**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего профессионального образования**

**"Национальный исследовательский университет
"Высшая школа экономики"**

Институт развития образования

Магистерская программа «Измерения в психологии и образовании»

**Магистерская диссертация**

На тему

**«Анализ кросс-культурной сопоставимости результатов тестирования
 на основе данных SAM»**

 Студентка группы № 701

Брун Ирина Викторовна

Научный руководитель

Кандидат физико-математических наук, доцент

Карданова Елена Юрьевна

Москва, 2014 г.

Оглавление

[Введение 3](#_Toc390170914)

[Глава 1. Методы установления эквивалентности в мировой практике кросс-культурных исследований 6](#_Toc390170915)

[1.1 Эквивалентность 6](#_Toc390170916)

[1.1.1 Конструктная эквивалентность 7](#_Toc390170917)

[1.1.2 Эквивалентность метода 8](#_Toc390170918)

[1.1.3 Эквивалентность заданий 10](#_Toc390170919)

[1.2 Стандарты в области кросс-культурных исследований 15](#_Toc390170920)

[1.3 Типы кросс-культурных исследований 19](#_Toc390170921)

[1.4 Инструмент SAM 23](#_Toc390170922)

[Выводы 25](#_Toc390170923)

[Глава 2. Исследование сопоставимости результатов тестирования SAM в Новгородской области и Таджикистане 26](#_Toc390170924)

[Введение 26](#_Toc390170925)

[2.1 Подготовительный этап 28](#_Toc390170926)

[2.1.1 Выборка тестирования 28](#_Toc390170927)

[2.1.2 Шкалирование результатов тестирования 28](#_Toc390170928)

[2.1.3 Создание выборки Новгородской области 31](#_Toc390170929)

[2.1.4 Психометрический анализ данных 33](#_Toc390170930)

[2.2 Содержательный этап 39](#_Toc390170931)

[2.2.1 Установление Конструктной эквивалентности 39](#_Toc390170932)

[2.2.2 Психометрический анализ объединённых данных 48](#_Toc390170933)

[2.2.3 Эквивалентность заданий 49](#_Toc390170934)

[2.2.4 Построение единой шкалы 52](#_Toc390170935)

[Выводы 58](#_Toc390170936)

[Заключение 60](#_Toc390170937)

[Список использованных источников: 62](#_Toc390170938)

[Приложение 1. Наиболее ожидаемые типы ошибок 65](#_Toc390170939)

[Приложение 2. Разделы содержания целевого конструкта SAM 66](#_Toc390170940)

[Приложение 3. Карта заданий и испытуемых 67](#_Toc390170941)

# Введение

Количество кросс-культурных исследований со временем увеличивается. Политиков и работников в сфере образования интересует вопрос, насколько хорошо дети в стране знают математику? Или в какой стране школьники больше всего читают? Для того, чтобы ответить на подобные вопросы, необходимы измерительные инструменты, которые позволили бы сопоставлять и ранжировать разные стран по какому-либо признаку. На сегодняшний день существует множество кросс-культурных исследований, особенно в области образовательных достижений (Hambleton, de Jon, 2003). Какой бы признак ни измеряли эти исследования, их суть одинакова – обеспечить возможность сравнения результатов разных стран. Однако, до сих пор не существует единого принятого ряда мер, соблюдение которых гарантировало бы правомерность сопоставления результатов разных стран. В кросс-культурных исследованиях можно выделить 2 типа: исследования, которые изначально создавались для применения в нескольких странах, исследования, которые создавались для применения в одной стране, но затем были адаптированы для использования в других странах. Что касается первого типа, то здесь существуют довольно полные указания и предписания по созданию единой шкалы для разных стран, поскольку в таких исследованиях есть возможность заложить специальные процедуры на этапе разработке инструмента. Доказательство сопоставимости результатов тестирования, которое не задумывалось как кросс-культурное описано в литературе гораздо хуже. Некоторые исследователи придерживаются точки зрения, что результаты разных стран по одним и тем же адаптированным опросникам достижений (а также интеллекта) не сравнимы в принципе (см. Greenfield, 1997).

В основном кросс-культурные исследования используют в качестве измерительного инструмента какие-либо стандартизированные методики. Поскольку результаты страны по итогам измерения планируется сообщать и сравнивать с результатами других стран, необходимо сделать измерение массовым, т.е. отражающим общую тенденцию страны. Когда измерительный инструмента (тест) создается в одной стране, а применяется в другой, в другом культурном контексте, необходимо доказать инвариантность психометрических характеристик теста (включая надёжность и валидность). При этом недостаточно будет просто показать, что тест имеет хорошие психометрические свойства в каждой отдельной стране, необходимо также доказать, что баллы, полученные за тест, значат *одно и то же* в разных странах (Vijver, Poortinga, 1997). Также для возможности сравнения результатов необходимо создать единую шкалу, на которую могут быть «положены» результаты различных стран (Poortinga, 1989). Когда тест проводится в одной стране (назовём её страной А), результаты (то есть информация о некоторой латентной переменной, которую стремится измерить исследователь) помещаются на определённую шкалу (назовём её шкалой А). Размерность этой шкалы зависит от множества факторов, включающих и модель, в рамках которой создавался тест. Представим себе, что этот же тест проведен в стране Б и полученные здесь результаты помещаются на шкалу Б. Сам факт сопоставления результатов этих двух стран означает, что шкалы А и Б равны, то есть одинаковы. Сюда хорошо подходит простой пример: предположим, исследователь хочет сравнить длину ладони в странах А и Б. Для этого измеряются ладони в стране А (в сантиметрах) и в стране Б (в дюймах). Очевидно, что исследуемый конструкт – длина ладони одинаковый в обоих странах, но он также должен выражаться в одних и тех же единицах в обеих странах. Поскольку сравнение длины ладони в дюймах с длиной ладони в сантиметрах даст некорректные результаты. Безусловно, данный пример несколько гипертрофирует ситуацию, в которой оказывается исследователь, поскольку разница единиц измерения здесь болезненно очевидна, а соотношение единиц известно и можно добиться сопоставимости результатов. Однако, эта очевидность пропадает, когда исследователь занимается измерением какого-либо латентного параметра (Poortinga, 1989). Поэтому исследователям, занимающимся кросс-культурными сравнениям в области социальных наук следует быть особенно внимательными в вопросе доказательства сравнимости результатов. Стоит оговориться, что в данной работе под кросс-культурным исследованием понимается исследование, результаты которого сравниваются в двух культурах, в которых тестирование было проведено на ***разных языках****,* поскольку язык является одним из определяющих свойств культуры (Ellis, 1989).

**Целью** данной работы является сопоставление результатов кросс-культурного исследования, которое изначально не разрабатывалось как кросс-культурное на данных SAM в Новгородской области и Таджикистане. **Исследовательский вопрос**, который мы ставим перед собой в данной работе: каким образом можно добиться сопоставимости результатов тестирования SAM и создать единую шкалу для Новгородской области и Таджикистана?

Для того чтобы достичь поставленной цели необходимо решить ряд **задач**:

1. Проанализировать мировой опыт установления сопоставимости результатов кросс-культурных исследований,
2. Выделить ряд мер установления сопоставимости, подходящих для опросника SAM,
3. Проанализировать результаты тестирования SAM в Новгородской области и Таджикистане с тем, чтобы установить возможность сопоставления.
4. Дополнить ряд мер, которые можно применять для установления сопоставимости результатов тестирований.

**Объектом** данного исследования являются результаты тестирования SAM, полученные в Таджикистане и Новгородской области. **Предметом** исследования является сопоставимость (эквивалентность) результатов SAM в указанных странах. Основная **гипотеза** исследования: в Новгородской области и Таджикистане невозможно достичь сопоставимости результатов.

Обратимся к более подробному рассмотрению комплекса мер, позволяющих говорит о сопоставимости результатов исследований, полученных в различных культурах. Ключевым здесь является понятие эквивалентности.

# Глава 1. Методы установления эквивалентности в мировой практике кросс-культурных исследований

## 1.1 Эквивалентность

Как указывалось ранее, одна из важнейших задач кросс-культурных исследований – обеспечение сопоставимости результатов, полученных в различных культурах и языках. Для исследователей это означает, что тесты на разных языках должны быть ***эквивалентны*** (AERA, APA & NCME, 1999). Понятие эквивалентности зеркально понятию ошибки (bias). Ошибка появляется, когда различия индикаторов измеряемого конструкта не связаны с разницей способностей (или другого целевого конструкта).

Приведем простой пример. Предположим, в опроснике спрашивается о том, какой город является столицей Польши. Целевым конструктом здесь является знание по географии о столице Польши. Однако, можно предположить, что при ответе на этот вопрос будет также играть роль и географическое местоположение опрашиваемого: так, жители Европы скорее всего будут отвечать на этот вопрос лучше, нежели жители Африки. Таким образом, процент людей, знающих, что столица Польши – Варшава в Европе и Африке будет различаться не только из-за знаний географии, но и из-за места жительства опрашиваемых (Vijver, Tanzer, 2004). Это и есть ошибка.

Исследователи выделяют три типа ошибки (Vijver, Hambleton, 1996):

* Ошибка конструкта (construct bias),
* Ошибка метода (method bias),
* Ошибка заданий (item bias).

Рассмотрим каждый из видов ошибок более подробно, а также определим статистические процедуры, используемые для выявления различных типов ошибок.

### 1.1.1 Конструктная эквивалентность

Конструктная эквивалентность достигается за счет доказательства отсутствия ошибки конструкта. Ошибка конструкта появляется тогда, когда измеряемый, целевой конструкт различается в разных странах (Vijver, Hambleton, 1996). Этот тип ошибок корениться в операционализации конструктов. Поэтому такая ошибка чаще всего появляется в исследованиях, которые изначально планировались как локальные, и заранее не предполагалось, что данный инструмент будет использоваться в различных культурах. Например, в одной из работ Хо было показано, что понятие «быть хорошей дочерью/сыном» в Китае гораздо шире, нежели в большинстве европейских стран, поэтому использование одного и того же измерительного инструмента в данном случае даст сдвинутые оценки, и, как следствие, приведет к невалидной интерпретации результатов. Этот тип исследований будет подробно рассмотрен в разделе «типы кросс-культурных исследований».

Для того чтобы избежать появления ошибок этого рода, можно применять различные процедуры. До того, как начать полевой этап можно использовать группу экспертов из различных культур, которые должны будут оценить, насколько схожим является операционализация конструкта в исследуемых культурах. По сути, этот процесс обратен операционализации. Используя задания теста, эксперты должны выстроить «карту конструкта».

Если же у исследователей нет возможности исследовать конструкт в различных культурах до тестирования, то такое исследование можно произвести и пост фактум. В литературе выделяется 4 статистических процедуры:

1. Эксплораторный факторный анализ (exploratory factor analysis),
2. Конфирматорный факторный анализ (confirmatory factor analysis),
3. Многомерное шкалирование (multidimensional scaling),
4. Сравнение номологических карт (comparison of nomological networks) (Vijver, Tanzer, 2004)

Первые три процедуры позволяют доказать эквивалентность структуры конструкта. Однако, эквивалентность структуры является необходимым, но не достаточным условием конструктной эквивалентности. Нужно также показать, что эта структура функционирует одинаково в разных культурах. Для того, чтобы доказать конструктную эквивалентность, необходимо также провести исследование связи между тестовыми баллами исследуемого инструмента в разных странах и какими-либо внешними переменными, о которых известна гипотетическая связь с исследуемым конструктом.

Стоит отметить, что на сегодняшний день опубликовано мало исследований, в которых приводится подробное описание исследования конструктной эквивалентности.

### 1.1.2 Эквивалентность метода

Как указывалось ранее, эквивалентность – понятие обратное ошибке, поэтому эквивалентность метода означает отсутствие ошибок метода. Ошибки метода представить себе очень легко. Это всё, что относится к администрированию (проведению) теста на выборке. Сюда входят различия в социальной желательности, знакомство со стимулом и форматом вариантов ответа, различия в коммуникации между интервьюерами и респондентами и т.д. Обычно, если появляется ошибка метода, то она влияет на все задания теста, вызывая суммарный «сдвиг» в оценках культурных групп.

Выделяют 3 типа ошибок, препятствующие достижению эквивалентности метода:

* Ошибка выборки (sample bias)
* Ошибка инструмента (instrument bias)
* Ошибка проведения (administration bias) (Vijver, Tanzer, 2004)

Под ошибкой выборки понимаются различные характеристики выборки, которые не связаны с измеряемым конструктом (например, различия в мотивации или социально-экономическом статусе). Ошибку выборки можно выявить с помощью ряда процедур, среди которых ковариационнный и регрессионный анализы, и randomized-block design.

Ошибка инструмента появляется в ситуации, когда различные выборки в различной степени знакомы с форматом проводимого тестирования. Ошибка проведения включает в себя ситуации, в которых нарушается процедура проведения тестирования (например, в одной из групп администраторы неправильно поняли инструкции, имели место проблемы с оборудованием и т.п.). Для того чтобы оценить наличие ошибки инструмента и проведения существует как минимум 3 процедуры:

* Исследование одного конструкта разными методами (monotrait-multimethod study)
* Использование дополнительных конструктов, связанных с исследуемым (collateral information)
* Исследование изменений (Vijver, Tanzer, 2004).

Monotrait-multumethod studies подразумевают, что один и тот же целевой конструкт измеряется разными методами и затем оценивается согласованность полученных оценок. Основным недостатком такой процедуры можно считать большую затратность (по сути, требуется найти такой измерительный инструмент, о котором есть надежные психометрические данные, и который также будет измерять тот же самый конструкт) и невозможность проведения постфактум. Collateral information используется схожим образом, и при этом не требует наличия измерительных инструментов для того же конструкта; требуется наличие измерительного инструмента для любого конструкта, с которым целевой конструкт связан определённым образом. Третья процедура предлагает повторный опрос испытуемых тем же методом. При невысокой надёжности проведённых замеров можно говорить о наличии ошибки инструмента или проведения (Vijver, Tanzer, 2004).

Стоит отметить, что все процедуры для выявления ошибки инструмента и проведения являются очень затратными и требуют специальной подготовки, которая должна быть реализована до сбора данных.

### 1.1.3 Эквивалентность заданий

Большинство исследований под эквивалентностью заданий понимают одинаковое функционирование заданий в разных культурных группах. Другими словами – отсутствие DIF (differential Item Functioning) (Ellis, 1989). DIF появляется в ситуации, когда испытуемые из одной группы имеют более высокую вероятность правильно ответить на задание, чем испытуемые из другой группы с тем же уровнем способностей (Zumbo, 2007, Ellis 1989). В нашем случае это определение можно конкретизировать: DIF появляется в ситуации, когда испытуемые, принадлежащие к одной культуре имеют более высокую вероятность ответить на задание правильно, по сравнению с испытуемыми из другой культуры, при том, что у этих испытуемых одинаковых уровень подготовленности (или одинаковый уровень выраженности признака, если речь идёт о не когнитивных опросниках).

