时间限制或复杂刺激呈现的高要求实验，例如快速刺激呈现和判断、测量反应时细微的变化、视觉刺激的准确感知等，在线实验都可以获得与传统实验室质量相当的实验数据。

图1.Web实验（带有浅色菱形的深色竖条）和传统实验室（带有深色菱形的浅色竖条）结果，
菱形代表mean performance水平。

Crump等人直接使用在线实验复制了一系列实验心理学经典任务，例如Stroop, Switching, Flanker, Posner cueing, attentional blink, category learning tasks等，结果发现对于毫秒级别的反应时实验、快速刺激呈现、以及具有复杂指令的学习任务，在线实验同样能够取得质量令人满意的数据结果。

但是在所进行的7项经典反应时任务中，唯一没能复制的实验就是Masked Priming，在实验要求以64ms以下的时间间隔对刺激呈现时间进行非常精准的控制时，结果并没有复制前人的研究，而64ms及以上则有积极的结果（图2）。这也暴露出来了在线实验现有的问题——由于技术限制，暂时无法实现需要非常精准控制的实验，例如心理物理学实验。除此之外，对于需要专门设备的实验，例如专门的反应盒、光学运动追踪设备、眼动仪等，在线实验也无法满足。

图2. Masked Priming实验结果

值得一提的是，随着在线实验的普及，研究人员对技术和方法的开发，越来越多的实验能够在线上进行。例如，有研究人员开发了一种Virtual Chinrest的方法，Virtual Chinrest能够使用网页浏览器测量被试的屏幕分辨率，以及通过盲点精准测量被试的观看距离（图3），实现自动调整刺激的大小和到被试的观看距离的位置，使基于网络的人类视觉感知心理物理实验成为可能（Li, Q., Joo, S., Yeatman, J., & Reinecke, K., 2020）。



图3. (a)卡片任务：被试将一张银行卡或同等大小的卡片放在屏幕上，滑动滑块使得屏幕上的卡片图像大小与现实相匹配，因此可以计算逻辑像素密度(logical pixel density; LPD)；(b)盲点任务：被试闭上右眼，注视静止的黑色方块，同时红点反复从右向左移动，要求被试在看不见红点时按下空格。

图4.利用盲点和三角函数计算被试的距离

类似的，一些研究者使用智能手机的摄像头（Valliappan, N et al., 2020）或触摸屏代替眼动仪来记录人的眼动（Lio, G et al., 2019），相信随着在线实验的广泛使用和科学技术的发展，会有更多“不可能”的线上实验变成“可能”。

图5.通过在线探索任务中眼球中央凹和手指的同步移动来使用触屏记录眼球运动

## 2.实验的样本量多大才合适呢？

很好，如果你发现自己的实验可以在线上进行，并且决定使用在线实验收集数据，也已经设计好实验，下一步我们需要做的就是确定样本量大小。在线实验的优点之一就是能够在更大的地区范围、群体范围收集更多数量的被试数据。我们都知道小样本量会带来结果假阳性等问题，所以在研究中研究者通常期望样本量越大越好；但是由于经费和时间的限制，我们必须确定一个合适的样本量大小。然而很多人被问到样本量是如何确定时都是一脸懵圈的：“啥？这玩意难道不是参考前人研究的吗/大家不是都这么做的吗？”。再或者转移矛盾中心，将问题抛给导师，逃避虽然可耻但是有用，~~出问题子也有导师扛着~~（×）。

图6.谢谢导师√

然而很多时候前人研究中样本量的选择并不可靠。Psychological Science上发表的一篇文章《False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant》提到，大量心理学研究的不可重复很大程度上是由于原研究的假阳性过高，这种过高的假阳性与研究方法和实验的不严谨密切相关。其中比较突出的就是研究人员在研究过程中进行P值操纵（p-hacking），采用不合理的手段达到统计上的显著（$p<0.05$），例如，有条件地选择样本量，即在收集数据的同时分析数据，若数据结果达到统计上的显著就停止收集数据，这样会大大降低研究的统计检验力（Statistic power），减小效应量（Effect size），降低研究的可重复性；或者采用多个小样本，进行低统计检验力的研究，选择其中的显著结果报告，而非进行一个大样本的高统计检验力的研究，使得研究结果难以重复（Ioannidis，2008; John et al., 2012）。

图7.P值会随着样本量而变化

那么我们该如何确定可靠的实验样本量呢？

在心理学的统计检验方法假设检验(Hypothesis testing)的理论框架中存在相互关联的四个变量：统计检验力（Statistic power）、显著性水平（Alpha）、效应量（Effect size）、样本量（Sample size）。在这四个变量中，若确定了其中三个变量，以及确定了统计模型（T检验、ANOVA或是其他统计方法），那么就可以推导出第四个变量的值。我们可以通过这一网站直观地看出四个变量间的关系（图8）。

图8.可视化统计检验力、显著性水平、效应量、
样本量之间的关系

统计检验力即假设检验中正确地拒绝虚无假设的概率，也就是实验有多大的把握能将真实存在的效应检测出来，通常用1－β来表示，其中β代表假设检验中犯 II 型错误的概率。在进行正式实验前我们通常需要确定一个能够达到显著性水平、并具有合格统计检验力的样本量大小。显著性水平alpha通常选择0.05或者更低的数值，而统计检验力的选择在理论上是越高越好，其选取可以参考相关研究的前人文献，通常高于80%。在显著性水平和统计检验力都确定的情况下，只需要再确定效应量，就可以估计实验所需的样本量了。

