


# Шаблон дизайна АБ теста

Ответственный продакт MANAGER	@Иван Иванов
Ответственный аналитик ANALYST	@Вася Пупкин
Аппрувящий аналитик SR. ANALYST	@Синьор Синьоров
Даты проведения теста	с 1 авг. 2023 г. по 15 авг. 2023 г.
Трафик в тесте	N% трафика в тесте Пользователи, которые сделали ____ / Пользователи, которые зашли на экран ____ / Пользователи, которые имеют атрибут ____ / etc
Платформы	App: iOS, Android Web: Desktop, Mobile
Срезы	new/repeated ios/android/web
Ссылка на PROJ	<div> PROJ-1 - Проект Jira не существует, либо недостаточно прав на его просмотр.</div>
Ссылка на ФТ	

- 0. Кто за что ответственен
- 1. Какую гипотезу проверяем и какой результат хотим получить.
  - Гипотеза
  - Описание фичи
  - Что видят пользователи каждой из групп?\*
  - Что мы ожидаем от проведения теста?
  - Как принимаем решение о выкатке
  - Длительность теста
  - Трафик
- 2. Метрики / срезы / пересечения
  - Метрики
  - Срезы
  - Пересечения с другими тестами
- 3. Техническая реализация теста
  - Где реализован тест
  - В какой момент пользователь попадает в тест / Фильтрующее событие
  - Параметры эксперимента
- 4. Результаты
  - Ключевые результаты
  - Дополнительные исследования
  - Выводы и принятые решения
- 5. Чек лист

## 0. Кто за что ответственен



ПОМНИТЕ: Основным владельцем А/В-теста, как и фичи, является продакт менеджер. Поэтому только он ответственен за то, что все будет задизайнено, запущено и посчитано вовремя.

Разделы, которые необходимо заполнить **продакт менеджеру** отмечены тэгом MANAGER

Разделы, которые необходимо заполнить **аналитику** отмечены тэгом ANALYST

## 1. Какую гипотезу проверяем и какой результат хотим получить.



Перед заполнением пунктов 1-2 рекомендуется изучить гайд по АВ-тестированию в Сбермаркете: [Guide to AB](#)

### Гипотеза

MANAGERANALYST

Что проверяем в рамках данного теста.

⚠️ Хороший пример: Если внедрим фичу X, это решит важную проблему Y, и эти изменения можно измерить с помощью метрики W.

⚠️ Аналитику - внимательно чекать гипотезу менеджера и челленджить её при необходимости.

### Описание фичи

MANAGER

Функциональное описание поведения фичи

Добавили \_\_\_\_\_, чтобы \_\_\_\_\_.

### Что видят пользователи каждой из групп?\*

MANAGER

Приложить ссылку на макет/моки/скриншоты/схемы/ФТ

### Что мы ожидаем от проведения теста?

MANAGER

**i** Тут может быть один из нескольких вариантов:

1. Мы хотим подтвердить гипотезу путем **значимого улучшения метрик**. В случае серого или красного эксперимента, мы не выкатываем этот вариант, а возвращаемся на этап генерации гипотез. Решение принимается по росту целевых метрик и **отсутствию падения** ключевых бизнес метрик.
2. Мы в любом случае должны выкатить какое-либо **продуктовое** изменение. Тогда нам нужно убедиться, что эксперимент не портит ключевые метрики. **Выкатываются серые или зеленые эксперименты**.
3. **Технический эксперимент** - по сути аналогичен варианту выше, но тестируется техническое изменение, которое не отражается на пользовательском опыте.
4. Ухудшающий или **обратный эксперимент**. Этот тип эксперимента подразумевает отключение уже внедренной фичи с целью проверки / актуализации оценок эффекта.

Пишем, какой результат ожидаем от теста. Желательно также **указать ожидаемый размер эффектов** на метриках. Это пригодится для оценки длительности эксперимента.