В настоящее время разработано множество методов выявления DIF. Одна из наиболее распространённых классификаций делит методы на параметрические, непараметрические и многомерные (Zumbo, 2007). Под параметрическими методами в данном случае понимаются методы, использующие логику и параметризацию IRT (включая Rasch) (Wang, Su, 2004; McNamara, Roever, 2006; Karami, 2012; Magis, Beland, Tuerlinckx, De Boeck, 2010). Несмотря на то, что DIF-анализ проводится уже более 50 лет, нет единого мнения о том, какой метод является наиболее эффективным, поскольку различные методы выявления DIF реагируют на различные параметры, такие, как: длина теста, количество заданий, демонстрирующих DIF, величина DIF, и размер выборки, и соотношение размеров групп, в отношении которых проверяется DIF (Rogers, Swaminathan, 1993). В данной работе мы будем использовать 4 наиболее популярных метода выявления DIF, которые относятся к группе непараметрических:

1. Mantel-Haenszel,
2. Standardization,
3. Logistic Regression,
4. t-statistic.

Рассмотрим кратко каждый из методов с указанием их специфики.

**Mantel-Haenszel**

Данный метод, по сути, основывается на трехмерных таблицах сопряженности. Все испытуемые делятся на группы в зависимости от общего балла за тест, затем по правильности ответа на исследуемое задание, и, наконец, по принадлежности к референтной группе (Millsap, Everson, 1993). Тестируется гипотеза Но: DIF=0, оценка вероятности справедливости Но гипотезы происходит на основе статистики хи-квадрат с одной степенью свободы (Uttaro, Millsap, 1994).

Для оценки величины DIF применяется ещё показатель , который измерен на метрической шкале Educational Testing Service’s delta scale. Значение 0 на этой метрической шкале указывает на отсутствие DIF. Отрицательное значение указывает на то, что данное задание хуже решается испытуемыми из фокальной группы. Положительное значение говорит о том, что для испытуемых из референтной группы вероятность ответить правильно на данное задание меньше, чем для фокальной группы (Chiu; Millsap, Everson,1993).

Если показатель , то считается, что размер DIF достаточно мал и им можно пренебречь.

Если показатель  то DIF считается среднего размера.

Если показатель  то величина DIF считается большой (difR package manual)

По приведённым формулам видно, что данный метод достаточно прост в вычислении, что обуславливает его распространённость. Это свойство является важным достоинством данного метода, также как и то, что Мантель-Ханцель показывает надежные результаты на относительно небольших выборках (около 200 испытуемых) (Clauser, Mazor, 1998; Scheuneman, Bleistein, 1989).

**Standardization**

Dorans & Kulick (1986) разработали простой математический метод, позволяющий судить о наличии DIF по вероятности правильного ответа на задание двух подгрупп испытуемых. Вся логика данного метода исходит из того, что можно утверждать, что DIF отсутствует тогда, когда вероятность правильно ответить на задание для одной из подгрупп испытуемых равна вероятности правильно ответить на то же задания для другой подгруппы испытуемых (при контроле уровня подготовленности в обеих группах) (Dorans, Kulick 1986; Karami, 2012). На основе разницы вероятностей строится индекс квадратный корень из взвешенной разницы (Root Mean Weighted Squared Difference - RMWSD) (более подробно см. Dorans, Kulick, 1986, стр 359; Dorans, Schmitt, 1988). RMWSD - колеблется в интервале от -1 до 1, где отрицательное значение говорит о том, что испытуемые в референтной группе имеют преимущество перед испытуемыми из фокальной группы. Положительное значение, соответственно, является признаком того, что у испытуемых из фокальной группы вероятность правильно ответить на исследуемое задание выше. Если , то принято считать, что задание не демонстрирует DIF (Millsap, Everson 1993). В случае, когда индекс колеблется в пределах , задание следует дополнительно проверить, во всех остальных случая принято считать, что задание демонстрирует DIF. Стоит отметить, что у данного метода выявления DIF есть недостатки, наиболее важный из них – у данного метода отсутствует статистический критерий значимости, в отличие от статистики Мантель-Ханцель (Karami, 2012; Clauser & Mazor 1998).

**Logistic Regression**

Данный метод появился вследствие публикаций исследований, доказывающих неэффективность статистики Мантель-Ханцель при выявлении неоднородного DIF. Авторы данного метода (Swaminathan and Rogers, 1990) ставили перед собой цель создания такого инструментария, который был бы также же эффективен, как Мантель-Ханцель при выявлении однородного DIF, обладал мощностью для выявления неоднородного DIF и при этом не использовал параметры IRT (Gómez-Benito, Hidalgo, Padilla, 2009). Общая идея проста: строится уравнение логистической регрессии, куда принадлежность к референтной или фокальной группе включается как предиктор (также в качестве предикторов выступают трудность задания и дискриминативность).

Тестируется несколько гипотез, которые позволяют сделать вывод о наличии и направлении DIF (какая из групп получает необоснованное преимущество), а также определить какой DIF присутствует: однородный или неоднородный (используется распределение хи-квадрат с двумя степенями свободы). Поскольку понятие однородности DIF приводится в данной работе впервые, приведем определение. Если испытуемым одной из групп труднее отвечать правильно на задание (то есть различается только параметр трудности), то такой DIF называют однородным (uniform) (Clauser, Mazor, 1998; Zumbo, 2007). Неоднородный DIF (nonuniform), возникает в ситуации, когда между группами наблюдаются различия по параметру дискриминативности (и, возможно, трудности) – рис.2 (Clauser, Mazor, 1998; Zumbo, 2007; Magis, De Boeck, 2012, стр. 300).

**t-statistic**

Последний метод выявления DIF рассчитывался нами вручную, поэтому описание расчётов содержится во второй главе данной работы.

Если тест разрабатывался в рамках парадигмы Rasch моделирования (или IRT моделирования, в данном случае различия между этими подходами не неважны) то исследования item bias начинается с построения с построения ICC (Item Characteristic Curve) отдельно для каждого задания и для каждой страны. ICC в общем виде представляет собой график зависимости между способностями (подготовленностью $– θ$) и вероятность правильного ответа на задание (или вероятности выбрать ту или иную категорию, если используются шкалы типа Ликерта). ICC монотонно возрастает и имеет S-образную форму.

Затем, ICC сравниваются между странами. Если кривые совпадают, то можно говорить о том, что задание в странах функционирует одинаково, если кривые различаются – присутствует DIF.

Однако, напрямую сравнивать характеристические кривые по странам нельзя. Для того, чтобы можно было сравнивать ICC нужно сначала положить оценённые способности и характеристики заданий в разных выборках на одну шкалу. Лорд и Стокинг (1983) пишут, результаты одного и того же теста, проведённого на разных выборках будут различны (оценённые параметры будут различаться), поскольку метрическая шкала параметров создается отдельно для каждой выборки, т.е. параметры калибруются на основе данных по выборке. Поэтому прежде, чем сравнивать результаты требуется создать единую шкалу, на которую бы помещались параметры, полученные на разных выборках. Другими словами, параметры двух тестов могут считаться инвариантными, когда они основываются на одной шкале способностей испытуемых (Ellis, 1989).

Для теста, проведённого на двух выборках, будут различаться параметры заданий (если модель однопараметрическая – то трудность, если двухпараметрическая – то трудность и дискриминативность), но вероятность правильного ответа на задание останется прежней. По сути, параметры заданий теста, полученные на разных выборках, будут иметь линейную зависимость, поскольку различаются только «центры» шкалы (т.е. относительный 0 шкалы) (Stocking, Lord, 1983). В общем, процедура приведения к одному основанию заключается в поиске двух констант (назовём их А и В) линейной функции, которая отражает перевод параметра задания, полученного на первой выборке, к параметру задания, общему для двух выборок. То же самое должно быть проделано и для оценки подготовленности испытуемых. Казалось бы, это можно сделать просто: поместить два параметра задания, полученных на разных выборках на плоскость, провести между ними линию и вычислить А и В. И так действительно можно было бы сделать, если бы параметры, между которыми проводится регрессионная прямая были бы истинными, то есть оценёнными без ошибки. Однако, это не так, и в руках исследователя оказываются оценки параметров, включающие ошибку измерения.

Лорд и Стокинг (1983) разработали две процедуры поиска А и В, позволяющих нивелировать недостатки процедур, предложенных ранее (см. стр. 203). Это процедуры «robust mean» и «sigma method» (Stocking, Lord, 1983). Подробный план применения этих методов с описанием их основных преимуществ изложен в статье Лорда и Стокинга (1983), поэтому здесь мы не будем подробно останавливаться на этом.

После проведения процедуры уравнивания параметров, построение характеристических кривых заданий в одном пространстве становится правомерным. Анализ функционирования характеристических кривых обычно применяется исследователями в комплексе с расчетом других показателей DIF (как параметрических. Так и не параметрических), поскольку ICC наглядно показывают, совпадают или различаются кривые, но при этом не дают информации о том, в пользу какой из выборок функционирует задание , а также не дают информации о статистической значимости различий выборок.

## 1.2 Стандарты в области кросс-культурных исследований

В 90-ых годах XX века были разработаны стандарты психологического и образовательного тестирования. Примерно в одно и то же время появились два различных стандарта по созданию, проведению, валидизации и адаптации теста.

В 1999 году American Educational Research Association, American Phychological Association, National Council on Measurement in Education (AERA, APA & NCME, 1999) совместно создали Стандарты тестирования в области образования и психологии (Standards for Educational and Psychological Testing). Данные стандарты содержат рекомендации не только по адаптации тестов на другие языки, но и по разработке теста в целом. Остановимся более подробно на главе, посвящённой тестированию индивидов в различной лингвистической среде (AERA, APA & NCME, 1999, стр. 91). В данных стандартах указывается, что для того, чтобы перевести тест на другой язык, недостаточно выполнить перевод заданий. Нужно также показать эквивалентность в содержании и трудности заданий, а также в надёжности и валидности тестовых форм на разных языках (AERA, APA & NCME, 1999, стр. 92). Интересно, что двойной слепой перевод НЕ советуют применять при переводе теста на другой язык. Более подходящим считается итерационный подход, при котором эксперты-билингвы стараются не наиболее точно перевести слово, а подобрать наиболее эквивалентное слово: по смыслу, частоте употребления (распространённости в языке), числу слогов и т.д. Также предлагается выбирать наиболее подходящие типы вопросов, варианты ответов, и время заполнения теста для языка, на который осуществляется перевод (AERA, APA & NCME, 1999, стр. 92).

Стандарт **9.4** описывает необходимость для разработчиков теста показывать необходимость внесённых лингвистических изменений. Внесенные в перевод изменения должны учитываться при интерпретации тестового балла.

Стандарт **9.5** рекомендует разработчикам теста давать специальные рекомендации при интерпретации тестовых баллов, если было выявлено, что версии на разных языках неэквивалентны.

Стандарт **9.2** предписывает проводить специальное психометрическое исследование на выборках каждого языка отдельно, если было выявлено, что версии на разных языках неэквивалентны.

Стандарт **9.7** указывает на необходимость публиковать все данные об установлении эквивалентности между версиями теста на разных языках (AERA, APA & NCME, 1999, стр.98-99).

Стандарт **9.9** указывает на то, что если тесты на разных языках создаются с интенцией сравнения результатов между языковыми группами, то разработчики теста должны предоставить доказательство возможности сравнения данных тестов.

В 1992 году Internationl Test Commision (ITC) начала проект по разработке рекомендаций для перевода и адаптации тестов на другой язык. В разработке этих рекомендаций принимали участие 7 организаций. Впоследствии эти рекомендации были апробированы и одобрены ITC (ITC 2010, Hambleton, de Jon, 2003). Эти рекомендации можно разделить на 4 категории:

1. контекст,
2. рекомендации по переводу и адаптации,
3. администрирование
4. документация и интерпретация результатов.

Рассмотрим эти рекомендации более подробно.

1. **Контекст:**

**C 1.** Исследование специфических культурных эффектов, которые не связаны непосредственно с измеряемым конструктом (или не важны); их минимизация.

**C 2.** Оценка того, насколько «пересекаются» измеряемые конструкты в различных культурах (в которых будет применяться данный тест).

1. **Перевод и адаптация опросника:**

**D 1.** Создатели теста должны учитывать языковые и культурные особенности групп испытуемых, которые могут проявиться в заданиях теста.

**D 2.** Создатели теста должны продемонстрировать, что язык (терминология) должен быть приемлемым для всех культур, в которых планируется использовать данный тест.

**D 3.** Создатели теста должны продемонстрировать, что формат заданий, техники и процедуры, также правила тестирования знакомы всем культурным группам, которые будут тестироваться.

**D 4.** Создатели теста должны предоставить свидетельства того, что содержание заданий и другие стимулы известны различным культурным группам, в которых планируется проводить тестирование.

**D 5.** Создатели теста должны предоставить свидетельства эквивалентности теста в разных культурных группах.

**D 6.** Создатели теста должны быть уверенными, что техники сбора данных позволят в дальнейшем оценивать эквивалентность заданий в разных культурных группах.

**D 7.** Создатели теста должны применять подходящие статистические процедуры, которые позволяют установить эквивалентность заданий и выявить компоненты, которые различаются в различных культурных группах.

**D 8.** Создатели теста должны предоставлять данные о валидности методики во всех культурных группах, в которых планируется использовать этот тест.

**D 9.** Создатели теста должны предоставить статистические доказательства эквивалентности заданий во всех культурных группах, в которых планируется использовать тест.

**D 10.** Если были выявлены задания, которые не являются эквивалентными во всех популяциях, эти задания не должны использоваться для построения единой шкалы баллов для всех популяций, а также не должны быть использованы при сравнение популяций.

1. **Администрирование:**

**A 1.** Создатели теста и те, кто его проводят должны использовать все имеющиеся средства для того, чтобы избежать проблем при подготовке материалов и инструкций.

**A 2.** Группа специалистов, проводящих тестирование должны быть особенно внимательны к факторам, которые могут повлиять на восприятие стимульного материала, и процедуру проведения тестирования, и, таким образом, поставить под угрозу валидность получаемых данных.

**A 3.** Все средовые характеристики (характеристики места проведения тестирования) должны быть как можно более близкими между исследуемыми популяциями.

**A 4.** Инструкции экзаменаторам должны быть составлены таким образом, чтобы минимизировать влияние побочных факторов на ответы испытуемых.

**A 5.** Все аспекты процедуры тестирования, которые могут повлиять на результаты, должны быть чётко прописаны в мануале теста.

