A/B 测试项目：优化产品设计

业务背景

- 产品：在线学习网站
- 功能优化：添加“学习时长提示”功能
- 变更前：

网站首页有“开始免费试学”和“访问课程资料”两个选项，如果学生点击“开始免费试学”，系统将要求其绑定支付方式，并在 14 天后自动收费，除非学生提前取消。或者可以直接访问课程资料，但不包括付费服务。

- 变更后：

若学生点击了“开始免费试学”，系统将新增“提示学习时间安排的建议”，我们的期望是学生因此能提前规划时间，最终坚持完成 14 天的免费试学并完成付费的用户数会增加。

试验设计

数据采集

分组单位选 cookie，数据采集流程图：见附件《ab-test-data-collection.pdf》

指标定义

任何提及“唯一 cookie”的地方，其唯一性按天决定。（在两个不同日期进行访问的同一个 cookie 将计算两次。）用户 id 自动唯一，因为网站不允许同一个用户 id 参与两次。d 指实际显著性（由经验得出）。

- cookie 的数量：即访问课程概述页面的唯一 cookie 的数量。（d 最小=3000）
- 用户 id 的数量：即参与免费试学的用户数量。（d 最小 =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）。（d 最小 =240）
- 点进概率：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（d 最小=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。（d 最小 =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。（d 最小 =0.01）
- 净转换率：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。（d 最小 =0.0075）

度量选择

指标	指标用途	理由	预期
Cookie 的数量	不变指标	该数据采集时间在用户发现变更前	
用户 id 的数量	未采用	● 该数据采集时间在用户发现实验变更后，所以将受影响，满足作为评估指标的基本要求。 ● 但该指标的风险是，作为分组单元的 cookie，在每一次分流上的不确定性可能导致两组 id 数量相差很大，所以该指标不	

		适合论证实验效果。 ● 其他比例类型的指标不受分组单元分流不确定性的影响，是更好的选择。	
点击次数	不变指标	该数据采集时间在用户发现变更前	
点进概率	不变指标	该数据采集时间在用户发现变更前	
总转化率	评估指标	● 该数据采集时间在用户发现实验变更后，所以将受影响，满足作为评估指标的基本要求。 ● 分析单元与分组单元相同。 ● 通过计算有效样本量，计算实验运行时间符合要求。	时间预期不足的学生将不再继续登录参加免费试学，因此采集到的“参与免费试学的 id 数量”将减少，而“点击次数”不受影响，因此预期总转化率将显著减小。
留存率	未采用	● 该数据采集时间在用户发现实验变更后，所以将受影响，满足作为评估指标的基本要求。 ● 经计算需要较大样本量（页面浏览量），而受系统每日流量限制，实验运行时间需要好几个月。（经调研一般要求不超过一个月）	
净转化率	评估指标	● 该数据采集时间在用户发现实验变更后，所以将受影响，满足作为评估指标的基本要求。 ● 分析单元与分组单元相同。 ● 通过计算有效样本量，计算实验运行时间符合要求。	无法判断“完成试学的 id 数量”是否明显减少，所以假设不变或增加，而“点击次数”不受影响，因此预期净转化率没有显著变化或显著增加。

测量标准偏差

若每日分组单位为 400（对应的页面量 5000），用分析法（即比例的抽样分布的偏差公式）估计标准偏差如下。

指标标准偏差	分析估计与经验估计	理由
总转化率标准偏差：0.0202	认为相似	分析单元和分组单元都是 cookie
净转化率标准偏差：0.0156	认为相似	分析单元和分组单元都是 cookie

规模

样本数量和功效

不采用 Bonferroni 校正

（原因：本实验只有两个指标，总体 P（FP=0）=0.95*0.95=0.9025，可接受；2.实验要求同时看到两个指标符合期望，如果使用 Bonferroni 校正将使结果过于保守无法参考。）
页面浏览量：685325
（参见下方“计算步骤”第一步到第三步。另外，分别计算指标“总转化率、留存率、净转化率”所需页面浏览量后发现，“留存率”所需实验持续时间太长超过 30 天，该指标不可取；“总转化率、净转化率”所需时间可接受，最后取两者中所需页面浏览量较大的值。）

