Дмитрий Аббакумов

Семинар «Психометрика онлайн-обучения»

ВШЭ, 15 сентября 2016 год

Дмитрий Аббакумов, руководитель Центра психометрических исследований в онлайн-образовании НИУ ВШЭ, докторант KU Leuven (Бельгия), автор курса «Теория и практика компьютерного тестирования».

Содержание семинара:

1. Базовые принципы проектирования измерений в онлайн-курсах. Разработка плана измерений для курса. Разработка спецификации измерения для темы.
2. Задания закрытой формы: области применения, особенности конструирования, типичные трудности и ошибки.
3. Как выйти из творческого тупика при создании заданий? Практические советы.
4. Задания открытой формы: области применения, особенности конструирования, типичные трудности и ошибки.
5. Как перехитрить хитрого студента? Приемы создания заданий, в которых невозможно угадать ответ.
6. Экспертиза. Доработка заданий на основе результатов экспертизы.
7. Особенности апробации измерений в условиях массовых открытых онлайн-курсов. Оптимизация измерительных инструментов по результатам апробации.
8. Что такое «работающий тест» и как это проверить? Основы классической теории тестирования и анализ результатов измерений.
9. Практические вопросы: определение оптимального количества попыток, установление пороговых баллов (оценок), анализ параллельности и выравнивание вариантов.

В НИУ ВШЭ действует единственная в России магистратура «Измерения в психологии и образовании» в партнерстве с Массачусетским университетом, который считается одним из лидеров в этом направлении.

# Зачем тесты нужно оптимизировать?

Сейчас тесты часто создаются по остаточному принципу, не учитываются и не используются (или используются неэффективно) все возможности платформы, в том числе, разные типы заданий, квизы, «проверь себя» и т.д.

Хорошо оптимизированные и разнообразные тесты позволяют сократить количество заданий, необходимых для выставления оценки студенту с учетом ошибки измерения.

Наконец, хорошо оптимизированные тесты повышают точность оценивания. Хотя компьютерное тестирование не может полностью заменить другие формы оценивания, оно может быть эффективным.

# Базовые принципы проектирования измерений в онлайн-курсах. Разработка плана измерений для курса. Разработка спецификации измерения для темы

Речь на этом семинаре пойдет о составлении тестовых заданий. Как правило, составлению заданий уделяется мало внимания, а отношение многих преподавателей можно описать фразой «А давайте быстренько набросаем несколько вопросов?». К созданию тестов обычно относятся формально. На самом деле, измерения всегда подразумевают систему и предполагают несколько итераций правок, которые позволяют оптимизировать задания и сделать их наиболее эффективными.

## Проектирование измерений в МООК:

1. Планирование, составление экспликации
2. Создание вопросов и компиляция теста
3. Содержательная экспертиза
4. Релиз-1 (апробация (в случае с Курсерой – первый запуск курса) или бета-тестирование с хорошей выборкой (не менее 50 человек с уровнем предполагаемой целевой аудитории курса)
5. Психометрическая экспертиза
6. Оптимизация
7. Релиз-2

Для каждой новой сессии желательно менять до 70% заданий. Понятно, что для точных наук аналогичные задания проще придумать, чем для гуманитарных: достаточно просто поменять условие задачи. Исследования показывают, что в вопросах на гуманитарных курсах иногда даже формулировка вопроса, даже то, что стоит в конце – точка или вопросительный знак – может привести к изменению трудности вопроса.

Чем выше значимость сертификата, который выдается за успешное прохождение курса, тем важнее менять задания, так как на курс могут записываться слушатели предыдущих сессий, помнящих правильные ответы.

На практике внедрение этой системы часто сталкивается с непониманием со стороны преподавателей, которые, во-первых, не осознают, зачем это надо, во-вторых, не хотят изменять привычный стиль работы или не доверяют тестам в принципе и, в-третьих, зачатую болезненно реагируют на вмешательство третьих лиц (психометристов).

Поэтому предлагается использовать подход «У вас все хорошо, прекрасные задания, но вот можно немного откорректировать, чтобы они стали еще лучше».

## Планирование

Планирование тестов для курса проводится исходя из цели оценивания (где и зачем будет применяться тест) и педагогических целей (т.е. что будет измерять тест).

Для практической работы предлагается опираться на три основные цели заданий:

1. Assesment **for** learning (formative assessment)

Оценивание внутри курса, результаты которого используются как преподавателем, так и студентами для повышения качества обучения и может не учитываться в итоговой оценке. К такой форме оценивания относятся задания типа «Проверь себя», викторины, quizы, аннотирование и так далее. Например, это вопрос на понимание после видеолекции.

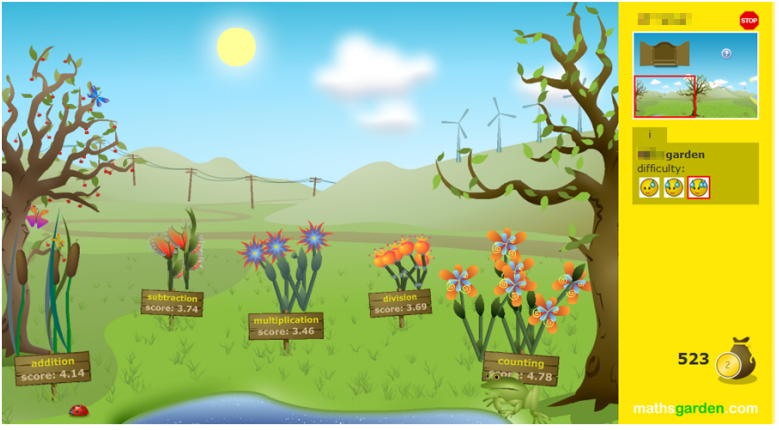
Студентам помогает:

* определить свои сильные и слабые стороны
* оценить собственную динамику в рамках курса
* понять формат и требования итогового оценивания. Поэтому если в итоговом тесте предполагаются какие-то нестандартные задания, надо, чтобы аналогичные вопросы встречались и в течение курса (модуля), и студенты привыкли к форме и требованиям ввода ответа. Например, если есть задание, содержащее определяемое слово и ошибочную дефиницию, студент должен понять, что именно ему надо сделать – исправить определение или написать верное определяемое слово. В противном случае он может невнимательно прочитать или не понять задание и допустить ошибку в оцениваемом тесте.

Призвано помочь преподавателю:

* выявить трудности, возникающие у студентов
* (в идеальном мире) скорректировать содержание курса при необходимости

1. Assesment **of** learning (summative assessment) – оценивание в конце темы, модуля или курса, по итогам которого делается вывод об освоении содержания. Результаты этого вида оценивания используются для формирования итоговой оценки.
2. Assesment **as** learning, или использование заданий непосредственно для обучения (тренажер). Например, Нидерландский проект <https://www.mathsgarden.com/>, который использует тестовые задания для обучения детей счету: чем больше примеров решает ребенок, тем больше цветов появляется на полянке и т.п.



### Педагогические цели

Для определения целей курса можно использовать традиционную таксономию Блума (Знание→Понимание→Использование→Анализ→Синтез→Оценка). Но на практике лучше применять ее укороченную версию Знание→Понимание→Применение. Остальные компоненты включаются в «Применение». Так проще общаться с преподавателями, которые редко настроены на сотрудничество.

1. Задания, педагогической целью которых является «**Знание**» проверяют, как студент припоминает факты, даты, имена, термины, определения, концепции. Очень часто большинство тестов проверяют именно знание, и не затрагивают достижение двух других целей.
2. **Понимание** – преобразование информации из одной формы в другую: сравнение, приведение в порядок, группировка, предсказывание последствий и так далее.
3. **Применение** – перенос и использование информации в новом контексте: решение задачи или кейса, разработка проекта.

Например, знание формулы площади треугольника – это **знание,** умение решить простую геометрическую задачу – **понимание,** а способность найти площадь газона сложной формы, который можно разбить на несколько треугольников – **применение.**

Тесты хорошо подходят для проверки достижения целей «знание» и «понимание», но иногда их можно адаптировать и для третьей цели. Например, слушателю предлагается решить какую-либо сложную задачу и записать ответ.

Конечно, невозможно полностью заменить тестовыми заданиями собеседование с экспертом. И эксперта, оценивающего знания и навыки студента тесты не заменят, но и хорошего эксперта клонировать невозможно. А хорошо оптимизированные задания могут решить проблему.

## Инструмент измерения – не просто тест, это система:

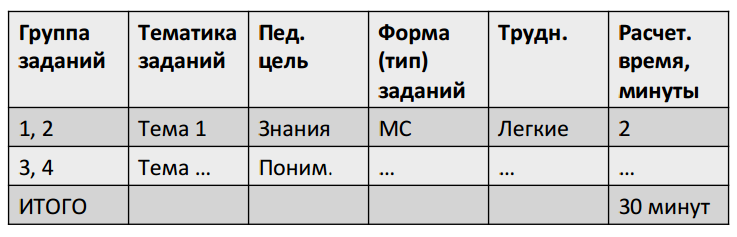
1. Спецификация

2. Задания + Ключи

3. Инструкция для испытуемого

4. Правила интерпретации результатов

Спецификация – это документ, определяющий содержание теста. Спецификация должна отражать связь заданий с содержанием онлайн-курса, формируемыми компетенциями и планируемыми результатами обучения.



