Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

"Национальный исследовательский университет
"Высшая школа экономики"

**Институт образования**

**Магистерская программа «Измерения в психологии и образовании»**

**МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ**

на тему
**«Экспертно-ориентированный подход к разработке теста»**

Студентка группы № 701:

Илюшина Наталия Вадимовна

Научный руководитель:

Карданова Елена Юрьевна

Москва, 2014г.

Оглавление

[ВВЕДЕНИЕ 2](#_Toc390160427)

[1. ЭКСПЕРТНЫЕ ОЦЕНКИ В ТЕСТИРОВАНИИ 5](#_Toc390160428)

[1.1. МЕТОДОЛОГИЯ РАЗРАБОТКИ ТЕСТОВ 5](#_Toc390160429)

[1.1.1 Этапы создания теста 5](#_Toc390160430)

[1.1.2 Способы создания теста на основе имеющихся заданий 9](#_Toc390160431)

[1.2. ПРИМЕНЕНИЕ ЭКСПЕРТНЫХ ОЦЕНОК В ТЕСТИРОВАНИИ 12](#_Toc390160432)

[1.2.1 Способы использования экспертных оценок в тестировании 12](#_Toc390160433)

[1.2.2 Искажения, связанные с экспертными оценками 14](#_Toc390160434)

[1.2.3 Способы обнаружения искажений 18](#_Toc390160435)

[1.2.4 Выводы 23](#_Toc390160436)

[2. РАЗРАБОТКА МЕТОДОЛОГИИ ЭКСПЕРТНО-ОРИЕНТИРОВАННОГО ПОДХОДА К СОЗДАНИЮ ТЕСТОВ 24](#_Toc390160437)

[2.1. ОЦЕНКА ГОТОВНОСТИ К ОБУЧЕНИЮ НА ИНЖЕНЕРНЫХ СПЕЦИАЛЬНОСТЯХ ВУЗОВ 24](#_Toc390160438)

[2.1.1 Сравнительный анализ вступительных экзаменов по математике на технические специальности вузов в странах БРИК 27](#_Toc390160439)

[2.1.1. Выводы 30](#_Toc390160440)

[2.2. ПРОГРАММА ИССЛЕДОВАНИЯ 30](#_Toc390160441)

[2.2.1 Методологический раздел программы исследования 30](#_Toc390160442)

[2.2.2 Процедурный раздел программы исследования 32](#_Toc390160443)

[2.3. ОБНАРУЖЕНИЕ ИСКАЖЕНИЙ В ЭКСПЕРТНЫХ ОЦЕНКАХ 34](#_Toc390160444)

[2.3.1 Демонстрация искажений в симуляционных исследованиях 34](#_Toc390160445)

[2.3.1 Анализ искажений в реальных данных 41](#_Toc390160446)

[2.3.3 Выводы 48](#_Toc390160447)

[2.4. РАЗРАБОТКА ИНСТРУМЕНТА 50](#_Toc390160448)

[2.4.1 Методология разработки теста с помощью экспертно-ориентированного подхода 50](#_Toc390160449)

[2.4.2 Выводы 54](#_Toc390160450)

[ЗАКЛЮЧЕНИЕ 55](#_Toc390160451)

[Список литературы 57](#_Toc390160452)

[*Приложение 1*. Образец инструкции и бланка для оценки заданий 59](#_Toc390160453)

[*Приложение 2*. Образец инструкции и бланка для оценки тем 62](#_Toc390160454)

# ВВЕДЕНИЕ

В настоящее время разработке тестов уделяется достаточно много внимания. Все больше ресурсов тратится на то, чтобы создать качественную методику, стандарты качества постоянно корректируются, требования к тестам возрастают. Вместе с этим возрастают и затраты на создание одного теста, и очень часто эти затраты не оправдываются целями исследования.

Наличие банка заданий является одним из факторов, который может существенно уменьшить цену разработки теста, так как задания являются одной из главных составляющих теста, на которую уходит большая часть сил разработчика.

В системе образовательного тестирования такое может встречаться довольно часто: кто-то уже разрабатывал тестовые задания в данной предметной области и нередко это неплохие задания с хорошими психометрическими свойствами. Однако возникает вопрос – как создать хороший инструмент на основании имеющихся заданий? Какие задания включить в тест, чтобы он был валиден по содержанию, не требуя при этом очень много времени на выполнение?

Нередко ответы на эти вопросы находятся достаточно быстро и без опоры на теорию, в результате чего разработчик получает большой набор заданий, который утомляет респондента и искажает информацию о его реальных способностях. Другая крайность - наоборот, стремясь сделать тест компактным, разработчик жертвует качеством и оставляет какие-то области полностью не охваченными. При этом, он по-прежнему имеет в своем распоряжении большой банк хороших заданий, но не умеет ими грамотно распоряжаться.

Так как наличие банка заданий может значительно снизить цену разработки теста, важно иметь технологию, которая позволит сделать новый тест на базе существующих заданий. Мы предполагаем, что в данной ситуации самым оптимальным решением будет использование экспертных оценок для отбора заданий. При этом, метод экспертных оценок обладает рядом трудностей, которые не свойственны стандартным методам разработки теста.

Данная работа посвящена разработке экспертно-ориентированного подхода к созданию тестов, его возможностей и ограничений. Целью данной работы, таким образом, является формулировка основных методологических принципов, лежащих в основе экспертно-ориентированного подхода к разработке тестов.

*Объект исследования*: экспертные оценки в тестировании.

*Предмет исследования*: применение экспертных оценок для создания теста.

*Общая гипотеза исследования*: экспертные оценки можно использовать для создания нового теста на базе существующих.

В работе использован метод многопараметрического анализа результатов экспертной оценки, проанализированный с точки зрения экспертных оценок заданий, а не ответов испытуемых. Кроме этого, был использован метод опроса экспертов с целью проверки функционирования методов анализа на небольшой выборке.

# 1. ЭКСПЕРТНЫЕ ОЦЕНКИ В ТЕСТИРОВАНИИ

# МЕТОДОЛОГИЯ РАЗРАБОТКИ ТЕСТОВ

## 1.1.1 Этапы создания теста

Разработка методики оценивания практически в любой сфере обязательно должна проходить через ряд этапов. Это позволяет достичь определенного качества создаваемой методики. Этапы не являются строго упорядоченными, и нередко они применяются в другом порядке, но, в той или иной степени, создание любой методики проходит через них.

При разработке теста принято опираться на схему, состоящую из 12 шагов (Downing, 2006):

1. Составление общего плана теста
2. Определение содержания
3. Составление спецификации теста.
4. Разработка заданий.
5. Разработка дизайна теста.
6. Выпуск теста.
7. Проведение теста.
8. Проверка психометрических характеристик.
9. Установление пороговых значений (проходных баллов).
10. Предоставление результатов.
11. Формирование банка заданий.
12. Составление технического отчета по тесту.

Некоторые из этих этапов являются специфичными для тестов достижений (например, установление проходных баллов), однако в большинстве случаев каждый из этих этапов в той или иной степени должен быть пройден при разработке хорошей методики.

Опишем эти этапы подробнее.

1. *Составление общего плана теста.*

В этот этап входит разработка основной информации о тесте, сюда же включается проработка методов оценки валидности, основные возможные интерпретации результатов, психометрические модели, контроль качества и т.д. Самый главный вопрос, который нужно решить на первом этапе разработки теста – это вопрос о том, какой конструкт нужно измерять. Также важно определиться, какие типы вопросов будут использоваться, как будут интерпретироваться результаты тестирования, как будет проводиться тестирование (в бланковой или компьютерной форме).

Этот этап является одним из самых главных этапов при разработке теста, так как именно от него зависят все дальнейшие шаги и качество итогового инструмента. Точное определение целей тестирования и его основных параметров дает возможность ясно представить результат своей работы и избежать многих возможных ошибок.

1. *Определение содержания*

Этот этап также является ответом на основополагающий вопрос любой тестовой методики: какой конструкт должен быть измерен? Это особенно важный вопрос для тестов достижений и педагогических тестов, который определяет возможную интерпретацию тестовых баллов, однако и психологические методики также нуждаются в определении содержательной базы. Если содержание теста определено неточно, все дальнейшие этапы работы над тестом будут сделаны неверно.

1. *Составление спецификации теста*

Создание спецификации теста подразумевает определение не только базовых параметров будущего теста, что было сделано на первом этапе, но и достаточно точное описание всего теста, подразумевающее полную проработку концепции инструмента. Основными частями спецификации являются: форма тестовых заданий, их число и их точный тип (если это задание с выбором одного правильного ответа, то сколько предполагается вариантов ответов и т.д.), таксономия для оценки когнитивного уровня (например, таксономия Блума), наличие или отсутствие визуальных стимулов в тестовых материалах, правила оценивания выполнения заданий, правила интерпретации результатов, затраты времени на каждый вопрос. В более простом варианте спецификация может содержать только количество тестовых вопросов по каждой содержательной области и уровню таксономии. Детализированность спецификации связана с тем, какие последствия будут иметь результаты тестирования. В целом, третий этап является результатом обобщения решений, принятых на этапах 1 и 2, и по его результатам будет составлен итоговый план будущего теста. Спецификация впоследствии может подвергаться изменениям, однако ее основные пункты остаются неизменными, и на них базируются все следующие этапы создания теста.

На этом этапе могут использоваться экспертные оценки - при переходе от списка необходимых содержательных элементов, определенных на этапе 2, к спецификации. Экспертные оценки в данном случае используются для оценки степени важности темы и уровня ее освоения, чтобы впоследствии можно было распределить тестовые вопросы в соответствии с рекомендациями.

1. *Разработка заданий*

Разработка заданий является очень важным этапом создания теста. При разработке заданий необходимо учитывать ряд параметров, которые могут повлиять на результаты тестирования, например, формулировки тестовых вопросов, вариантов ответа и т.д. При разработке тестовых вопросов может возникнуть ряд трудностей, делающих процесс разработки теста еще более трудоемким. Эти сложности будут описаны в следующей главе.

1. *Составление дизайна теста*

Следующим важным шагом после разработки заданий становится разработка дизайна теста. Хорошие задания не всегда складываются в хороший тест, и валидность окончательного инструмента зависит от того, как были скомпонованы полученные вопросы. Общий дизайн теста, выполненный на первом этапе разработки, представляет собой основу для деятельности на этом этапе, так как дизайн теста зависит от его целей и возможностей интерпретации. Большое влияние также оказывает и форма проведения тестирования (к примеру, будет ли это бумажная или компьютерная форма, или будет ли тестирование обычным или адаптивным и т.д.). На этом же этапе происходит корректировка заданий с целью облегчения чтения и уменьшения когнитивных усилий, не связанных с измеряемым конструктом. Кроме этого, необходимо сбалансировать ключи (оптимальным является равномерное распределение номеров правильных ответов).

1. *Выпуск теста*

В этот этап входят такие действия как распечатка теста при бланковой форме тестирования, или подготовка компьютерной версии. На этом этапе обычно производится последняя проверка тестовых материалов. Также проверяется качество печати и другие технические параметры, которые, при некачественном исполнении, также могут повлиять на результаты тестирования.

1. *Проведение теста*
2. *Проверка психометрических характеристик*
3. *Установление пороговых значений*
4. *Предоставление результатов*
5. *Формирование банка заданий*
6. *Составление технического отчета по тесту*

В проведении теста (этап 7) и последующей работе с ним существует множество тонкостей, в том числе, административных, но в рамках данного обзора мы не будем на них останавливаться, так как это не связано с предметом настоящего исследования.

Одним из самых трудоемких этапов работы является работа над созданием и совершенствованием тестовых вопросов, так как они составляют основу для всего теста.

## 1.1.2 Способы создания теста на основе имеющихся заданий

При наличии уже имеющегося банка заданий остается открытым вопрос о том, как «собрать» из этих заданий качественный тест. В данном случае стоит руководствоваться спецификацией теста, которая должна исчерпывающе описывать тест. Однако, в некоторых случаях (например, в психологических опросниках, измеряющих один конструкт) спецификация не всегда поможет отобрать нужное количество правильных вопросов, а предъявлять респондентам весь имеющийся банк заданий не представляется оптимальным.

Существуют две разные ситуации, в которых необходимо решать проблему отбора заданий в тест: это отбор из откалиброванного банка (например, изначально созданного под один тест) и комбинация нового теста из заданий других, уже существующих тестов.