**A 6.** Взаимодействие тех, кто проводит тестирование и испытуемых должно быть минимизировано. Все возможные аспекты взаимодействий должны быть чётко прописаны в мануале.

1. **Документация и интерпретация баллов:**

**I 1.** При переводе теста для использования в другой культурной среде все изменения должны быть задокументированы, так же, как и доказательство эквивалентности полученных тестовых форм.

**I 2.** Разницу баллов, полученных в различных культурах нельзя интерпретировать как разницу в достижениях разных культурных групп. Обязанностью исследователей является проверка того, что разница баллов действительно отражает разницу в целевом конструкте.

**I 3.** Сравнение результатов тестирования в разных культурах может быть проведено, только при наличии эквивалентности тестов в этих культурах.

**I 4.** Создатели теста должны предоставить информацию о том, в какой мере культурные особенности популяций могут повлиять на результаты тестирования, а также провести исследование того, каким образом эти особенности могут быть учтены при интерпретации результатов (ITC 2010, Hambleton, de Jon, 2003).

Эти рекомендации довольно новые, в 2010 году были внесены последние поправки (в раздел о переводе и адаптации). В целом, эти рекомендации ITC охватывают три типа эквивалентности, которые были рассмотрены в данной работе. На наш взгляд, наиболее ценная информация состоит в том, что эквивалентность тестовых форм в различных культурах должна быть не только доказана, но и опубликована. Также эти рекомендации дают исследователям из различных стран «общее основание» для создания кросс-культурных исследований, способствуя, таким образом, интеграции знаний в области образовательных достижений и психологического тестирования.

Подводя итог анализу стандартов в области образовательного и психологического тестирования, необходимо отметить, что все стандарты и пояснения к ним написаны чётко, но оставляют возможности для интерпретации. Например, в комментариях к стандарту 9.9 написано, что такое доказательство должно включать в себя показатели надежности, валидности, а также сходства конструктов, но не должны ограничиваться этими показателями. Таким образом, процедура установления эквивалентности остается на усмотрение исследователей, что подтверждает актуальность поднимаемой в работе темы.

## 1.3 Типы кросс-культурных исследований

В зависимости контекста создания кросс-культурных опросников можно выделить две категории:

1. Методики, которые изначально создавались для того, чтобы применять в различных культурах (например, PIRLS).
2. Методики, которые изначально создавались для одной популяции, но после создания методики появилась необходимость применять её в другом культурном контексте (Hambleton, de Jon, 2003).

Для опросников, относящихся к первой категории основная цель состоит в том, чтобы сравнивать результаты в различных странах (культурных контекстах). Такие тесты имеют преимущество перед представителями второй группы: здесь все формы теста (на разных языках) создаются параллельно, есть возможность учитывать культурные особенности на этапе разработки инструментария. При создании кросс-культурного исследования у разработчиков есть все возможности сделать тест эквивалентным в целевых популяциях, при помощи guide lines разработанных ITC.

Для опросников, относящихся ко второй категории, справедливо утверждение, что они создавались для оценки какого-либо параметра в одной популяции и не предполагали изначально сравнения результатов за пределами этой популяции. Это является недостатком по сравнению с опросниками из первой категории, поскольку здесь уже есть известные психометрические свойства, которые впоследствии должны быть воспроизведены в другой популяции. Это накладывает отпечаток на то, какие процедуры могут быть использованы для адаптации таких опросников.

Поскольку такой тип создания кросс-культурных исследований наиболее популярен, обратимся к более подробному анализу того, какие из методов проверки эквивалентности могут быть проведены, и есть ли здесь какие-либо особенности.

Первый вопрос, который становится перед разработчиками – действительно ли тест требует адаптации (Geisinger, 1994)? Если какой-либо инструмент планируется переводить на другой язык, то, очевидно, что требуется перевод заданий и скорее всего адаптация, поскольку другой язык во многом подразумевает другой культурный контекст. Однако, если же опросник планируется адаптировать к применению в популяции, говорящей на том же языке (как, например, MMPI, который был разработан для детей, а затем адаптирован для диагностики взрослых), то над вопросом необходимости адаптации стоит задуматься.

Если популяции действительно сильно различаются между собой в терминах культуры и языка, то в первую очередь необходимо задуматься о конструктной эквивалентности. Поскольку эквивалентности метода и заданий могут устанавливаться одинаково в обоих типах исследований. До 90-ых годов XX века это вообще не считалось проблемой (Geisinger, 1994). Эквивалентность конструктов постулировалась исходя из того, что в новой популяции применяется старый тест, а старый тест уже подразумевает определённую структуру конструкта, которая будет воспроизведена в новой популяции. Таким образом, структура конструкта воспринималась как относящаяся к опроснику, а не к реальности, в которой он применяется. Однако, довольно быстро исследователи пришли к выводу, что структуры конструкта могут различаться в различных популяциях.

Создание общей для нескольких популяций карты конструкта уже не является доступной опцией, поэтому необходимо установить степень соответствия структуры конструкта в изначальной популяции и сравнить её со структурой конструкта в новых популяциях (Mylonas, Furnham, 2013). Для этого нужно построить карту измеряемого конструкта в новых популяциях и сравнить ее с картой конструкта в старой популяции. Это уже было описано в соответствующем разделе данной работы (эквивалентность конструкта), поэтому мы не будем подробно на этом останавливаться. Отметим только, что при определении конструктной эквивалентности важно использовать и работу с экспертами, и статистические процедуры в комплексе. Это также является частью требований ITC к адаптации исследований.

Также важным является вопрос о том, что делать с обнаруженной разницей в конструктах? По стандартам ITC для международного сравнения должны использоваться только общие для разных стран «части конструкта», а специфические концепты могут применяться только для национальных оценок. Ещё один подход предлагает не сравнивать различные страны, если их концепты не полностью совпадают (Mylonas, Furnham, 2013).

Представим себе пример, которые мы уже использовали в данной работе:



***Рис.1.*** *Пример разницы в содержании конструкта.*

Если у измеряемого конструкта бОльшая часть дисперсии не является общей для всех стран, то можно ли делать сравнение всех стран только по маленькой части конструкта, которая является общей? Насколько валидным будет сравнение, при котором некоторые части концепта (которые могут быть существенными, как в примере с Китайским восприятием того, кто считается хорошей дочерью/сыном) упускаются? В имеющейся на сегодняшний момент литературе ответа на этот вопрос нет. Однако, мы предполагаем, что данный вопрос может быть разрешен с помощью конфирматорного фактора анализа, который даст оценку объяснённой и необъяснённой дисперсии (чтобы более подробно об этом говорить, необходимо произвести хотя бы первичные расчёты).

Американский исследователь Гайсенджер разработал общий алгоритм по адаптации теста, если он был создан для оценки параметра только в одной популяции:

1. Перевод и адаптация инструмента. Это может быть *перевод по заданиям*, или по *операционализации конструкта*. Обычно первый тип перевода используется, когда популяции очень похожи и контекст, включенный в описание заданий является общим для обеих популяций. Если популяции сильно отличаются, то адаптация опросника происходит на уровне концепта. Например, если в оригинальном тесте есть вопрос о том, что испытуемый любит делать по вечерам, ходить в кино или на танцы, то для какой-нибудь развивающейся африканской страны этот типично «западный» вопрос не будет функционировать адекватно. Для такой ситуации подойдет перевод на уровне концепта. Перевод должен осуществляться группой экспертов билингвов, имеющих одинаково хорошие представления о двух культурах, затем должен использоваться обратный перевод. Некоторые исследования показатели, что когда эксперты знают о том, что будет произведен обратный перевод, они специально подбирают слова таким образом, чтобы обратный перевод был более всего похож на первоначальный вариант (Hambleton, 1993).
2. Оценка качества переведенных вариантов. Гайсенджер предлагает вместо обратного перевода использовать ещё одну группу экспертов-билингвов, которые просматривают перевод, сделанный первыми экспертами. Сначала вторые эксперты индивидуально просматривают все задания и пишут письменный отзыв по каждому заданию, затем работают группой и приходят к окончательному варианту формулировок заданий.
3. Пилотное тестирование инструмента. Проверка функционирования инструментария на небольшом числе испытуемых из генеральной совокупности (т.е. новой культурной группы). Здесь выявляется также качество разработанных инструкций, приемлемость времени, числа слов в заданиях. Характеристики, полученные на пилотном исследовании не должны отличаться значимо от характеристик, полученных на первичной популяции. Если пилотажное исследование выявило какие-либо проблемы, они должны быть устранены. Успешным завершением пилотного тестирования завершается этап адаптации инструментария.

## 1.4 Инструмент SAM

SAM (Student Achievement Monitoring) является инструментом оценки учебно-предметных компетенций учащихся начальной школы (Нежнов, Карданова, 2011; Nezhnov, 2011).Он разработан на основе теории культурного развития Л.С.Выготского, по которой предполагается, что «существует три качественных уровня овладения культурным способом действия», которые можно назвать формальным, рефлексивным и функциональным (Нежнов, Карданова, Эльконин, 2011).

На первом уровне – формальном, ученики решают задачи, ориентируясь на образец, повторяют его внешние характеристики. Они могут применять полученные знания только в типовых ситуациях, по сути, применяя пройденный алгоритм действия в стандартной ситуации. Здесь решение опирается на ассоциативные связи.

На втором уровне – рефлексивном, ученики действуют более осознанно, выделяют существенный признак и с его помощью решают задание. Поэтому при таком уровне освоения материала ученики могут решать весь класс заданий, который можно решить данным способом, вне зависимости от того, описывается ли задача в стандартной форме, или сформулирована нестандартно. «Ориентировка второго уровня опирается на умственную структуру, которая фиксирует существенное отношение объектной ситуации. В психологии такие структуры обозначаются термином «гештальт», а учителя в таких случаях говорят, что ребенок начал понимать предметный материал» (Нежнов, Карданова, 2011).

Третий уровень – функциональный, отражает свободное владение материалом. При таком уровне овладения материалом ученик понимает какой способ решения задачи является наиболее подходящим, отличает подходящие способы решения от неподходящих (Нежнов, Карданова, Эльконин, 2011).

В рамках методологии SAM разработаны тесты по математике и русскому языку. Тесты разработаны для учеников, заканчивающих начальную школу (4-ый класс). Тесты имеют блочную структуру: 1-ое задание каждого блока отражает первый уровень освоения материала, 2-ое задание – второй уровень, 3-е задание – третий уровень. Такое построение заданий имеет восходящую трудность, каждый блок заданий отражает определенную тему/раздел материала. Всего каждый вариант теста содержит 15 блоков и, таким образом, 45 заданий. Благодаря такой структуре, инструмент SAM измеряет успешность освоения различных разделов учебной программы, а также диагностирует уровень усвоения материала.

Данный инструмент разрабатывался в Rasch модели. Задания оцениваются дихотомически: 0 за неправильный ответ, 1 – за правильный. Таким образом, максимальный балл составляет 45.

Разработаны тесты SAM по математике и русскому языку. Эти тесты были переведены на таджикский, казахский и киргизский языки и апробированы. Российская версия SAM прошла международный аудит и имеет полное психометрическое сопровождение.

В данной работе будут анализироваться тесты по математике (вариант 1) на русском и таджикском языках. Таджикская версия была выбрана исходя из того, что результаты таджикских учеников очень сильно отличаются от русских и встаёт вопрос о том, возможно ли сравнивать результаты SAM в этих странах. Более подробно разница в результатах будет рассматриваться во 2-ой главе данной работы.

## Выводы

Итак, в данной главе мы подробно рассмотрели вопрос о том, как добиться сопоставимости результатов кросс-культурного исследования между странами. Выделяется три типа эквивалентности, обеспечение которых дает исследователям основания считать тест одинаковым в различных культурах: это конструктная эквивалентность, эквивалентность метода и эквивалентность заданий. Также мы описали общемировые правила перевода и адаптации различных тестов.

Конструктная эквивалентность является наиболее сложной для определения, поскольку она включает в себя суждение о том, что ненаблюдаемые характеристики функционируют одинаковым образом в различных странах. Здесь выделяется ряд процедур, которые необходимо провести исследователям как до, так и после тестирования. Мы также отдельно рассмотрели вопрос о том, как показать конструктную эквивалентность в случае, когда исследование не создавалось изначально для межстранового сравнения. В таком случае исследователям остается применять статистические процедуры и рассматривать различные аспекты функционирования заданий, поскольку работа с экспертами часто оказывается недоступной. Стоит отметить, что методы доказательства эквивалентности метода и заданий разработаны куда лучше, чем методы установления конструктной эквивалентности; и в конечном итоге то, каким образом доказывать сопоставимость результатов исследований, остается на усмотрение разработчиков.

Также в первой главе данной работы мы рассмотрели инструмент SAM. Существенная разница в достижении учеников в России и Таджикистане подтолкнула нас к использованию этого инструмента для определения необходимых мер по установлению сопоставимости результатов.

# Глава 2. Исследование сопоставимости результатов тестирования SAM в Новгородской области и Таджикистане

## Введение

Инструмент SAM относится к типу кросс-культурных исследований, которые изначально разрабатывались для оценки способностей в России, но потом появилась необходимость перевести опросник на другие языки. Поскольку часть процедур адаптации не была заложена заранее, выполнение полного ряда мер, рассмотренных в 1-ой главе данной работы, не представляется возможным. Также, коррективу в процесс установления соответствия между российской и таджикской версиями опросника вносит и характер данных.

В области установления конструктной эквивалентности невозможно выполнение конфирматорного и эксплораторного факторных анализов (ЭФА и КФА). Это обусловлено тем, что в Таджикских данных слишком много пропусков (35%).

На рисунке 2 представлено распределение учеников из Таджикистана по правильным ответам (ось Х) и пропускам ответа (ось Y).

***Рис.2*** *Распределение таджикских детей по правильным ответам и пропускам ответа в заданиях.*

Как видно по рисунку, в Таджикистане отсутствуют ученики, которые дали бы больше, чем 27 правильных ответов, однако, есть дети, которые пропустили практически все задания. С учётом того, что задания оценивались дихотомически, вариации в данных недостаточно для проведения ЭФА и КФА.

Эквивалентности метода, которая предполагает эквивалентность выборки, инструмента и проведения, в случае с опросником SAM доказать невозможно, поскольку все известные процедуры установления методной эквивалентности должны быть заложены исследователем заранее.

Что касается эквивалентности заданий, то здесь есть возможность полной проверки данных SAM в двух странах на эквивалентность заданий.

Как указывалось ранее, **целью** данной работы является установление сопоставимости (эквивалентности) результатов тестирования в России и Таджикистане. Установление сопоставимости означает, что результаты SAM в этих странах можно сравнить, т.е. **результаты могут быть положены на одну шкалу**.