В идеале, спецификация первична по отношению к тесту и может выступать как техническое задание для разработки, в том числе, если проверочные задания делают сторонние эксперты, а не автор курса. Но и уже разработанные к курсу задания можно расписать по этому принципу и проанализировать. Так можно понять, каких заданий не хватает, возможно, где-то надо изменить трудность и т.д.

Общее время, которое студент тратит на выполнение задания (любого), не должно превышать полтора часа, потом у него падает концентрация, и результат будет ниже, чем мог бы.

Эксперт (преподаватель) вначале разрабатывает спецификацию и отмечает, сколько заданий нужно на каждую тему (модуль, блок), достижение каких целей они должны проверять и в какой пропорции. Например, в начале курса часто важно **знание,** так как во вводных модулях слушателя знакомят с терминологией, датами, подходами и т.д., а к концу курса более приоритетным станет **применение** и заданий на проверку достижения этой цели станет больше.

Спецификация полезна при обновлении заданий перед перезапуском курса (рекомендуется обновлять около 70% заданий): по ней видно, что можно оставить, а какие задания следует заменить на аналогичные.

Также спецификация необходима, если проводятся психометрические измерения, т.к. только из нее можно понять, что именно измеряется.

# Задания закрытой формы: области применения, особенности конструирования, типичные трудности и ошибки

1. Задания на выбор одного правильного ответа из нескольких
2. Выбор всех правильных ответов из числа предложенных
3. Задания на установление соответствия
4. Задания, предполагающие вписывание правильного ответа

Задания должны сопровождаться подробной инструкцией: сколько попыток, за какое время, один ли правильный ответ в вопросе или несколько, как надо записать ответ и куда нажать и т.д. Особенно внимательно надо составлять инструкции к нестандартным вопросам и заданиям с развернутым ответом.

## Задания закрытой формы (multiple-сhoice). Советы по составлению

1. Избегайте в тексте задания двусмысленности или неясности формулировок (никогда: «как вы думаете»).
2. Больший текст следует включать в основную часть задания, оставляя для ответа только наиболее важные, ключевые слова.
3. Старайтесь равномерно распределять частоту выбора номера правильного ответа в тесте, либо номер правильного ответа выбирайте в случайном порядке (эта рекомендация актуальна для бланковых тестов, но не для онлайн-курсов, где ответы перемешиваются).
4. Соблюдайте правила русского языка (орфография, знаки препинания, начало слов с заглавной буквы) при формулировании заданий теста. Важно вычитывать задания после правок: после корректировок могут измениться число и падеж у отдельных слов, и предложение станет несогласованным.
5. При формулировке дистракторов (вариантов ответов близких к правильному, но таковым не являющихся) не рекомендуется использовать без необходимости слова: все, ни одного, никогда, всегда и т.д. и выражения «ни один из перечисленных», «все перечисленные» и т.д. Последние перестают работать при перемешивании ответов, ведь такой вариант должен стоять в конце перечня, а не в середине.

Варианты ответа должны быть равновероятными, то есть, такими, чтобы студенты действительно могли их выбрать. Лучше, чтобы ответных опций было не менее четырех. Больше зачастую придумать сложно, даже четыре равновероятных ответа не всегда легко придумать. Почему четыре? Допустим, варианта ответа только три. Тогда если студент может отбросить одну опцию, опираясь на свои знания. Тогда выбор правильного ответа превращается в угадывание с вероятностью попадания 50%. Именно так советуют поступать западные пособия по решению тестов – убирать заведомо неверные ответы, а затем – если уж верного ответа не знаешь, – угадывать. Конечно, даже если предложено четыре опции, студент может отказаться от двух вариантов, но вероятность этого меньше.

Нет однозначного мнения, стоит ли использовать «смешные» варианты ответов. С одной стороны, они вроде бы могут снять напряжение, развеселить студента, с другой – их все равно, скорее всего, забракуют. Чем важнее тест, тем меньше вероятность выбора «смешного» ответа. Иногда это можно использовать как «ловушку» (о них - дальше), когда правильный ответ смотрится забавным на фоне остальных. Тогда студент, не знающий его, будет склонен выбрать другой вариант.

### Шесть принципов композиции заданий закрытой формы:

#### Принцип противоречия

Ответы подбираются так, чтобы второй отрицал первый.

Органическое стекло – это пластмасса, состоящая из термопластической смолы

*a) с наполнителями.*

*b) без наполнителей.*

Обратите внимание на знаки препинания: в конце вопроса нет ни точки, ни вопросительного знака, а варианты, хоть и начинаются со строчных букв, заканчиваются точками. То есть сочетание вопроса и ответа дает законченное предложение.

#### Принцип противоположности

Ответы допускают возможность потенциального существования других переходных понятий и состояний.

Функция, имеющая в интервале положительную производную, в этом интервале

*a) возрастает.*

*b) сохраняет постоянное значение.*

*c) убывает.*

#### Принцип однородности

Подбираются такие ответы, которые относятся к одному роду, виду, или отображают различные стороны, грани явления.

Конституционный контроль во Франции осуществляет

*a) Государственный совет.*

*b) Конституционный совет.*

*c) Конституционный суд.*

#### Принцип кумуляции

Содержание каждого ответа вбирает в себя (аккумулирует) содержание предыдущего.

Толерантным называется отношение

*a) рефлексивное.*

*b) рефлексивное и симметричное.*

*c) рефлексивное, симметричное и транзитивное.*

Понятно, что этот вопрос можно переформатировать во «флажки», это можно рассматривать как альтернативный вариант вопроса.

#### Принцип сочетания

Используется сочетание слов (знаков) по два, три, четыре в каждом ответе.

Механическая работа, совершенная телом при поступательном движении, зависит от

*a) силы и времени.*

*b) силы и перемещения.*

*c) силы и пройденного пути.*

Этот вопрос можно изменить, добавив слово «силы» в сам вопрос.

#### Принцип двойной альтернативы

При конструировании ответов используется противопоставление, построенное на принципе противоречия или принципе противоположности.

Для развитых стран характерны

*a) низкая рождаемость и низкая смертность.*

*b) высокая рождаемость и низкая смертность.*

*c) высокая рождаемость и высокая смертность.*

*d) низкая рождаемость и высокая смертность.*

## Задания с вводом короткого ответа

1. Вопрос: представляет собой конкретный вопрос, на который можно ответить одним словом, фразой, числом или символом.
2. Задание на дополнение: представляет собой неполное утверждение, которое необходимо дополнить одним словом, фразой, числом или символом.

Требования к вопросам:

1. Не должно быть некорректных утверждений, а правильный ответ должен быть единственным (Плохой пример: «Чем знаменит Александр Пушкин?»\_\_\_\_\_).
2. Не допускается слишком много пропусков, чтобы не терялся смысл задания.
3. Необходимо избегать невольных подсказок (когда формулировка вопроса или утверждения наводит на нужные род/ число и т.д.).

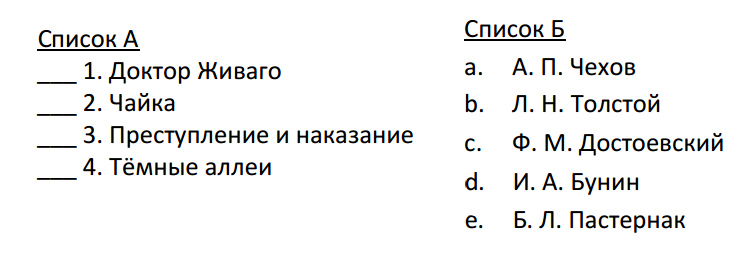
## Задания на установление соответствия

Задания этого типа требуют очень четких инструкций, чтобы студенту было очевидно, на основании какого признака надо производить соответствие, все ли варианты обязательно должны использоваться, может ли один вариант использоваться дважды. На некоторых платформах выбранный один раз вариант нельзя отметить еще раз, таким образом, последняя проблема снимается.

Раньше было обязательное требование, чтобы в одном из столбцов количество вариантов было больше, чем во втором, потому что, в противном случае, последняя пара устанавливается автоматически. Сейчас это необязательно.

Оптимальное задание содержит от 4 до 10 условий. Их можно организовать в логической последовательности, например, по алфавиту или в хронологической последовательности.

Например:



# Как выйти из творческого тупика при создании заданий? Практические советы

Как отмечалось раньше, обычно самыми распространенными становятся вопросы на **знание.** Зачастую это фактологические вопросы (отвечают на вопросы «Что?», «Где?» или «Когда?») или персонологические («Кто?»).

Поэтому важно указать другие типы вопросов для использования в тестах. Иногда, чтобы выйти из «творческого тупика» достаточно предложить ее автору вопросов.