曝光比例和持续时间
曝光比例：100%
（理由：该实验不会造成明显不好的影响。主要考虑用户习惯、学习周期、系统性能、信息安全、道德。）

评估项	释义	结论
最低风险	包括身体、心理、情感、社会和经济方面。	本次实验变更是一种信息提示，信息内容意在表达学习课程需要投入足够的时间，仅此而已：用户身体健康不会受损；不存在性质恶劣的心理暗示；就算用户反感实验变更，也不会对用户情感造成伤害；不影响用户在社会中既有的生存状态；也不涉及个人财务。
有益性	即使风险很小，研究结果是否有意义。	本次实验变更作为一种信息提示，可以帮助用户自己评估怎样的预期投入学习时间是对最终完成课程有利的，这对用户是有用的。
备选方案	用户是否有其他选择，实验变更是否是强制性的，做其他选择是否有转换成本。	用户可以在看到实验变更后，可继续完成他们之前的决定，并不强制改变用户原来的决定，也不限制接下来的操作，不产生额外成本。
敏感性	包括用户对保密性的期望、内部保密措施（保密性、控制访问、安全性、监控和审计等）、数据额外用途、最终公式范围。	实验数据仅用于本次，仅限实验相关人员访问，且不会对外公开。 信用卡号是敏感信息，但是“输入信用卡号”不是由实验设计导致的，原来就有该操作步骤。
不确定性	是否影响系统稳定性（数据库），；如果最终未采用更改，是否困扰用户。	不影响系统稳定性，用户也不会因为“每周学习时间”这样的提示有或者没有而产生困惑影响使用（不需要用户重新适应）。
道德	是否有悖社会道德规范。	没有

持续时间：18 天
（参见下方“计算步骤”第四步）

计算步骤如下

第一步：计算所需分析单元规模。根据“基准转化率、最小探测效应、绝对值/相对值、统计

功效、显著水平临界值”，计算得出指标“总转化率、留存率、净转化率”所需有效样本量（分析单元数量）。
第二步：计算所需分组单元规模。根据“指标所需有效分析单元数量、分析单元与分组单元比率基准值”，计算得出所有指标所需有效分组单元（cookie）数量。
第三步：计算实验规模。因为网页计算器上提示计算结果旁边提示结果为“per variation”，也就是一个实验组所需样本量，所以将所有“有效分组单元（cookie）数量乘以 2”，得到所需实验总样本量。
第四步：根据“每次流量基准值”，计算所需实验周期。（取指标“净转化率”，有小数直接加 1，所以为 18 天）

step 1			
	gross conversion	retention	net conversion
baseline conversion rate	0.20625	0.53	0.1093125
minimum detectable effect	0.01	0.01	0.0075
absolute/relative	absolute	absolute	absolute
statistic power 1-β	80%	80%	80%
significance level α	0.05	0.05	0.05
	cookies to click	ids to complete checkout	cookies to click
sample size per variation: unit of analysis	25835	39115	27413
step 2			
	Click-through-probability on "Start free trial"	Enrollments per day **divided by** Unique cookies to view page per day	Click-through-probability on "Start free trial"
baseline values	0.08	0.0165	0.08
	cookies to view	cookies to view	cookies to view
sample size per variation：unit of diversion	322937.5	2370606.061	342662.5
step 3			
sample size needed	645875	4741212.121	685325
step 4			
Unique cookies to view page per day	40000	40000	40000
fraction of traffic exposed			1
days estimated	16.1	118.5	17.1

试验分析

合理性检查

不变量指标“cookies 数量，点击次数，点进概率”均通过检查。

	Lower bound	Upper bound	Observed	Passes
• Number of cookies	0.4988	0.5012	0.5006	☑
• Number of user-ids				☐
• Number of clicks on "Start free trial"	0.4959	0.5041	0.5005	☑
• Click-through-probability on "Start free trial"	-0.0013	0.0013	0.0001	☑

	cookies=pageviews=N	clicks on "Start free trial"	click-through-probability on "Start free trial"
Control	345543	28378	**0.082125814**
Experiment	344660	28325	**0.082182441**
total	690203	56703	**0.082154091**
lower	0.4988	0.4959	-0.0013
upper	0.5012	0.5041	0.0013
observed	0.5006	0.5005	0.0001

结果分析

效应大小检验

		Pageviews	Clicks	Enrollments	Payments
Control	count	212163	17293	3785	2033
Experiment	count	211362	17260	3423	1945
	sum	423525	34553	7208	3978
				gross conversion	net conversion
	d			-0.0206	-0.0049
	p-pool			0.2086	0.1151
	SE-pool			0.0044	0.0034
	z-score			1.96	1.96
	m			0.0086	0.0067
	lower			-0.0291	-0.0116
	upper			-0.0120	0.0019
	统计显著性			是	
	实际显著性			是	