Вторая глава нашего исследования состоит из двух частей: предварительной и основной. Предварительная часть заключается в подготовке результатов исследования SAM, а именно:

* Создании выборки российских данных,
* Психометрическом анализе российских и таджикских данных в рамках КТТ и IRT,
* Шкалировании российских и таджикских данных по отдельности,
* Проверке российских и таджикских данных на согласие с моделью Rasch.

Основная часть второй главы посвящена установлению эквивалентности. Она включает в себя:

* Анализ «карт конструктов» в российской и таджикской версиях,
* Анализ перевода заданий на таджикский язык,
* DIF-анализ,
* Психометрический анализ объединённых российских и таджикских данных, включая согласие с моделью Rasch,
* Установление общих заданий для построения единой шкалы,
* Шкалирование результатов тестирования,
* Описательная статистика.

## 2.1 Подготовительный этап

### 2.1.1 Выборка тестирования

Анализ проводился на результатах тестирования SAM (математика, вариант 1) в Таджикистане и Новгородской области. Размер выборки в Таджикистане составил 408 ученика, в Новгородской области – 2216. В Таджикистане отбор респондентов осуществлялся методом кластерного отбора (заключительный отчет по локализации и адаптации SAM в Таджикистане). В Новгородской области была опрошена генеральная совокупность учеников 4-го класса.

База результатов тестирования SAM в Таджикистане создавалась на основе данных, представленных таджикской стороной. Готовая база, предоставленная сотрудниками CICED, не использовалась, поскольку в ней были выявлены ошибки кодирования (отсутствовало 4 испытуемых, все отказы от ответа кодировались не как пропущенные значения, а как неправильный ответ, что дает искажения при оценке испытуемых).

### 2.1.2 Шкалирование результатов тестирования

**Таджикистан**

После создания базы было проведено шкалирование результатов, чтобы проверить распределение учеников по уровням освоения материала. Для этого необходимо рассчитать пороговые значения перехода с одного уровня на другой, которые являются баллами на 1000-балльной шкале. Для того чтобы произвести необходимые расчеты требуются данные о средней подготовленности испытуемых и дисперсии подготовленности, а также средней трудности заданий по уровням освоения материала. Сначала все оценки переводятся в стандартизированные (z-шкала), затем рассчитываются пороги на 1000-балльной шкале. Используется следующая формула для перевода в z-шкалу:

$z\_{i}=\frac{\overbar{δ}-\overbar{θ}}{S\_{θ}}$, где формула 1

$\overbar{δ}$ – средняя трудность заданий по уровню,

$\overbar{θ}$ – средняя подготовленность испытуемых по выборке,

$S\_{θ}$ – стандартное отклонение подготовленности.

Средний уровень подготовленности испытуемых составил -2,06 логита, стандартное отклонение 1,06 (исключены 8 человек, которые не ответили ни на один вопрос).

***Таблица 1.*** *Трудность заданий по уровням освоения материала*

|  |  |
| --- | --- |
|  | **Средняя трудность** |
| **уровень 1** | -1,78 |
| **уровень 2** | 0,55 |
| **уровень 3** | 2,19 |

ZI = (-1,78-(-2,06)/1,06= 0,26

ZII = (0,55-(-2,06)/1,06= 2,46

ZIII = (2,19-(-2,06)/1,06= 4

Для расчёта порогов используется следующая формула:

B=X+50\*Z, формула 2

где X – средний балл по выборке на 1000-балльной шкале.

Х рассчитывается по формуле:

Х=500-50\*ZII формула 3

Х=500-50\*2,46

Х=377

В=377+50\*z формула 4

В1=377+50\*0,26=377+13=390

В2=377+50\*2,46=500

В3=377+50\*4=577

Для того, чтобы достичь 1-го уровня освоения материала требуется набрать как минимум 390 баллов, для того, чтобы перейти на второй уровень требуется набрать 500 баллов и для того, чтобы перейти на третий уровень освоения материала требуется набрать 577 или больше баллов.

После расчётов пороговых значений было проанализировано распределение подготовленности испытуемых по уровням освоения материала.

***Таблица 2****. Распределение учеников по уровням освоения материала*

|  |  |  |
| --- | --- | --- |
| **уровень** | **частота** | **доля** |
| 0 | 243 | 0,6 |
| 1 | 163 | 0,4 |
| 2 | 2 | 0,0 |
| 3 | 0 | 0,0 |
| Всего | 408 | 1 |

Как видно по представленной таблице, в Таджикистане ни один ученик не достиг третьего уровня освоения материала. На втором уровне оказались всего 2 ученика. Большая часть детей находятся на нулевом уровне освоения материала, то есть не могут применить знания, которые им объясняли на уроках даже в типичной ситуации.

**Новгородская область**

Для шкалирования результатов тестирования в Новгородской области была применена та же процедура, что и для шкалирования результатов тестирования SAM в Таджикистане.

Средний уровень подготовленности испытуемых составил 0,75 логита, стандартное отклонение 1,24 (исключены 8 человек, которые не ответили ни на один вопрос). Средние трудности заданий по уровням представлены в таблице 3.

***Таблица 3****. Трудность заданий по уровням освоения материала*

|  |  |
| --- | --- |
|  | **Средняя трудность** |
| **уровень 1** | -1,89 |
| **уровень 2** | 0,13 |
| **уровень 3** | 1,77 |

ZI = (-1,89-0,75)/1,24= -2,13

ZII = (0,13-0,75)/1,24= - 0,5

ZIII = (1,77-0,75)/1,24= 0,82

B=X+50\*ZII, формула 5

где X – средний балл по выборке на 1000-балльной шкале.

Х рассчитывается по формуле:

Х=500-50\*ZII формула 6

Х=500-50\*(-0,5)

Х=525

В=525+50\*z формула 7

В1=525+50\*(-2,13)=418,5

В2=525+50\*(-0,5)=500

В3=525+50\*0,82=566

***Таблица 4.*** *Распределение учеников по уровням освоения материала*

|  |  |  |
| --- | --- | --- |
| **уровень** | **частота** | **доля** |
| 0 | 46 | 0,02 |
| 1 | 624 | 0,28 |
| 2 | 1098 | 0,5 |
| 3 | 448 | 0,2 |
| всего | 2216 | 1 |

Как видно по таблице 4 распределение учеников по уровням в Новгородской области сильно отличается от распределения в Таджикистане. Меньше одной трети учеников к концу 4-го класса обладают первым уровнем усвоения материала. В Новгородской области большинство учеников находятся на втором уровне освоения материала. По теории Выготского предполагается, что половина учеников к концу начальной школы находятся на рефлексивном уровне освоения материала. Этот постулат подтверждается новгородскими данными. Третьего уровня достигли 20% опрошенных. На нулевом уровне находятся всего 2% учеников.

### 2.1.3 Создание выборки Новгородской области

Как указывалось в первой главе данной работы, для проведения DIF анализа соотношение размеров групп является критически важным. Поскольку большая разница между количеством испытуемых в группах может вызвать ложное срабатывание некоторых статистик, мы приняли решение сделать выборку из учеников Новгороской области с тем, чтобы количество испытуемых в Таджикистане и Новгородской области совпадало. Для создания выборки использовался квазислучайный отбор в пакете SPSS 20. Таким образом, была создана выборка, состоящая из 408 испытуемых, проходивших тестирование в Новгородской области. Даже при построении выборки методом случайного/квазислучайного отбора нет гарантии того, что выборка является несмещенной относительно генеральной совокупности. Проблема здесь заключается в том, что нет определённого параметра, по которому можно было бы оценить несмещённость. С учётом того, что тест разрабатывался в рамках Rasch моделирования, несмещённость оценивалась по параметрам распределения подготовленности испытуемых и трудностей заданий. Для этого были построены гистограммы распределения тестовых баллов на генеральной совокупности и выборке, а также были проанализированы трудности (IRT) и решаемости (КТТ) заданий, корреляции с тестовым баллом.

Генеральная совокупность: Выборка:



***Рис. 3.*** *Гистограммы распределения подготовленностей испытуемых по генеральной совокупности и выборке*

Как видно по представленному рисунку, распределение подготовленностей выглядит сходим образом. Средние значения подготовленностей по выборке и подвыборке совпадают, стандартные отклонения различаются на 0,001.

Генеральная совокупность: Выборка:

 

***Рис.4*** *Гистограммы распределения тестовых баллов по генеральной совокупности и выборке*

По гистограмме тестовых баллов можно также говорить о том, что не наблюдается серьёзных различий по генеральной совокупности и выборке.

***Таблица 5.*** *Трудность и решаемость заданий в Новгородской области и Таджикистане*

| **№** | **Генеральная совокупность** | **Выборка** | **t-статистика** |
| --- | --- | --- | --- |
| **трудность** | **ошибка** | **Корреляция** | **решаемость** | **трудность** | **ошибка** | **корреляция** | **решаемость** |
| 1 | -2,3 | 0,08 | 0,24 | 0,92 | -2,26 | 0,19 | 0,22 | 0,92 | -0,19 |
| 2 | 0,03 | 0,05 | 0,5 | 0,64 | 0,07 | 0,12 | 0,51 | 0,64 | -0,31 |
| 3 | 2,63 | 0,06 | 0,42 | 0,20 | 2,82 | 0,15 | 0,43 | 0,17 | -1,18 |
| 4 | -3,79 | 0,15 | 0,15 | 0,98 | -3,65 | 0,33 | 0,16 | 0,98 | -0,39 |
| 5 | -0,56 | 0,06 | 0,41 | 0,74 | -0,67 | 0,13 | 0,45 | 0,76 | 0,77 |
| 6 | 1,95 | 0,05 | 0,35 | 0,29 | 2,07 | 0,13 | 0,35 | 0,27 | -0,86 |
| 7 | -1,79 | 0,07 | 0,39 | 0,89 | -1,83 | 0,17 | 0,41 | 0,89 | 0,22 |
| 8 | -0,57 | 0,06 | 0,46 | 0,74 | -0,54 | 0,13 | 0,47 | 0,74 | -0,21 |
| 9 | 1,36 | 0,05 | 0,46 | 0,39 | 1,48 | 0,13 | 0,5 | 0,37 | -0,86 |
| 10 | -1,72 | 0,07 | 0,35 | 0,88 | -1,41 | 0,15 | 0,4 | 0,85 | -1,87 |
| 11 | -0,17 | 0,05 | 0,51 | 0,68 | -0,13 | 0,12 | 0,53 | 0,67 | -0,31 |
| 12 | 0,99 | 0,05 | 0,46 | 0,46 | 1,17 | 0,12 | 0,47 | 0,43 | -1,38 |
| 13 | -0,71 | 0,06 | 0,46 | 0,76 | -0,84 | 0,13 | 0,5 | 0,78 | 0,91 |
| 14 | 0,81 | 0,05 | 0,55 | 0,49 | 0,66 | 0,11 | 0,57 | 0,52 | 1,24 |
| 15 | 0,95 | 0,05 | 0,43 | 0,48 | 1 | 0,12 | 0,44 | 0,48 | -0,38 |
| 16 | -1,84 | 0,07 | 0,28 | 0,89 | -1,77 | 0,17 | 0,29 | 0,88 | -0,38 |
| 17 | 0,39 | 0,05 | 0,59 | 0,57 | 0,36 | 0,12 | 0,59 | 0,58 | 0,23 |
| 18 | 2,36 | 0,06 | 0,45 | 0,23 | 2,51 | 0,14 | 0,44 | 0,21 | -0,98 |
| 19 | -0,52 | 0,05 | 0,54 | 0,73 | -0,81 | 0,13 | 0,52 | 0,77 | 2,08 |
| 20 | 0,67 | 0,05 | 0,54 | 0,52 | 0,66 | 0,12 | 0,51 | 0,52 | 0,08 |
| 21 | 1,34 | 0,05 | 0,43 | 0,40 | 1,2 | 0,12 | 0,41 | 0,42 | 1,08 |
| 22 | -4,2 | 0,18 | 0,09 | 0,99 | -4,19 | 0,42 | 0,01 | 0,99 | -0,02 |
| 23 | -0,29 | 0,05 | 0,4 | 0,70 | -0,23 | 0,12 | 0,4 | 0,69 | -0,46 |
| 24 | 2,67 | 0,06 | 0,41 | 0,19 | 2,66 | 0,15 | 0,37 | 0,19 | 0,06 |
| 25 | -1,25 | 0,06 | 0,32 | 0,83 | -1,18 | 0,14 | 0,35 | 0,82 | -0,46 |
| 26 | 0,83 | 0,05 | 0,44 | 0,49 | 0,87 | 0,11 | 0,42 | 0,48 | -0,33 |
| 27 | 2,51 | 0,06 | 0,29 | 0,20 | 2,71 | 0,14 | 0,28 | 0,18 | -1,31 |
| 28 | -1,19 | 0,06 | 0,32 | 0,82 | -1,3 | 0,15 | 0,31 | 0,83 | 0,68 |
| 29 | 0,48 | 0,05 | 0,46 | 0,56 | 0,5 | 0,12 | 0,46 | 0,55 | -0,15 |
| 30 | 1,52 | 0,05 | 0,48 | 0,36 | 1,55 | 0,13 | 0,56 | 0,36 | -0,22 |
| 31 | -1,19 | 0,06 | 0,44 | 0,82 | -1,18 | 0,14 | 0,39 | 0,82 | -0,07 |
| 32 | -0,67 | 0,06 | 0,41 | 0,76 | -0,63 | 0,14 | 0,39 | 0,75 | -0,26 |
| 33 | 1,82 | 0,06 | 0,4 | 0,31 | 1,77 | 0,14 | 0,47 | 0,32 | 0,33 |
| 34 | -2,78 | 0,1 | 0,27 | 0,95 | -2,84 | 0,24 | 0,3 | 0,95 | 0,23 |
| 35 | -0,21 | 0,05 | 0,46 | 0,68 | -0,34 | 0,13 | 0,48 | 0,70 | 0,93 |
| 36 | 2,72 | 0,06 | 0,29 | 0,18 | 2,67 | 0,14 | 0,22 | 0,18 | 0,33 |
| 37 | -1,77 | 0,07 | 0,34 | 0,88 | -1,72 | 0,17 | 0,37 | 0,88 | -0,27 |
| 38 | 0,04 | 0,05 | 0,38 | 0,64 | -0,03 | 0,12 | 0,37 | 0,65 | 0,54 |
| 39 | 1,41 | 0,05 | 0,4 | 0,39 | 1,39 | 0,12 | 0,39 | 0,39 | 0,15 |
| 40 | -1,36 | 0,07 | 0,37 | 0,84 | -1,33 | 0,15 | 0,36 | 0,84 | -0,18 |
| 41 | 0,21 | 0,05 | 0,49 | 0,61 | 0,09 | 0,12 | 0,46 | 0,63 | 0,92 |
| 42 | 0,82 | 0,05 | 0,57 | 0,50 | 0,83 | 0,13 | 0,56 | 0,50 | -0,07 |
| 43 | -1,96 | 0,08 | 0,39 | 0,90 | -2,41 | 0,21 | 0,37 | 0,93 | 2,00 |
| 44 | 0,9 | 0,05 | 0,55 | 0,48 | 0,84 | 0,12 | 0,52 | 0,49 | 0,46 |
| 45 | 1,43 | 0,05 | 0,51 | 0,38 | 1,44 | 0,12 | 0,52 | 0,38 | -0,08 |

Для оценки разницы показателей трудности (IRT) использовалась t-статистика. По таблице 5 видно, лишь для двух заданий t-статистика она достигает критического значения (задания 19 и 43). Разница решаемостей (КТТ) колеблется в пределах от 0 до 0,04 (по модулю). Корреляции с тестовым баллом, рассчитанные для каждого из заданий по выборке и подвыборке также показывают очень близкие значения, максимальная разница составила 0,08.