Чтобы сделать задания разнообразнее, интереснее и (в некоторых случаях) изменить их трудность, первоначальную версию задания можно переформатировать. Важно использовать эту типологию как опору, а не пытаться включить в тест все типы вопросов.

## Типы вопросов:

* Причинно-следственный вопрос «Почему?», «Что следует?»
* Функционально-целевой вопрос «Для чего/зачем? С какой целью?»
* Структурный вопрос «Из чего?»
* Инструментальный вопрос «Как? Каким образом?»
* Обстоятельственный вопрос «При каких условиях?»
* Словарно-концептуальный вопрос «Как называется? Что значит термин?»
* Логико-дедуктивный вопрос «Что следует из данных посылок?»
* Концептуально-ассоциативный вопрос «Заполни пропуск»
* Концептуально-аналитический вопрос «Проведи аналогию»
* Задание на восстановление последовательности или структуры «Продолжи ряд» или «В какой последовательности?»
* Концептуально-семантическое задание для установления смысловой эквивалентности или близости
* Проблемно-операциональное задание «Реши задачу»
* Исключение лишнего и вопросы с «НЕ»

# Задания открытой формы: области применения, особенности конструирования, типичные трудности и ошибки

Задания с развернутым ответом

В задание введены ограничения, относящиеся к форме и элементам содержания ответа, но при этом сам ответ может быть достаточно разнообразным.

Плюсы: позволяют оценить навыки интерпретации, анализа, оценки, применительно к конкретным ситуациям.

Минусы: возможны искажения при оценивании (rater effects). Так, опыт ВШЭ в анализе собственных курсов, запущенных на Coursera, показывает, что в целом студенты относятся друг к оценке работ лояльно или снисходительно, но есть отдельные студенты, которые всем ставят сильно заниженные оценки или нули. По-хорошему, таких студентов надо исключать из числа оценивающих, но возможности платформы не позволяют это сделать. Также есть студент, которые ставят оценки «От балды», поэтому важно предусматривать возможность перепроверки задания экспертом, если оценки работы сильно расходятся.

## Общие правила:

Использовать только в тех случаях, где оценка другими, более объективными способами, неприменима. Важно четко определить задание:

* 1. Для заданий с ограниченным объёмом: задать объём и элементы содержания. Например: «приведите три примера», «укажите столько-то концепций»
  2. Для заданий со свободным объёмом: дать рекомендации по выполнению и оформлению, а также минимальные требования к содержанию.
  3. Важно избегать выражений типа: «по желанию добавьте таблицу», «как возможный вариант, изобразите схематически», так как непонятно, как такое задание оценивать, давать ли за это дополнительные баллы.
  4. Установить время выполнения (дедлайн).

Рекомендуемое время выполнения задания не должно превышать полутора часов.

## Задания на взаимное оценивание:

Взаимное оценивание предполагает, что работы студентов оцениваются самими студентами. Проверка задания также может являться обучающим элементом. Иногда для реализации этой цели можно использовать дробную систему оценивания (см. п.4).

Инструкцию по оцениванию работы студенты должны получать непосредственно перед этим этапом.

1. Разработайте рубрики, по которым студенты будут оценивать работы друг друга.
2. Сделайте так, чтобы все элементы задания подлежали оцениванию.
3. Убедитесь в том, что количество баллов, присваиваемое за выполнение различных элементов задания, является эквивалентным. Это даст стимул студенту к выполнению всего задания, а не тех его частей, которые дают наибольшее количество баллов в итоге. То есть не должно быть так, чтобы за выполнение одного параметра давали 5 баллов, а другого – 25. Если какой-то параметр для конкретно этого задания не важен, его можно не учитывать при оценивании.
4. Убедитесь в том, что рубрики просты и понятны для студентов. Не используйте сложных схем оценивания. Оптимальной является дихотомическая (0 – не выполнено, 1 – выполнено), или дробная (0 – не выполнено, 1 – выполнено частично, 2 – выполнено полностью) оценка для каждой рубрики.
5. Не забудьте включить пример выполненного задания, а также пример того, как это задание оценивалось на основе разработанных вами рубрик. Наличие примера и рубрик перед выполнением задания позволит студентам оценивать работы друг друга более оперативно и объективно.

Чтобы не давать студенту шаблон для написания работы, в качестве примера можно использовать образец, написанный на другую тему.

1. Процедуру оценивания рекомендуется организовать так, чтобы каждая работа была оценена не менее чем тремя студентами. Важно предусмотреть возможность перепроверки работ преподавателем или ассистентом в тех случаях, когда оценки, данные студентами, существенно расходятся.

## Оценка заданий на взаимное оценивание

Анализ оценок таких заданий показывает, что оценку можно рассматриваться как набор эффектов:

• эксперта (проверяющего), строг он или снисходителен

• студента, который выполняет задания

• задания, его формулировки и четкости критериев

# Как перехитрить хитрого студента?

Чем точнее автор может имитировать то, как будет трактовать задание студент, не знающий сути предмета, но старающийся угадать правильный ответ, тем удачнее он (автор) будет формулировать задания, а в особенности – дистракторы (ложные, отвлекающие ответные опции).

Психологика испытуемых – совокупность представлений, правил оперирования ими и стратегий поведения, которые конкретный испытуемый склонен проявлять в психологическом эксперименте. В том числе – студент при ответе на экзаменационный вопрос (подробнее – в книге «Практическая тестология» А.Г. Шмелева).

1. Позиция. Этот вариант не работает с теми платформами, где предусмотрено автоматическое перемешивание ответов. В бланковых тестах, если студент замечает, что в первых подряд идущих вопросах правильный вариант стоит на одном и том же месте (например, «а»), он может предположить, что в этом тесте все правильные ответы находятся на этом же месте. В некоторых случаях работает обратный механизм: «Не могут все правильные ответы идти под одной буквой».
2. Длина.
3. Наукообразность. Студент, не знающий правильного ответа, склонен выбирать самый длинный вариант или написанный наиболее сложно, «научно». Считается, что это наиболее справедливо для тестов по медицине.
4. «Кодовые слова». Студенты склонны выбирать вариант, где присутствует слово, связанное с темой лекции или какое-либо «любимое» слово преподавателя.
5. Пересечение слов. Склонность выбрать вариант ответа, который включает какие-то части из других ответных опций.
6. Нечто среднее. Например, среднюю дату, среднее значение, среднее количество. Испытуемый предполагает, что автор задания дает ему возможность сделать два типа ошибок «недооценка» и «переоценка», и с меньшей вероятностью испытуемый полагает, что правильный ответ может быть крайним.

Эти закономерности можно использовать при составлении заданий, чтобы исключить вероятность «угадывания» ответа. Такие правила действуют для людей, начиная со старшей школы, у испытуемых младше проявляются другие закономерности. Конечно, некоторые учащиеся знают про эти хитрости и, наоборот, будут склонны выбирать из крайних дат, самые простые варианты и т.д.

# Экспертиза (содержательная). Доработка заданий на основе результатов экспертизы

После разработки заданий они должны проходить содержательную экспертизу. Эксперты оценивают:

* качество вопроса: корректность формулировки и ответных опций, понятность формулировок, лаконичность, достаточность;
* предположительную трудность для целевой аудитории;
* соответствие теме теста.

На основании этой экспертизы они предлагают корректировки.

В качестве экспертов могут привлекаться коллеги преподавателя, студенты старших курсов, специализирующиеся тематике курса. Желательно, чтобы эксперты знали основы конструирования тестов. Рекомендуется, чтобы одно задание оценивали минимум три эксперта.

Считается, что один эксперт за день работы может проверить 100-150 тестовых заданий с выбором одного варианта ответа (два-три сеанса по 2-3 часа с перерывом между ними)

Некачественным считается задание, если:

* оно не связано с темой теста;
* указанный автором ответ спорен;
* есть неточности в формулировках, грамматические, орфографические, пунктуационные ошибки.

Пример 1: Этот летчик был первым человеком, который летал в космическом пространстве Земной орбиты в техногенном космическом аппарате, лучшем в эпоху освоения Космоса. Его космический корабль «Восток» облетел орбиту Земли 12 апреля 1961. О ком идет речь?

А. Сергей Королев

Б. Алексей Леонов

В. Юрий Гагарин

Г. Герман Титов

Вопрос многословный, его можно сократить:

Кто был первым советским летчиком, совершившим полет в космическое пространство?

При этом, в принципе, вопрос может быть и длинным. Например, если он включает кейс, который надо оценить/разобрать студентам.

Пример 2: Какие из приводимых ниже характеристик могут быть отнесены к финансовому праву как отрасли (разделу) науки?

А) относится к публично-правовым наукам

Б) относится к цивилистическим наукам

В) относится к точным наукам

Г) относится к техническим наукам

Вопрос можно упростить и сократить: Финансовое право относится к А) публично-правовым наукам. Б) цивилистическим наукам. В) точным наукам. Г) техническим наукам.

Также можно перенести повторяющееся слово из ответных опций в сам вопрос.

Задания с развернутым ответом:

Почему символы так важны в произведении «451 градус по Фаренгейту»?