符号检验

		gross conversion			net conversion		
	Date	cont	exp	sign	cont	exp	sign
1	Sat, Oct 11	0.195050946	0.153061224	1	0.101892285	0.049562682	1
2	Sun, Oct 12	0.188703466	0.147770701	1	0.089858793	0.115923567	0
3	Mon, Oct 13	0.183718372	0.164027149	1	0.104510451	0.089366516	1
4	Tue, Oct 14	0.186602871	0.166868198	1	0.125598086	0.111245466	1
5	Wed, Oct 15	0.19474313	0.168269231	1	0.07646356	0.112980769	0
6	Thu, Oct 16	0.167679222	0.163705584	1	0.09963548	0.077411168	1
7	Fri, Oct 17	0.195187166	0.162820513	1	0.101604278	0.056410256	1
8	Sat, Oct 18	0.174050633	0.144171779	1	0.110759494	0.095092025	1
9	Sun, Oct 19	0.189580318	0.172166428	1	0.08683068	0.110473458	0
10	Mon, Oct 20	0.191637631	0.177906977	1	0.112659698	0.113953488	0
11	Tue, Oct 21	0.226066897	0.165509259	1	0.121107266	0.082175926	1
12	Wed, Oct 22	0.193317422	0.15980025	1	0.109785203	0.087390762	1
13	Thu, Oct 23	0.190977444	0.190031153	1	0.084210526	0.105919003	0
14	Fri, Oct 24	0.326894502	0.278335725	1	0.18127786	0.134863702	1
15	Sat, Oct 25	0.254703329	0.189835575	1	0.185238784	0.121076233	1
16	Sun, Oct 26	0.22740113	0.220779221	1	0.146892655	0.145743146	1
17	Mon, Oct 27	0.306982872	0.276264591	1	0.163372859	0.154345006	1
18	Tue, Oct 28	0.20923913	0.220108696	0	0.123641304	0.163043478	0
19	Wed, Oct 29	0.265223275	0.27647868	0	0.116373478	0.132049519	0
20	Thu, Oct 30	0.227520436	0.284340659	0	0.102179837	0.092032967	1
21	Fri, Oct 31	0.246458924	0.252077562	0	0.14305949	0.170360111	0
22	Sat, Nov 1	0.22907489	0.204316547	1	0.136563877	0.143884892	0
23	Sun, Nov 2	0.297258297	0.251381215	1	0.096681097	0.142265193	0
	successes			**19**			**13**
	total			**23**			**23**
	two-tail P			**0.0026**			**0.6776**

汇总

没用 Bonferroni 校正。

（原因：本实验只有两个指标，总体 P（FP=0）=0.95*0.95=0.9025，可接受；2.实验要求同时看到两个指标符合期望，如果使用 Bonferroni 校正将使结果过于保守无法参考。）

	总转化率	净转化率
假设检验结果	具备统计显著性 具备实际显著性	不具备任何显著性
符号检验结果	具备统计显著性	不具备统计显著性
描述差异与原因	结果一致	结果一致

建议

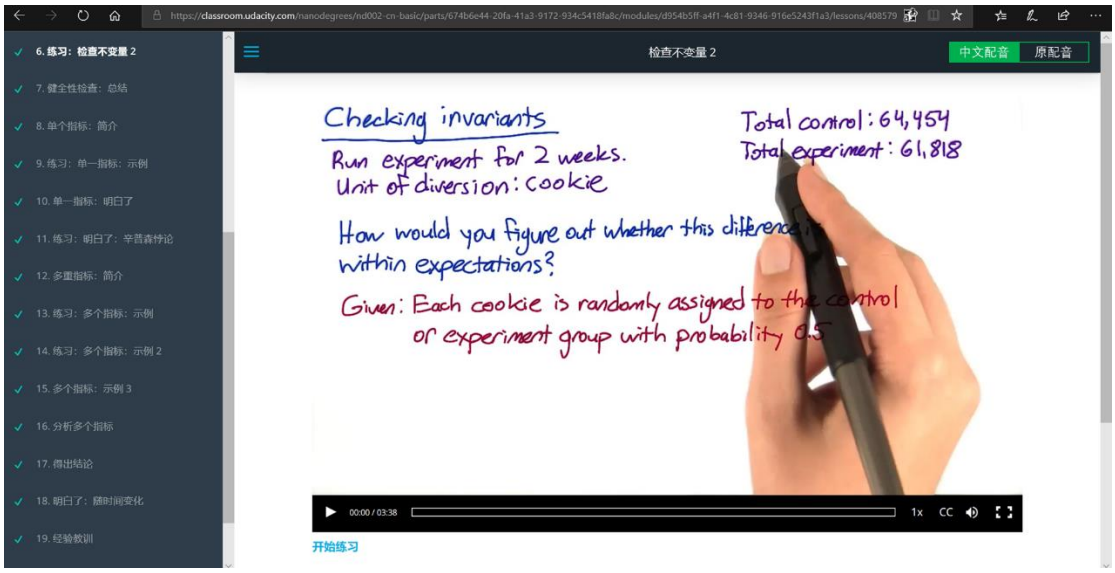
建议暂不启动并深入研究。

通过实验可知，“系统提问学生课程投入时间”的变更，一方面，将明显减少继续登录并选择免费试学的用户数量，这符合实验预期；但另一方面，无法证实完成整个 14 天免费试学的学生数量不会显著减少，因为该指标的实验结果不具备任何显著性。因此，建议不要马上发布该项变更，可以通过用户体验调研来深度了解用户对该项变更的看法并发掘替代举措，我可能从下面几个领域开始调查：

