A/B 测试：为网站优化提供数据化决策

By Grit, 2018.05

1 试验设计

- 项目背景——某个在线教育公司正谋划优化网站的措施，以改进整体的用户体验，将有限的资源用于支持能够完成课程的学生。其中一项提案“免费试学筛选器”，被用于 A/B 测试，以验证其可行性。
- 典型用户的体验流程——公司主页上有“开始免费试学”和“访问课程资料”两个选项：
  - 如果学生点击前者，系统将要求其输入信用卡信息，然后引导其进入付费课程版本的免费试学。14 天后，系统将对他们自动收费，除非他们在此期限结束前取消试用。
  - 如果学生点击后者，他们将能够观看课程视频，并免费进行小测试，但不会获得导师指导或认证证书，也无法提交最终项目报告来获取反馈。
- 优化方案——若学生点击“开始免费试学”，将被询问有多少时间投入到课程中。
  - 如果学生表示每周将投入 5 小时或更多时间，就可以按常规程序进行登录。
  - 如果学生难以保证至少 5 小时的学习时间，就将被告知“课程通常需要更多的时间投入才能顺利完成，你可免费访问课程资料。在这里，学生可选择继续进行免费试学，或免费访问课程资料。
- 试验假设
  - 零假设——“免费试学筛选器”并不能减少因为没有足够时间而离开“免费试学”，并因此受挫的学生数量。即使减少了这类学生的数量，却也将减少继续通过“免费试学”和最终完成课程的学生数量。
  - 备择假设——“免费试学筛选器”将为学生预先设定明确的期望，从而减少因为没有足够时间而离开“免费试学”，并因此受挫的学生数量，同时不会在很大程度上减少继续通过“免费试学”和最终完成课程的学生数量。
- 适用性评估
  - 本试验是基于网站现状做出的优化调整案，有明确的参照。
  - 可以用具体明确的指标加以评估，如下文将提到的总转化率和净转化率等指标。
  - 从学生进入主页到“免费试学”结束，共计 14 天的时间，可见数据获取的周期在合理范围内。

1.1 选择及描述指标

1.1.1 指标描述

首先给出待评估的指标列表，并给出具体的细节描述：

指标	描述或定义	d_{min}
Cookie 的数量	即访问课程概述页面的唯一 cookie 的数量。	3000
用户 id 的数量	即参与免费试学的用户数量。	50
点击次数	即点击“开始免费试学”按钮的唯一 cookie 的数量（在“免费试学筛选器触”发前发生）。	240
点入概率	即点击“开始免费试学”按钮的唯一 cookie 的数量除以查看课程概述页的唯一 cookie 的数量所得的比率。	0.01
总转化率	即完成登录并参加免费试学的用户 id 的数量除以点击“开始免费试学” 按钮的唯一cookie 的数量所得的比率。	0.01
留存率	即在 14 天的期限过后仍参加课程（因此至少进行了一次付费）的用户 id 数量除以完成登录的用户 id 的数量。	0.01
净转化率	即在 14 天的期限后仍参与课程的用户 id 的数量（因此至少进行了一次付费）除以点击了“开始免费试学”按钮的唯一 cookie 的数量所得的比率。	0.0075

1.1.2 指标选择

指标	类型	选择依据	期望结果
Cookie 的数量	不变指标	在试验发生之前就被采集，不受试验影响	在对照组和试验组中的容量一致，顺利通过完整性检验

指标	类型	选择依据	期望结果
用户 id 的数量	既非不变指标，亦非评估指标	首先，点击“开始免费试学”的用户是先触发筛选器，再进行登录，所以用户 id 的数量是在试验发生之后被采集，不能作为不变指标；其次，虽然试验会直接影响到用户 id 的数据采集，但是因为未作归一化处理，也就说本指标可能还会受到浏览人数的影响，而下面三个对用户 id 的数量进行归一化处理的指标，显然更为合适。	N/A
点击次数	不变指标	在试验发生之前就被采集，不受试验影响	在对照组和试验组中的容量一直，顺利通过完整性检验
点入概率	不变指标	点击“开始免费试学”和查看课程概述页动作均发生在筛选器之前	不受试验的直接明显的影响，顺利通过完整性检验
总转化率	评估指标	用户在触发筛选器之后才最终确定是否参加免费试学，所以本指标会受到试验的影响，且经过归一化处理，具有较高的评估性	期望显著降低。表明筛选器可以显著过滤掉那部分想参加试学但是没有充足投入时间的用户，以便资源得以集中，为有意继续学习的用户提供更好的服务。
留存率	评估指标	用户在14天期限之后仍参与课程的用户 id 的数量，显然会受到试验的影响，且留存率经过归一化处理，具有较高的评估性	期望显著增加。即便没有增加，也不至于在很大程度上降低。