Исходя из формулировки задания непонятно, что именно надо сделать студенту, в каком объеме, и какие критерии будут применяться в оценивании работы. Предлагаемая корректировка:

Символы очень важны в произведении «451 градус по Фаренгейту». Какие три, по Вашему мнению, являются наиболее важными в нём? Опишите их и выделите их роль и смысл в этом произведении.

# Основы классической теории тестирования и анализ результатов измерений

После содержательной экспертизы происходит релиз курса (или апробация), собирается статистика и проводится психометрическая экспертиза еще одна корректировка заданий.

Для апробации нужна выборка не менее 50 студентов, совпадающих по характеристикам с целевой аудиторией. В случае ВШЭ – это просто первый запуск курса на Coursera.

Апробация позволяет выявить ошибки, которые пропустили на этапе экспертизы, а также точно оценить реальную трудность задания для обучающихся.

Психометрика (англ. – psychometrics) фокусируется на высокоточных измерениях, анализе данных и математическом моделировании педагогических и психологических процессов.

Психометрика развивается более 100 лет. Один из вопросов, на который она отвечала: как сравнить двух человек по определенным признакам, чтобы верность сравнения была доказана.

Для этого использовались два основных инструмента:

1. Тесты
2. Математический аппарат

В последнее время на первый план выходит разработка инструментов (тестов), а анализ измерительных инструментов по уровню приближается к анализу данных, особенно это важно в случае МООК, где обучаются тысячи студентов.

Конечной целью психометрической экспертизы является оптимизация тестов (по количеству заданий, трудности), увеличение точности результатов тестирования.

Психометрика как научная область начала формироваться в конце 19 – начале 20 века. В том числе, внимание этому направлению уделялось и в Советском союзе, но в 1936 году после постановления «О педологических извращениях в системе Наркомпросов» работу в этом направлении прекратили.

Изначально психометрическая теория была связана только с созданием и усовершенствованием тестов, но сейчас стало понятно, что эти модели могут использоваться не только в тестировании, но и в онлайн-образовании, в том числе – оценки контента, адаптивном обучении.

Классическая теория тестирования (КТТ) вводит три концепции:

* тестового балла (балл, который получает студент по результатам тестирования)
* истинного балла (среднее значение наблюдаемых оценок, полученных по бесконечному числу повторяющихся тестирований при использовании одного и того же теста. В действительности недостижим)
* ошибки измерения (разность между тестовым баллом и истинным баллом)

Ошибки измерения бывают двух типов:

Систематические – ошибки, порождаемые недостаточным качеством теста и/или неправильными условиями его проведения (например, плохая аппаратура).

Случайные – ошибки, порождаемые особенностями поведения испытуемых (усталость, невнимательность) и/или нарушениями процедуры тестирования (например, подсказки).

В рамках этого семинара в психометрической экспертизе заданий предлагается остановиться на анализе двух основных характеристик – трудности задания (коэффициенте решаемости) и дифференцирующей способности задания (дискриминативности).

## Цели психометрической экспертизы:

* Получить статистические характеристики заданий, оценить их качество с целью дальнейшего использования.
* Идентифицировать некачественные задания (например, задания с низкой дискриминативностью или задания с неработающими дистракторами (вариантами ответов) и т.д.). Такие задания необходимо переработать и повторно апробировать.
* Идентифицировать задания, имеющие проблемы функционирования, например, задания, демонстрирующие различное функционирование по отношению к различным группам испытуемых.

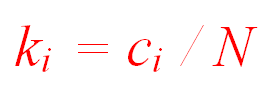
## Анализ измерительных инструментов курса:

1. Трудность
2. Различительная способность
3. Достаточность
4. Параллельность

Достаточно анализа и корректировок по трудности и различительной способности, чтобы составить хороший банк заданий.

### Трудность задания

Для дихотомических заданий определяется как доля испытуемых, получивших 1 балл за выполнение задания (выполнивших задание верно):



где *ki* – коэффициент решаемости задания *i*, *ci* – доля испытуемых, выполнивших задание *i* верно, *N* – общее количество испытуемых.

Трудность может изменяться от 0 до 1 и чем этот показатель выше, тем задание легче

|  |  |
| --- | --- |
| Значение | Интерпретация |
| Более 0,9 | Очень простое задание |
| 0,8-0,9 | Простое задание |
| 0,5 | Средняя трудность |
| Менее 0,2 | Трудное |
| Менее 0,1 | Очень сложное задание |

Задания, у которых коэффициент трудности равен 1, подлежат удалению, так как 100% студентов выполняют их правильно, и оно бесполезно для них. Время студентов тратится впустую. Если задание слишком сложное, оно перегружает студентов, поэтому задания с трудностью 0 тоже надо удалять или корректировать. Считается, что хорошее задание после первой попытки правильно решает около 50% студентов. Поэтому важно анализировать статистику первого прохождения тестов. Трудность меряется именно по результатам первого прохождения.

Начиная со второй попытки, количество решивших правильно должно увеличиваться, если этого не происходит, надо разбираться, что с этим заданием не так, возможно, студенты не могут решить задание и угадывают ответ.

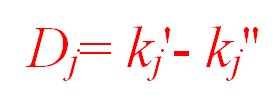
При этом если студенту все время предлагать только трудные задания, у него падает мотивация, поэтому значительная часть заданий должна быть не очень сложной.

### Дискриминативность

Различительная (дифференцирующая) способность или коэффициент дискриминативности.

Различительная (дифференцирующая) способность показывает, насколько хорошо задание разделяет испытуемых на сильных и слабых в рамках измеряемой дисциплины. Наиболее распространенной мерой различительной способности является коэффициент дискриминативности. Этот коэффициент учитывает коэффициенты решаемости заданий и рассчитывается следующим образом:

Пусть *kj'* - коэффициент решаемости *j*-го задания лучшей четвертью испытуемых (25% участников, имеющих высокие баллы), *kj''* - коэффициент решаемости *j*-го задания худшей четвертью испытуемых (25% участников, имеющих низкие баллы). Тогда



#### Интерпретация значений

Эти принципы интерпретации основываются на практическом опыте и не имеют никакой теоретической основы

* Если Dj ≥ 0,4, то задание функционирует удовлетворительно
* Если 0,30 ≤Dj ≤0,39, то требуется небольшая коррекция задания
* Если 0,20 ≤Dj ≤0,29, то задание нуждается в пересмотре
* Если Dj ≤0,19, то задание надо исключить из теста или полностью переделать.

Значение Dj может быть отрицательным (т.е. сильные студенты, которые обычно все решают правильно, с заданием не справились, а «двоечники», наоборот, преуспели). Тогда, во-первых, надо проверить, как задание занесено на платформу, возможно, ошибка в системе оценивания. Во-вторых, причина может быть в некорректной формулировке задания.

Как правило, принципу дискриминативности отвечает 1/6-1/5 от всех более или менее хорошо подготовленных заданий.

Считается, что на основании этих двух параметров можно составить хороший банк заданий и формировать варианты тестов.

Чем выше дискриминативность тестов и разнообразнее задания (по типу и педагогическим целям), тем меньше их нужно для адекватной оценки студента.

Оптимальным являются тесты, в которых не менее 15 хороших вопросов. Этого достаточно, чтобы оценить знания студента и нивелировать ошибки измерения.

Пример психометрической экспертизы:



Ответы к заданиям, отмеченные «+» – правильные. Из таблицы видно, что есть ответные опции, которые никто не выбирает. При корректировке заданий их надо заменить, так как сейчас они не работают.

Также внимание надо уделить вопросам №№ 4 и 5, так как у них отрицательная дискриминативность.

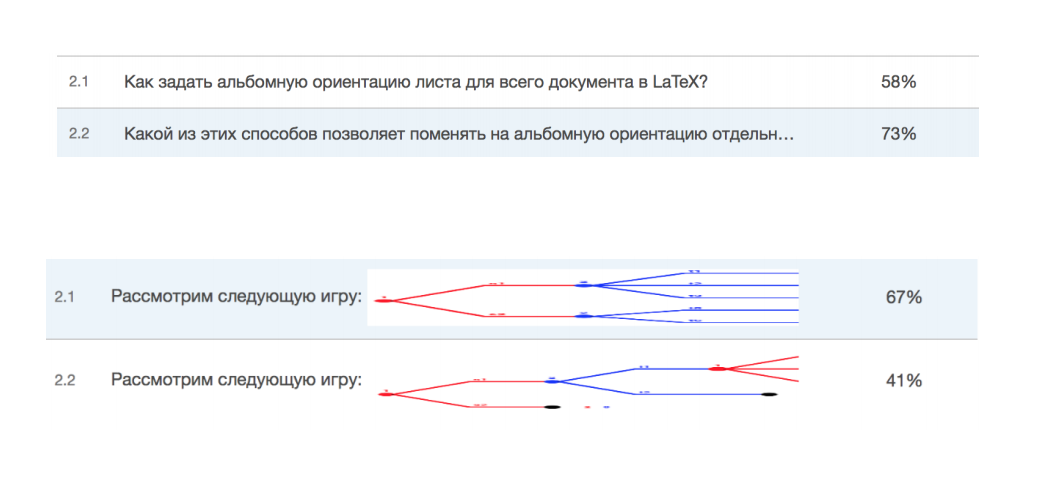
Из таблицы видно, что сложные задания (3 и 4) некоторые студенты пропускают, возможно, они не понятны студентам.