Для выбранной метрики нам нужно определить минимальный размер эффекта, который мы хотели бы задетектировать в тесте. Есть два основных варианта:

- Отталкиваться от значимого для бизнеса эффекта на метрику. Например, если мы считаем успехом прокрас метрики на 1%, а меньшие эффекты нас не интересуют, то минимальный размер эффекта = 1%.
- Подбирать эффект, ориентируясь на величину прокрасов метрики в прошлых тестах. Их можно посмотреть [здесь](#). А [здесь](#) можно найти название интересующей метрики на английском.

## Как принимаем решение о выкатке

MANAGER

**i** При каком условии раскатываем на всех и почему. Желательно постараться описать основные сценарии «если <получили такие-то изменения метрик>, то <что делаем>». При каких результатах готовы раскатить тест? При каких условиях возвращаемся на этап генерации гипотез и/или доработку решения?

Смотри дополнительно **Критерии принятия решения** в [Принципы проведения А/Б](#)

## Длительность теста

ANALYST

MANAGER

**i** Минимальный срок длительности эксперимента - **2 недели**. Возможны случаи, когда требуется проводить тест дольше (3-4 недели), например, если эффекты малы и за две недели необходимый объем трафика не наберется.

Что такое MDE и как его можно оценить:

- [краткое описание MDE](#)
- [\[OUTDATED\] Калькулятор расчёта MDE](#)
- Ссылка на дашборд для расчета MDE: MDE

В сложных случаях обращаемся к аналитику.

### Ожидаемый эффект:

На Целевую метрику 1: X%

На Целевую метрику 2: Y%

На Целевую метрику 3: Z%

### Предполагаемая длительность эксперимента:

X дней/недель

### Обоснование оценки:

⚠ Приложить артефакт, обосновывающий оценку длительности!

Тест планируется катить \_\_\_\_ дней, так как MDE Целевой метрики \_\_\_\_\_

## Трафик

MANAGER

ANALYST



На какую долю аудитории запускаем тест? Не всегда тест можно запустить на 100% пользователей: например, если в случае ухудшающего теста или, если у какой-то еще команды в те же даты планируется тест в той же части продукта (этапа воронки / страницы), то лучше развести тесты по трафику, то есть запустить тесты на непересекающихся фрагментах трафика.

- Тесты должны быть синхронизированы по соли. Что это значит: алгоритм сплитования будет использовать одну и ту же соль при разделении пользователей, и, тем самым, один и тот же пользователь в этих тестах будет попадать в одни и те же бакеты. К примеру, если тесты X и Y имеют одну соль, пользователь с определенным идентификатором N, попадающий в бакет 399 в тесте X, попадет в бакет 399 также и в тесте Y.
- Тесты должны быть разнесены по разным бакетам. Например, если тест X занимает бакеты 0-500, то тест Y должен занимать **другие** бакеты, не пересекаясь с тестом X (например, 500-1000, 800-900 и другие вариации)

## 2. Метрики / срезы / пересечения

### Метрики

MANAGER

ANALYST



Все метрики в эксперименте делятся на три группы:

#### 1. **Базовые** метрики бизнеса.

##### a. Сейчас сюда входят:

- GMV per User
- Average Check
- Conversion
- Gross Profit
- Average orders per User
- Косты

b. Эти метрики должны быть посчитаны в **каждом эксперименте** (и считаются авторасчетом для всех экспериментов). При формулировании гипотез желательно ориентироваться на их улучшение или как минимум отсутствие падения

#### 2. **Целевые** метрики для принятия решения.

- Целевые метрики выбираются менеджером и аналитиком в процессе дизайна эксперимента и их набор определяется в первую очередь тем, на какие метрики и как влияет тестируемый функционал.
- Целевых метрик не должно быть больше 3. Эти метрики учитываются при принятии решения о цвете эксперимента.