Из всего вышеизложенного можно заключить, что данная выборка, состоящая из 408 испытуемых, является адекватной репрезентацией генеральной совокупности из 2216 испытуемых. Весь дальнейших анализ будет выполняться с использованием выборки.

### 2.1.4 Психометрический анализ данных

***Таблица 6.*** *Согласие данных с моделью Rasch*

| **№** | **задание** | **Таджикистан** | **Новгород** |
| --- | --- | --- | --- |
| **INFIT** | **OUTFIT** | **Корр.** | **INFIT** | **OUTFIT** | **Корр.** |
| **MNSQ** | **ZSTD** | **MNSQ** | **ZSTD** | **MNSQ** | **ZSTD** | **MNSQ** | **ZSTD** |
| 1 | M-C-01-1-1 | 1,06 | 1,5 | 1,07 | 1 | 0,4 | 1,08 | 0,6 | 1,52 | 1,6 | 0,22 |
| 2 | M-C-01-1-2 | 0,98 | 0 | 1,91 | 1,9 | 0,17 | 0,95 | -0,9 | 0,9 | -1,1 | 0,51 |
| 3 | M-C-01-1-3 | 0,98 | 0 | 0,51 | -1 | 0,24 | 0,95 | -0,5 | 0,78 | -1,1 | 0,43 |
| 4 | M-C-03-1-1 | 1,01 | 0,1 | 1,19 | 1,3 | 0,4 | 1,04 | 0,2 | 1,11 | 0,4 | 0,16 |
| 5 | M-C-03-1-2 | 1,01 | 0,2 | 0,96 | -0,5 | 0,42 | 0,97 | -0,4 | 1,04 | 0,3 | 0,45 |
| 6 | M-C-03-1-3 | 0,98 | 0 | 0,81 | -0,4 | 0,23 | 1,1 | 1,6 | 1,2 | 1,5 | 0,35 |
| 7 | M-M-02-1-1 | 1,09 | 1,2 | 1,31 | 2,1 | 0,24 | 0,9 | -0,9 | 0,77 | -0,9 | 0,41 |
| 8 | M-M-02-1-2 | 1,02 | 0,4 | 2,76 | 2 | -0,02 | 0,96 | -0,5 | 0,92 | -0,6 | 0,47 |
| 9 | M-M-02-1-3 | 0,9 | 0,1 | 0,11 | -1,8 | 0,22 | 0,95 | -0,9 | 0,99 | -0,1 | 0,5 |
| 10 | M-M-03-1-1 | 1,02 | 0,6 | 1,03 | 0,5 | 0,42 | 0,96 | -0,3 | 0,9 | -0,5 | 0,4 |
| 11 | M-M-03-1-2 | 1,07 | 1,1 | 1 | 0,1 | 0,33 | 0,89 | -2 | 0,87 | -1,3 | 0,53 |
| 12 | M-M-03-1-3 | 1,04 | 0,3 | 3,16 | 2 | -0,05 | 0,99 | -0,2 | 1,09 | 1 | 0,47 |
| 13 | M-M-06-1-1 | 1,12 | 2,1 | 1,16 | 1,6 | 0,29 | 0,88 | -1,8 | 0,85 | -1 | 0,5 |
| 14 | M-M-06-1-2 | 0,9 | -1 | 0,77 | -1,3 | 0,41 | 0,87 | -2,9 | 0,84 | -2,2 | 0,57 |
| 15 | M-M-06-1-3 | 1,02 | 0,3 | 1,2 | 1,4 | 0,34 | 1,03 | 0,7 | 1,02 | 0,2 | 0,44 |
| 16 | M-M-11-1-1 | 0,92 | -1,6 | 0,88 | -1,6 | 0,5 | 1,08 | 0,7 | 1,26 | 1,1 | 0,29 |
| 17 | M-M-11-1-2 | 0,96 | -0,2 | 0,89 | -0,3 | 0,33 | 0,83 | -3,5 | 0,79 | -2,7 | 0,59 |
| 18 | M-M-11-1-3 | 0,94 | 0 | 0,31 | -1,2 | 0,24 | 0,95 | -0,6 | 0,94 | -0,3 | 0,44 |
| 19 | M-R-02-1-1 | 0,99 | -0,1 | 1,5 | 2,3 | 0,3 | 0,86 | -2 | 0,8 | -1,4 | 0,52 |
| 20 | M-R-02-1-2 | 0,84 | -0,6 | 0,39 | -1,8 | 0,35 | 0,96 | -0,9 | 0,91 | -1,1 | 0,51 |
| 21 | M-R-02-1-3 |  |  |  | 0 | 1,08 | 1,5 | 1,33 | 3,3 | 0,41 |
| 22 | M-R-05-1-1 | 1,06 | 0,4 | 1,13 | 0,5 | 0,22 | 1,08 | 0,3 | 2,49 | 2,1 | 0,01 |
| 23 | M-R-05-1-2 | 0,87 | -1,3 | 0,76 | -1,3 | 0,46 | 1,07 | 1,2 | 1,14 | 1,3 | 0,4 |
| 24 | M-R-05-1-3 |  |  |  | 0 | 0,99 | 0 | 1,2 | 1,1 | 0,37 |
| 25 | M-G-01-1-1 | 0,92 | -1 | 0,8 | -1,4 | 0,41 | 1,07 | 0,8 | 1,01 | 0,1 | 0,35 |
| 26 | M-G-01-1-2 | 1,09 | 0,4 | 1,82 | 1,6 | 0,03 | 1,07 | 1,5 | 1,11 | 1,4 | 0,42 |
| 27 | M-G-01-1-3 | 1,11 | 0,4 | 2,62 | 2,3 | -0,07 | 1,09 | 1,1 | 2,15 | 4,8 | 0,28 |
| 28 | M-D-03-1-1 | 0,93 | -1,5 | 0,9 | -1,4 | 0,47 | 1,11 | 1,2 | 1,29 | 1,5 | 0,31 |
| 29 | M-D-03-1-2 | 0,94 | -0,1 | 0,41 | -1,3 | 0,25 | 1,01 | 0,3 | 1,01 | 0,2 | 0,46 |
| 30 | M-D-03-1-3 | 1,12 | 0,8 | 1,19 | 0,7 | 0,14 | 0,88 | -2,2 | 0,78 | -2,2 | 0,56 |
| 31 | M-D-05-1-1 | 0,86 | -3,2 | 0,82 | -2,7 | 0,52 | 0,96 | -0,4 | 1,39 | 2 | 0,39 |
| 32 | M-D-05-1-2 | 1,07 | 1,4 | 1,13 | 1,7 | 0,34 | 1,08 | 1,1 | 1,09 | 0,7 | 0,39 |
| 33 | M-D-05-1-3 |  |  |  | 0 | 0,96 | -0,6 | 1,37 | 2,8 | 0,47 |
| 34 | M-D-08-1-1 | 0,97 | -0,3 | 0,92 | -0,6 | 0,44 | 0,96 | -0,1 | 0,7 | -0,7 | 0,3 |
| 35 | M-D-08-1-2 | 0,91 | -1,2 | 0,86 | -1 | 0,44 | 0,97 | -0,4 | 0,87 | -1,1 | 0,48 |
| 36 | M-D-08-1-3 | 1,12 | 0,4 | 2,09 | 1,6 | -0,01 | 1,18 | 2,1 | 1,71 | 3,2 | 0,22 |
| 37 | M-R-03-1-1 | 1,08 | 1,1 | 1,01 | 0,1 | 0,33 | 1 | 0 | 0,83 | -0,7 | 0,37 |
| 38 | M-R-03-1-2 | 1,15 | 1,9 | 1,52 | 3,3 | 0,22 | 1,14 | 2,4 | 1,18 | 1,8 | 0,37 |
| 39 | M-R-03-1-3 | 0,91 | -0,4 | 0,72 | -0,8 | 0,37 | 1,1 | 1,9 | 1,21 | 2,1 | 0,39 |
| 40 | M-C-05-1-1 | 1,05 | 0,9 | 1,03 | 0,3 | 0,35 | 1,02 | 0,2 | 1,29 | 1,5 | 0,36 |
| 41 | M-C-05-1-2 | 0,9 | -0,5 | 0,68 | -1 | 0,35 | 1,01 | 0,2 | 1,07 | 0,8 | 0,46 |
| 42 | M-C-05-1-3 | 0,87 | -0,4 | 0,54 | -0,9 | 0,31 | 0,89 | -2,2 | 0,84 | -2 | 0,56 |
| 43 | M-M-08-1-1 | 0,97 | -0,5 | 0,97 | -0,3 | 0,43 | 0,92 | -0,5 | 0,59 | -1,4 | 0,37 |
| 44 | M-M-08-1-2 | 0,93 | -0,2 | 1,01 | 0,2 | 0,23 | 0,94 | -1,3 | 0,91 | -1,2 | 0,52 |
| 45 | M-M-08-1-3 | 1 | 0,1 | 1,09 | 0,4 | 0,24 | 0,93 | -1,3 | 0,88 | -1,2 | 0,52 |

Согласие с моделью Новгородских и Таджикских данных оценивалось на основе 6-ти показателей, с особенным вниманием к значениям взвешенных статистик согласия, поскольку они взвешиваются на дисперсию и меньше подвержены случайным колебаниям. INFIT MNSQ - взвешенная статистика согласия, OUTFIT MNSQ - общая статистика согласия, INFIT ZSTD - стандартизированная взвешенная статистика согласия, OUTFIT ZSTD - стандартизированная общая статистика согласия. Показатели MNSQ должны находится в пределах от 0,7 до 1,3, эти показатели отражают то, насколько модель хорошо предсказывает дисперсию данных. Значения этих показателей больше 1, говорят о том, что в данных в данных присутствует дисперсия, не объясняемая моделью. Показатели меньше 1 говорят о том, что данные слишком хорошо предсказываются моделью, в них недостаточно дисперсии, что может привести к завышению показателей надёжности (<http://www.winsteps.com/a/winsteps-manual.pdf>, стр. 600). Статистики ZSTD показывают вероятность того, что статистики MNSQ покажут согласие данных с моделью. Для 95% доверительной вероятности значение статистик ZSTD должно находится в пределах от -2 до 2, чтобы можно было делать вывод о том, что задание демонстрирует согласие с моделью Rasch (<http://www.winsteps.com/a/winsteps-manual.pdf>, стр. 600-601).

По совокупности показателей согласия данных с моделью, рассчитанных в пакете Winsteps (<http://www.winsteps.com/>) можно говорить о том, что ***все*** задания (и Таджикские и Новгородские) находятся в согласии с моделью Rasch.

Обратимся к анализу решаемостей заданий в Таджикистане и Новгородской области. Поскольку на российской выборке был проведен полный психометрический анализ опросника SAM, в данной работе Российская версия будет рассматриваться как эталонная (Нежнов, Карданова, 2011), весь проводимый анализ преследует цель – привести Таджикскую версию опросника SAM по математике к сопоставимости с Российской версией.