Также в пересмотре нуждаются задания с низкой дискриминативностью (3 и 6).

## О параллельности заданий и вариантов теста

Проблема в том, что большинство заданий, особенно для гуманитариев, получаются похожими по форме, но неравносложностными. Получается, что, решая разные варианты одного и того же теста, студенты зачастую получают задания разной трудности.

Это видно на предложенных примерах, в процентах указано количество решивших задание правильно:



Хотя изначально эксперты оценили эти задания как равносложностные, на практике (особенно во втором варианте) разница в трудности очень большая.

# Интерпретация результатов измерений, или как установить проходной балл

## Два похода к интерпретации

Нормативно ориентированный

• Задача – определение сравнительного места каждого из тестируемых в общей группе испытуемых

• Подход: введение норм, отражающих результаты выполнения теста. Нормы устанавливаются на большой репрезентативной выборке испытуемых.

Критериально ориентированный

• Задача – сопоставление учебных достижений каждого ученика с планируемым к усвоению объемом знаний, умений и навыков

• Подход: установление проходного балла, отделяющего тех, кто освоил проверяемый материал, от тех, кто не освоил

## Наиболее популярные методы интерпретации:

Сначала несколько традиционных, затем – хорошая

### Метод Angof

Один из самых широко используемых методов.

Существует традиционный и модифицированный метод, который был изменен с учётом недостатков традиционного метода.

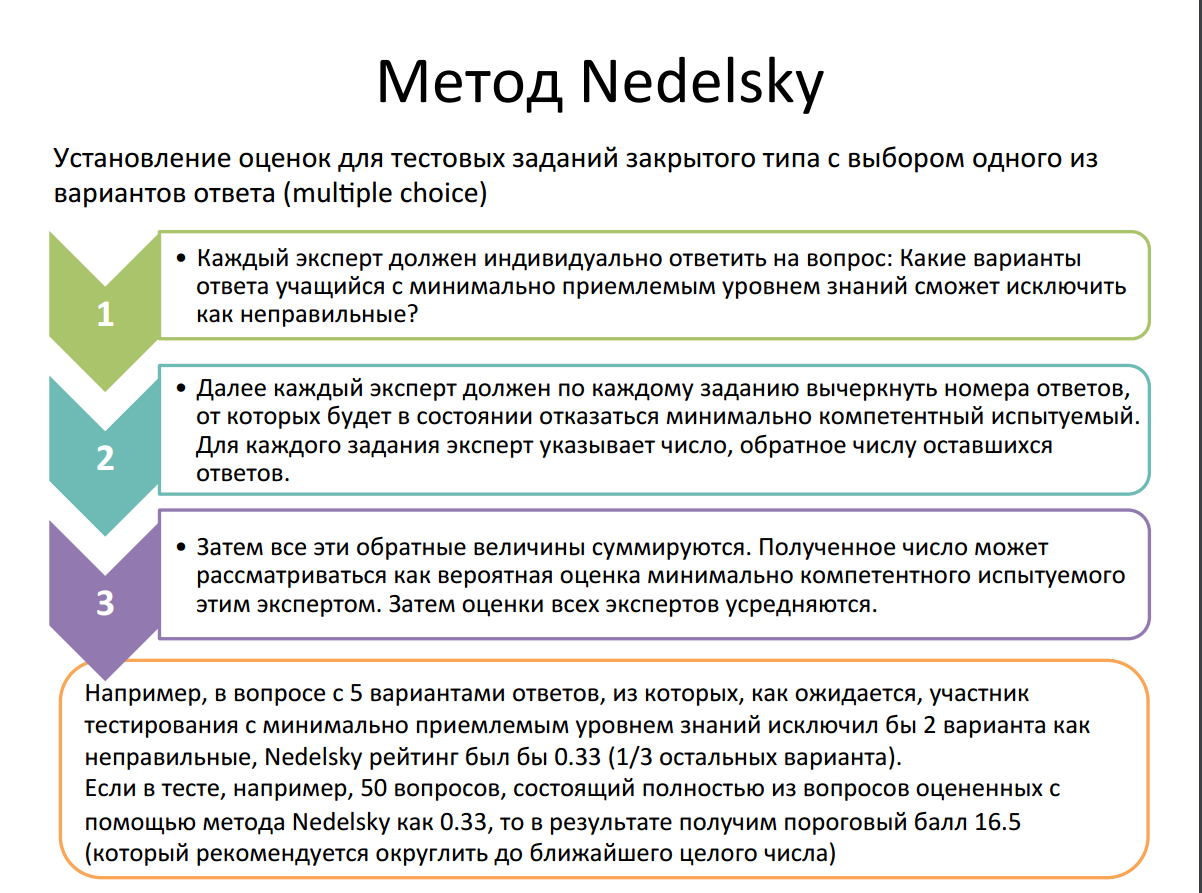
1. Экспертов просят представить себе группу минимально компетентных испытуемых
2. Для каждого задания нужно оценить вероятность того, что минимально компетентный испытуемый ответит на задание правильно
3. Данные вероятности складываются по каждому эксперту и усредняются по всем экспертам

### Модифицированный метод Angof

1. Экспертов просят представить себе группу минимально компетентных испытуемых
2. Эксперты проверяют каждое задание, чтобы ответить на вопрос: «Сможет ли кандидат с минимальным уровнем компетенций решить данное задание верно?»
3. Заданиям, на которые предполагается, что кандидат с минимальными компетенциями будет способен правильно ответить, присваивается вес 1, а заданиям, на которые предполагается, что такие кандидаты не смогут ответить правильно, присваивается значение 0.
4. Сумма оценок за вопросы будет равна первичному баллу, полученному испытуемым с минимально приемлемым уровнем компетенций.
5. Далее производится решение о пороговом балле путём усреднения оценок или с помощью медианы.

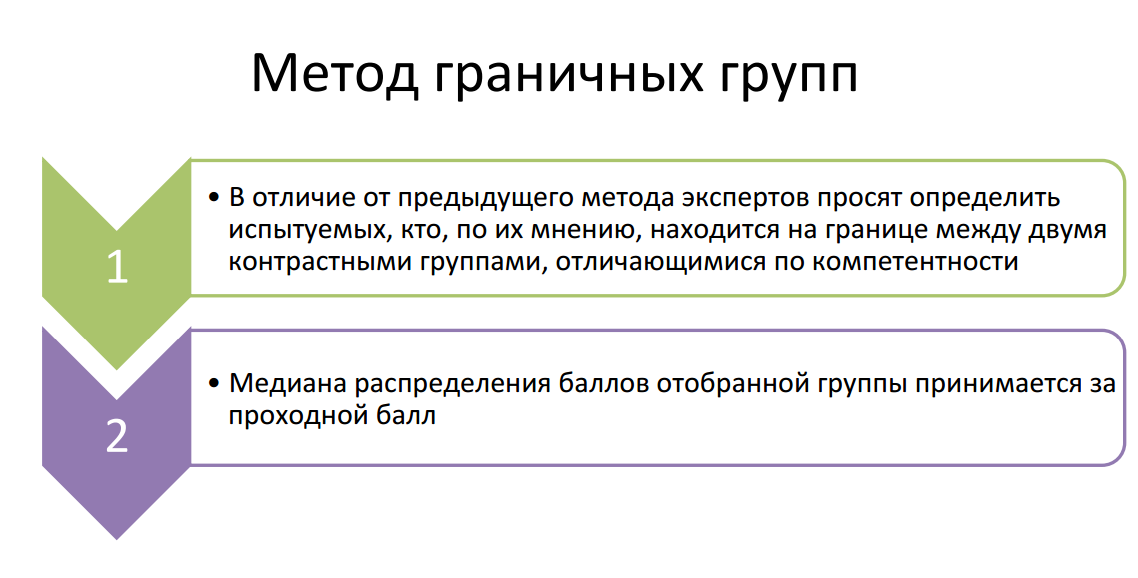
Обычно процедура модифицированного метода совершается в несколько этапов, чтобы позволить экспертам пересмотреть свои оценки. В один из этапов (или несколько) экспертам предоставляют данные об эмпирической трудности заданий.

Как правило, такой метод используется для вопросов закрытого типа с выбором одного из вариантов ответа.









При этом методе график будет выглядеть так:



Этот метод хуже метода контрастных групп, т.к. сильных и слабых учеников определить проще, чем «пограничных».