#### 3. **Информационные** метрики

- Все дополнительные метрики, которые аналитик и менеджер считают нужным смотреть в эксперименте.
- Не следует ориентироваться** на них при принятии решений, но используем в качестве дополнительного источника гипотез или проверки на то, что все проходит нормально.

Уровень метрики	Название метрики	Ожидаемое поведение	Комментарий // метрики из авторасчёта
Базовая	GMV per User		
Базовая	AOV		
Базовая	Conversion в заказ		
Базовая	Gross Profit		
Целевая			

Информационная			

## Срезы

MANAGER

- ☐ Плановая/быстрая доставка
- ☐ Новые/повторные пользователи
- ☐ iOS/Android
- ☐ web

## Пересечения с другими тестами

MANAGER



Указать, что делаем, если есть пересечения с другими тестами, то есть если есть тесты, которые затрагивают ту же функциональность и есть подозрение в неаддитивности:

- Рекомендуемый безопасный вариант (1) – разводить такие тесты во времени и запускать их друг после друга
- Рекомендуемый безопасный вариант (2) – разводить такие тесты по трафику и запускать одновременно на непересекающихся фрагментах трафика. При использовании не всего трафика особое внимание рекомендуется уделить оценке необходимого трафика для теста (см п.6)
- Возможный, но не рекомендуемый вариант – проводить тесты одновременно и не разводить по трафику. После окончания теста оценивать, исказило ли пересечение результаты. В зависимости от размера искажения оба теста будут корректными / оба некорректными, но их можно будет полечить без перезапуска / оба нужна будет перезапустить.

Посмотреть планируемые тесты для конкретного места в продукте можно в [календаре в А/В платформе](#).

## 3. Техническая реализация теста

### Где реализован тест

MANAGER



Подробнее: [Guide to AB/Имплементация AB/Виды имплементации AB](#)

На бэкенде / на клиенте

### В какой момент пользователь попадает в тест / Фильтрующее событие

MANAGER

ANALYST

Пользователь первый раз взаимодействует с фичей на экране \_\_\_\_\_



При желании можно пофилтровать трафик, чтобы убрать из эксперимента пользователей, на которых эксперимент не повлиял. Делается это с помощью фильтрации пользователей по факту наличия некоторого события за период эксперимента ('фильтрующего события'). Это можно сделать как с использованием авторасчета (указав фильтрующее событие при заведении теста в А/В платформе, так и при ручной оценке эксперимента.

При необходимости фильтрующего события, подбираем его вместе с аналитиком.

(Если нужно) фильтруем пользователей по событию \_\_\_\_\_

## Параметры эксперимента

MANAGER

Платформа	Трафик в группах	Метка (label)	Контрольная группа, параметр	Тестовая группа, параметр	Ссылка на Stage	Ссылка на Prod
APP		Обязательно: feature_state для App	feature_state = 0	feature_state = 1		
WEB		Обязательно: name для Web	name = 0	name = 1		

#### 4. Результаты

ANALYST

!

Комментарий

Были технические проблемы при проведении теста?

Есть ли подозрение, что на результаты теста повлияло пересечение с другим тестом?

Могли ли попасть на период теста какие-то праздники / другие источники аномального поведения пользователей?

	App	Web
Даты проведения теста		
Длительность теста		
Размер групп		
Ссылка на тест в AB платформе		
Ссылка на дашборд с авторасчетом		
Метод подсчета: ручной или авторасчет		

#### Ключевые результаты

i

- Таблица с эффектами по ключевым метрикам эксперимента в абсолютном и относительном выражении, p-value. Очень желательно использовать соответствующий color-код.
- Как объясняем прокрасы метрик? Обязательно для зеленых тестов. Можно отдельным столбцом в таблице с эффектами

Метрика	Значение в контрольной группе	Значение в тестовой группе	p-value	Разница по отношению к контролю, %	Комментарий / интерпретация
					Поведение метрики ожидаемое: да/нет  Изменение этой метрики может нам говорить о том, что _____

#### Дополнительные исследования

i

Полная табличка с метриками, эффектами и p-value. Дополнительные расчеты и графики.  
Если тест считался вручную, прикладываем рабочий ноутбук.