***Таблица 7****. Решаемость и Трудность заданий в Новгородской области и Таджикистане*

| **№** | **задание** | **Таджикистан** | **Новгород** |
| --- | --- | --- | --- |
| **Выполнили** | **Решаемость** | **Трудность** | **Выполнили** | **Решаемость** | **Трудность** |
| 1 | M-C-01-1-1 | 222 | 0,54 | -3,27 | 373 | 0,91 | -2,26 |
| 2 | M-C-01-1-2 | 20 | 0,05 | 0,36 | 248 | 0,61 | 0,07 |
| 3 | M-C-01-1-3 | 3 | 0,01 | 2,37 | 61 | 0,15 | 2,82 |
| 4 | M-C-03-1-1 | 265 | 0,65 | -3,81 | 396 | 0,97 | -3,65 |
| 5 | M-C-03-1-2 | 68 | 0,17 | -1,13 | 288 | 0,71 | -0,67 |
| 6 | M-C-03-1-3 | 5 | 0,01 | 1,84 | 105 | 0,26 | 2,07 |
| 7 | M-M-02-1-1 | 70 | 0,17 | -1,17 | 359 | 0,88 | -1,83 |
| 8 | M-M-02-1-2 | 1 | 0,00 | 3,48 | 294 | 0,72 | -0,54 |
| 9 | M-M-02-1-3 | 1 | 0,00 | 3,48 | 127 | 0,31 | 1,48 |
| 10 | M-M-03-1-1 | 173 | 0,42 | -2,67 | 341 | 0,84 | -1,41 |
| 11 | M-M-03-1-2 | 67 | 0,16 | -1,11 | 261 | 0,64 | -0,13 |
| 12 | M-M-03-1-3 | 19 | 0,05 | 0,42 | 166 | 0,41 | 1,17 |
| 13 | M-M-06-1-1 | 134 | 0,33 | -2,18 | 316 | 0,77 | -0,84 |
| 14 | M-M-06-1-2 | 49 | 0,12 | -0,7 | 209 | 0,51 | 0,66 |
| 15 | M-M-06-1-3 | 41 | 0,10 | -0,48 | 174 | 0,43 | 1 |
| 16 | M-M-11-1-1 | 199 | 0,49 | -2,99 | 359 | 0,88 | -1,77 |
| 17 | M-M-11-1-2 | 41 | 0,10 | -0,48 | 227 | 0,56 | 0,36 |
| 18 | M-M-11-1-3 | 1 | 0,00 | 3,48 | 77 | 0,19 | 2,51 |
| 19 | M-R-02-1-1 | 31 | 0,08 | -0,14 | 311 | 0,76 | -0,81 |
| 20 | M-R-02-1-2 | 7 | 0,02 | 1,49 | 206 | 0,50 | 0,66 |
| 21 | M-R-02-1-3 | 26 | 0,06 | 0,06 | 150 | 0,37 | 1,2 |
| 22 | M-R-05-1-1 | 184 | 0,45 | -2,81 | 400 | 0,98 | -4,19 |
| 23 | M-R-05-1-2 | 34 | 0,08 | -0,25 | 273 | 0,67 | -0,23 |
| 24 | M-R-05-1-3 | 1 | 0,00 | 3,48 | 69 | 0,17 | 2,66 |
| 25 | M-G-01-1-1 | 66 | 0,16 | -1,09 | 329 | 0,81 | -1,18 |
| 26 | M-G-01-1-2 | 2 | 0,00 | 2,78 | 194 | 0,48 | 0,87 |
| 27 | M-G-01-1-3 | 2 | 0,00 | 2,78 | 72 | 0,18 | 2,71 |
| 28 | M-D-03-1-1 | 159 | 0,39 | -2,5 | 339 | 0,83 | -1,3 |
| 29 | M-D-03-1-2 | 14 | 0,03 | 0,75 | 217 | 0,53 | 0,5 |
| 30 | M-D-03-1-3 | 5 | 0,01 | 1,84 | 126 | 0,31 | 1,55 |
| 31 | M-D-05-1-1 | 76 | 0,19 | -1,28 | 333 | 0,82 | -1,18 |
| 32 | M-D-05-1-2 | 78 | 0,19 | -1,32 | 282 | 0,69 | -0,63 |
| 33 | M-D-05-1-3 | 1 | 0,00 | 3,48 | 104 | 0,25 | 1,77 |
| 34 | M-D-08-1-1 | 187 | 0,46 | -2,84 | 381 | 0,93 | -2,84 |
| 35 | M-D-08-1-2 | 68 | 0,17 | -1,13 | 277 | 0,68 | -0,34 |
| 36 | M-D-08-1-3 | 5 | 0,01 | 1,84 | 71 | 0,17 | 2,67 |
| 37 | M-R-03-1-1 | 85 | 0,21 | -1,45 | 353 | 0,87 | -1,72 |
| 38 | M-R-03-1-2 | 81 | 0,20 | -1,38 | 257 | 0,63 | -0,03 |
| 39 | M-R-03-1-3 | 8 | 0,02 | 1,35 | 143 | 0,35 | 1,39 |
| 40 | M-C-05-1-1 | 126 | 0,31 | -2,07 | 336 | 0,82 | -1,33 |
| 41 | M-C-05-1-2 | 21 | 0,05 | 0,31 | 251 | 0,62 | 0,09 |
| 42 | M-C-05-1-3 | 9 | 0,02 | 1,23 | 165 | 0,40 | 0,83 |
| 43 | M-M-08-1-1 | 96 | 0,24 | -1,63 | 370 | 0,91 | -2,41 |
| 44 | M-M-08-1-2 | 15 | 0,04 | 0,68 | 189 | 0,46 | 0,84 |
| 45 | M-M-08-1-3 | 3 | 0,01 | 2,37 | 134 | 0,33 | 1,44 |

Как видно по представленной таблице 7, оценка трудностей для России и Таджикистана как в рамках КТТ, так и в рамках IRT сильно различается, поэтому было принято решение сравнивать паттерны трудностей по уровням заданий. Основная идея состоит в том, чтобы сравнить задание по трудности с предыдущим и следующим. Если задание является труднее предыдущего и последующего, то такое расположение должно сохраняться и в России и в Таджикистане. Для этого были построены графики трудностей в рамках КТТ и рассчитаны корреляции заданий по уровням.

Сравнение трудностей заданий в России и Таджикистане по уровням (КТТ)

***Рис. 5.*** *Решаемости заданий 1-го уровня в Новгородской области и Таджикистане*

Корреляция трудностей для заданий 1-го уровня составила 0,67.

***Рис. 6.*** *Решаемости заданий 2-го уровня в Новгородской области и Таджикистане*

Корреляция трудностей для заданий 2-го уровня составила 0,5.

***Рис. 7.*** *Решаемости заданий 3-го уровня в Новгородской области и Таджикистане*

Корреляция трудностей для заданий 3-го уровня оценивать нельзя, слишком мало учеников в Таджикистане решили задания этого уровня.

Как видно из представленных рисунков 5-7 и таблицы 7, паттерны трудностей для заданий 3-го уровня в России и Таджикистане сильно различаются, поэтому было принято решение не учитывать задания 3-го уровня при построении одной шкалы для России и Таджикистана.

Задания 2-го уровня в целом имеют схожий паттерн, однако есть задание, решаемость которого в связке с другими заданиями в Таджикистане сильно отличается от решаемости в России (M-M-02-1-2) Данное задание должно быть легче предыдущего и следующего за ним, однако, в Таджикистане это задание оказалось сложнее предыдущего и следующего. Его решаемость оказалась схожей с заданиями 3-го уровня, его выполнил верно всего 1 ученик из 408.

Что касается заданий 1-го уровня, то здесь также наблюдается схожий паттерн решаемости заданий в двух выборках. Также как и в группе заданий 2-го уровня «выпадает» одно задание, которое относится к той же предметной области: M-M-02-1-1.

## 2.2 Содержательный этап

### 2.2.1 Установление Конструктной эквивалентности

Для установления конструктной эквивалентности мы использовали карты конструктов, поскольку ЭФА и КФА, как было указано во введении к данной главе, невозможно провести на имеющихся данных. Карты конструктов были представлены в заключительном отчёте по локализации и адаптации опросника SAM в Таджикистане, и во фреймворке по разработке SAM в России (заключительный отчет по адаптации и локализации SAM в Таджикистане; Нежнов, Карданова, 2011).

Карта конструкта в обоих документах выглядит одинаково (рис. 8)



***Рис 8.*** *Карта конструкта «математическая грамотность» в России и Таджикистане.*

Приведём описание выделенных на рисунке 8 подконструктов (разделов предметного содержания теста):

***«Числа и вычисления***

*Раздел включает содержание, относящееся к формальной стороне понятия натурального числа (позиционная запись чисел, стандартные алгоритмы действий над числами, порядок выполнения действий, свойства действий). Сюда же отнесен учебный материал, связанный с представлением чисел на координатной прямой. Последнее важно для понимания действительного числа и освоения координатного метода.*

***Измерение величин***

*В раздел включен учебный материал, связанный собственно с действиями прямого и косвенного измерения. Сюда же отнесены геометрические измерения.*

*Что касается собственно прикладного аспекта данного раздела, связанного с конкретными измерениями и представлением их результатов в виде таблиц и диаграмм («анализ данных»), то он в большей степени может быть отнесен к учебному предмету «Окружающий мир».*

***Закономерности***

*Содержание раздела связано с построением числовых и геометрических последовательностей и других структурированных объектов, а также с определением их количественных характеристик. Эта линия важна для развития математического мышления (в первую очередь – алгоритмического и комбинаторного).*

***Зависимости***

*Содержание раздела связано с выделением и описанием математической структуры отношений между величинами, обычно представляемых текстовыми задачами.*

***Элементы геометрии***

*Раздел охватывает геометрический материал, связанный с определением пространственных форм и взаимным расположением объектов.*

*Содержание математического теста представлено в виде матрицы (табл. 2.1.), в которую включены а) разделы предметного содержания и б) математические средства (понятия, представления, принципы, правила, формулы, схемы и проч.), овладение которыми лежит в основе математической компетентности*» (заключительный отчет по адаптации и локализации SAM в Таджикистане; Нежнов, Карданова, 2011).

В документах по разработке российской версии SAM также присутствует более подробное описание того, что входит в каждый из разделов (приложение 2). Также в документах разработчиков SAM в России содержится описание того, сколько заданий и блоков направлены на проверку каждого из разделов (приложение 2).

Стоит отметить, что в документе по адаптации таджикской версии SAM не содержится такой подробной операционализации конструкта «математическая грамотность», там лишь содержится описание разделов, полностью повторяющее описание разделов российской версии. Таким образом, судить о конструктной эквивалентности или неэквивалентности по карте конструкта затруднительно. С учётом того, что перевод SAM на таджикский язык осуществлялся по заданиям, анализ формулировок заданий может помочь определить, совпадают ли области содержания, проверяемые таджикской версией с областями содержания, проверяемыми российской версией.

Для решения этой задачи был проведен анализ тетрадей заданий с привлечением специалиста, знающего русский и таджикский языки, а также документов по двойному слепому переводу, полученных от таджикских специалистов. Результаты анализа перевода представлены в таблице 8.

***Таблица 8.*** *Изменения формулировок заданий*

| **Номер** | **Код заданий** | **Изменения в формулировке****Российский вариант=таджикский вариант** | **Решение об изменении** |
| --- | --- | --- | --- |
| 1 | M-C-01-1-1 | *Без изменений* |  |
| 2 | M-C-01-1-2 | Заменяют = являются | Изменение оставлено в итоговом варианте |
| 3 | M-C-01-1-3 | Разные буквы заменяют разные цифры =каждая буква заменяет одну цифру | Изменение отменено, итоговый вариант как русский (разные буквы..) |
| 4 | M-C-03-1-1 | Найдите самое большое из следующих чисел = найдите наибольшее число | Изменение оставлено |
| 5 | M-C-03-1-2 | Какое число= какую цифру | Изменение отменено, итоговый вариант как русский (какое число) |
| 6 | M-C-03-1-3 | Расположение чисел | Изменение отменено, числа располагаются аналогично русской версии |
| 7 | M-M-02-1-1 | ПОЛНАЯ ЗАМЕНА ЗАДАНИЯ (вместо расчета площади – расчет суммы длин сторон «объезд ремонта дороги» | Отредактирована формулировка, смысл задания сильно изменен |
| 8 | M-M-02-1-2 | ПОЛНАЯ ЗАМЕНА ЗАДАНИЯ (вместо площади – сумма длин трёх сторон) | Аналогично заданию 7 |
| 9 | M-M-02-1-3 | ПОЛНАЯ ЗАМЕНА ЗАДАНИЯ | Аналогично заданию 7 и 8 |
| 10 | M-M-03-1-1 | Ломаная=линия | Изменение сохранено (убрано слово ломаная, поскольку в программе его нет) |
| 11 | M-M-03-1-2 | Ломаная=линия | То же что и в вопросе 10 |
| 12 | M-M-03-1-3 | Ломаная=линия | То же что и в вопросе 10 |
| 13 | M-M-06-1-1 | *Без изменений* |  |
| 14 | M-M-06-1-2 | *Без изменений* |  |
| 15 | M-M-06-1-3 | Ломаная=линия | То же что и в вопросе 10 |
| 16 | M-M-11-1-1 | Сторона=стороны | Изменение отменено, оставлено «сторона» |
| 17 | M-M-11-1-2 | Найди=найдите | Изменение отменено, оставлено слово «найди» |
| 18 | M-M-11-1-3 | Разрезание=разделение | Изменение отменено, оставлено «разрезали» |
| 19 | M-R-02-1-1 | Гном=ученик | Изменение оставлено, в таджикском нет слова «гном» |
| 20 | M-R-02-1-2 | Гном=ученик | Изменение оставлено, в таджикском нет слова «гном» |
| 21 | M-R-02-1-3 | Гном=ученик | Изменение оставлено, в таджикском нет слова «гном» |
| 22 | M-R-05-1-1 | Располагаются=расположены | Изменение оставлено |
| 23 | M-R-05-1-2 | *Без изменений* |  |
| 24 | M-R-05-1-3 | Скобки с пояснением (Сколько в нем всего фигур) перенесено в конец задания | Изменение оставлено |
| 25 | M-G-01-1-1 | Квадрат=четырёхугольник | Изменение отменено, в тетради оставлено как в русском варианте – квадрат |
| 26 | M-G-01-1-2 | *Без изменений* |  |
| 27 | M-G-01-1-3 | *Без изменений* |  |
| 28 | M-D-03-1-1 | *Без изменений* |  |
| 29 | M-D-03-1-2 | *Без изменений* |  |
| 30 | M-D-03-1-3 | *Без изменений* |  |
| 31 | M-D-05-1-1 | *Без изменений* |  |
| 32 | M-D-05-1-2 | Коробки=банка, бидон и ведро | Изменение отменено, оставлены 3 коробки |
| 33 | M-D-05-1-3 | *Без изменений* |  |
| 34 | M-D-08-1-1 | *Без изменений* |  |
| 35 | M-D-08-1-2 | *Без изменений* |  |
| 36 | M-D-08-1-3 | *Без изменений* |  |
| 37 | M-R-03-1-1 | Цвета написаны в квадратах в таджикском варианте полностью, в русском – сокращение цветом, снизу расшифровкаВ текст задания включено слово «гирлянда» как на таджикском «силсила», так и на русском (пояснено в скобках)Съехали места для ответов | Изменение оставлено |
| 38 | M-R-03-1-2 | Цвета написаны полностью, «гирлянд» вместо «гирлянда» - отличие от предыдущего вопроса!Съехали места для ответов | То же что в задании 37 |
| 39 | M-R-03-1-3 | Цвета написаны полностью, «гирлянда» вместо «гирлянд» | То же что в задании 37 |
| 40 | M-C-05-1-1 | *Без изменений* |  |
| 41 | M-C-05-1-2 | *Без изменений* |  |
| 42 | M-C-05-1-3 | *Без изменений* |  |
| 43 | M-M-08-1-1 | *Без изменений* |  |
| 44 | M-M-08-1-2 | *Без изменений* |  |
| 45 | M-M-08-1-3 | *Без изменений* |  |

Поскольку для построения общей шкалы для России и Таджикистана было решено не использовать задания 3-го уровня, здесь и далее будут анализироваться только задания 1-го и 2-го уровня.

Без изменений остались следующие задания:

1. M-C-01-1-1
2. M-C-03-1-2
3. M-M-06-1-1
4. M-M-06-1-2
5. M-M-11-1-1
6. M-M-11-1-2
7. M-R-05-1-2
8. M-G-01-1-1
9. M-G-01-1-2
10. M-D-03-1-1
11. M-D-03-1-2
12. M-D-05-1-1
13. M-D-05-1-2
14. M-D-08-1-1
15. M-D-08-1-2
16. M-C-05-1-1
17. M-C-05-1-2
18. M-M-08-1-1
19. M-M-08-1-2

Незначительные изменения были внесены изменения в следующие задания:

1. M-C-01-1-2
2. M-C-03-1-1
3. M-M-03-1-1
4. M-M-03-1-2
5. M-R-02-1-1
6. M-R-02-1-2
7. M-R-05-1-1

Существенные изменения были внесены в следующие задания:

1. M-M-02-1-1
2. M-M-02-1-2
3. M-R-03-1-1
4. M-R-03-1-2

Под существенными изменениями понимаются такие изменения в формулировке, которые могли бы привести к изменению функционирования задания. Например, замена слова «гном» на «ученик» в таджикском варианте (M-R-02-1) была сделана исходя из того, что в таджикской мифологии отсутствуют гномы. Такая замена не должна привести к существенному изменению функционирования задания, поэтому такое изменение можно считать незначительным. С другой стороны, в задании M-R-03-1 в таджикском опроснике цвета написаны в квадратах полностью, без сокращений, что может сделать вопрос более простым для таджикских учеников, поскольку в российском опроснике цвета написаны сокращённо, а полные названия цветов приводятся в полях снизу. То есть российским ученикам нужно сделать лишнее «действие» в уме – сопоставить цвета, написанные в квадратах с полным вариантом, приводимом чуть ниже.

