

A/B-тест доказывает, что изменение помогло, вместо доверия мнению. Сделайте пять вещей правильно: выберите ОЕС (watch time и удержание, не клики), задайте guardrails по качеству, способные отменить релиз, рассчитайте размер выборки, определите правило остановки до старта (не подглядывайте) и выберите метод. Видимые изменения тестируйте достаточно долго, чтобы новизна угасла. Цифры Netflix меняются — сверяйтесь перед планированием.

1 · ВЫБЕРИТЕ ОЕС (что значит успех)

- ☐ **Оптимизируйте watch time и удержание** — удержание месяц-к-месяцу и часы стриминга, не клики и не запуски.
- ☐ **Клики обманывают** — тап с отскоком за 30 секунд — это провал под видом победы; кликбейт убивает доверие.
- ☐ **Быстрые прокси, медленная проверка** — двигайтесь по прокси вовлечённости, важное сверяйте с удержанием.
- ☐ **Запишите ОЕС заранее** — одна согласованная метрика до старта, а не дашборд из двадцати после факта.

2 · ЗАДАЙТЕ GUARDRAILS (что не должно ухудшиться)

- ☐ **Guardrails по качеству (QoE)** — время старта видео, доля ребуферинга, частота крашей, ошибки приложения.
- ☐ **Бизнес-guardrails** — отписки, сбои оплат, обращения в поддержку; победа по ОЕС может вредить в другом месте.
- ☐ **Ухудшение guardrail отменяет релиз** — вовлечённость ценой худшего потока — не настоящая победа.
- ☐ **Проверьте sample-ratio mismatch** — если группы разделились не по плану, эксперимент сломан; стоп и почините.

3 · РАЗМЕР И НЕ ПОДГЛЯДЫВАТЬ (дисциплина)

- ☐ **Размер $\approx 16 \times \text{дисперсия} / (\text{размер эффекта})^2$** — меньший эффект стоит квадратично больше зрителей и времени.
- ☐ **Правило остановки — до старта** — фикс-горизонт с одним взглядом ИЛИ sequential-тест для мониторинга.
- ☐ **Не останавливайте фикс-тест рано** из-за «значимости» — подглядывание раздувает ложные срабатывания.
- ☐ **Sequential-тест для живого мониторинга** — always-valid результаты дают убить вредный релиз за часы.

4 · ВЫБЕРИТЕ МЕТОД И ИЗБЕГАЙТЕ ЛОВУШЕК

- ☐ **A/B-тест** — решающая проверка любого изменения (UI, цена, плеер, регистрация) по удержанию; нужна большая выборка.
- ☐ **Interleaving** — сравнение двух ранжирований при 100+ раз меньшем числе зрителей; отсеьте, затем A/B по выжившим.
- ☐ **Квази-эксперимент** — сравнение по региону/времени, когда нельзя рандомизировать (смена CDN, цена); слабее, запас.
- ☐ **Победите novelty effect** — видимые изменения тестируйте дольше, чтобы всплеск угас; CUPED может вдвое сократить выборку.

ПОРЯДОК ДЕЙСТВИЙ — ВЫБЕРИТЕ МЕТРИКУ, ЗАДАЙТЕ GUARDRAILS, РАССЧИТАЙТЕ РАЗМЕР, ЗАФИКСИРУЙТЕ ОСТАНОВКУ, ЗАТЕМ ЗАПУСКАЙТЕ

На масштабе стриминга один процент улучшения — это миллионы часов, поэтому угадывать нельзя, а контролируемый эксперимент — единственный надёжный способ отличить реальное улучшение от шума. По порядку. Первое: выберите Overall Evaluation Criterion (ОЕС) — единственный исход, определяющий успех, привязанный к watch time и удержанию месяц-к-месяцу, а не к кликам или запускам, потому что изменение под клики выигрывает по кликам и проигрывает по часам просмотра, которые и платят по счетам. Второе: задайте guardrail-метрики, которые не должны ухудшаться, — время старта видео, долю ребуферинга, частоту крашей, отписки — и позвольте любому их ухудшению отменить релиз даже при победе по ОЕС, ведь вовлечённость ценой худшего потока — не настоящая победа. Третье: рассчитайте размер: нужное число зрителей растёт примерно как 16-кратная дисперсия метрики, делённая на квадрат искомого эффекта, поэтому меньший целевой эффект стоит квадратично большей выборки и времени, а снижение дисперсии (CUPED) способно примерно вдвое его сократить. Четвёртое: определите правило остановки до старта — либо зафиксируйте длительность и посмотрите один раз в плановом конце, либо примите sequential-тест для постоянного мониторинга — и никогда не останавливайте фикс-тест рано лишь потому, что он пересёк линию значимости, ведь такое подглядывание фабрикует ложные победы; именно sequential-тест позволяет смотреть на раскатку вживую и убить вредное изменение за часы. Наконец, выберите метод под изменение: классический A/B-тест для решающего замера по удержанию, interleaving для сравнения ранжирований при 100+ раз меньшем числе зрителей как быстрый первый фильтр и квази-эксперимент как запас, когда рандомизировать по зрителю нельзя. Видимые изменения тестируйте достаточно долго, чтобы угас novelty effect. Цифры и платформы Netflix меняются — сверяйтесь перед планированием.