指标	类型	选择依据	期望结果
净转化率	评估指标	用户在14天期限之后仍参与课程的用户 id 的数量，显然会受到试验的影响，且净转化率经过归一化处理，具有较高的评估性	期望显著增加。即便没有增加，也不至于在很大程度上降低。

1.2 测量可变性

根据评估指标的可变性，我们可以更好地设计试验。比如选择试验的规模，计算评估指标的置信区间，以及判断各指标的显著性等等。

标准偏差可以很好地度量可变性。本案中提及的评估指标全部属于“是/否”类型，符合二项分布，所以我们先对各指标进行分析变异估计：

> 各评估指标的标准方差的计算过程详见 Google Spreadsheets - Final Project Baseline Values

以上，由分析变异得到的估计结果的准确性有待验证，即通过分析变异和经验变异的匹配度来进行判断。而分析差异性和经验差异性相匹配的条件是，分析单元=分流单元，其中本试验的分流单元为 Cookie。

于是，选定的三个评估指标的标准偏差及其匹配情况如下：

指标	标准偏差	分析单元	分析差异性和经验差异性是否相匹配	判断依据
总转化率	0.0202	Cookie	是	分析单元=分流单元
留存率	0.0549	用户 id	否	分析单元≠分流单元
净转化率	0.0156	Cookie	是	分析单元=分流单元

鉴于留存率的分析差异性与其经验差异性不相匹配，所以在时间允许的情况下，有必要进行经验估计。（在接下来的分析中，我们发现留存率所需的页面流量过高，将被剔除作为评估指标的资格。）

1.3 规模

> 计算过程详见 Google Spreadsheets - Final Project Baseline Values

1.3.1 样本数量和功效

对于多评估指标分析，存在显著性被放大的问题。常见的解决方法有 Bonferroni 校正，但其适用于 n 次独立检验，而选定的三个评估指标之间具有较高的关联性，所以不宜使用 Bonferroni 校正，否则分析结果会过于保守。三个指标对应所需的页面流量统计如下：

指标	所需页面流量	曝光时间 (假设转入全部流量，以天为单位)
总转化率	645875	16.15
留存率	4741212	118.53
净转化率	685325	17.13

从上表判断，首先可以排除掉所需流量过于庞大的指标——留存率，以当前每日流量 40000 计，需要 119 天来完成数据采集，这对 A/B 测试来说，显然时间过于漫长了。

关于统计留存率指标所需的页面流量，即在 14 天的期限后仍参与课程的用户 id 的数量所对应的流量。计算过程如下：

> 1) 利用在线计算器来确定样本容量。其中已知条件为：基准转换率（即 Probability of payment, given enroll）= 0.53，最低可探测效果（即 d_{min}）= 0.01, Alpha = 0.05, Beta = 0.2 。可得至少需要 39115 个用户 id ，才足以支撑对留存率的统计分析。

> 2) 根据转化关系：(Unique cookies to view course overview page) * (Click-through-probability on "Start free trial") * (Probability of enrolling, given click) = 39115.

> 代入已知条件 Click-through-probability on "Start free trial" = 0.08 和 Probability of enrolling, given click = 0.20625，可得 Unique cookies to view course overview page = 39115 / 0.20625 / 0.08 = 4741212，即统计留存率所需的页面流量。（统计其他两个指标所需流量的计算方法同理）

在剩下的两个指标中，所需流量接近，从运行试验的角度出发，两个指标都可以作为评估指标。至于本试验最终所需的页面流量，选择相对值较高的，即 685325，就可以满足需求。

1.3.2 持续时间和曝光比例

1) 这是一个低风险试验。即便用户认为自己没有足够的投入时间，也可以选择继续进行免费试学，或访问课程资料。而且试验不涉及数据库安全和道德风险。

2) 曝光比例=1.因为试验风险低，而所需页面流量大，为了尽量压缩试验周期，可以考虑将全部流量转入试验。

3) 持续时间=18天。由 685325/40000=17.13，向上舍入得 18，这是一个比较合理的时间长度。

2 试验分析

试验已经设计完毕，但是在进入实施阶段之前，必须分析其实施的可行性。

2.1 完整性检查

> 计算过程详见 Google Spreadsheets - Fianl Project Results: SanityChecks Sheet

由于试验中诸多环节可能出错（例如试验分配错误，导致对照组和试验组之间不具有可比性），需要通过不变指标来检查试验的完整性。

由计算可得选定的三个不变指标的 95% 置信区间及观察值，从而判断其是否通过完整性检查：

指标	下限	上限	观察值	是否通过完整性检查
Cookie 的数量	0.4988	0.5012	0.5006	是

指标	下限	上限	观察值	是否通过完整性检查
“开始免费试用”的点击次数	0.4959	0.5041	0.5005	是
“开始免费试用”的点入概率	0.0812	0.0830	0.0822	是