**Отбор заданий из откалиброванного банка** часто осуществляется на основании спецификации, по содержанию и психометрическим характеристикам заданий. Однако, при наличии объемного банка это достаточно трудно сделать не-эксперту, даже если в заданиях прописаны темы.

Одним из наиболее продуктивных способов отбора заданий из большого банка является адаптивное компьютерное тестирование, позволяющее подбирать индивидуальный набор заданий для каждого респондента, способный обеспечить высокое качество измерений. Однако, разработка адаптивного алгоритма является достаточно затратной, и это не всегда оказывается оправданным, особенно в рамках психологических методик.

Для того, чтобы использовать преимущества адаптивного алгоритма, потратив меньше средств на разработку, в таких ситуациях используют создание специализированного ПО, направленного на анализ функционирования банка заданий и отбор комбинаций заданий, наиболее удачно функционирующих вместе.

Существуют работы, в которых указывается, что отбор заданий в тест достижений на основании только их дискриминативности и трудности часто ошибочен (Liu et al., 2012). Основанием для такого суждения является предположение о том, что функционирование теста не складывается напрямую из функционирования тестовых заданий, и различные комбинации хороших заданий могут по-разному сказываться на функционировании теста в целом. Авторы предлагают решать эту проблему с помощью компьютерного алгоритма, анализирующего ответы респондентов. Итогом их работы является разработка двухэтапного метода компоновки качественно откалиброванных заданий в новый тест. Первым этапом в данном случае является анализ вопросов из архивных тестов. После этого производится отбор тех заданий, которые вместе функционируют лучше всего.

Такой алгоритм при отборе заданий может оказаться востребованным и полезным при наличии достаточно большого банка решенных тестов, который позволит проанализировать совместное функционирование вопросов. В противном случае авторы говорят о том, что стоит использовать экспертные оценки для оценивания параметров заданий. В данном случае авторы не говорят о работе с какими-либо критериями, кроме примерной трудности и дискриминативности заданий.

Похожую методику используют и в другом исследовании (Hwang et al., 2006). В данном случае также речь идет о разработке алгоритма, позволяющего отбирать задания из большого откалиброванного банка. Авторы акцентируют внимание на масштабных инструментах (в качестве примера рассматривается GRE), которые проводятся много лет подряд, в связи с чем требуется постоянная генерация новых вариантов.

В исследованиях, предусматривающих разработку инструмента на базе существующих вопросов, разработка ПО становится все более востребованной. В данном случае, вероятно, это действительно имеет смысл, так как в ситуациях масштабного тестирования любые другие способы конструирования большого количества вариантов будут более затратными.

Вторая ситуация (**отбор заданий из разных тестов**) чаще встречается в психологических опросниках, чем в тестах достижений. В таких случаях, чаще всего, не имеется откалиброванного банка заданий из разных методик. Максимум, что можно сделать – это провести тестирование всех задействованных методик на одной выборке, однако это связано с большими затратами.

Например, в исследовании, посвященном нарушениям сна у детей, создавался новый опросник на базе двух существующих (Biggs et al., 2012). Авторы использовали два опросника (общее число вопросов составило 111) на выборке 628 детей, после чего применили к полученным данным эксплораторный факторный анализ. Окончательный вариант теста состоял из 33 вопросов, распределенных по 8 факторам. Для проверки надежности методики исследователи проводили ретест на 109 случайных пациентах, промежуток между тестированиями составил от 18 месяцев. Большинство факторов показали хороший коэффициент надежности (больше 0,47). При этом, в обсуждении указывалось, что при создании итоговой версии опросника авторы руководствовались не только результатами факторных анализов (их проводилось достаточно много), но и своими медицинскими знаниями.

Так как авторы не применяли экспертных оценок, то в данном случае трудно заключить, был ли окончательный отбор заданий верным с точки зрения содержания, и действительно ли отобранные авторами факторы измеряют то, что требуется (авторы указывают, что при адаптации на других языках количество факторов менялось). В связи с этим, возникает вопрос о пригодности факторного анализа для конструирования опросника. Этот метод, несомненно, достаточно прост и недорог, однако им нельзя ограничиваться при создании опросника. Если принимать решение о включении вопроса в тест только на основании результатов факторного анализа, то велика вероятность того, что полученный инструмент будет неверно функционировать, пострадает его валидность, либо будут охвачены не все необходимые темы. Так или иначе, на определенном этапе отбора вопросов оказывается задействованным экспертное мнение (даже если это только авторское мнение). Однако важно понимать, что использование только мнения автора в качестве экспертного также ставит под сомнение содержательную валидность методики, хотя в данном случае автор приводит достаточно обширные доказательства своей точки зрения по каждому фактору и по спорным вопросам.

# 1.2. ПРИМЕНЕНИЕ ЭКСПЕРТНЫХ ОЦЕНОК В ТЕСТИРОВАНИИ

## 1.2.1 Способы использования экспертных оценок в тестировании

По словам А.И. Орлова: «Методы экспертных оценок - это методы организации работы со специалистами-экспертами и обработки мнений экспертов» (Орлов, 2002). То есть, в самом определении метода не обозначается узконаправленная область его применения. Однако, в тестировании и разработке тестов экспертные оценки используются, чаще всего, на этапе оценивания респондентов. Одним из самых привычных способов использования экспертных оценок в России – это оценка результатов части С в ЕГЭ. Основным параметром для отнесения задания к категории «экспертное оценивание» является политомический характер оценки, чаще всего это относится к заданиям со свободно конструируемым ответом.

Однако, экспертные оценки используются не только для анализа ответов респондентов на открытые задания. Другим популярным методом использования экспертных оценок является выставление пороговых баллов в рамках классической теории тестирования. Для этого существует несколько методов, одним из которых является модифицированный метод Ангофф (Crocker, Algina, 2008). Кроме этого, известно использование экспертных оценок для определения DIF, содержательной валидности (Zhu, 1998). При этом, нужно учитывать, что не каждая область, в которой используют экспертные оценки, действительно в этом нуждается, и не всегда использование экспертов является необходимым или самым оптимальным решением, так как экспертные оценки имеют как свои преимущества, так и недостатки, о которых будет сказано позже.

В частности, экспертные оценки используются для выравнивания вариантов теста разных лет (Bramley, 2005). В данном случае использование экспертных оценок уменьшает затраты на выравнивание, так как, по мнению автора, отпадает необходимость в общих заданиях или испытуемых. В данном исследовании эксперты ранжировали работы респондентов от лучшей к худшей (не зная их баллов), и результаты показали, что такой метод позволяет сопоставлять баллы 2003 и 2004 годов тестирования.

Кроме этого, известно использование экспертных оценок для оценки трудности заданий, что близко к проблеме нашего исследования (Hamp-Lyons et al., 1994). По результатам этого исследования было выяснено, что эксперты дают согласованную оценку трудности заданий (на примере теста по языку MELAB) без предварительного обучения оцениванию заданий. Эксперты показали высокие корреляции и значимую связь с результатами тестирования. Авторы статьи не применяли психометрических методов для анализа деятельности экспертов, но их анализ показал, что, несмотря на высокую согласованность, эксперты неверно оценивают трудность задания. Предполагалось, что с возрастанием трудности задания полученный балл будет снижаться, однако в полученных результатах связь была не обратной, а прямой. Причинами этому могут быть маленькая выборка исследования (4 эксперта), принципы оценивания (для оценки трудности задания брали сумму баллов экспертов). Другие причины могли бы быть выявлены в ходе психометрического анализа. Авторы предлагают несколько возможных причин таких результатов, связанных, в основном, с оцениванием студенческих работ, однако в дальнейшей работе их необходимо учитывать и проверять.

Экспертные оценки в тестировании также используются для сокращения опросников, преимущественно на основании суждения о репрезентативности набора заданий исследуемому конструкту. При этом, применяются различные техники использования уже полученных экспертных оценок (Hardesty, 2004). Авторы статьи выделяют три основных используемых метода: использование суммарного балла (чаще всего, 3 – задание репрезентативно, 1 – не репрезентативно) для отбора заданий, подсчет экспертов, которые оценили задание как полностью репрезентативное и подсчет экспертов, которые оценили задание как полностью нерепрезентативное.

Одной из особенностей экспертной оценки как метода оценивания респондентов или заданий является большая вероятность возникновения побочных эффектов. В настоящее время существуют разные классификации этих искажений, однако, в любом случае, такие ошибки вносят сложности в экспертное оценивание.

С точки зрения оценивания экспертных оценок ответов респондентов существует достаточно большой опыт, включающий в себя, в том числе, методы многопараметрического анализа IRT, дающие исчерпывающие результаты качества экспертной оценки. Эти методы позволяют не только оценить степень согласованности экспертных оценок, но и математически выявить эффекты и искажения, возникающие в процессе оценки.

## 1.2.2 Искажения, связанные с экспертными оценками

Так как метод экспертных оценок предполагает непосредственное оценивание заданий экспертом по открытым критериям, высока вероятность возникновения искажений, связанных с личностью эксперта. Эти искажения классифицируются на следующие классические виды (Myford, 2003):

1. Эффект строгости/снисходительности
2. Эффект центральной тенденции
3. Эффект гало
4. Эффект ограничения спектра

Существуют также и более дробные классификации возможных эффектов. В рамках данной работы мы также рассмотрим эффект случайности. Стоит отметить, что эффекты строгости/снисходительности и центральной тенденции представляют собой частные случаи эффекта ограничения спектра. Рассмотрим представленные выше эффекты более подробно, согласно обзорной статье Кэрол Майфорд (Myford, 2003).

**Эффект строгости/снисходительности**

Существует множество определений эффекта строгости / снисходительности, однако можно говорить о том, что, в целом, они схожи между собой. Самое первое определение эффекта снисходительности принадлежит Книланду и звучит как тенденция эксперта "ставить оценку выше среднего по используемой шкале" (Kneeland, 1929, p. 356). Позднее Гилфорд добавил к этому определению свойство, согласно которому снисходительный эксперт будет снисходительным вне зависимости от черты, которую он оценивает. Позднее к характеристике снисходительности начали относиться в большей степени как к личностной характеристике эксперта. Этот аспект впоследствии рассматривался с разных сторон, в том числе, экспертам приписывалось нежелание ставить низкие баллы знакомым, и в данном случае это не было личностной чертой.

Для того, чтобы помочь экспертам избежать этого эффекта, применяются разные техники, в том числе:

* Четко определить категории используемых шкал
* Обозначить по несколько категорий с положительной и отрицательной стороны
* Ознакомить экспертов с возможностью наличия такого эффекта
* Заранее ограничить количество респондентов на каждую категорию оценки или попросить ранжировать респондентов
* Использовать несколько экспертов для оценки одного респондента
* Использовать статистические методы для сопоставления баллов у строгих и не строгих экспертов

**Эффект центральной тенденции**

Чаще всего в определении эффекта центральной тенденции говорится о том, что эксперт чрезмерно часто использует средние баллы на шкале, избегая крайних категорий. Такое поведение часто считается типичным в ситуации оценивания малознакомых респондентов, однако оно ведет к снижению надежности и валидности за счет сужения спектра оценок.

Диагностика эффекта центральной тенденции чаще всего происходит посредством сравнения среднего балла эксперта с центром шкалы.

Для уменьшения эффекта центральной тенденции применяется ряд стратегий:

* Четко определить категории используемых шкал
* Ввести более четкие различия между категориями шкал, особенно расположенными ближе к центру
* Не делать крайние категории слишком нетипичными
* Рассказать экспертам о наличии эффекта
* Заранее ограничить количество респондентов на каждую категорию оценки или попросить ранжировать респондентов

**Эффект гало**

Считается, что эффект Гало является самым изученным из всех классических эффектов. Общим определением этого эффекта можно считать следующее: "Существует внутренняя тенденция оценивать за общие заслуги, когда необходимо оценить по какому-то конкретному показателю, и позволять общему впечатлению об индивидууме повлиять на его результат по шкале" (Cooper, 1981, стр. 218). Впоследствии было принято много определений этого эффекта, но, в целом, все они подразумевают наличие связи между впечатлением от респондента и оценкой за задание, не обусловленной оцениваемыми характеристиками. При этом, также существует точка зрения, говорящая о том, что существует не только "ложное" гало, но и "истинное", которое оценивает нужную характеристику по связанным с ней чертам респондента.