Все 3 задания блока M-M-02-1 были существенно изменены при проведении опроса на таджикском языке. Была изменена единица содержания, проверяемая данными заданиями. В России эти задания направлены на выявление знаний о том, что такое площадь и способность её рассчитать

В Таджикистане эти задания проверяют способность учеников рассчитать сумму длин сторон (фактически – периметра фигуры).

Обратимся к рассмотрению того, как задания оформлены в таджикских и российских тетрадях. Способ представления заданий испытуемым, так же как и содержание заданий, может обуславливать разницу в функционировании заданий на разных выборках. Поэтому для кросс-культурного исследования важно, чтобы задания в разных странах представлялись одинаковым или, если это невозможно, максимально схожим образом.

***Таблица 9****. Различия в визуальном представлении заданий*

| **Номер** | **Код заданий** | **Отличия в представлении задания в Таджикской версии** |
| --- | --- | --- |
| 1 | M-C-01-1-1 | *Нет отличий* |
| 2 | M-C-01-1-2 | *Нет отличий* |
| 4 | M-C-03-1-1 | Варианты ответа расположены не в одну строку, а в два столбца |
| 5 | M-C-03-1-2 | *Нет отличий* |
| 7 | M-M-02-1-1 | Даются варианты ответа |
| 8 | M-M-02-1-2 | - |
| 10 | M-M-03-1-1 | Незначительное уменьшение масштаба изображения (1,1 см в российской тетради 1,0 см в таджикской) |
| 11 | M-M-03-1-2 | Более чёткие линии на картинке, изменено расположение вариантов ответа (два столбика) |
| 13 | M-M-06-1-1 | Уменьшение масштаба изображения (1,6 см в российской тетради, 1,0 см в таджикской) |
| 14 | M-M-06-1-2 | Увеличение масштаба изображения (0,6 см в российской тетради, 1,0 см в таджикской) |
| 16 | M-M-11-1-1 | *Нет отличий* |
| 17 | M-M-11-1-2 | *Нет отличий* |
| 19 | M-R-02-1-1 | Незначительное уменьшение масштаба изображения (изоб.не используется для расчетов), изменено расположение вариантов ответа (два столбика, варианты ответа «съехали» вверх относительно полей для ответа) |
| 20 | M-R-02-1-2 | Задание написано без абзацев |
| 22 | M-R-05-1-1 | Более чёткие линии изображения |
| 23 | M-R-05-1-2 | Незначительное уменьшение масштаба изображения |
| 25 | M-G-01-1-1 | Уменьшение масштаба изображения |
| 26 | M-G-01-1-2 | 2 из 5-ти изображений зеркально развернуты, все фигуры пронумерованы, введены поля для ответа с нумерацией (в российской тетради ученикам не давалась нумерация, нужно было отметить фигуры) |
| 28 | M-D-03-1-1 | *Нет отличий* |
| 29 | M-D-03-1-2 | Варианты ответа «съехали» вверх относительно полей для ответа |
| 31 | M-D-05-1-1 | *Нет отличий* |
| 32 | M-D-05-1-2 | *Нет отличий* |
| 34 | M-D-08-1-1 | *Нет отличий* |
| 35 | M-D-08-1-2 | *Нет отличий* |
| 37 | M-R-03-1-1 | Варианты ответа «съехали» относительно полей, нет сокращений цветов (в российской тетради цвета сокращены, + приводится расшифровка сокращений) |
| 38 | M-R-03-1-2 | Варианты ответа «съехали» относительно полей, нет сокращений цветов (в российской тетради цвета сокращены, + приводится расшифровка сокращений) |
| 40 | M-C-05-1-1 | *Нет отличий* |
| 41 | M-C-05-1-2 | *Нет отличий* |
| 43 | M-M-08-1-1 | Увеличение масштаба изображения (не связано с расчетами, которые нужно произвести) |
| 44 | M-M-08-1-2 | Увеличение масштаба изображения (не связано с расчетами, которые нужно произвести) |

Систематизируя информацию, представленную в таблице 9 можно сказать, что основные отличия в представлении заданий сводятся к:

1. *Изменению масштабов рисунков в сторону упрощения расчетов*. Изменены масштабы тех рисунков, которые используются для получения ответов в задании (которые используются для расчетов). Во всех заданиях, где ученику даётся «линейка» с делениями для измерения требуемого объекта, деления этой линейки приведены в соответствие реальным сантиметрам.



***Рис 9.*** *Пример задания с изменением масштабов рисунка*

1. *Изменению формы записи правильного ответа*.
* Расположение вариантов ответа в две строки, вместо одной строки
* Введение дополнительной нумерации объектов, которой не было в российской тетради.





***Рис 10.*** *Пример заданий с изменением формы записи вариантов ответа*

1. *Все изображения имеют более чёткие линии, по сравнению с российской тетрадью*.
2. *Поля для ответов и варианты ответа, им соответствующие, не располагаются на одной строке*

По результатам описанного выше анализа мы исключили из дальнейшей работы все задания 3-его уровня и блок заданий M-M-02-1, поскольку они проверяют различные темы в России и Таджикистане. Дальнейшая работа будет проводиться со следующими заданиями:

1. M-C-01-1-1
2. M-C-01-1-2
3. M-C-03-1-1
4. M-C-03-1-2
5. M-M-03-1-1
6. M-M-03-1-2
7. M-M-06-1-1
8. M-M-06-1-2
9. M-M-11-1-1
10. M-M-11-1-2
11. M-R-02-1-1
12. M-R-02-1-2
13. M-R-05-1-1
14. M-R-05-1-2
15. M-G-01-1-1
16. M-G-01-1-2
17. M-D-03-1-1
18. M-D-03-1-2
19. M-D-05-1-1
20. M-D-05-1-2
21. M-D-08-1-1
22. M-D-08-1-2
23. M-R-03-1-1
24. M-R-03-1-2
25. M-C-05-1-1
26. M-C-05-1-2
27. M-M-08-1-1
28. M-M-08-1-2

Прежде, чем перейти к объединению российских и таджикских данных нам хотелось бы обозначить логику исследования перевода и визуального представления заданий.

1. Исследование документов по локализации и адаптации теста: выделение заданий, имеющих существенную и незначительную разницу в формулировках, а также заданий, которые с точки зрения языка переведены точно.
2. Привлечение специалиста, владеющего обоими языками для проверки пункта 1. Для этого специалиста просили заново перевести задания из таджикской тетради на русский язык. Изменения классифицировались по трём типам, сравнивались с результатами, полученными в ходе анализа документов.
3. Для оценки разницы в визуальном представлении заданий использовались только анкеты опросника (тетради). Для каждого задания любые изменения в визуальном представлении задания фиксировались в таблицу, затем выделялись группы параметров, по которым отличаются российская и таджикская тетради.

### 2.2.2 Психометрический анализ объединённых данных

Данные по 28-ми заданиям из России и Таджикистана были объединены в одну выборку, которая была проверена на соответствие модели Раша. Общий размер выборки составил 816 учеников.

***Таблица 10.*** *Согласие данных с моделью объединённых данных*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| № | код задания | трудность | число учеников | ответили верно | ошибка | INFIT. MNSQ | INFIT. ZSTD | OUTFIT. MNSQ | OUTFIT. ZSTD |
| 1 | M-C-01-1-1 | -1,43 | 797 | 589 | 0,1 | 1,07 | 1,50 | 1,39 | 2,12 |
| 2 | M-C-01-1-2 | 1,31 | 683 | 262 | 0,1 | 0,85 | -3,15 | 0,86 | -1,20 |
| 3 | M-C-03-1-1 | -2,96 | 803 | 715 | 0,13 | 1,02 | 0,28 | 1,26 | 0,94 |
| 4 | M-C-03-1-2 | -0,2 | 664 | 403 | 0,1 | 1,13 | 2,54 | 1,14 | 1,28 |
| 5 | M-M-03-1-1 | -0,74 | 745 | 497 | 0,1 | 1,07 | 1,46 | 1,05 | 0,45 |
| 6 | M-M-03-1-2 | 0,46 | 686 | 346 | 0,1 | 1,15 | 2,92 | 1,12 | 1,20 |
| 7 | M-M-06-1-1 | -0,03 | 748 | 426 | 0,09 | 1,05 | 0,98 | 1,06 | 0,67 |
| 8 | M-M-06-1-2 | 1,41 | 700 | 259 | 0,1 | 0,99 | -0,27 | 0,96 | -0,33 |
| 9 | M-M-11-1-1 | -1,41 | 734 | 563 | 0,11 | 1,07 | 1,30 | 1,48 | 2,39 |
| 10 | M-M-11-1-2 | 1,34 | 683 | 258 | 0,1 | 0,85 | -3,01 | 0,92 | -0,66 |
| 11 | M-R-02-1-1 | 0,44 | 722 | 359 | 0,1 | 0,80 | -4,28 | 0,88 | -1,28 |
| 12 | M-R-02-1-2 | 1,76 | 673 | 219 | 0,1 | 0,85 | -2,91 | 0,66 | -2,73 |
| 13 | M-R-05-1-1 | -3,62 | 616 | 593 | 0,23 | 1,09 | 0,57 | 2,64 | 2,97 |
| 14 | M-R-05-1-2 | 0,69 | 613 | 315 | 0,1 | 1,02 | 0,34 | 1,02 | 0,21 |
| 15 | M-G-01-1-1 | 0,19 | 808 | 406 | 0,09 | 0,87 | -3,05 | 0,77 | -2,77 |
| 16 | M-G-01-1-2 | 1,99 | 720 | 204 | 0,1 | 1,04 | 0,71 | 1,10 | 0,70 |
| 17 | M-D-03-1-1 | -0,55 | 719 | 472 | 0,1 | 1,04 | 0,74 | 1,19 | 1,54 |
| 18 | M-D-03-1-2 | 1,68 | 694 | 225 | 0,1 | 0,89 | -2,12 | 0,69 | -2,47 |
| 19 | M-D-05-1-1 | -0,52 | 703 | 460 | 0,1 | 0,96 | -0,80 | 1,29 | 2,26 |
| 20 | M-D-05-1-2 | -0,24 | 634 | 396 | 0,1 | 1,29 | 5,17 | 1,46 | 3,56 |
| 21 | M-D-08-1-1 | -2,29 | 721 | 615 | 0,12 | 1,00 | 0,03 | 0,86 | -0,47 |
| 22 | M-D-08-1-2 | 0,5 | 684 | 343 | 0,1 | 0,97 | -0,60 | 0,92 | -0,88 |
| 23 | M-R-03-1-1 | -0,34 | 687 | 434 | 0,1 | 0,92 | -1,61 | 0,76 | -2,25 |
| 24 | M-R-03-1-2 | 0,65 | 671 | 325 | 0,1 | 1,29 | 5,26 | 1,62 | 5,42 |
| 25 | M-C-05-1-1 | -0,32 | 682 | 427 | 0,1 | 1,02 | 0,32 | 1,10 | 0,92 |
| 26 | M-C-05-1-2 | 1,16 | 642 | 272 | 0,1 | 0,89 | -2,20 | 0,77 | -2,23 |
| 27 | M-M-08-1-1 | -0,8 | 652 | 461 | 0,11 | 0,81 | -3,70 | 0,61 | -3,09 |
| 28 | M-M-08-1-2 | 1,87 | 588 | 199 | 0,11 | 0,94 | -1,20 | 0,90 | -0,60 |

Выявлено 2 задания, не согласующихся с моделью Раша:

* M-D-05-1-2
* M-R-03-1-2

Данные задания не будут включаться в построение единой шкалы.

### 2.2.3 Эквивалентность заданий

Для того чтобы можно было положить Российские и Таджикские данные на одну шкалу, необходимо проверить задания на DIF. Поскольку в научной среде нет консенсуса относительно того, какой метод выявления DIF наиболее эффективен, в данной работе мы будем использовать 4 метода:

1. Мантель-Ханцель,
2. Стандартизацию,
3. Логистическую регрессию,
4. t-статистику.

Мы будем считать, что задание демонстрирует DIF, если 3 или 4 статистики показывают, что в данном задании присутствует DIF. Поскольку все методы выявления DIF в той или иной мере демонстрируют ошибку II рода (ложное срабатывание), нельзя ограничиваться показаниями только одной статистики. В первой главе данной работы описано более подробно, от чего зависит функционирование статистик DIF-анализа.

3 статистики DIF-анализа были рассчитаны в пакете difR. Задание отмечалось как демонстрирующее DIF, если наблюдался large или moderate effect.

T-статистика рассчитывалась в Excel. Для того, чтобы рассчитать t-статистику, требовалось выделить несколько (не менее 4 заданий), которые функционируют максимально схожим образом в России и в Таджикистане и которые потенциально свободны от DIF. Было выбрано 5 заданий:

1. M-C-01-1-1
2. M-C-03-1-1
3. M-D-03-1-1
4. M-C-05-1-1
5. M-D-08-1-1

Эти задания использовались как «якорные». Их трудность фиксировалась на том уровне, который демонстрировали задания на российской подвыборке, трудность остальных заданий переоценивалась для Таджикской выборки. Затем производился расчет статистики. Критические значения для неё были выбраны традиционные: если t-статистика показывала значения по модулю превышающее 2, задание отмечалось как демонстрирующее DIF в соответствии с данной статистикой (Wang, Su, 2004).

Если 3 или 4 статистики показывали, что задание демонстрирует DIF, мы считали что DIF присутствует.