1. 用户的学习时间有何规律，是否可以分组？各组的课程完成率差异是否明显？因为对于中学生、大学生、留学生、上班族、教育从业者而言，安排时间的习惯都有很大差异，他们可能认为自己每周能够拿出 5 小时时间，关键是有些组的日常作息导致他们很难坚持到底。本次变更可能只对某组用户有所效果。
2. 另外，我想调查，完成免费试学的用户的学习时间有什么特点，是下班后两小时开始学习，还是每天半夜学习；是利用周末集中学习，还是每天都学一点。因为除了每周至少花费 5 小时，花费方式对坚持完成课程也很重要。了解这些，我可以改善我的变更举措。
3. 最后我想了解用户选择免费课程的目标是什么。因为如果对整个课程还没有深入了解，那么在试学的过程中，发现课程并不符合自己的目标，即便每周坚持固定时间学习，也会中途退出。这个问题需要采用调研问卷等方法处理。

后续试验

实验概述：母语配音开启按钮
在进行实验时，试学课程视频页面，每个视频标题栏右侧增加了“母语配音开启按钮”，默认是关闭的，用户可以点击开启，则课程配音将采用母语配音；按钮状态决定了每个视频的配音设置。
在此实验中，系统不会对用户做任何多余的提示，仅仅是在设计“母语配音开启按钮”时选择了醒目的颜色，比如绿色，只有被随机分配到实验组的用户才能看到此按钮。



我们假设母语配音会让学生对课程内容有更快的反应，从而减少学生因为英文配音语速过快而受挫离开免费试学。如果这个假设最后为真，优达学城将对全部课程视频添加母语配音以及开启按钮。
分组单位 id，尽管学生参加的是免费试学，但在登陆后他们的用户 id 便被跟踪。同一个用户 id 不能两次参加免费试学。

度量

指标	类型	理由	期望
注册 id 数：注册免费试学课程的 id 数量	不变量	注册 id 数：数据采集在实验变更之前；与分组单元一致。	它应该被随机分配到对照组和实验组。
完成率：完成免费试学的用户 id 数除以注册 id 数	评估指标	完成率：数据采集在实验变更之后；分析单元与分组单元一致，有利于避免所需实验规模过大。	如果假设为真，那么完成免费试学的用户应明显增加，分母不变，该指标会增大。

分组单位：用户 id
（因为要求用户体验一致。）

转移单位：1
（理由）

1. 实验成本：用户来自各个国家，免费试学课程不止一个，如果都配备母语配音，则实验成本太高，所以选择数量更多的中文用户，并通过软件将母语字幕转化为母语配音。
2. 实验规模：根据之前实验结果，我们可能需要至少约 27413 的样本规模，每天注册数约 660，估计实验需要持续至少 42 天。时间有点长，但是可以接受，所以需要使用 100% 数据。
3. 风险评估：只有一点，若实验结束后未发布变更，用户需要重新适应原来的配音。这一点是通过控制母语范围只包括中文，来控制的，认为风险可以得到控制，又权衡了对样本规模的需求，所以决定采用中文用户所有流量。

评估项	释义	结论
最低风险	包括身体、心理、情感、社会和经济方面。	用户身体健康不会受损；不存在性质恶劣的心理暗示；新按钮的增加对页面设计风格和操作体验的改变很小，不会对用户情感造成伤害；不影响用户在社会中既有的生存状态；也不涉及个人财务。
有益性	即使风险很小，研究结果是否有意义。	母语配音可增加亲切感，有助于学生专注于课程内容本身。
备选方案	用户是否有其他选择，实验变更是否是强制性的，做其他选择是否有转换成本。	用户可以在看到实验变更后，可继续完成他们之前的决定，并不强制改变用户原来的决定，也不限制接下来的操作，不产生额外成本。
敏感性	包括用户对保密性的期望、内部保密措施（保密性、控制访问、安全性、监控和审计等）、数据额外用途、最终公式范围。	实验数据仅用于本次，仅限实验相关人员访问，且不会对外公开。
不确定性	是否影响系统稳定性（数据库），；如果最终未采用更改，是否困扰用户。	不影响系统稳定性；但若未启用变更，用户需要重新适应原来的配音。
道德	是否有悖社会道德规范。	没有