### Метод Bookmarks

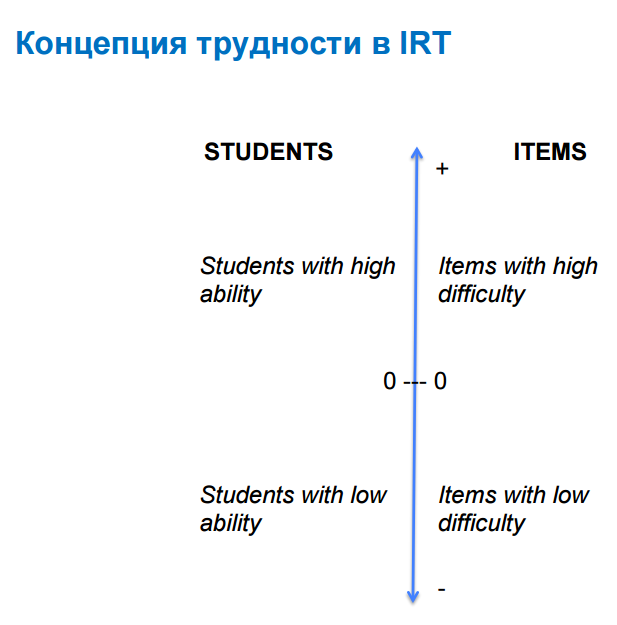
Применяется для установления пороговых оценок. Это хороший метод.

1. Экспертам предоставляется информация о трудности заданий (которая рассчитывается после апробации), и задания ранжируются по уровню трудности от наименее до наиболее трудного.
2. Эксперты выбирают самый трудный вопрос, на который с наибольшей вероятностью испытуемый с минимальными компетенциями мог бы ответить правильно. Этот вопрос помечается закладкой.
3. Эксперты обсуждают установленные закладки или пороговые точки.
4. Затем экспертам необходимо установить вторую закладку, которая может быть одной и той же или отличной от первой закладки. Вторая закладка определяет пороговый балл по результатам суммирования числа вопросов до закладки для каждого эксперта и усреднения по экспертам.

## Психометрический анализ трудности контента относительно способностей студентов

IRT – Item Response Theory – эта теория позволяет проанализировать сразу и трудность заданий, и уровень студентов. Данные можно отобразить на одном графике. Это современное направление в психометрике.

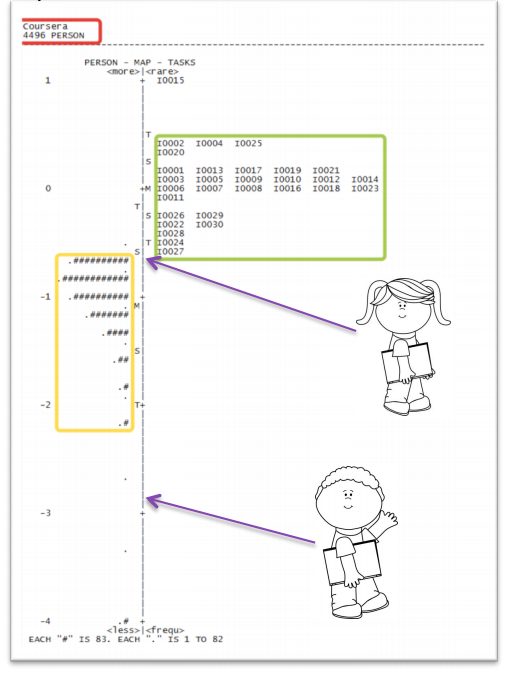
Для этого можно использовать [модель Раша](https://en.wikipedia.org/wiki/Rasch_model), которая одновременно отображает трудность задания и уровень подготовленности студентов и позволяет спрогнозировать результат решения задания (верно/неверно). Модель Раша часто используется для подбора оптимальных по трудности заданий в компьютерном адаптивном тестировании.



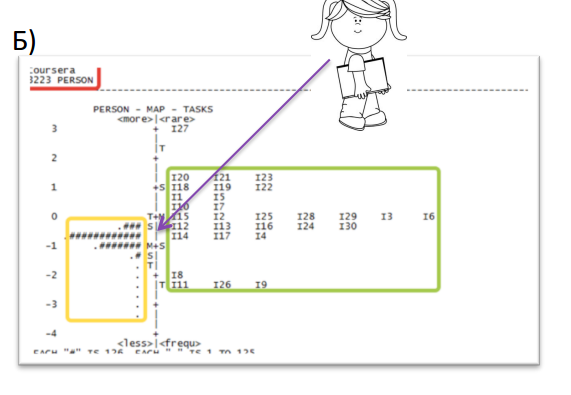
В идеале большинство показателей группируется в середине, примерно так:



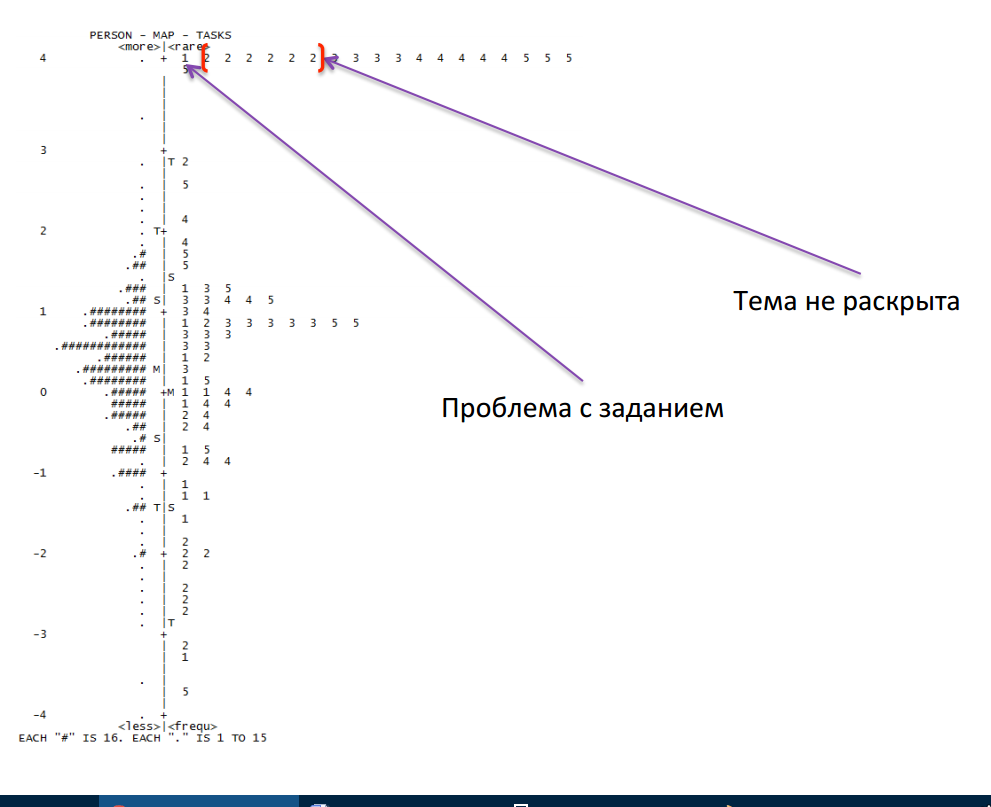
Пример:



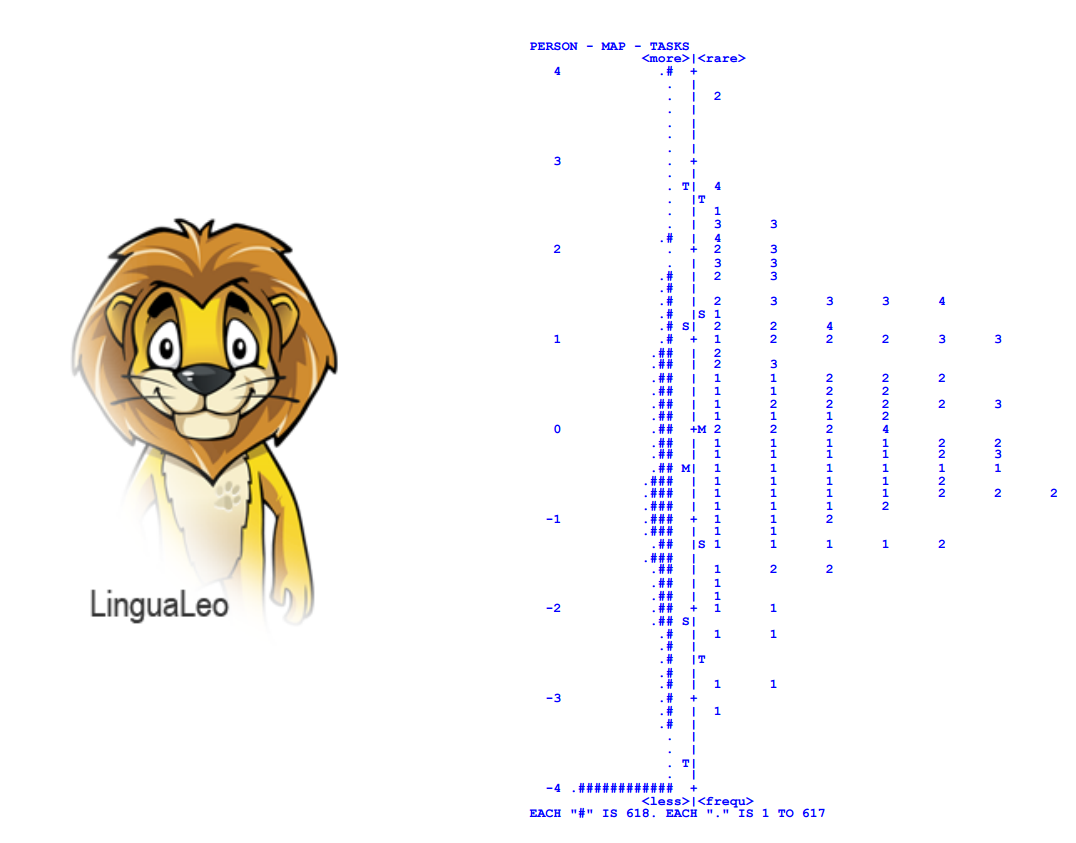
На представленном графике задания слишком сложные для студентов, так как находятся выше по оси. Сейчас на курсе обучается 4496 человек, через некоторое время их останется 3323, и модель придет в равновесие. Студенты, для которых задания были слишком сложными, покинули курс:



