A/B测试最终项目：免费试学筛选器

试验设计

将实验对象随机分为数量相近的实验组和对照组，对实验组采用免费试学筛选器，对照组保持原样，进行为期37天的实验，观察每一天的页面浏览量，点击“开始免费试学”的唯一Cookie的数量，以及前23天的试学用户和最终付费用户的数量，进行分析。

度量选择

不变度量：cookie数量、点击数量、点进概率

评估度量：总转化率，存留率和净转换率

Cookie数量

Cookie数量是该试验的引流单位，并且在实验组和对照组都拥有相同的分布，因此是非常合适的不变度量。

用户ID数量

用户ID数量是会受到试学筛选器的影响，我们认为用户群体经过筛选，数量会有变化。但是点进概率会是一个更好的评估度量，因此可以略过。

点击数量

由于试验并没有更改用户在浏览页面->点击“开始免费试学”过程中的体验，因此点击量不会发生显著变化，因此这也是一个非常好的不变度量，而不是评估度量。

点进概率

点进概率可以作为一个不变度量，因为它不会受到试学筛选器的影响，因为查看课程概述和点击“开始免费试学”都在试学筛选器被触发之前执行。因此作为不变度量是合适的。

总转化率

总转化率会直接受到“免费试学筛选器”的影响，因为用户可能会因为筛选器的提示而放弃免费试学，因此总转化率是一个非常好的评估度量而不是不变度量。

存留率

存留率同样会受到“免费试学筛选器”的影响，因为用户在通过筛选器的筛选后，继续试学直到付费的比例是有可能变化的，因此该度量也是一个非常好的评估度量而非不变度量。

净转换率

和以上两者同理，受到筛选器影响的用户行为可能会有不同的表现，因此是非常好的评估度量而非不变度量。

对评估度量的期望结果

- 总转化率：降低。因为免费试学筛选器的功能就是筛选用户，将潜在的难以完成课程的用户筛去。
- 存留率：提高。因为经过免费试学筛选器筛选后的用户更有可能付费完成课程。
- 净转换率：不降低。我们期望从用户群体中筛选掉难以完成的用户，而本身的用户数量没有增加，因此期望是不降低的。

测量标准偏差

列出你的每个评估度量的标准偏差。（这些应是来自“计算标准偏差”小测试中的答案。）

- 总转换率：0.0202
- 存留率：0.0549
- 净转换率：0.0156

对于总转换率和净转换率，我认为分析估计和经验变异是类似的，因为分析单位和引流单位相同，都是Cookie。
对于存留率，我认为分析估计和经验变异是不同的，因为分析单位和引流单位不同，前者是用户ID，后者是Cookie，因此在时间允许的情况下要进行经验估计。

规模

样本数量和支持

在分析阶段不使用Bonferroni校正。
支持访问网页数：685325
需要说明的是，该访问页数是选择“净转换率”指标计算得出的。选择不同的指标会计算出不同的访问页数，例如：

- 选择“存留率”得出的结果为4741212，数字过大而导致试验进行所需时间较长，因此不采用；
- 选择“总转化率”得出的结果为645875，小于685325，因此不采用。

持续时间和风险暴露

我选择将**100%的流量**转入此试验，需要**18天**来运行。
选择原因：该试验风险较低，可以使用大比例流量进行试验，有助于缩短试验周期。
我认为该试验的风险是很小的，因为用户不太可能因为增加了试学筛选器就大幅度减少或者增加付费学习的意愿，同时也不涉及敏感的用户隐私。

试验分析

合理性检查

- Cookie数量：
    - 期望观察值的置信区间: (0.4988, 0.5012)
    - 实际观察值: 0.5006
    - 是否通过完整性检验: 是

- 点击数量：
  - 期望观察值的置信区间: (0.4959, 0.5041)
  - 实际观察值: 0.5005
  - 是否通过完整性检验: 是
- 点进概率：
  - 期望观察值的置信区间: (0.0812, 0.0830)
  - 实际观察值: 0.0822
  - 是否通过完整性检验: 是

结果分析

效应大小检验

- 总转化率：
  - 置信区间: (-0.0291, -0.0120)
  - 最低敏感度： 0.01
  - 统计显著: 是
  - 实际显著: 是
- 存留率：
  - 置信区间: (0.0081, 0.0540)
  - 最低敏感度: 0.01
  - 统计显著: 是
  - 实际显著: 否
- 净转换率：
  - 置信区间: (-0.0116, 0.0019)
  - 最低敏感度: 0.0075
  - 统计显著: 否
  - 实际显著: 否

符号检验

- 总转换率：
  - p值: 0.0026
  - 统计显著: 是
- 存留率：
  - p值: 0.6776
  - 统计显著: 否
- 净转换率：
  - p值: 0.6776
  - 统计显著: 否

汇总

对“存留率”指标而言，符号检验和效应值检验的结果不吻合，原因可能有：符号检验的效能低于效应值检验，因为它对样本不作任何假设。又或者，该指标在周末的差异具有显著性而在工作日没有，而在符号检验中无法体现这一点，从而导致符号检验和效应值检验的结果不符。

本次试验未使用Bonferroni校正，原因：评估度量只有3个，数量较少，不足以造成过多的假阳性情况。

建议

基于A/B测试的结论，我的建议是：不发布“试学筛选器”功能，对试验作进一步的评估。 根据三个评估度量的置信区间计算，我们可以发现实验组：

- 总转换率降低了，即表明试学筛选器确实能使一些无法投入足够时间的用户不选择进行试学，这也符合我们的期望；
- 存留率虽然提高了，但仅仅具有统计显著性而不具有实际显著性，即这个指标的提高并没有达到我们预期的效果，因此就这个指标的变化而言，这次试验并不成功；
- 净转换率无论在统计意义还是实际意义上均没有显著变化，基本达到了实验目的。

后续试验

我认为，该试验的目的是为了给能充分投入学习的学生更加多的支持，改进学生体验，那么可以设计如下的后续试验：

学时提醒。即记录每个试学用户的ID，在试用期（2周）的时间内，记录他们的学习时间，在第7天和第14天时发送提醒，使用户了解他们每周的学习时间，并给出建议（一周投入不足5小时的，建议免费观看课程而非付费）。将样本分成实验组和对照组，观察不同条件（有无“学时提醒”）下，存留率和净转换率的变化。

因此，显而易见，可以确定的引流指标是用户ID，选择不变度量如下：

- 参加免费试学的用户ID。因为这是引流指标，所以是非常好的不变度量选择。
- 点击次数（点击“免费试学”按钮的唯一Cookie数量）。因为点击次数不会受“学时提醒”影响，该试验针对的是参加免费试学的用户，因此这个度量是合适的不变度量。

选择评估度量如下：

- 存留率：存留率预计会受到“学时提醒”的影响，用户通过知道自己的学习时间，可以判断自己是否适合成为付费用户。因此这个度量可以作为评估度量；
- 净转换率：同理，由于选择付费的用户数量可能发生变化，那么净转换率也可以成为一个合理的评估度量选择