Для уменьшения эффекта гало применяются различные техники:

* Четко определить каждую характеристику и разделить характеристики между собой
* Если критериев несколько, периодически менять порядок следования категорий
* Ознакомить экспертов с возможностью наличия такого эффекта
* Использовать процедуру, когда один эксперт оценивает всех респондентов по одной характеристике в один промежуток времени, а потом переходит к другой
* Уменьшить количество критериев, либо использовать несколько экспертов для оценки одного респондента

**Эффект ограничения спектра**

Эффект ограничения спектра связан с эффектом центральной тенденции. Различия между ними заключаются в том, что ограничение спектра подразумевает не только выбор средних категорий, но и предпочтение высоких или низких категорий.

При этом, важно учитывать, что наличие такого эффекта не всегда является искажением - так, существует исследование, в котором демонстрируется, что японские и тайванские студенты используют середину шкалы чаще, чем американские и канадские, в то время как американские студенты используют крайние категории чаще других (Chen et al., 1995). Таким образом, эффект ограничения спектра может быть обусловлен не только личностными, но и социально-культурными особенностями эксперта, и такую возможность необходимо учитывать, если будет обнаружено ограничение спектра на групповом уровне.

Для определения наличия этого эффекта исследуют дисперсию баллов, полученных всеми респондентами по одному критерию (чем меньше дисперсия, тем больше эффект), распределение частотности баллов по критерию (остроконечное распределение может свидетельствовать о наличии эффекта) и анализируют несоответствие между респондентом, экспертом и критерием.

Для уменьшения эффекта используются те же стратегии, что и при работе с эффектом центральной тенденции.

**Эффект случайности**

Под эффектом случайности подразумевается тенденция эксперта применять один или несколько критериев в манере, отличной от других экспертов. Такой эксперт непостоянен в использовании шкалы и демонстрирует более случайное распределение, чем должно быть на основании его оценок. Причиной этого может быть неверная интерпретация шкалы, которая ведет к ее отличному от других использованию.

Для уменьшения эффекта случайности применяются следующие стратегии:

* Четко определить категории используемых шкал
* Ознакомить экспертов с возможностью наличия такого эффекта
* Использовать несколько экспертов для оценки одного респондента

В рамках данной работы мы будем анализировать обнаружение эффектов строгости / снисходительности, случайности и центральной тенденции.

## 1.2.3 Способы обнаружения искажений

В предыдущем разделе мы обозначили основные эффекты, которым могут быть подвержены эксперты при оценке респондентов, а также, в некоторых из них, обозначили принятые методы обнаружения и предотвращения. Однако, в большинстве случаев, применяются различные методы для разных эффектов, которые не позволяют определить (или заподозрить) наличие какого-либо эффекта по одному анализу. Для того, чтобы оптимизировать и упростить анализ деятельности экспертов применяется многопараметрический анализ в рамках моделей Раша (Myford, 2003). Такой анализ позволяет ответить на ряд важных вопросов: отличаются ли эксперты по уровню строгости, насколько эффективно эксперты разграничивают респондентов по характеристикам.

Кроме этого, такой анализ позволяет анализировать деятельность отдельных экспертов в рамках единой группы. Это позволяет не только определить наличие эффекта в группе, но и выделить эксперта, демонстрирующего этот эффект - определить более строгого эксперта, чем другие, найти эксперта, который использует шкалу оценки отлично от других, или эксперта, который не может эффективно разграничивать респондентов или критерии оценки.

Также важно учитывать, что анализ в рамках моделей Раша позволяет проанализировать различные уровни функционирования категорий и экспертов, так как для многих ситуаций можно индивидуально подобрать модель оценки. Это позволяет находить ответы на вопросы о том, инвариантна ли строгость эксперта во всех группах респондентов, на протяжении всего времени, во всех группах оценивающих, по всем основаниям.

Для анализа деятельности экспертов по оценке политомических заданий обычно применяют следующую модель (Myford, 2003, Карданова, 2004):



*1. RSM для анализа деятельности экспертов*

где *Pnilk* - вероятность того, что испытуемый *n* получит *k* баллов за выполнение задания *i* при оценке экспертом *l*; *Pnil(k-1)* - вероятность того, что испытуемый *n* получит *k*-1 баллов за выполнение задания *i* при оценке экспертом *l*; *θn* - уровень подготовленности испытуемого *n*; *δik*- уровень трудности выполнения *k*-го шага в задании *i*; *ξl*- уровень строгости эксперта *l*. В отечественных исследованиях эта модель применялась для анализа деятельности экспертов при оценивании заданий части С ЕГЭ (Карданова, 2005).

Для оценивания деятельности экспертов используется несколько статистик (эти статистики выдаются программами, позволяющими провести многопараметрический анализ, в частности, ConQuest и Facets). В данной работе мы будем проводить анализ искажений с использованием программы ConQuest, поэтому при описании способов выявления искажений мы будем акцентировать внимание на тех статистиках, которые вычисляет эта программа.

* Оценка уровня строгости эксперта в логитах (Measure)
* Статистики согласия экспертов с используемой моделью измерения.

Статистики согласия описывают степень согласия реальных данных с ожидаемыми модельными данными. Ожидаемый балл - это балл, предсказанный моделью на основании анализа уровня строгости эксперта, баллов других экспертов и т.д. Преимущественно для анализа деятельности экспертов используют статистику MNSQ (среднеквадратичный индекс) с математическим ожиданием 1 во взвешенном и невзвешенном вариантах. Значение статистики меньше 1 показывают сверхсогласие с моделью, недостаточную дисперсию, а больше 1 – недостаточное согласие с моделью. Приемлемые значения статистик не являются строго заданными, самые «мягкие» границы обозначены Линакром и составляют (0,6; 1,4). Чаще всего применяют достаточно строгие границы (0,8; 1,2), их мы и будем придерживаться в данной работе.

* Критерий Хи-квадрат

Критерий Хи-квадрат направлен на оценку того, различаются ли эксперты по уровню строгости. Нулевая гипотеза критерия хи-квадрат состоит в том, что эксперты не отличаются по уровню строгости. Значимый показатель хи-квадрат позволяет говорить о том, что как минимум два эксперта отличаются по уровню строгости. При этом, важно учитывать, что критерий хи-квадрат чувствителен к объему выборки, и при большом количестве экспертов может показывать высокую значимость даже в том случае, когда различия в строгости невелики.

* Индекс Separation (его надежность в случае ConQuest)

Индекс Separation (индекс отделимости) представляет собой число статистически различных уровней выполнения задания (то есть, при индексе равном 3, можно говорить о том, что респондентов можно разделить на три группы по результатам). Индекс получается путем перевода показателя Separation по формуле: H = (4G+1)/3, где H - индекс отделимости, G - показатель отделимости.

ConQuest показывает только надежность этого индекса, которой, в целом, достаточно. Надежность предоставляет информацию о том, как четко отделены элементы внутри фасета для того, чтобы определить надежность фасета. Значение надежности, меньшее 0.5, говорит о том, что различия между мерами строгости экспертов не существенны, лежат в пределах погрешности измерения. Для экспертов высокая надежность отражает потенциально нежелательное разделение оценивающих по уровню строгости.

**Эффект строгости/снисходительности**

Эффект строгости/снисходительности с точки зрения его проявления схож с эффектом ограничения спектра, так как он заключается в том, что эксперты склонны ставить более высокие или более низкие баллы, поэтому разграничить их часто бывает сложно. При этом, нужно учитывать, что эффект строгости не всегда является искажением, часто это просто склонность эксперта оценивать респондентов строже. Искажение в измерениях возникает, когда одного респондента оценивают, например, два строгих эксперта.

При наличии эффекта строгости у некоторых экспертов критерий *хи-квадрат* будет демонстрировать высокий уровень значимости и достаточно высокое значение. Надежность индекса *separation* будет больше 0,5, что будет говорить о том, что экспертов можно разделить на несколько категорий по строгости. Статистики согласия в данном случае, скорее всего, не будут демонстрировать отклонений от реальных данных, поэтому заключение о чрезмерной строгости эксперта должно делаться на основании групповых индикаторов, приведенных выше, и анализа уровня строгости (Measure).

**Эффект центральной тенденции**

Статистические индикаторы на групповом уровне будут демонстрировать значения, схожие со значениями в случае эффекта строгости/снисходительности: критерий *хи-квадрат* будет демонстрировать высокий уровень значимости и достаточно высокое значение, надежность индекса *separation* будет больше 0,5. На индивидуальном уровне *статистики согласия* (как взвешенная, так и невзвешенная) будут демонстрировать значения значительно меньшие 1 (сверхсогласие с моделью, вариация меньше ожидаемой). При этом, сверхсогласие с моделью возможно также в том случае, если эксперт поставил нескольким респондентам очень сходные баллы по всем характеристикам, поэтому для выявления этого эффекта нужно дополнительно смотреть на ответы эксперта.

**Эффект случайности**

 При наличии эффекта случайности критерий *хи-квадрат* будет демонстрировать высокий уровень значимости и достаточно высокое значение, индекс *separation* будет демонстрировать достаточно низкое значение. *Статистики согласия* будут демонстрировать значение значительно больше 1 (недостаточное согласие с моделью, слишком высокая дисперсия баллов).

В качестве дополнительного средства анализа деятельности экспертов можно использовать графический анализ ответов экспертов, а также корреляционные исследования.

Обобщая результаты анализа, можно говорить о том, что значения обеих статистик согласия, меньшие 1, как правило, свидетельствуют о наличии у эксперта эффекта центральной тенденции. С другой стороны, значения обеих статистик, большие 1, как правило, свидетельствуют о наличии у эксперта эффекта случайности.

## 1.2.4 Выводы

Анализ литературы показал, что существуют разные способы использования экспертных оценок, однако, в случае с конструированием теста на основе имеющихся заданий, чаще всего используют те или иные программные решения, которые впоследствии подлежат дополнительному анализу с помощью экспертных оценок. Таким образом, можно предположить, что конструирование теста на основании одних экспертных оценок также может быть полезным и возможным, если понять, как можно диагностировать возможные искажения в экспертных оценках.

# 2. РАЗРАБОТКА МЕТОДОЛОГИИ ЭКСПЕРТНО-ОРИЕНТИРОВАННОГО ПОДХОДА К СОЗДАНИЮ ТЕСТОВ

# 2.1. ОЦЕНКА ГОТОВНОСТИ К ОБУЧЕНИЮ НА ИНЖЕНЕРНЫХ СПЕЦИАЛЬНОСТЯХ ВУЗОВ

Последние тридцать лет можно по праву считать "Образовательной революцией" - приток людей в вузы, особенно в развитых странах, постоянно возрастает, в образовании проявляется тенденция к глобализации и интернационализации - многие студенты получают высшее образование за пределами своей родной страны, и это число постоянно увеличивается (Altbach, 2009). Существует много различных предположений о причинах такого роста, однако, независимо от причин, высшее образование приобретает все большую популярность. В частности, инженерное образование интенсивно развивается во многих странах мира - если тридцать лет назад основное число инженеров обучали в развитых странах, таких как США или Япония, то в последнее время большое число инженеров обучают также в странах БРИК - Бразилии, России, Индии и Китае (Altbach, 2009 The Giants Awake, Gereffi, 2008). Эти страны привлекли внимание исследователей в области экономики и образования благодаря тому, что с начала 21 века в них сосредоточена очень большая часть населения земного шара, и, кроме этого, они производят около четверти мирового ВВП (O'Neil, 2001). В области высшего инженерного образования эти страны также заслуживают внимания: в 2004 году в США выпустились 137 000 бакалавров по инженерному профилю, тогда как в Китае и Индии 351 000 и 112 000, соответственно(Xin, 2008).

Однако, несмотря на большое количество инженеров, качество их подготовки остается спорным, а запрос на квалицифированный персонал постоянно возрастает. Независимо от того, насколько прогрессивным является техническое образование в странах БРИК, от работодателей постоянно поступает информация о недостаточном качестве подготовки выпускников. Об этом же свидетельствуют и данные исследований (Loyalka, 2012). В данном контексте важно обозначить одну из ведущих тенденций, господствующих в образовательных системах стран БРИК - это постоянно возрастающая дифференциация между "элитными", ориентированными на мировой уровень образования и науки, и "неэлитными", принимающими большую часть абитуриентов, университетами (Carnoy, 2013). Несмотря на активную политику, направленную на сокращение этой дифференциации, она продолжает возрастать, что, несомненно, сказывается на качестве образования большей части выпускников (Loyalka, 2012).