三个不变指标全部通过完整性检查，表明试验架构和设置无误，可以展开下一步的分析。

2.2 结果分析

2.2.1 效应值检验

> 计算过程详见 Google Spreadsheets - Fianl Project Results: EffectSize Sheet

筛选器是否能显著改变相关指标呢？变化的方向是否符合我们的预期呢？这些问题需要通过效应值检验来分析，包括统计显著性和实际显著性分析，分别计算两个指标的 95% 置信区间，再判断显著性：

指标	下限	上限	d_{min}	是否具有统计显著性	是否具有实际显著性
总转化率	-0.0291	-0.0120	0.01	是	是
净转化率	-0.0116	0.0019	0.0075	否	否

1) 由上表可知，筛选器的使用对降低总转化率具有统计显著性，也就是说，我们有 95% 的把握认为变更将减少因为没有足够的投入时间而离开免费试学，并因此受挫的用户数量。又因其置信区间和[-d_{min}, d_{min}]没有交集,所以，变更对降低总转化率还具有实际显著性。

2) 而对于净转化率，变更带来的效应不具备统计显著性，当然实际显著性也无从谈起。 从置信区间 [-0.0116, 0.0019] 来看，净转化率可能有细微的提升，但是也可能有大幅的降低。

2.2.2 符号检验

> 计算过程详见 Google Spreadsheets - Fianl Project Results: SignTests Sheet

在参数检验之后，利用非参数检验（最常见的是符号检验）作进一步的检验。经计算可得各评估指标的 p 值：

指标	p-value	是否具有统计显著性
总转化率	0.0026	是
净转化率	0.6776	否

由上表可知，符号检验的结果与参数检验的结果一致，表明总转化率的变化不具有随机性，而净转化率变化的随机性较强。

2.2.3 汇总

在上述分析过程中，并未使用 Bonferroni 校正。因为 Bonferroni 校正主要适用于 n 次独立检验，然而本试验中的总转化率和净转化率是相互关联的，该校正方法可能导致分析结果过于保守，所以应避免使用。

2.3 结论与建议

1）因为变更对降低总转化率同时具有统计显著性和实际显著性，所以如果单从总转化率这个指标来看，试试变更是可行且有利的。

2) 但是，本次试验的商业考量是，在显著降低总转化率的同时，也要保证净转化率不会大幅降低。从效应值检验中对净转化率的分析来看，第二个条件不能满足。

综上所述， 不建议启用变更 。

3 后续试验

- 概览：试学结果评估器

  在此试验中，参与免费试学的用户在试学 14 天后，系统将对他们自动收费，除非他们在此期限结束前取消试用。这里，我们可以测试一项变更，即在试学 7 天后，邀请用户评估自己的试学结果，在评估结束后给出两个选项：
  - “试学效果良好，计划于试学结束后继续学习”
  - “试学效果仍待观察，继续试学”

- 假设

  我们假设评估器会协助和督促用户及时评估自己的试学结果，并提高付费用户的数量。

- 选择指标

指标	类型	选择依据	期望结果
Cookie 的数量	不变指标	在试验发生之前就被采集，不受试验影响	在对照组和试验组中的容量一致，顺利通过完整性检验
用户 id 的数量	不变指标	用户只有登录 id 后，才可能评估自己的试学结果，所以不受试验影响	在对照组和试验组中的容量一致，顺利通过完整性检验
点击次数	不变指标	在试验发生之前就被采集，不受试验影响	在对照组和试验组中的容量一直，顺利通过完整性检验
点入概率	不变指标	点击“开始免费试学”和查看课程概述页动作均发生在评估器之前	在对照组和试验组中的容量一致，顺利通过完整性检验

指标	类型	选择依据	期望结果
总转化率	不变指标	只有参加试学的用户才可能评估试学结果，所以数据采集在试验之前，不受试验影响	在对照组和试验组中的容量一致，顺利通过完整性检验
留存率	评估指标	用户在14天期限之后仍参与课程的用户 id 的数量，显然会受到试验的影响，且留存率经过归一化处理，具有较高的评估性	期望显著增加
净转化率	评估指标	用户在14天期限之后仍参与课程的用户 id 的数量，显然会受到试验的影响，且净转化率经过归一化处理，具有较高的评估性	期望显著增加

- 分流单元

  “试学结果评估器”是由参与免费试学的用户触发的（也就是说，在用户登录 id 之后），所以在此我们选择用户 id 作为分流单元，这样即便同一个用户使用不同的设备登录时，仍不至于被重复计算，从而保证引流的一致性。

参考资料： http://discussions.youdaxue.com/