## 3.效应量的选择？

效应量用于测量总体中两个变量间关系的强度，是衡量效应大小的指标（摘自维基百科）。与显著性检验不同，这些指标不受样本容量影响，还可以进一步说明p值无法表明的差异和相关程度大小。它可以是原始单位，也可以是标准化后的量。伴随着心理学研究的可重复性危机，效应量这一统计指标越来越受重视，许多研究要求在报告P值的同时报告效应量及其置信区间（Confidence Interval, CI）。因为研究中很可能出现P值显著，但是效应量却特别小的情况，报告效应量能让研究人员更正确地理解统计结果，同时，研究者可以应用多个研究的效应量及其置信区间进行元分析，以得到对真实效应更加准确的估计。

通常在心理学研究中，我们需要采用标准化的效应量对不同条件下测量到的效应进行研究。Lakens(2013)总结了Cohen's d family和R family两类最广泛应用的标准化效应量指标以

及其推荐用途。ANOVA中常用到的partial eta squared也归类到R family中。



表1. Cohen's d的各种指标、标准化效应量和推荐用途（翻译自Lakens(2013)）



表2. η2的各种指标和推荐用途（翻译自Lakens(2013)）

那么我们该如何获得具体的效应量值呢？可能会有以下这些情况：

1） 如果研究问题已经有了相应的研究和效应量的元分析，那么可以直接采用元分析的效应量结果，先验分析实验所需的样本量；

2） 如果研究问题没有元分析，但是有部分前人的研究，那么可以自行对这些研究进行元分析，例如使用固定效应模型或随机效应模型进行元分析（可以参考Goh, J（2016）的文章进行mini meta-analysis）；

3） 如果是一个全新的研究问题，没有前人的研究可以借鉴，按照往常的做法，一般会先做一个较小样本的预实验来估计效应量，但是这种预实验的结果（Pilot data）通常会产生各种偏差，不建议使用这种方法来估计效应量，以及使用该效应量来设计后续研究（Albers, &Lakens, 2018）。Albers和Lakens（2018）给出的建议是，可以基于理论论据确定最小兴趣效应量大小（smallest effect size of interest ; SESOI），并在先验分析中使用SESOI，这可以使得主要研究有一个预先确定的统计检验力去检测或拒绝被认为值得研究的SESOI。例如，研究人员确定了SESOI为中等效应量η2=0.0588，那么一项两组各有87名

被试的研究将有90%的统计检验力来检测SESOI，或者在等效性测试（equivalence test）中拒绝（Lakens，2017）。选择SESOI可以让研究人员精确控制他们所关心的效应量大小的II型错误率（Lakens, 2014）；

4） 相比于SESOI，更有效的方法是采用序列分析（Sequential analyses），这允许研究人员多次（例如，在收集了50、100、150个被试数据后）分析数据并同时控制 I 型错误。序列分析可以采用传统频率方法（Lakens, 2014），或者使用贝叶斯方法进行序列假设检验（Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017）。

而对于效应量大小参考指标，早在上世纪60年代初就开始推荐使用效应量的大佬Cohen给出了建议。Cohen(1962，1969，1988)提出，d=0.2、d=0.5、d=0.8分别对应于小、中、大的效应量；=0.010、=0.059、=0.138分别对应于效应量的小、中、大；用于多元回归的f2效应量指标，f2 =0.02、f2 =0.15、f2=0.35分别对应效应值的小、中、大。

同时，Cohen指出不可盲目使用这一标准，如果类似于P值是否大小于0.05那样参考此标准，则又陷入了会导致类似P值操纵的二分思维，并且在某些研究领域，有时即使是非常小的效应量也是很重要的，因此Cohen建议对效应量大小的解释最好还是参照以往的研究结果或结合实际情况进行。

## 4.样本量的计算

在确定了统计检验力、显著性水平和效应量后，可以用G*Power软件来进行先验分析（priori analysis），估计出所需要的样本量。G*Power是专门用于统计检验力（包括样本量）计算的免费软件，在　官网就可以下载，其使用也十分简单和方便，只需要按要求点点点即可。

图9. G*POWER界面

在G*Power上设定要求的统计方法、已知的以及想要求得的参数，就可以得到相应的数值。如图9所示，统计方法为独立样本T检验，在效应量大小为0.5（中等），显著性水平为0.05，并且统计检验力为80%时，需要每组有64个样本量。

## 参考文献

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. Journal of Experimental Social Psychology, 74, 187-195

Crump, M., McDonnell, J., & Gureckis, T. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PloS One, 8(3), E57410.

Germine, L., Nakayama, K., Duchaine, B., Chabris, C., Chatterjee, G., & Wilmer, J. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. Psychonomic Bulletin & Review, 19(5), 847-857.

Goh, J., Hall, J., & Rosenthal, R. (2016). Mini Meta-Analysis of Your Own Studies: Some Arguments on Why and a Primer on How. Social and Personality Psychology Compass, 10(10), 535-549.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science : A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 1-12.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. European Journal of Social Psychology, 44(7), 701-710.

Lakens, D. (2017). “Equivalence Tests : a Practical Primer for t Tests, Correlations, and Meta-Analyses.” Social psychological & personality science,8(4), 355–362, Web.

Li, Q., Joo, S., Yeatman, J., & Reinecke, K. (2020). Controlling for Participants’ Viewing Distance in Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. Scientific Reports, 10(1), 904.

Schönbrodt, F., Wagenmakers, E., Zehetleitner, M., & Perugini, M. (2017). Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. Psychological Methods, 22(2), 322-339.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359-1366.

刘雨晨|作者
刘良宇|校对
汪寅、高晓雪|审阅
付小敏|排版

People who liked this content also liked

脑岛平台PsychoPy实用妙招+自救手册（第一期）

脑岛科研平台