И если в отношении начального школьного образования в странах БРИК уже существуют исследования, анализирующие их достижения и причины таких достижений (Chaudhary, 2012), то в области высшего образования исследований стран БРИК, в частности, России, пока еще достаточно мало. В связи с высоким количеством выпускников в странах БРИК, а также с большим количеством запросов на качественно подготовленных выпускников, возникает необходимость в оценке качества образования (в частности, инженерного), осуществляемого в каждой из стран БРИК.

Для России проблема инженерного образования и его качества имеет и самостоятельный интерес в силу известных проблем с инженерным образованием в стране. Падает престиж инженерного образования, вузы (за исключением небольшого числа элитных) вынуждены принимать для обучения на технические специальности абитуриентов по остаточному принципу, не поступивших на другие специальности. Вузам важно знать, насколько их студенты готовы к обучению по данной специальности, чтобы в случае необходимости выработать стратегию по работе с неподготовленным в достаточной мере контингентом. Очевидно, что используя только ЕГЭ, нельзя оценить готовность к обучению на технических специальностях в вузе. Особенно это справедливо в отношении математики, т.к. экзамен по математике является обязательным и вынужден совмещать в себе разные функции.

Таким образом, на данный момент мы не можем с уверенностью утверждать, что высшее образование в странах БРИК достаточно качественно. Немногочисленные исследования в этой области не позволяют непосредственно оценить, насколько будущие студенты готовы к обучению в вузе и какое образование они получают в итоге. Необходима более систематическая экспертиза качества высшего образования — основанная на прямой оценке качества готовности к обучению и качества подготовки студентов, в университетах стран БРИК.

Главной проблемой, актуальной для всех стран БРИК, в частности, для Китая и России, производящих достаточно большое количество инженеров, является отсутствие инструментов оценки, как готовности студентов к обучению в вузе, так и качества их подготовки по завершении определенного этапа обучения или по окончании вуза. Поэтому невозможно оценить, улучшается или ухудшается качество подготовки студентов вузов (в частности, технических специальностей) - нет специальных исследований, нет инструментов, не проводился анализ факторов, влияющих на качество подготовки. В условиях постоянно увеличивающегося числа выпускников технических специальностей вузов задача оценки качества их подготовки является чрезвычайно актуальной.

Одним из вариантов решения данной проблемы стала идея о создании международного теста, позволяющего оценить студентов-инженеров на всех этапах обучения - на первом курсе, сразу после поступления, после второго курса и после четвертого курса. Тестирование предполагается проводить по трем областям: математика, физика (проверяемый уровень освоения зависит от курса) и критическое мышление.

На этапе обсуждения возникла проблема с отбором тестовых заданий, так как в распоряжении разработчиков были большие банки заданий по математике и физике из Китая и России, и провести их все на одной выборке, чтобы положить на одну шкалу, не представлялось возможным. Так как готового способа для действий в такой ситуации найти не удалось, было предложено использовать экспертные оценки.

В качестве первого этапа было решено создать тест по математике. Однако, предварительно было проведено сравнение вступительных экзаменов в странах БРИК, чтобы можно было понять, насколько отличаются абитуриенты этих стран по уровню освоения математики.

## 2.1.1 Сравнительный анализ вступительных экзаменов по математике на технические специальности вузов в странах БРИК

В рамках данной работы был проведен сравнительный анализ вступительных экзаменов в вузы по математике в странах БРИК: ЕГЭ в России, JEE и AIEEE (Индия), CEE (Китай) и ENEM (Бразилия). Как отмечалось выше, Индия и Китай имеют специализированные экзамены для студентов, поступающих на инженерные (Индия) или естественно-научные (Китай) специальности. Более того, в Индии экзамены для поступления на инженерные специальности вузов различаются в зависимости от того, в какой вуз будет поступать абитуриент.

Результаты сравнения вступительных экзаменов по математике приведены в таблице 1.

*Таблица 1. Сравнение вступительных экзаменов в странах БРИК*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Параметр** | **ЕГЭ****(Россия)** | **JEE****(JEE Advanced с 2013 г.)****(Индия)** | **AIEEE****(JEE Main с 2013 г.)****(Индия)** | **CEE (Китай)** | **ENEM****(Бразилия)** |
| Модель экзамена | Один экзамен для выпуска из школы и поступления в вуз | Для поступления в вуз, сдавать могут только лучшие 150 000 по JEE Main | Для поступления в вуз | Для поступления в вуз | Один экзамен для выпуска из школы и поступления в вуз |
| Количество уровней трудности | 1 уровень | 1 уровень | 2 типа экзаменов для разных специализаций технических университетов | 1 уровень (каждый регион может изменять экзамен) | 1 уровень |
| Необходимость дополнительной подготовки | Нет | Да | Да | Нет |  |
| Количество проведений в год и количество попыток сдачи | 1 раз1 попытка | 1 раз | 2 раза (онлайн и оффлайн), 3 попытки | 1 раз, 1 попытка (могут пересдать через год) | 1 раз (октябрь-ноябрь), 1 попытка |
| Время экзамена | 4 часа | 2 блока вопросов по 3 часа каждый | 3 часа | 2 дня (120 минут для теста по математике) | 2 дня, всего 10 часов |
| Количество заданий | 20 | Всего 60 в каждой из двух частей (20 по математике) | Всего 90 (30 по математике) | 20-22 по математикеОколо 12-14 по физике | Всего 180 (45 по математике) |
| Тип заданий | С кратким ответом (14) ис развернутым ответом(6) | С выбором одного (14), или нескольких правильных ответов (6 вопросов) | С выбором одного ответа (выбор из 4 альтернатив) | Множественный выбор из 4 альтернатив (1 или несколько правильных ответов), заполнение пропусков, открытые вопросы | С выбором одного ответа (выбор из 5 альтернатив) |
| Содержание | Алгебра, уравнения и неравенства, функции, математический анализ, геометрия, статистика и теория вероятности  | Алгебра, тригонометрия, аналитическая геометрия и стереометрия, дифференциальное исчисление, интегралы, векторы  | Прогрессии, функции, комплексные числа и квадратные уравнения, матрицы, математическая индукция, биномиальная теорема, пределы, интегралы, дифференциальные уравнения, геометрия, векторы, статистика, тригонометрия | Векторы на плоскости, прогрессии, основы логики, функции, неравенства, тригонометрические функции, последовательности чисел, формулы линии и круга, конические уравнения, геометрия, биномиальная теорема, статистика и теория вероятности, пределы, производная комплексные числа |  |
| Максимальный первичный балл | 32 |  | 360 | Обычно 750 за все четыре предмета: 150 -английский, 150 -математика, наука (физика и химия) - 300, китайский -150 |  |
| Необходимость экспертной проверки заданий | Да | Нет | Нет | Да | Нет |
| Оценивание участников экзамена | В соответствии с системой перевода в 100-балльную шкалу(КТТ) | Сумма баллов по двум частям будет суммарным баллом студента (КТТ) | Итоговый балл за экзамен рассчитывается на основании экзамена (40%) и балла за JEE Main (60%)  | Оценкой является первичный балл по четырем предметам(КТТ) | IRT |
| Шкала сообщения результатов экзамена | 100 баллов | Процентиль (100) |  | 750 баллов | 1000 баллов |
| Пороговые значения |  | Устанавливаются перед экзаменом | Нет | Устанавливаются после экзамена |  |
| Статус сертификата (внутренний/международный) | Внутренний | Внутренний (для ряда университетов) | Внутренний (для ряда университетов) | Внутренний | Внутренний |
| Цена | Бесплатно | Цена зависит от балла за JEE Main | Цена зависит от типа экзамена и ряда других факторов | Цена зависит от региона | 17$ |

 В целом, по результатам анализа можно говорить о том, что вступительные экзамены в вузы по математике в странах БРИК очень различны и по форме, и по структуре, и по трудности, и по содержанию.

## 2.1.1. Выводы

На основании анализа литературных данных стало известно, что в России нет теста, позволяющего отбирать абитуриентов на инженерные специальности вузов, в отличие от большинства стран БРИК. В связи с этим, возникают подозрения, что качество образования (и, в частности, степень готовности абитуриентов к обучению на технических специальностях вузов) в этих странах выше, так как у них идет более тщательный отбор абитуриентов. На основании этого можно говорить о том, что возникает необходимость инструмента, который бы позволил оценить степень готовности к обучению студентов инженерных специальностей в разных странах и сопоставить качество приходящих абитуриентов. Так как существуют банки заданий с известными психометрическими характеристиками по математике из разных стран, создание данного теста можно рассматривать как способ формулировки методологии экспертно-ориентированного подхода.

# 2.2. ПРОГРАММА ИССЛЕДОВАНИЯ

## 2.2.1 Методологический раздел программы исследования

*Проблема исследования*:

Известно, что процесс создания тестов является достаточно трудоемким, и много усилий разработчиков уходит на создание тестовых вопросов. Однако, при создании образовательных тестов часто уже имеется некоторый набор заданий (например, из других материалов), которые можно было бы использовать для создания нового инструмента.

Однако, не известно, чем нужно руководствоваться при отборе заданий в таком случае, так как не существует разработанной методологии для работы с такими материалами. Использование экспертных оценок заданий представляется хорошим решением проблемы, однако необходимо понять, какие возможны искажения экспертных оценок при работе с заданиями, и как можно организовать работу экспертов, чтобы получить максимально эффективный процесс разработки теста.

В качестве примера решения этой проблемы в данной работе будет рассматриваться создание теста по математике для России и Китая.

*Цель исследования*:

Разработка методологии создания теста на основании экспертных оценок имеющихся вопросов.

*Задачи*:

1. провести анализ литературы по методам создания теста и возможным применениям экспертных оценок в тестировании
2. провести анализ имеющихся данных о конструировании тестов достижений путем комбинации существующих тестов
3. провести экспертную оценку имеющихся заданий
4. провести анализ деятельности экспертов
5. адаптировать модели IRT для анализа деятельности экспертов при оценке заданий, а не респондентов
6. разработать методологию создания теста на основании экспертных оценок на основании работы с тестом по математике

*Объект исследования*: экспертные оценки в тестировании

*Предмет исследования:*применение экспертных оценок для создания теста.

*Исследовательские вопросы*:

1. Как можно использовать экспертные оценки при создании нового теста?
2. Какие шаги необходимо предпринять для создания теста на базе существующих?

*Уточнение основных понятий:*

Под экспертными оценками в данной работе будут пониматься оценки, которые эксперты дают заданиям, руководствуясь предоставленными им критериями (например, трудность задания).

Строгость экспертов в данной работе понимается как латентная переменная, характеризующая склонность эксперта давать более низкие баллы по сравнению с другими экспертами.

**Гипотезы**

*Общая гипотеза:* экспертные оценки можно использовать для создания нового теста на базе существующих.

*Гипотезы-следствия:*

* Существует стратегия работы с экспертными оценками, позволяющая упростить процесс создания теста (по сравнению с созданием теста "с нуля").

## 2.2.2 Процедурный раздел программы исследования

*Стратегический план исследования*: поисковый план. Применение поискового плана в данном случае обосновано отсутствием разработанных ранее методов работы с экспертными оценками при создании нового теста на базе существующих.

*Выборка исследования*:

Генеральная совокупность - эксперты в области различных предметов

Данное исследование включает в себя пилотный опрос экспертов с целью возможности обнаружения искажений и подтверждения результатов, полученных в результате симуляционных исследований искажений. В пилотном опросе приняли участие 4 эксперта - преподавателя математики у первокурсников инженерных специальностей в вузах. Эксперты принадлежат к вузам разной направленности (1 классический, 3 технических) и разного расположения (Екатеринбург, Санкт-Петербург, Псков, Великий Новгород).

Тестирование проводилось на 177 студентах первого курса инженерных специальностей из четырех университетов России (47 студентов из Санкт-Петербурга, 60 из Пскова, 40 из Екатеринбурга, 30 из Великого Новгорода). Студенты заполняли тест в онлайн режиме, у них было ограничено время на выполнение заданий, тестирование было завершено 8 декабря 2013 года.