С помощью этой модели можно видеть проблемные задания. Задания, которые остаются внизу, ниже уровня студентов, слишком легкие, и их следует убрать. Верхняя строка показывает задания, которые оказались самыми сложными для студентов. Цифрами обозначены модули, из которых эти задания. Если из одного модуля слишком много нерешенных заданий, значит, есть основания предполагать, что тема плохо раскрыта или задания ей не соответствуют. Еще возможен вариант, что подготовлены слишком трудные задания для этой аудитории.



Эта модель позволяет анализировать и сами задания, например, в ЛингваЛео для первичного тестирования студентов применяется ветвящаяся структура заданий: если человек справляется с заданиями уровня 1, ему дают задания следующего уровня.



Из схемы видно, что большинство заданий уровня 1 находится среди «легких» или «средних», но и одно самых сложных заданий (вверху шкалы) было отнесено экспертами к заданиям 1 уровня.

## Анализ студентов. Прогнозирование. Психометрический прокторинг

Современные психометрические теории позволяют анализировать действия отдельных студентов, определять распространенные паттерны поведения, строить индивидуальные модели и помогать в прокторинге. Это, скорее, объяснительный прокторинг.

Каждому студенту присваивается идентификатор, по которому можно отследить его путь на курсе.



Каждая цифра – это задание, 1 – решено верно, 0 – ответ неверный. Верхняя линейка показывает действия среднестатистического ученика: сначала он делает все правильно, потом задания становятся сложнее, и самые сложные он без ошибок решить не может.

Вторая сверху линейка показывает паттерн невнимательного ученика: даже простые задания вызывают у него затруднения, хотя потом, несмотря на единичные ошибки, ученик показывает хорошие результаты.

При этом важно отметить, выше приведены только предположения: возможно, Миша не разобрался в теме из-за нехватки времени или еще из-за каких-то причин, а, может быть, проблема в изложении материала.

Также восприятие материала особенности решения задач могут зависеть от культурных особенностей, возраста, пола и т.д.

Анализ поведения студента во время курса позволяет предположить, какие результаты в этом курсе он будет получать в дальнейшем. Сравнение вероятностного и реального профиля студентов называется «психометрический прокторинг». Он не рассматривается как альтернатива традиционным методам прокторинга (то есть контроля во время тестов за тем, чтобы их выполняли честно), но если студент сильно отклоняется от предлагаемой для него модели, это может стать поводом пристального внимания к его действиям.

Подозрения может вызывать схема, где студент вначале допускает ошибки, но в конце при выходе на финальный тест, решает все правильно.

Например:

Начало теста: 001011

Вероятностный профиль продолжения: - - - - - - 10110101010000

Фактический профиль продолжения: - - - - - - 10110101111111

Также для анализа успеха учащихся и задания можно использовать модифицированную модель Эло ([Elo Rating System](https://en.wikipedia.org/wiki/Elo_rating_system)), которая используется для расчета рейтинга игроков в шахматы. Расчёт рейтинга игрока по результатам турнира основан на сравнении количества набранных им очков с ожидаемым, предсказанным на основе его рейтинга, количеством очков. Если оба игрока играют в шахматы впервые и не имеют рейтинга, соответственно, оба имеют равные силы и ожидаемое количество очков для них равно. Если по итогам турнира количество набранных очков оказывается больше, чем предсказанное значение, то рейтинг данного игрока возрастает. Если меньше – рейтинг уменьшается. (http://www.edutainme.ru/post/adaptive-magic/).

В случае с применением в психометрике один игрок – это ученик, а второй – задание, то есть ученик соревнуется с заданием. Таким образом, можно непрерывно уточнять рейтинг задания и учащихся.

# Адаптивное обучение и анализ данных. Доклад

Презентация: <https://yadi.sk/i/DLX4pk1pvASAo>

<https://yadi.sk/i/PHxyChdUvASGp>

Авторы работ по адаптивному обучению часто ссылаются на работы по адаптивному тестированию. Суть адаптивного тестирования в том, что из банка заданий собирается тест для конкретного студента с учетом уровня его подготовки на основании того, как он решал тесты в этом курсе ранее. Это позволяет сделать тест относительно коротким и проконтролировать, чтобы студент не получал ни заведомо слишком легких заданий, которые для него бесполезны и тратят его время впустую, ни слишком сложных, которые он решить не может. Проводится на основании модели [IRT](https://en.wikipedia.org/wiki/Item_response_theory) (Item response theory).

Изначально предполагалось, что адаптивное обучение – это создание образовательной траектории (подбор контента, оптимального по трудности) по аналогичному принципу для создания оптимальных условий обучения и повышения результативности.

В последнее время возникла мысль, что эта модель не очень хорошо работает, так как не учитывает опыт студента. Например, видно, что студенту было сложно решать какой-то тест, но было ли ему интересно?

Поэтому надо учитывать не только статические параметры, но и фидбэк от студента.

Способ анализа, о котором пойдет речь, называется i3d-анализ контента

I – интерес

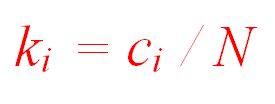
D – статистическая трудность

D – воспринимаемая трудность

D – восприятие воспринимаемой трудности («мне трудно, но мне это нравится», «мне легко, и меня это полностью устраивает»)

## Трудность-1 (статистическая)

Для дихотомических заданий (решил/не решил) определяется как доля испытуемых, получивших 1 балл за выполнение задания (выполнивших задание верно) и рассчитывается по формуле:



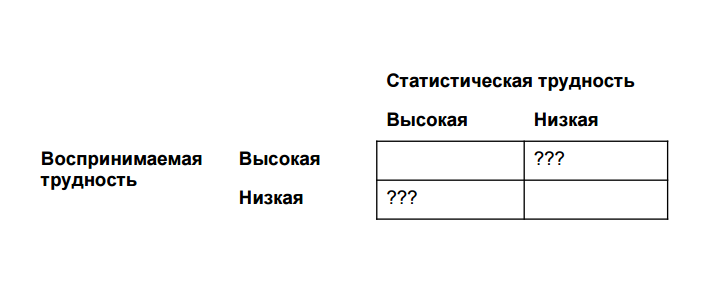
Где *ki* – коэффициент решаемости задания *i*, *ci* – доля испытуемых, выполнивших задание *i* верно, *N* – общее количество испытуемых.

Коэффициент решаемости может принимать значения от 0 до 1: чем его значение выше, тем легче тестовое задание. Задания с коэффициентом решаемости до 0,3 являются трудными, от 0,3 до 0,7 – средней трудности и выше 0,7 – легкими. Встречаются задания с коэффициентами 0 (их не решил никто) и 1 (их решили все). Обычно такие задания удаляют после апробации, так как они не различают испытуемых.

## Трудность-2 (воспринимаемая)

Можно говорить не только о трудности задания (лекции) как таковой (, но и воспринимаемой трудности. В двух словах, если задача трудная с первой попытки ее решает мало студентов, а воспринимая трудность – это фидбэк от студентов.

Например:



Если статистическая трудность высокая, а воспринимаемая – низкая, возможно, материал изложен так, что под мнимой легкостью изложения скрыто хорошее объяснение сложной темы. А еще бывает, что студент поначалу думает, что задача легкая, а решить ее не может.

## Трудность-3. Восприятие воспринимаемой трудности

А еще бывает восприятие воспринимаемой трудности, то есть то, как студент относится к сложности задания, интересно ли ему. Так, некоторым студентам нравится, когда задачи трудные, а другие предпочитаю задания полегче и постепенное увеличение нагрузки.

## Интересность контента

Поэтому следующий параметр – интерес. Этот параметр используется и при решении задач и при оценке видеоконтента.