#### Выводы и принятые решения

ANALYST

MANAGER

На основе полученных результатов делаем вывод о том, что \_\_\_\_\_.

Решение по раскатке: \_\_\_\_\_

⚠ **Влияние на GMV при 100% раскатке - *gmV* в год** ⚠  
(плюс ссылка на калькулятор, в котором был рассчитан эффект)

## 5. Чек лист

MANAGER

ANALYST

Этап	Ответственный	Задача	Комментарии
До запуска теста	MANAGER	<input checked="" type="checkbox"/> Описание тестируемой гипотезы в дизайне эксперимента	
	MANAGER	<input type="checkbox"/> Согласование сроков, бакетов, разведение тестов, донесение информации до аналитика	
		<input type="checkbox"/> Первичное описание метрик эксперимента	
		<input type="checkbox"/> Оценка ожидаемого эффекта (дэшборд для оценки <a href="#">MDE</a> , <a href="#">краткое описание MDE</a> )	
	ANALYST	<input type="checkbox"/> Помощь в оценке ожидаемого эффекта	
		<input type="checkbox"/> Дополнительное исследование выбранных метрик, дополнение метрик эксперимента при необходимости	
		<input type="checkbox"/> Расчет размера необходимой выборки для теста	
		<input type="checkbox"/> Оценка достаточности событий для теста, описание событий, которых не хватает, добавление их в ФТ	
		<input type="checkbox"/> Указание в ФТ меток и параметров эксперимента	
		<input type="checkbox"/> Добавление новой событийки в json-схему	
	MANAGER	<input type="checkbox"/> Контроль выполнения разработчиками задачи на доработку событий (при необходимости)	
	ANALYST	<input type="checkbox"/> Описание и постановка задач на доработку метрик авторасчета	
	MANAGER	<input type="checkbox"/> Заполнение шаблона эксперимента в АБ платформе	
	MANAGER	<input type="checkbox"/> Добавление метрик авторасчета в тест (АБ платформа)	

	MANAGER	<input type="checkbox"/> Поддержка тикета на доске TESTS (при необходимости), сам тест создается автоматически (подробнее <a href="#">тут</a> )	
	MANAGER	<input type="checkbox"/> Передача заведенного теста аналитику на проверку	
	ANALYST	<input type="checkbox"/> Проверка корректности заведенного теста	
Тест запущен	MANAGER	<input type="checkbox"/> Планирование, запуск теста	
	ANALYST MANAGER	<input type="checkbox"/> Добавление себя в тестовую группу (посмотреть свой anonymous_id и понять, как добавить себя в группу, можно <a href="#">по ссылке</a> )	
	MANAGER	<input type="checkbox"/> Тестирование функционала	
	ANALYST MANAGER	<input type="checkbox"/> Наблюдение за тестом на дашборде экспериментов (проверить, появился ли тест в на дашборде с результатами)	
	MANAGER	<input type="checkbox"/> Принятие решения об остановке теста (при необходимости)	
После теста	ANALYST	<input type="checkbox"/> Подведение итогов теста, описание их в конфле	
		<input type="checkbox"/> Проведения постанализа по результатам теста (при необходимости)	
		<input type="checkbox"/> Валидация результатов АБ платформы (только в крайней необходимости)	
	MANAGER	<input type="checkbox"/> Добавление информации об остановке теста в календарь АБ тестов	
	MANAGER ANALYST	<input type="checkbox"/> Описание результатов теста для команды и информирование стейкхолдеров о результатах	
	MANAGER	<input type="checkbox"/> Принятие решения о раскатке одного из вариантов теста	
	ANALYST MANAGER	<input type="checkbox"/> Один из вариантов теста раскатан на 100% через АБ платформу	