*Метод и методики исследования*:

В данном исследовании в качестве тестовых вопросов использовались задания из тестов Индии (JEE Main 2012) и Китая (CEE 2012), предоставленные научным консультантом Лоялкой П.

*Этапы проведения исследования*:

**Первым этапом** проведения исследования стал анализ опроса преподавателей на предмет наличия искажений. Задачей преподавателей было оценить предоставленные задачи по математике по четырем критериям. Предварительный анализ показал, что в данных могут присутствовать искажения.

**Вторым этапом** проведения исследования стал анализ литературных данных на предмет способов выявления искажений в экспертных оценках, прежде всего, при оценке ответов, а не заданий. С целью проверки функционирования данных способов в ситуации экспертной оценки заданий на этом этапе были проведены симуляционные исследования, показавшие корректное функционирование статистик при обнаружении искажений.

**Третьим этапом** исследования стал более глубокий анализ опроса экспертов, с целью проверки функционирования статистик на реальных данных и сравнения результатов работы статистик на симулированных данных и на реальных данных. Кроме этого, был также проведен опрос студентов из этих же вузов выбранными вопросами с целью проверить работу экспертов. Инструкция и бланки для экспертов приведены в приложении 1. Также на третьем этапе происходила разработка методологии создания теста на основании экспертных оценок.

# 2.3. ОБНАРУЖЕНИЕ ИСКАЖЕНИЙ В ЭКСПЕРТНЫХ ОЦЕНКАХ

## 2.3.1 Демонстрация искажений в симуляционных исследованиях

Искажения, возникающие при экспертном оценивании ответов, а также методы обнаружения этих искажений были описаны выше. Однако, важно отметить, что все эти методы давали положительный результат при анализе ряда экспертов, оценивающих достаточно большое количество респондентов. Так как мы не можем утверждать, что все заявленные выше методы анализа и обнаружения искажений будут работать в ситуации оценивания небольшого количества элементов (т.е. заданий, а не ответов респондентов), то было решено провести симуляционные исследования, демонстрирующие работу статистик в нужных нам условиях. Симуляции проводились на уровне матрицы. Анализ проводился в программе ConQuest.

В реальных данных, полученных нами, каждый эксперт получал 30 заданий и 4 критерия для оценки каждого из них. Такой вариант кажется нам наиболее приближенным к реальности в случае работы по созданию теста этим методом. Однако, с учетом того, что критерии отличались и по направленности шкалы, и по степени возможной согласованности экспертов, в симуляциях использовался один критерий для каждого эксперта, а не четыре. Таким образом предполагалось добиться наиболее объективных результатов, так как, при анализе критериев как совместно, так и попарно появлялось много отклонений в статистиках ввиду обозначенных выше причин. При работе с реальными данными, где эксперты анализировали задания с помощью четырех критериев, каждый критерий предлагается анализировать отдельно, и в дальнейшем сравнивать полученные результаты, так как, к примеру, эксперт, демонстрирующий эффект случайности, должен показывать эту тенденцию при работе с каждым критерием.

Для анализа данных использовалась модель RSM, описанная в первом разделе данной работы, однако использовалась другая интерпретация параметров. В формуле модели:



используются следующие категории: - вероятность того, что задание n получит k баллов по i-му критерию при оценке экспертом l; - вероятность того, что задание n получит k-1 баллов по i-му критерию при оценке экспертом l; - характеристика трудности задания n или релевантность содержания (в зависимости от критерия в данном параметре описывается латентный конструкт задания); - уровень трудности выполнения k-го шага в критерии i; - уровень строгости эксперта l.

Таким образом, видоизмененная формула 1 будет выглядеть следующим образом:



*2. Видоизмененная RSM*

где  - безусловная вероятность того, что задание  получит  баллов по критерию  при оценке экспертом .

**Первым этапом** симуляционных исследований было создание "идеальной" матрицы, где все эксперты статистически различаются по уровню строгости и соответствуют модели. Полученная матрица демонстрировала различия в уровне строгости экспертов (χ2 = 42,49, df = 9, p<0.001). Статистики полученных экспертов представлены в таблице 5.

*Таблица 2. Все эксперты соответствуют модели*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| 1 | -0.242 | 0.153 | 1.12 | 0.5 | 1.19 | 0.9 |
| 2 | -0.151 | 0.129 | 0.85 | -0.5 | 0.86 | -0.7 |
| 3 | 0.163 | 0.153 | 0.82 | -0.7 | 0.81 | -0.9 |
| 4 | 0.007 | 0.155 | 0.77 | -0.9 | 0.77 | -0.9 |
| 5 | -0.181 | 0.140 | 0.92 | -0.2 | 0.94 | -0.2 |
| 6 | -0.524 | 0.146 | 1.04 | 0.3 | 1.06 | 0.3 |
| 7 | 0.372 | 0.148 | 0.73 | -1.1 | 0.74 | -1.2 |
| 8 | 0.222 | 0.131 | 1.16 | 0.7 | 1.07 | 0.4 |
| 9 | -0.140 | 0.150 | 1.07 | 0.4 | 1.08 | 0.4 |
| 10 | 0.473 | 0.435 | 1.03 | 0.2 | 1.03 | 0.2 |

Приемлемые значения статистики MNSQ попадают в промежуток (0,8;1,2), т.е. можно говорить о том, что в данной матрице все эксперты соответствуют модели и не демонстрируют искажений.

Все следующие этапы работы с симуляциями включали в себя изменение *первого* эксперта в соответствии с требуемым искажением, в то время как остальные эксперты оставлялись неизменными. Для демонстрации эффектов моделировались экстремальные ситуации эффектов, каждый из которых описан в соответствующем разделе ниже.

**Эффект строгости и эффект снисходительности**

Как уже было сказано выше, при проявлении эффекта строгости или эффекта снисходительности эксперт не будет демонстрировать отсутствие согласия с моделью, и уровень строгости будет виден в таблице. Также отследить этот эффект можно графически. Для симуляции эффекта снисходительности первому эксперту приписывались максимальные баллы (оценка 5 для 27 заданий, оценка 4 для трех заданий).

В таблице ниже показан результат симуляции эффекта снисходительности, жирным шрифтом выделен измененный эксперт.

*Таблица 3. Эффект снисходительности*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| **1** | **-1.734** | **0.182** | **1.61** | **2.1** | **1.11** | **0.4** |
| 2 | 0.014 | 0.157 | 0.76 | -0.9 | 0.81 | -0.9 |
| 3 | 0.334 | 0.158 | 0.78 | -0.8 | 0.77 | -1.1 |
| 4 | 0.163 | 0.143 | 0.62 | -1.6 | 0.61 | -1.8 |
| 5 | -0.029 | 0.149 | 0.89 | -0.3 | 0.91 | -0.3 |
| 6 | -0.378 | 0.151 | 1.16 | 0.7 | 1.21 | 0.9 |
| 7 | 0.550 | 0.134 | 0.87 | -0.5 | 0.89 | -0.4 |
| 8 | 0.400 | 0.153 | 1.39 | 1.4 | 1.24 | 1.2 |
| 9 | 0.025 | 0.455 | 0.98 | 0.0 | 1.02 | 0.1 |
| 10 | 0.656 | 0.132 | 1.11 | 0.5 | 1.11 | 0.5 |

По таблице выше видно, что по взвешенным статистикам эксперт не демонстрирует сильных отклонений, однако его уровень строгости сильно отличается от уровня строгости других экспертов. В данном случае также значение критерия χ2 показало значимые различия по уровню строгости экспертов, и его значение оказалось гораздо выше, чем в "идеальной" ситуации (χ2 = 151,42, df = 9, p<0.001). Разница между уровнем строгости экспертов также может быть продемонстрирована графически (рисунок 1).

*Рисунок 1. Демонстрация эффекта снисходительности*

На рисунке также видно, что результаты эксперта 1 выделяются на фоне остальных экспертов.

Аналогичные результаты были получены при симуляции эффекта строгости (оценка 1 для 26 заданий, оценка 2 для 4 заданий). В таблице ниже показаны результаты симуляции эффекта строгости, жирным шрифтом выделен измененный эксперт. В данной ситуации эксперты также значимо различались по уровню строгости, и значение критерия хи квадрат было достаточно высоким (χ2 = 211,89, df = 9, p<0.001).

*Таблица 4. Эффект строгости*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| **1** | **2.274** | **0.182** | **1.06** | **0.3** | **1.01** | **0.1** |
| 2 | -0.435 | 0.158 | 1.10 | 0.4 | 0.98 | 0.0 |
| 3 | -0.109 | 0.159 | 0.84 | -0.6 | 0.84 | -0.7 |
| 4 | -0.284 | 0.144 | 0.58 | -1.8 | 0.57 | -2.0 |
| 5 | -0.478 | 0.150 | 1.35 | 1.3 | 1.23 | 1.0 |
| 6 | -0.830 | 0.153 | 1.06 | 0.3 | 1.08 | 0.4 |
| 7 | 0.110 | 0.135 | 0.86 | -0.5 | 0.91 | -0.3 |
| 8 | -0.043 | 0.154 | 1.76 | 2.5 | 1.61 | 2.4 |
| 9 | -0.423 | 0.458 | 0.97 | -0.0 | 1.02 | 0.1 |
| 10 | 0.217 | 0.133 | 1.13 | 0.6 | 1.19 | 0.8 |

В данном случае статистики согласия не выходят за принятые рамки даже в случае невзвешенных статистик, однако также видно, что эксперт 1 достаточно сильно выделяется на фоне остальных. В таком случае отследить наличие эффекта возможно только по оценке уровня строгости, что продемонстрировано в таблице и на рисунке ниже.

*Рисунок 2. Демонстрация эффекта строгости*

**Эффект центральной тенденции**

Выше было сказано, что при эффекте центральной тенденции в случае оценивания ответов респондентов будет демонстрироваться значение взвешенных и невзвешенных статистик меньше 1. Результаты симуляции эффекта центральной тенденции представлены ниже.

*Таблица 5. Эффект центральной тенденции*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| **1** | **-2.820** | **0.190** | **0.64** | **-1.5** | **0.93** | **0.2** |
| 2 | 0.133 | 0.158 | 0.80 | -0.7 | 0.81 | -0.8 |
| 3 | 0.459 | 0.160 | 0.82 | -0.6 | 0.82 | -0.8 |
| 4 | 0.284 | 0.145 | 0.63 | -1.6 | 0.62 | -1.7 |
| 5 | 0.088 | 0.151 | 0.91 | -0.3 | 0.90 | -0.3 |
| 6 | -0.267 | 0.153 | 1.21 | 0.9 | 1.25 | 1.0 |
| 7 | 0.676 | 0.135 | 0.73 | -1.0 | 0.75 | -1.1 |
| 8 | 0.522 | 0.155 | 1.39 | 1.4 | 1.28 | 1.3 |
| 9 | 0.140 | 0.462 | 1.13 | 0.6 | 1.17 | 0.8 |
| 10 | 0.786 | 0.134 | 1.14 | 0.6 | 1.13 | 0.6 |

Для симуляции эффекта центральной тенденции первому эксперту приписывался балл 3 за все задания, кроме одного. Уровень строгости экспертов статистически различался (χ2 = 308,33, df = 9, p<0.001).

В данном случае видно, что эксперт оказался достаточно снисходительным, однако это нельзя принять за эффект снисходительности, несмотря на низкие значения параметра строгости, за счет того, что обе статистики MNSQ показывают значения меньше 1. Таким образом, мы можем заключить, что способы выявления эффекта центральной тенденции работают и применительно к ситуации оценивания заданий.

Отдельно мы рассмотрели эффект *групповой* центральной тенденции. Для демонстрации этого эффекта всем экспертам были приписаны баллы 3 и 2. Результаты проверки по эффекту групповой центральной тенденции представлены в таблице ниже.