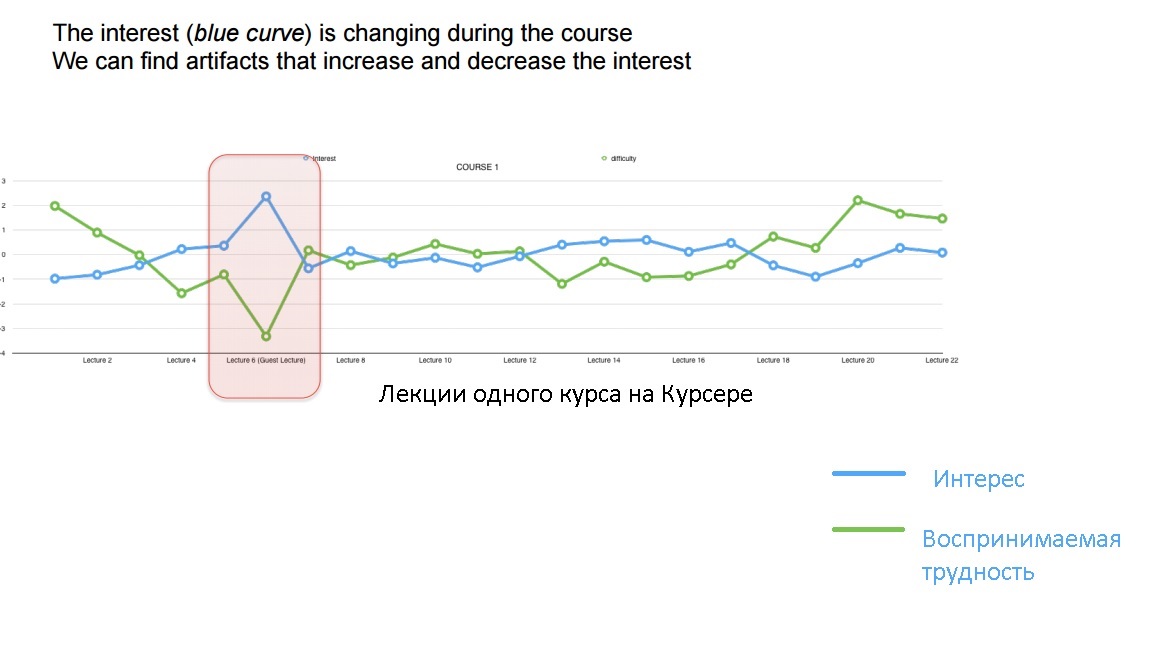
Гипотеза: если задача трудная, но студент считает ее интересной, это Challenge, а если трудность низкая, и задача интересная – это Edutainment.

# I3d на практике

Подобным образом можно оценивать видеолекции. Важно отметить, исследование проводилось на МООК-курсах ВШЭ на Курсере. Учащимся после просмотра лекции предлагалось по пятибалльной шкале оценить, интересно им было или нет. Аналогично замерялась воспринимаемая трудность. Понятно, что данные могут искажаться: студент может ответить «от балды» или пропустить вопрос (с первого раза отвечают 70-80% студентов). Или вся лекции была скучная, но в конце стало интересно, и в этот момент студенту задали вопрос, поэтому он ответил, что ему эта лекция показалась интересной, хотя это не совсем так.

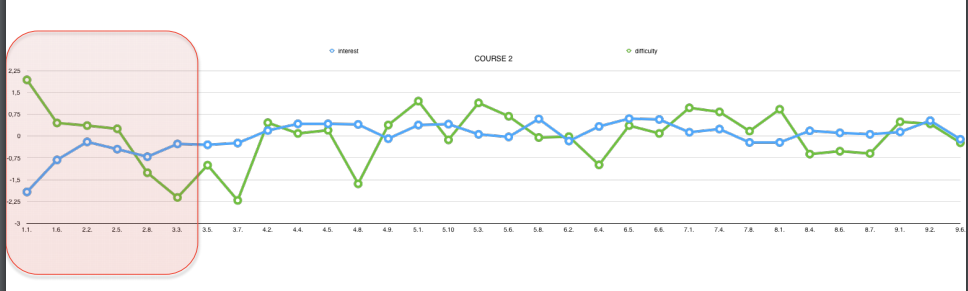
Результаты обрабатывались с помощью комплекса математических моделей, которые связывают ответы студентов, укладывая их на одну шкалу.

Есть еще другие способы измерения интереса (методики), например, с помощью «лайков» или опросов, но они не свободны от недостатков, на Курсере это пока не работает.



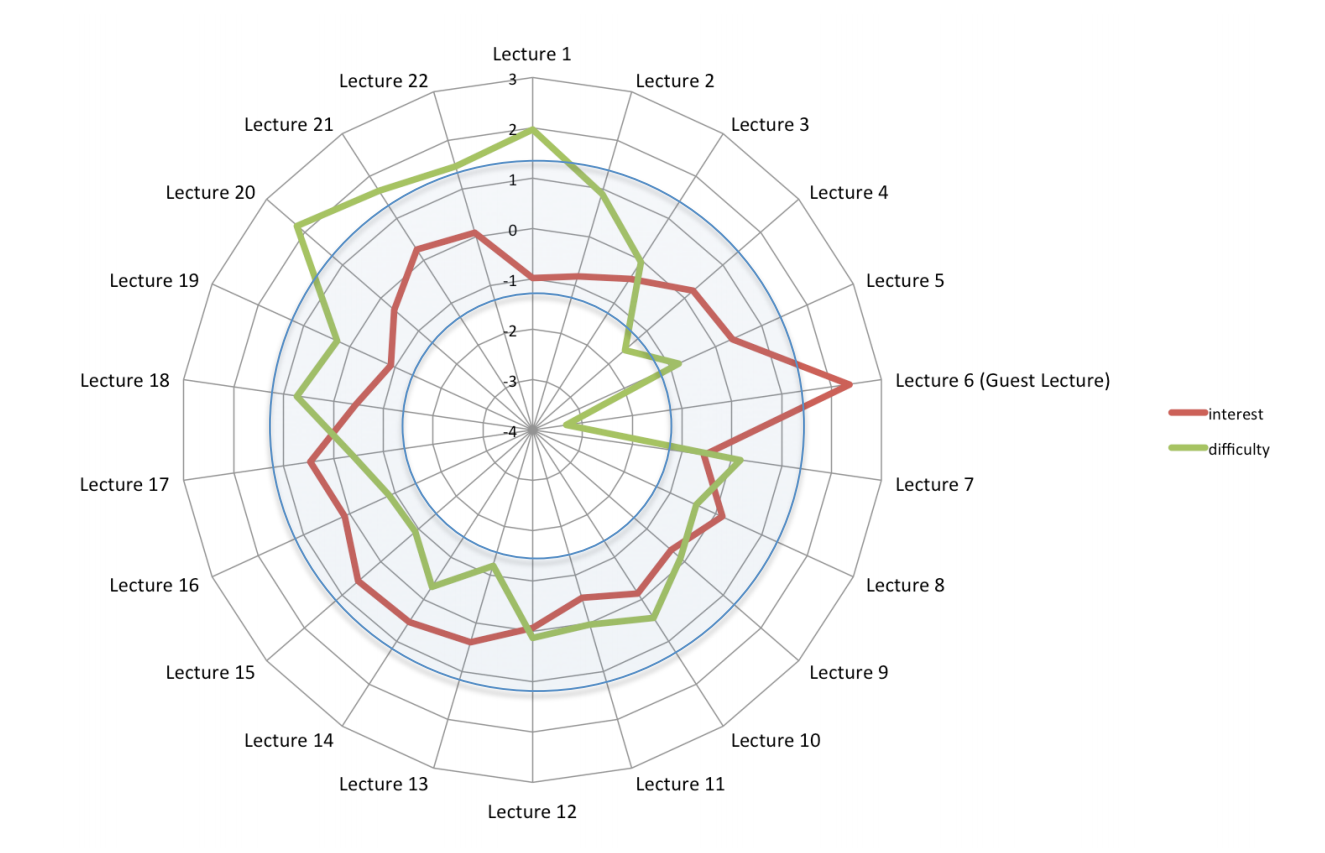
Пик интереса и максимально низкая воспринимая трудность приходятся на лекцию приглашенного преподавателя. Эти данные подлежат дальнейшей интерпретации: возможно, приглашенный преподаватель более харизматичный, а может быть, его тема выделялась.

Пример с другого курса:



Видно, что в начале курса студентам и сложно и неинтересно. Отзывы о курсе на форуме подтвердили это предположение, студенты писали, что «в начале курса хотелось уснуть».

[Интерес в обучении](https://iq.hse.ru/news/181243227.html) служит важным мотивационным рычагом, который позволяет управлять вниманием и способствует эффективному получению знаний. Экспертиза контента с точки зрения интереса студентов позволяет выявить артефакты, повышающие или снижающие привлекательность контента для студентов, а также помогает анализировать опыт взаимодействия студентов с контентом. (<http://www.therunet.com/articles/7462>). В итоге можно сбалансировать интересность и трудность курса.



Исследователям не стоит чрезвычайно увлекаться статистикой, потому что за этими данными зачастую не виден опыт студентов, который надо учитывать в алгоритмах адаптивного обучения. Это поможет и разработчикам и преподавателям.