*Таблица 6. Эффект групповой центральной тенденции*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| 1 | -0.045 | 0.520 | 1.15 | 0.6 | 1.04 | 0.3 |
| 2 | -0.045 | 0.520 | 1.06 | 0.3 | 1.04 | 0.3 |
| 3 | 0.402 | 0.470 | 0.95 | -0.1 | 1.01 | 0.2 |
| 4 | -0.045 | 0.520 | 1.01 | 0.1 | 1.02 | 0.3 |
| 5 | -0.045 | 0.520 | 1.02 | 0.1 | 1.03 | 0.3 |
| 6 | -0.045 | 0.520 | 1.00 | 0.1 | 1.02 | 0.2 |
| 7 | -0.045 | 0.520 | 1.31 | 1.2 | 1.06 | 0.3 |
| 8 | -0.045 | 0.520 | 1.18 | 0.7 | 1.04 | 0.3 |
| 9 | -0.045 | 1.544 | 0.89 | -0.4 | 1.01 | 0.2 |
| 10 | -0.045 | 0.520 | 0.96 | -0.1 | 1.02 | 0.2 |

В данном случае важно обратить внимание на то, что все статистики согласия показывают очень хорошие значения, и для того, чтобы выявить эффект групповой центральной тенденции, нужно смотреть на уровень строгости экспертов и на проверку по критерию хи квадрат. В данном случае эксперты не отличались по уровню строгости (χ2 = 0,79, df = 9, p=1,000), что может говорить о наличии групповых эффектов, например, ограничения спектра или центральной тенденции как частного случая ограничения спектра.

**Эффект случайности**

При наличии эффекта случайности в случае оценивания ответов респондентов значения обеих статистик MNSQ становятся больше 1. Для проверки функционирования статистик в ситуации оценивания заданий первому эксперту приписывались значения, полученные с помощью генератора случайных чисел. Результаты представлены в таблице ниже.

*Таблица 7. Эффект случайности*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| **1** | **0.505** | **0.112** | **1.34** | **1.3** | **1.22** | **1.1** |
| 2 | -0.228 | 0.148 | 0.99 | 0.0 | 0.94 | -0.3 |
| 3 | 0.074 | 0.149 | 0.92 | -0.2 | 0.93 | -0.3 |
| 4 | -0.072 | 0.134 | 0.83 | -0.6 | 0.84 | -0.6 |
| 5 | -0.248 | 0.140 | 1.01 | 0.1 | 1.01 | 0.1 |
| 6 | -0.579 | 0.143 | 1.01 | 0.1 | 1.03 | 0.2 |
| 7 | 0.271 | 0.125 | 0.91 | -0.3 | 0.94 | -0.2 |
| 8 | 0.129 | 0.144 | 1.25 | 1.0 | 1.17 | 1.0 |
| 9 | -0.216 | 0.408 | 0.86 | -0.5 | 0.90 | -0.4 |
| 10 | 0.364 | 0.123 | 0.93 | -0.2 | 0.94 | -0.2 |

Как видно из таблицы, оба значения статистики MNSQ значительно больше единицы, и сильно отличаются от статистик других экспертов. При этом, по строгости эксперты значимо различаются (χ2 = 56,95, df = 9, p<0.001). Таким образом, можно говорить о том, что в случае оценки эффекта случайности статистики работают так же, как при оценке ответов респондентов.

## 2.3.1 Анализ искажений в реальных данных

Реальные данные были получены при опросе 4 экспертов, как уже было сказано выше. Им предоставлялся бланк для оценки заданий, включающий в себя 30 заданий из тестов Индии и Китая, которые им было необходимо оценить по 4 критериям. Образец бланка экспертов представлен в приложении 1. Как мы отмечали выше, анализ деятельности экспертов по четырем критериям сразу не может дать объективной оценки деятельности экспертов, так как критерии имеют разную направленность и необходимую степень объективности, поэтому дальнейший анализ реальных данных будет проводиться по каждому из четырех критериев отдельно.

Принимая во внимание то, что эксперты могут давать согласованные, но неверные ответы по поводу трудности задания, как было показано в первой части работы, нами был дополнительно проведен опрос студентов первого курса с использованием тех же заданий, что предъявлялись экспертам. Так как в нашем случае эксперты оценивали трудность заданий по двум критериям разной направленности (трудность задания и ожидаемый процент решивших), то мы проводили корреляционный анализ между результатами опроса студентов и оценками экспертов. В качестве одной переменной брался суммарный балл за задание (в каждом вузе отдельно, за исключением последней строки), в качестве другой - балл эксперта за задание (в случае подсчета общего балла бралась суммарная оценка экспертов по данному заданию). Результаты анализа представлены в таблице ниже.

*Таблица 8. Корреляции между оценками экспертов и баллами студентов*

|  |  |  |
| --- | --- | --- |
| Эксперт | Корреляция Балл\*Трудность задания | Корреляция Балл\*Ожидаемый процент решивших |
| 1 | -0,53\*\* | 0,44\* |
| 2 | -0,12 | 0,21 |
| 3 | -0,16 | 0,15 |
| 4 | -0,09 | 0,19 |
| Общее | -0,32 | 0,35\* |

Как видно из таблицы, все эксперты правильно использовали две представленные шкалы - у всех экспертов с возрастанием трудности понижался суммарный балл за задание, и также у всех экспертов с повышением балла по шкале "Ожидаемый процент решивших" повышался суммарный балл. Таким образом, мы можем предположить, что эксперты правильно использовали представленную им шкалу и смогли определить примерную трудность задания и количество студентов из их вуза, которые смогут справиться с этим заданием.

**Критерий 1.**

Формулировка задания для экспертов:

*Важна или не важна тема, к которой относится это задание, для овладения всей программой подготовки специалистов в инженерной области?*

*(совсем не важна) 1 ………. 5 (очень важна)*

Данный критерий должен оценить важность тематической области задания с точки зрения эксперта. По результатам анализа данных были получены следующие статистики:

*Таблица 9. Критерий 1*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| 1 | -3.363 | 0.219 | 0.92 | -0.2 | 1.07 | 0.4 |
| 2 | 0.260 | 0.175 | 1.15 | 0.6 | 1.12 | 0.5 |
| 3 | 0.604 | 0.172 | 1.15 | 0.6 | 0.85 | -0.5 |
| 4 | 2.499 | 0.329 | 0.92 | -0.2 | 0.92 | -0.2 |

По результатам анализа таблицы, представленной выше, можно говорить о том, что все эксперты, в целом, демонстрируют согласие с моделью. При этом, мы видим также излишне высокое значение критерия хи квадрат (χ2 = 249,89, df = 3, p<0.001), что говорит о том, что может присутствовать искажение.

Важно отметить, что эксперт 1 представляется излишне лояльным, и, возможно, стоит обратить внимание на дисперсию его баллов. Действительно, данный эксперт демонстрировал яркий пример ограничения спектра в виде крайней лояльности - все его оценки были равны 5.

*Рисунок 3. Результаты эксперта 1 по критерию 1*

В то же время, мы видим, что результаты эксперта 4 могут показаться излишне строгими, однако, графически это не подтверждается, несмотря на то, что он действительно выглядит более строгим, чем другие эксперты, и такие статистические результаты были получены за счет того, что уровень строгости последнего эксперта вычислялся так, чтобы в сумме уровни строгости экспертов были равны 0.



*Рисунок 4. Результаты эксперта 4 по критерию 1*

По результатам анализа полученных данных можно говорить о том, что эксперт 1 является излишне лояльным, и его следует убрать из дальнейшего анализа.

Отдельным параметром для анализа согласованности экспертов можно считать корреляции между ними. В данном случае в таблице ниже видно, что результаты эксперта 1 имеют достаточно низкие корреляции с результатами других экспертов, что также можно считать основанием для предположения об искажениях в его оценках и последующего удаления его из анализа.

*Таблица 10. Корреляция экспертных оценок, критерий 1*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Критерий 1.** | **Эксперт 1** | **Эксперт 2** | **Эксперт 3** | **Эксперт 4** |
| Эксперт 1 | 1 |  |  |  |
| Эксперт 2 | -0,18 | 1 |  |  |
| Эксперт 3 | 0,29 | 0,30 | 1 |  |
| Эксперт 4 | 0,01 | 0,16 | 0,33 | 1 |

**Критерий 2.**

Формулировка задания для экспертов:

*Трудным или легким является данное задание для типичного выпускника российской школы, поступающего на техническую специальность вуза?*

*(очень легкое задание) 1 ………. 5 (очень трудное задание)*

Данный критерий должен демонстрировать достаточно объективную оценку трудности задания, так как речь идет не о студентах вуза, в котором преподает эксперт, но о любых поступающих на инженерные специальности вузов. Результаты анализа представлены в таблице ниже.

*Таблица 11. Критерий 2*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| 1 | -1.685 | 0.193 | 0.64 | -1.5 | 0.68 | -1.3 |
| 2 | -0.038 | 0.182 | 0.88 | -0.4 | 0.85 | -0.6 |
| 3 | 0.275 | 0.184 | 1.75 | 2.4 | 1.81 | 2.8 |
| 4 | 1.448 | 0.323 | 0.78 | -0.8 | 0.76 | -0.9 |

Во втором критерии эксперты также статистически различаются по уровню строгости (χ2 = 78,32, df = 3, p<0.001), однако оценка по критерию хи квадрат не так высока. В данном случае также выделяются результаты эксперта 1. Помимо того, что они по-прежнему являются достаточно лояльными, можно говорить о том, что при оценке задания по данному критерию у него проявился также эффект центральной тенденции. Этот эффект также проявился у двух других экспертов, но в меньшей степени, и, возможно, был обусловлен тем, что третий эксперт продемонстрировал эффект случайности.

При этом, анализ корреляций оценок экспертов по данному критерию показывает, что эксперты показывают достаточно высокую согласованность. Результаты эксперта 3 также достаточно высоко согласуются с результатами других экспертов, поэтому можно предположить, что эксперт 3 просто показал больший разброс оценок, чем другие, что привело к таким результатам анализа.

*Таблица 12. Корреляция экспертных оценок, критерий 2*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Критерий 2.** | **Эксперт 1** | **Эксперт 2** | **Эксперт 3** | **Эксперт 4** |
| Эксперт 1 | 1 |  |  |  |
| Эксперт 2 | 0,57\*\* | 1 |  |  |
| Эксперт 3 | 0,53\*\* | 0,37\* | 1 |  |
| Эксперт 4 | 0,35 | 0,21 | 0,40\* | 1 |

**Критерий 3.**

Формулировка критерия для экспертов:

*Какой процент студентов, зачисленных в целом на первый курс инженерных специальностей Вашего вуза, может решить это задание?*

*(очень малый, <20) 1 ………. 5 (большой, >80)*

Этот критерий является более субъективным, по сравнению с предыдущим, несмотря на их тематическую близость. Кроме этого, он обратно ориентирован.

*Таблица 13. Критерий 3*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| 1 | -2.223 | 0.205 | 0.96 | -0.1 | 0.92 | -0.2 |
| 2 | -0.056 | 0.177 | 1.04 | 0.3 | 1.09 | 0.4 |
| 3 | 0.123 | 0.178 | 1.27 | 1.0 | 1.30 | 1.1 |
| 4 | 2.157 | 0.325 | 0.65 | -1.5 | 0.54 | -1.9 |

Все эксперты в данном случае значимо различаются по уровню строгости (χ2 = 117,89, df = 3, p<0.001). По результатам анализа полученных результатов можно говорить о том, что четвертый эксперт демонстрирует достаточно ярко выраженный эффект центральной тенденции. Эксперты 1 и 2 показывают хорошее соответствие модели, хотя эксперт 1 и по результатам оценки этого критерия оказывается очень снисходительным. Эксперт 3 показывает склонность к эффекту случайности, однако его показатели не выходят за рамки нормы. В то же время, как видно из таблицы ниже, его результаты показывают наименьшую корреляцию с результатами других экспертов.

*Таблица 14. Корреляция экспертных оценок, критерий 3*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Критерий 3.** | **Эксперт 1** | **Эксперт 2** | **Эксперт 3** | **Эксперт 4** |
| Эксперт 1 | 1 |  |  |  |
| Эксперт 2 | 0,49\*\* | 1 |  |  |
| Эксперт 3 | 0,31 | 0,26 | 1 |  |
| Эксперт 4 | 0,42\* | 0,57\*\* | 0,48\*\* | 1 |

**Критерий 4.**

Формулировка критерия для экспертов:

*Насколько подробно тема, к которой относится данное задание, изучается в курсе высшей математики в вузе?*

*(совсем не изучается ) 1 ………. 5 (изучается очень подробно)*

Данный критерий является одним из самых объективных среди представленных, так как он опирается на восприятие экспертом университетской программы для студентов-инженеров, которая, в целом, достаточно схожа даже в различных вузах.

*Таблица 15. Корреляция экспертных оценок, критерий 4*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Критерий 4.** | **Эксперт 1** | **Эксперт 2** | **Эксперт 3** | **Эксперт 4** |
| Эксперт 1 | 1 |  |  |  |
| Эксперт 2 | 0,34 | 1 |  |  |
| Эксперт 3 | 0,11 | 0,77\*\* | 1 |  |
| Эксперт 4 | 0,19 | 0,5\*\* | 0,7\*\* | 1 |

По результатам анализа корреляций также можно говорить о том, что эксперт 1 показывает несогласованность с результатами других экспертов. Это видно по тому, что он демонстрирует достаточно низкие корреляции с каждым из экспертов, на фоне их корреляции друг с другом.

*Таблица 16. Критерий 4*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № эксперта | Строгость | Ошибка | Невзвешенные статистики | Взвешенные статистики |
| MNSQ | T | MNSQ | T |
| 1 | -1.117 | 0.146 | 0.96 | -0.1 | 0.73 | -0.6 |
| 2 | -0.112 | 0.129 | 1.07 | 0.3 | 1.13 | 0.6 |
| 3 | 0.382 | 0.128 | 1.36 | 1.3 | 1.53 | 1.8 |
| 4 | 0.846 | 0.234 | 0.81 | -0.7 | 0.85 | -0.5 |

В данном критерии эксперты также различны по уровню строгости (χ2 = 67,79, df = 3, p<0.001). Результаты психометрического анализа четвертого критерия позволяют говорить о том, что эксперт 1 по-прежнему является самым лояльным. Эксперт 3 в данном критерии также демонстрирует наличие эффекта случайности, однако, при этом, показывает достаточно высокие корреляции с другими экспертами, не показавшими эффекта случайности, из чего мы можем предположить, что в данном случае нет эффекта случайности.

## 2.3.3 Выводы

По результатам проведенных видов анализа можно говорить о том, что методы многопараметрического анализа можно использовать в ситуации анализа деятельности экспертов. Также, в качестве дополнения к этим методам, рекомендуется использовать корреляционные исследования и графическую репрезентацию данных, чтобы нивелировать возможные искажения в связи с небольшой выборкой заданий. При оценке большим количеством экспертов (как показано в симуляциях) большего количества заданий (в случае с тестом для оценки уровня готовности студентов) результаты многопараметрического анализа будут более точными, и дополнительные методы анализа не будут играть решающую роль.

Важно отметить, что критерии, направленные на оценку трудности задания (в нашем исследовании это критерии 2 и 3), функционируют лучше критериев, направленных на оценку релевантности темы. В частности, единственный критерий, предполагающий субъективную оценку эксперта (критерий 1) показывает самую плохую согласованностью экспертов, из чего можно заключить, что экспертную оценку стоит использовать для оценки объективных параметров заданий. Возможно, формулировку вопроса 1 стоит изменить, предложив вместо нее аналог дискриминативности (например: "Как вы считаете, будет ли это задание хорошо разделять сильных и слабых учеников?").

В результате анализа деятельности экспертов на реальных данных можно говорить о том, что среди экспертов не было выявлено искажений на групповом уровне. Можно порекомендовать убрать из дальнейшего анализа заданий эксперта 1, так как при оценке по всем критериям он демонстрирует эффект снисходительности. Мы рекомендуем также обратить внимание на эксперта 3, так как в трех критериях из четырех он оказался подвержен эффекту случайности. При этом, нужно учитывать, что его корреляции с оценками других экспертов достаточно высоки, поэтому наличие эффекта случайности в данном случае сомнительно. Другие эксперты тоже подвержены некоторым эффектам, однако эти искажения нельзя считать систематическими, так как они проявляются только в некоторых критериях, и могут оказаться недостатками в формулировке самих оценочных критериев. В связи с этим одной из рекомендаций может считаться доработка и более глубокий анализ самих критериев оценки, несмотря на то, что по результатам прошлогоднего исследования был сделан вывод о том, что шкалы во всех четырех критериях работают, и могут быть использованы для оценки заданий.

# 2.4. РАЗРАБОТКА ИНСТРУМЕНТА

## 2.4.1 Методология разработки теста с помощью экспертно-ориентированного подхода

Исследование, связанное с экспертными оценками качеств заданий, приведенное выше, является начальным этапом разработки методики по оценке студентов первого курса инженерных специальностей вузов в Китае и России. В рамках работы над данной методикой планируется использовать методы анализа, продемонстрированные выше, в связи с чем возникла необходимость не только проверить функционирование статистических методов, но и разработать схему работы с экспертами при наличии достаточно большого банка заданий (в данном случае банк составляет около 700 заданий по математике из каждой страны). Все задания должны иметь удовлетворительные психометрические характеристики.

Предложенная нами схема работы с экспертами включает в себя три основных этапа. Данная схема может быть использована при создании другого теста достижений с использованием экспертов.

**Первый этап** заключается в предоставлении небольшому числу экспертов (предположительно, двум или трем) списка тем, к которым относятся имеющиеся задания. Заданием экспертов становится оценка всех тем по уровню важности и степени необходимого минимального освоения данной темы для успешного обучения на инженерной специальности вуза. Для того, чтобы оценить степень важности темы, была предложена следующая формулировка:

*Насколько важна данная тема для овладения программой подготовки специалистов в инженерной области?*

*1 – Тема не важна*

*2 – Тема относительно невысокой важности*

*3 – Тема относительно высокой важности*

*4 – Тема очень важна*

Помимо необходимости оценить уровень важности темы, нам представляется существенным мнение экспертов о том, почему эта тема считается важной. При работе с другой образовательной дисциплиной представляется не менее важным оценивание степени важности темы. В зависимости от цели разработки теста формулировка вопроса изменяется, необходимость получения причин оценок экспертов зависит от целей исследования.

Для оценки минимального необходимого уровня освоения темы нами была разработана таксономия на основе таксономий, принятых в исследованиях TIMSS и PISA. Данная таксономия включает в себя три уровня, представленные ниже, и предлагается к использованию не только в рамках создания теста по математике, но и в рамках создания других образовательных тестов с использованием экспертно-ориентированного подхода, так как она является достаточно простой и исчерпывающей. Формулировка задания для экспертов включает в себя описание таксономии:

*Каков необходимый минимальный уровень освоения данной темы для успешного дальнейшего обучения в вузе? (отметьте подходящий вариант)*

*0 – Освоение темы не требуется*

*1 –* ***Знание*** *основных понятий (фактов, свойств, правил)*

*2 –* ***Применение*** *стандартных алгоритмов для решения задач.*

*3 –* ***Размышления****: математическая формулировка проблемы и составление алгоритма для ее решения.*

Данная таксономия является вспомогательным инструментом при отборе заданий, поэтому основными критериями при ее разработке были простота и понятность формулировок, а также возможность охватить весь спектр трудности заданий. На основании этого данная таксономия может рекомендоваться к использованию при разработке других тестов достижений с помощью экспертно-ориентированного подхода.

Кроме этого, экспертов также предлагается спросить о корректности формулировки тем, так как представленный нами банк заданий был категорирован не в соответствии с образовательным стандартом, несмотря на то, что он находится в рамках этого стандарта. Мы предлагаем сформулировать вопрос следующим образом:

*Насколько хорошо сформулирована тема и основные контролируемые элементы (в контексте целей данного исследования)? (отметьте подходящий вариант)*

*1 – Сформулированы хорошо*

*2 – Требуют уточнения*

В случае, если представленные темы требуют уточнения, в бланке оставлено место для коррекции темы. Предполагается, что проведение экспертной оценки тем и заданий будет проводиться в электронном виде с использованием ресурсов веб-анкетирования, в связи с чем нами была оформлена каждая тема отдельно. Инструкция и образец оформления одной темы представлены в приложении 2.

**Вторым этапом** работы с экспертами является просьба распределить все имеющиеся задания по ячейкам таксономии. Мы предлагаем экспертам (также двум или трем людям) оценить задания с точки зрения уровня таксономии, необходимого для выполнения этого задания. Образец таблицы представлен ниже.

*Таблица 17. Образец таблицы для второго этапа разработки*

|  |  |  |  |
| --- | --- | --- | --- |
| Трудность задания | Уровень 1 – знание | Уровень 2 – применение | Уровень 3 - размышления |
| Задание 1 |  | + |  |
| Задание 2 |  |  | + |
| … |  |  |  |

На основании полученных результатов и объединения информации, полученной на первых двух этапах, формируется *предварительный* банк заданий. Для этого отбираются задания необходимого уровня по важным темам, и направляются на оценку большему количеству экспертов.

Таким образом, **третий этап** работы с экспертами заключается в оценке определенного числа заданий, меньшего, чем в исходном банке, экспертами в количестве от 10 человек. Предполагается, что оценка заданий будет производиться по четырем критериям, использованным в пилотном опросе, с последующим психометрическим анализом, приведенным выше.

На основании результатов третьего этапа работы, по итогам анализа деятельности экспертов с учетом наличия различных искажений и согласованности ответов, должен быть составлен окончательный список заданий, которые будут использоваться в тестировании первокурсников.

Таким образом, схема конструирования теста может быть представлена следующим образом:

*Таблица 18. Схема конструирования теста*

|  |  |  |
| --- | --- | --- |
| **№ этапа** | **Количество экспертов** | **Действие** |
| 1 | 2-3 | Анализ тем: уровень важности и степень минимального освоения темы (по таксономии)  |
| 2 | 2-3 | Оценка задания с точки зрения уровня таксономии, необходимого для выполнения этого задания  |
| 3 | 10 и более | Оценка релевантных заданий нужного уровня трудности по отобранным критериям |
| 4 | - | Анализ деятельности 10 экспертов с помощью методов, представленных выше  |
| 5 | - | Конструирование теста на основании рекомендаций экспертов  |
| 6 | - | Апробация и анализ результатов, доработка теста |

## 2.4.2 Выводы

В результате анализа имеющихся данных и проведения симуляционных и реальных исследований по методам анализа экспертных оценок нами была разработана схема создания теста на основе имеющихся заданий с использованием экспертных оценок. Данная схема адаптирована под исследование степени готовности студентов инженерных специальностей вузов России и Китая к обучению в университете, однако она также может применяться в различных исследованиях, где возникает необходимость создания инструмента для оценки достижений.

# ЗАКЛЮЧЕНИЕ

В данной работе были рассмотрены методы конструирования теста, основанные на экспертных оценках. Так как создание теста предполагает ситуацию оценивания заданий по заданным критериям, то необходимо было разработать систему оценки полученных от экспертов данных, ввиду того, что эти данные могут быть подвержены искажениям.

Для того, чтобы проанализировать функционирование статистик IRT (как самых точных в выявлении искажений) в ситуации оценивания заданий, а не в привычной ситуации оценивания ответов респондентов, нами были проведены симуляционные исследования, показавшие правомерность использования многопараметрических моделей IRT для анализа такого рода ситуаций. После этого мы провели исследование реальных данных опроса четырех экспертов согласно проверенным методам, что позволило нам сделать вывод о том, что методы, предложенные нами, действительно работают.

Заключительным этапом работы стало формулирование методологии отбора заданий с использованием экспертных оценок. Данная методология позволит более эффективно отбирать задания для тестов достижений на основании существующих банков (в том числе, и не связанных между собой статистически).

Таким образом, данная работа освещает один из методов разработки тестов, который является менее затратным, чем разработка образовательного теста с нуля, так как в нем отсутствует самый трудоемкий этап формулировки вопросов, их анализа и доработки. Несмотря на то, что на определенном этапе может понадобиться видоизменение или коррекция вопросов на основании экспертных оценок, изначально эксперты работают с уже готовыми заданиями, и этот пункт можно считать одним из достоинств экспертно-ориентированного подхода к конструированию тестов.

В будущем может быть целесообразным более глубокая проработка симуляционных исследований, а также анализ большего количества эффектов с последующей проверкой функционирования модели на большем количестве экспертов (от 10). Вероятно, при таком количестве респондентов можно ограничиться только методами IRT, не задействуя дополнительные методы анализа, однако такая гипотеза требует дополнительной проверки в будущем.

# Список литературы

1. Altbach, P. (2009) The Giants Awake: The Present and Future of Higher Education Systems in China and India, *Higher Education to 2030*, OECD, pp. 179-204.
2. Biggs, S.N., Kennedy, J.D. et al. (2012) Psychometric properties of an omnibus sleep problems questionnaire for school-aged children, *Sleep Medicine,* 13, pp. 390-395
3. Bramley, T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgment, *Journal of Applied Measurement,* 6 (*2*), pp. 202-223
4. Carnoy, M., Loyalka, P., Froumin, I., Dossani, R, Jandhyala T., and Wang R. (2013*) Higher Education in the Global Knowledge Economy: Triumph of the BRICs?* Stanford, CA: Stanford University Press
5. Chaudhary, L., Musacchio, A. et al. (2012) Big BRICs, weak foundations: The beginning of public elementary education in Brazil, Russia, India, and China, *Explorations in Economic History*, 49, pp. 221–240
6. Chen, C., Lee, S. et al. (1995) Response style and cross-cultural comparisons of rating scales among East Asian and North American students, *Psychological Science*, 6, pp. 170-175
7. Cooper, W.H. (1981) Ubiquitous halo, *Psychological Bulletin*, 90, pp. 218-244
8. Crocker, L., Algina, J. (2008) Introduction to Classical and Modern Test Theory, Cengage Learning, USA
9. Downing, S.M. Twelve steps for effective test development, *Handbook of test development*, Lawrence Erlbaum, 2006, pp. 3-26
10. Hamp-Lyons, L., Mathias, S.P. (1994) Examining Expert Judgments of Task Difficulty on Essay Tests, *Journal of Second Language Writing,* 3 (*1*), pp. 49-68
11. Hardesty, D.M., Bearden, W.O. (2004) The use of expert judges in scale development. Implications for improving face validity of measures of unobservable constructs, *Journal of Business Research*, 57, pp. 98– 107
12. Hwang, G.-J., Lin, B.M.T. et al. (2006) An effective approach for test-sheet composition with large-scale item banks, *Computers & Education*, 46, pp. 122–139
13. Kneeland, N. (1929) That lenient tendency in rating, *Personnel Journal*, 7, pp. 356-366
14. Liu, Y.-C., Chen, P.-J. (2012) Discovering discriminative test items for achievement tests, *Expert Systems with Applications,* 39, pp. 1426-1434
15. Loyalka, P. et al. (2012) Getting the quality right: engineering education in the BRIC countries (Working paper)
16. Myford, C.M., Wolfe, E.W. (2003) Detecting and measuring rater effects using many-faceted Rasch measurement: part 1, *Journal of Applied Measurement*, 4 (*4*), pp. 386-422.
17. Myford, C.M., Wolfe, E.W. (2003) Detecting and measuring rater effects using many-faceted Rasch measurement: part 2, *Journal of Applied Measurement*, 5 (*2*), pp. 189-227.
18. O’Neill, J. (2001*). Building Better Global Economic BRICs*. New York: Goldman Sachs, Global Economics Paper No: 66
19. Xin, H., Normile, D. (2008) Gunning for the Ivy League, *Science*, 319, pp. 148-151
20. Zhu, W., Ennis, C.D., Chen, A. (1998), Many-faceted Rasch modeling expert judgment in test development, *Measurement in Physical Education and Exercise Science*, 2(*1*), pp. 21-39.
21. Карданова Е.Ю. Математические модели многофасетного анализа // Вопросы тестирования в образовании. – 2004,    № 11, с. 11-27.
22. Карданова Е.Ю. Применение многопараметрического анализа для исследования деятельности экспертов // Вопросы тестирования в образовании. 2005. №14. С. 6-31.
23. Орлов А. И. Экспертные оценки. Учебное пособие. М.: ИВСТЭ, 2002
24. Федеральная Служба Государственной Статистики (2012) *Социально-экономическое положение России*, Москва

# *Приложение 1*. Образец инструкции и бланка для оценки заданий

**Уважаемый эксперт!**

Предлагаем Вам принять участие в международном исследовании, имеющем целью анализ качества подготовки по математике и физике студентов технических специальностей российских вузов в сравнении со студентами аналогичных вузов стран БРИК (Бразилия, Россия, Индия и Китай). В России исследование проводит Институт образования Национального исследовательского университета «Высшая школа экономики».

Есть основания полагать, что качество образования, полученного студентами инженерных специальностей в некоторых странах БРИК, улучшилось за прошедшее десятилетие. Во-первых, Бразилия, Китай и Индия увеличили инвестиции в школьное среднее образование, и, таким образом, в образование потенциальных будущих инженеров. Во-вторых, правительства этих стран приложили большие усилия, чтобы улучшить качество инженерного образования: программы инженерного образования в высших учебных заведениях, и особенно в элитных, имеют большее финансирование в расчете на одного студента, более квалифицированный профессорско-преподавательский состав и больший фокус на исследования, чем когда-либо прежде. В-третьих, для поступления на инженерные специальности вузов школьники должны сдать специальный экзамен, для чего углубленно изучают математику и физику в старших классах школы.

Для России проблема инженерного образования и его качества имеет особый интерес в силу известных проблем с инженерным образованием в стране. Информация о том, насколько студенты готовы к обучению, важна как самим вузам (чтобы в случае необходимости выработать стратегию по работе с неподготовленным в достаточной мере контингентом), так и управляющим органам. Именно анализ готовности студентов к обучению на технических специальностях вузов является целью первого этапа исследования.

В рамках проводимого нами исследования планируется провести тестирование наиболее сильных студентов первого курса разных вузов РФ (по 50-100 студентов из каждого вуза). Тестовые задания отобраны из вступительных тестов по математике других стран и предназначены для оценки степени готовности к обучению в вузе на технической специальности. Цель тестирования – проверить функционирование заданий на российской выборке.

Просим Вас проанализировать отобранные задания и дать их оценку в соответствии с инструкцией.

Спасибо Вам за участие!

**Инструкция по оцениванию заданий**

На следующей странице представлены задания по математике. Оцените, пожалуйста, каждое из них в отдельности по четырем критериям, представленным ниже. Для оценки просим использовать предлагаемую пятибалльную шкалу.

1. Важна или не важна тема, к которой относится это задание, для овладения всей программой подготовки специалистов в инженерной области?

(совсем не важна) 1 ………. 5 (очень важна)

1. Трудным или легким является данное задание для типичного выпускника российской школы, поступающего на техническую специальность вуза?

(очень легкое задание) 1 ………. 5 (очень трудное задание)

1. Какой процент студентов, зачисленных в целом на первый курс инженерных специальностей Вашего вуза, может решить это задание?

(очень малый, <20) 1 ………. 5 (большой, >80)

1. Насколько подробно тема, к которой относится данное задание, изучается в курсе высшей математики в вузе?

(совсем не изучается ) 1 ………. 5 (изучается очень подробно)

*Примечание.* Мы будем благодарны, если Вы укажите задания, формулировки которых вызывают сомнения в их корректности и ясности для студентов, а также, если Вы сможете предложить альтернативу. Сделать это можно, используя графу «Комментарии» после каждого задания.

Желтым цветом помечен правильный ответ. Если вы считаете правильным другой вариант ответа, пожалуйста, укажите это в комментариях.

Три задания (№ 9, 15 и 30) выделены голубым цветом. Это – задания, включение которых в тест остается под вопросом.

В заключение мы просим Вас ответить на несколько вопросов анкеты эксперта.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | Задание | 1 | 2 | 3 | 4 |
| 1 | Пусть и - два единичных вектора. Если векторы = + 2 и = 5 - 4 перпендикулярны друг другу, то угол между и равен:1) π/3 2) π/4 3) π/6 4) π/2  |  |  |  |  |
| *Комментарии*  |
| 2 | Если: , то а = 1) 1 2) 2 3) -1 4) -2 |  |  |  |  |
| *Комментарии* |
| 3 | Дана функция .*Утверждение* 1: *f* '(4) = 0*Утверждение* 2: *f* (*x*) непрерывна на отрезке [2;5], дифференцируема на интервале (2;5), и *f* (2) = *f*(5).1. Утверждение 1 верно, утверждение 2 верно. Утверждение 1 не является следствием утверждения 2.
2. Утверждение 1 верно, утверждение 2 неверно.
3. Утверждение 1 неверно, утверждение 2 верно.
4. Утверждение 1 верно, утверждение 2 верно. Утверждение 1 является следствием утверждения 2.

 |  |  |  |  |
| *Комментарии* |
| 4 | При каком *k*, точка, делящая отрезок, соединяющий точки (1,1) и (2,4), в соотношении 3:2, принадлежит линии *2x + y = k*?1) 6 2) 11/5 3) 29/5 4) 5 |  |  |  |  |
| *Комментарии* |
| 5 | *Утверждение* 1. Прямая *y =* 2*x +*  является общей касательной параболы *y2 =* 16 *x* и эллипса 2*x2 + y2 =* 4*.* *Утверждение* 2. Если прямая *y = mx + (m* ≠ 0) является общей касательной параболы *y2 =* 16 *x* и эллипса 2*x2 + y2 =* 4, то *m* удовлетворяет уравнению *m4 +* 2*m2 =* 24. 1. Утверждение 1 верно, утверждение 2 верно. Утверждение 1 не является следствием утверждения 2.
2. Утверждение 1 верно, утверждение 2 неверно.
3. Утверждение 1 неверно, утверждение 2 верно.
4. Утверждение 1 верно, утверждение 2 верно. Утверждение 1 является следствием утверждения 2.
 |  |  |  |  |

# *Приложение 2*. Образец инструкции и бланка для оценки тем

**Исследование элементов содержания дисциплины «Математика» для проведения оценки готовности студентов инженерных специальностей к обучению в вузе.**

(форма эксперта)

Уважаемый эксперт,

приглашаем Вас принять участие в работе по разработке инструмента, позволяющего оценить готовность студентов инженерных специальностей к обучению в вузе. Данный инструмент предполагается использовать для проведения сравнительного исследования готовности к обучению в вузе студентов технических специальностей в странах БРИК, а также качества их подготовки по истечении двух лет обучения в вузе.

Разработчиками тестовых заданий был предложен список тем по математике на основе программы средней школы. На данном этапе мы просим Вас оценить степень важности этих тем и требуемый начальный уровень освоения материала для дальнейшего успешного обучения в вузе на технической специальности.

Каждую из перечисленных тем мы просим оценить по трём критериям и дать развёрнутое пояснение своей оценки. Все темы оцениваются по одним и тем же критериям.

Критерий 1. Насколько важна данная тема для овладения программой подготовки специалистов в инженерной области?

Критерий 2. Каков необходимый минимальный уровень освоения данной темы для успешного дальнейшего обучения в вузе?

Критерий №3. Насколько хорошо сформулирована тема и основные контролируемые элементы (в контексте целей данного исследования)?

**Тема 1. Степени и корни**

Контролируемые элементы

знать: понятие корня n-ой степени

уметь: выполнять тождественные преобразования с корнями и находить их значение

Критерий №1

*Насколько важна данная тема для овладения программой подготовки специалистов в инженерной области? (отметьте подходящий вариант)*

1 – Тема не важна

2 – Тема относительно невысокой важности

3 – Тема относительно высокой важности

4 – Тема очень важна

*Пожалуйста, дайте краткое объяснение тому, почему эта тема важна (или не важна) для подготовки специалистов в инженерной области?*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

Критерий №2

*Каков необходимый минимальный уровень освоения данной темы для успешного дальнейшего обучения в вузе? (отметьте подходящий вариант)*

0 – Освоение темы не требуется

1 – **Воспроизведение** основных понятий (фактов, свойств, правил)

2 – **Применение** стандартных алгоритмов для решения задач. Способность к самостоятельной формулировке проблемы задачи и распознаванию информации, необходимой для её решения.

3 – **Математические размышления**: составление математических моделей для решения задач. Решение сложных задач, требующих критически осмыслить проблемную ситуацию, математизировать ее, решить задачу и дать аргументированную интерпретацию решения.

Критерий №3

*Насколько хорошо сформулирована тема и основные контролируемые элементы (в контексте целей данного исследования)? (отметьте подходящий вариант)*

1 – Сформулированы хорошо

2 – Требуют уточнения

*Если формулировка темы и основные контролируемые элементы требуют уточнения, напишите, пожалуйста, какие именно уточнения требуются? Другими словами, что должен знать и уметь студент в рамках данной темы для успешного обучения в вузе?*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*