***Таблица 11.*** *Результаты DIF-анализа*

| **№** |  | **Код задания** | **M-H** | **LR** | **STND** | **t-stat.** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | M-C-01-1-1 |  |  |  | 0,00 |
| 2 | 2 | M-C-01-1-2 |  | + | + | 5,02 |
| 4 | 3 | M-C-03-1-1 |  |  |  | 0,00 |
| 5 | 4 | M-C-03-1-2 | + | + | + | -3,79 |
| 10 | 5 | M-M-03-1-1 |  |  |  | -1,25 |
| 11 | 6 | M-M-03-1-2 | + | + |  | -3,74 |
| 13 | 7 | M-M-06-1-1 |  | + |  | -0,52 |
| 14 | 8 | M-M-06-1-2 | + | + | + | -2,84 |
| 16 | 9 | M-M-11-1-1 | + | + | + | -3,22 |
| 17 | 10 | M-M-11-1-2 | + | + | + | 1,11 |
| 19 | 11 | M-R-02-1-1 |  | + |  | 4,53 |
| 20 | 12 | M-R-02-1-2 |  | + | + | 3,00 |
| 22 | 13 | M-R-05-1-1 |  | + |  | -0,43 |
| 23 | 24 | M-R-05-1-2 |  | + | + | 0,60 |
| 25 | 15 | M-G-01-1-1 | + | + |  | 4,78 |
| 26 | 16 | M-G-01-1-2 |  | + | + | 3,25 |
| 28 | 17 | M-D-03-1-1 |  |  | + | 0,00 |
| 29 | 18 | M-D-03-1-2 |  | + | + | 4,19 |
| 31 | 19 | M-D-05-1-1 | + | + | + | -1,56 |
| 32 | 20 | M-D-05-1-2 |  | + | + | -4,60 |
| 34 | 21 | M-D-08-1-1 |  |  |  | 0,00 |
| 35 | 22 | M-D-08-1-2 |  | + | + | -0,35 |
| 37 | 23 | M-R-03-1-1 | + | + |  | 4,50 |
| 38 | 24 | M-R-03-1-2 |  | + |  | -2,81 |
| 40 | 25 | M-C-05-1-1 |  | + | + | 0,00 |
| 41 | 26 | M-C-05-1-2 |  | + | + | 2,65 |
| 43 | 27 | M-M-08-1-1 |  | + |  | 5,11 |
| 44 | 28 | M-M-08-1-2 |  | + | + | 1,80 |

В соответствии с описанным выше критерием, было выделено 14 заданий, демонстрирующих DIF. 10 из них принадлежат ко 2-ому уровню освоения материала. Задания, демонстрирующие DIF:

1. M-C-01-1-2
2. M-C-03-1-2
3. M-M-03-1-2
4. M-M-06-1-2
5. M-M-11-1-1
6. M-M-11-1-2
7. M-R-02-1-2
8. M-G-01-1-1
9. M-G-01-1-2
10. M-D-03-1-2
11. M-D-05-1-1
12. M-D-05-1-2
13. M-R-03-1-1
14. M-C-05-1-2

Обратимся к рассмотрению причины появления DIF в данных заданиях.

**M-C-01-1-2**. Данное задание является более трудным для таджикских детей. Данное задание было переведено на таджикский язык с точностью до одного слова. Также задания представлены одинаково в обеих тетрадях. В таджикской тетради вместо \* стоит точка (знак умножения)

**M-C-03-1-2**. Данное задание является более сложным для российских детей. Присутствуют незначительные изменения при переводе («самое большое» заменено на «наибольшее»), а также в оформлении задания: варианты ответа в таджикской тетради располагаются не в одну строку, а в две.

**M-M-06-1-2**. Задание является более трудным для российских детей. Масштаб рисунка в таджикской тетради значительно больше.

**M-R-02-1-2**. Задание является более трудным для таджикских детей. Присутствует незначительное изменение формулировки задания (слово «гном» заменено на слово «ученик»), также, в таджикской тетради данное задание не разбито на абзацы.

**M-G-01-1-2**. Задание является более трудным для таджикских детей. Присутствуют существенные изменения в представлении задания в таджикской тетради: перевёрнуты 2 из 5 нарисованных фигур, введены отдельные поля для ответов и нумерация фигур, что усложняет процедуру записывания ответов.

**M-D-05-1-1**. Данное задание демонстрирует DIF в пользу таджикских учеников, для российских учеников оно является необоснованно сложным. Различий в формулировке и представлении данного задания обнаружено не было.

**M-M-03-1-2.** Данное задание является более трудным для Российских детей. В данном задании в таджикской тетради слово «ломаная» заменено на слово «линия», также изменено расположение вариантов ответа: в 2 строки вместо одной строки.

**M-M-11-1-1.** Данное задание демонстрирует DIF в пользу таджикских учеников, для российских учеников оно является необоснованно сложным. Различий в формулировке и представлении данного задания обнаружено не было.

**M-M-11-1-2.** Данное задание демонстрирует DIF в пользу российских учеников, для таджикских учеников оно является необоснованно сложным. Различий в формулировке и представлении данного задания обнаружено не было.

**M-G-01-1-1.** Данное задание является необоснованно более сложным для таджикских учеников. Единственное отличие данного задания в российской и таджикской тетрадях – в таджикской версии рисунок, сопровождающий задание, имеет меньший масштаб.

**M-D-03-1-2.** Данное задание является необоснованно более сложным для таджикских учеников. Здесь поля для вариантов ответа не лежат на одной строке с вариантами ответов.

**M-R-03-1-1.** Данное задание демонстрирует DIF в пользу российских учеников, для таджикских детей оно является необоснованно сложным. У данного задания имеются существенные отличия как по содержанию, так и по представлению: в таджикской версии цвета написаны полностью,поля для вариантов ответа не лежат на одной строке с вариантами ответов.

**M-C-05-1-2.** Данное задание является необоснованно более сложным для таджикских учеников. Различий в формулировке и представлении данного задания обнаружено не было.

В результате проведённого анализа можно говорить о том, что возможно достижение только частичной эквивалентности SAM в России и Таджикистане. Общую шкалу рекомендуется строить на основе следующих заданий: **1, 19, 22, 28, 34, 35, 40, 43, 44.**

Данные задания имеют схожие характеристики как в рамках классической теории тестирования, так и в рамках IRT. Также данные задания проверяют одну и ту же область математического знания, не имеют разницы в переводе и оформлении и не демонстрируют DIF.

### 2.2.4 Построение единой шкалы

Как следует из предыдущего раздела работы, мы выделили 9 заданий (20% длины теста), которые функционируют одинаково в рассматриваемых странах и свободны от DIF. На их основе строилась единая шкала в Таджикистане и Новгородской области.

На данный момент существует множество методов построения одной шкалы (Vale, 1986; Карданова, Нейман, 2003). Один из них – одновременная калибровка (simultaneous calibration) (Карданова, Нейман, 2003). Он предполагает следующую процедуру. Ответы новгородских и таджикских учеников на 9 заданий объединяются, ответы на остальные 22 задания (по 11 на каждую выборку) присоединяются отдельно, при этом база данных приобретает вид:



***Рис 11.*** *Схема построения единой шкалы для новгородских и таджикских данных*

Далее происходит оценка подготовленности испытуемых и трудностей заданий в рамках модели Rasch (для этого использовался пакет Winsteps). Карту заданий и испытуемых можно увидеть в приложении 3. Таким образом, поскольку имеется «пересечение» в данных, оценки подготовленности и трудностей заданий попадают на одну (и центрируются относительно общего среднего значения трудностей заданий).

После построения общей шкалы оценки испытуемых были переведены в 1000-балльную шкалу с помощью двухшагового преобразования:

1. Преобразование оценок подготовленности в z-шкалу с помощью формулы

$z\_{i}=\frac{x\_{i}-\overbar{x}}{S\_{x}}$, где формула 8

Xi – значение подготовленности одного испытуемого в логитах,

$\overbar{x}$ – средняя подготовленность,

Sx – стандартное отклонение подготовленности в логитах.

Средняя подготовленность для 816-ти испытуемых составила 0,04, стандартное отклонение – 2,02 лонгета.

1. Преобразование оценок подготовленности из z-шкалу в 1000-балльную шкалу по формуле:

$y\_{i}=500+100×z\_{i}$, где формула 9

Yi  - балл испытуемого на 1000-балльной шкале,

Zi – балл испытуемого в z-шкале.

Обратимся к анализу получившихся данных.

После получения оценок испытуемых на общей шкале было проведено повторное шкалирование результатов тестирования с тем, чтобы определить, изменилось ли распределение учеников по уровням освоения материала. Поскольку процедура шкалирования была описана в начале данной главы, мы не будем приводить расчёты еще раз, укажем лишь получившиеся пороговые значения. Стоит отметить, что пороговые значения устанавливались отдельно для двух стран, поскольку метод одновременной калибровки предполагает наличие уникальных для каждой страны заданий, которые дают уникальный вклад в оценки подготовленности испытуемых.

***Таблица 12.*** *Распределение пороговых баллов*

|  |  |  |
| --- | --- | --- |
|  | **Таджикистан** | **Новгород** |
| **порог 1** | 409 | 423 |
| **порог 2** | 500 | 500 |

Порог 1 показывает, какое количество баллов необходимо набрать ученику (на 1000-балльной шкале) чтобы перейти на 1-ый уровень освоения материала, порог 2 – необходимое количество баллов для перехода на 2-ой уровень освоения материала.

Обратимся к рассмотрению результатов установления единой шкалы для Таджикистана и Новгородской области.

На рисунке 12 представлено распределение баллов учеников из Новгородской области и Таджикистана на 1000-балльной шкале (со средним 500 и стандартным отклонением 100). Видно, что распределение напоминает бимодальное. Ниже среднего балла в основном сконцентрировались таджикские ученики, в районе среднего и выше – новгородские.



***Рис. 12****. Распределение баллов учеников из Таджикистана и Новгородской области*

Средняя подготовленность учеников в логитах очень близка к нулю, что говорит о том, что данный тест хорошо центрирован относительно общей группы испытуемых. Однако, это достигается за счёт того, что общими являются только 9 заданий, остальные задания (11 для России и 11 для Таджикистана) являются уникальными (см.Приложение 3). Разброс подготовленностей очень высокий, за счет соединения выборок.

|  |  |
| --- | --- |
| **средняя подготовленность** | 0,045502 |
| **дисперсия** | 4,090278 |
| **станд.отклон** | 2,022444 |

Сравнение по уровням освоения материала

Поскольку 3-ий уровень освоения материала не сформирован в Таджикистане и данные задания были удалены из анализа, сравнить распределение учеников по уровням освоения материала в Новгородской области и Таджикистане возможно только для 2-ух уровней

После установления одной шкалы распределение учеников по уровням в Новгородской области и Таджикистане изменилось.

***Таблица 13.*** *Распределение учеников по уровням освоения материала*

|  |  |  |
| --- | --- | --- |
| **уровень** | **Таджикистан** | **Новгород** |
| **0** | 0,55 | 0,03 |
| **1** | 0,44 | 0,26 |
| **2** | 0,02 | 0,72 |

***Рис. 13****. Распределение учеников по уровням освоения материала*

В Таджикистане на 0 уровне после установления единой шкалы оказалось меньше учеников (на 0,05 – всего 223 человека), количество учеников на первом уровне освоения материала увеличилось с 0,4 до 0,44 (стало 178 человек). Также увеличилось число учеников, демонстрирующих 2-ой уровень освоения материала (до установления общей шкалы было 2 человека, после – 7). По представленным данным можно заключить, что с установлением общей шкалы оценка подготовленности таджикских учеников сдвинулась вверх. Этот же вывод справедлив и для новгородских учеников, однако, здесь разница гораздо больше: если до установления общей шкалы на 2-ом уровне освоения материала находилось около 32% учеников, то после установления общей шкалы более 70% учеников оказались на 2-ом уровне освоения материала. Это объясняется разницей подготовленности учеников из Новгорода и Таджикистана. После установления общей шкалы средняя подготовленность уменьшилась, за счет включения слабых учеников из Таджикистана, в результате чего оценка подготовленности Новгородских учеников существенно повысилась.

Обратимся к нормативной интерпретации результатов с использованием процентилей.

10-ый процентиль – 359 баллов

25-ый процентиль – 401 балл

50-ый процентиль (медиана) – 458 баллов

90-ый процентиль – 558 баллов

10% наиболее слабых учеников имеют балл от 234 до 359. 25% самых слабо подготовленных учеников имеют балл 401 и ниже. Половина испытуемых имеет балл, лежащий в пределах от 234 до 458. И, наконец, 90% опрошенных учеников набрали 558 баллов и меньше. 10% самых сильных учеников имеют балл от 558 до 647.

Посмотрим на то, как распределились студенты по процентилям с учётом страны.

***Таблица 14****. Процентильное распределение по странам*

|  |  |  |  |
| --- | --- | --- | --- |
| **Процентиль** | **Балл** | **Таджикистан** | **Новгород** |
| **число учеников** | **число учеников** |
| **10-ый процентиль** | 359 | 81 | 1 |
| **25-ый процентиль** | 401 | 205 | 2 |
| **50-ый процентиль** | 458 | 370 | 36 |
| **90-ый процентиль** | 558 | 408 | 329 |

Ниже 359 баллов в Новгородской области получил всего 1 человек, балл, находящийся в промежутке между 359 и 401 баллами имеет также один человек. В то время как в Таджикистане в нижние 10% попали 81 человек, в следующие 15% еще 124 человека. В нижних 25% опрошенных располагаются более половины учеников из Таджикистана. 458 баллов и меньше получили 36 учеников из Новгородской области и 370 учеников из Таджикистана. В 10% учеников, имеющих наиболее высокий балл, не попал никто из Таджикистана, из Новгородской области больше 558 баллов получили 79 учеников.

## Выводы

Эмпирическая часть данного исследования состояла из двух этапов:

1. Подготовительного,
2. Содержательного.

На первом этапе была заново создана база с результатами исследования в Таджикистане, создана выборка для данных, полученных в Новгородской области с тем, чтобы уравновесить размер групп. Шкалирование результатов тестирования отдельно по странам выявило большую разницу в достижениях учеников в Новгородской области и Таджикистане. В Таджикистане ни один ученик не достиг 3-его уровня освоения материала, и всего 0,4% выборки достигли 2-го уровня. В Новгородской области 3-его уровня достигли 12% учеников, 2-го уровня – 20% опрошенных. Психометрическое исследований заданий (как в рамках КТТ, так и в рамках IRT) результатов тестирования в двух странах показало, что все задания находятся в согласии с моделью Rasch. Однако 3-ий уровень освоения материала не сформирован учениками из Таджикистана, поэтому построение общей шкалы возможно только на 30 заданиях (1-ый и 2-ой уровень освоения материала).

Содержательный этап заключался в установлении эквивалентности конструктов, метода и заданий. Исследование карт конструктов российской и таджикской версий не дало достаточно информации о том, чтобы можно было сделать вывод о том, эквивалентны конструкты или нет. Однако с помощью анализа формулировок заданий на таджикском языке удалось установить, что конструкты различаются по одному разделу из пяти. Таким образом, была установлена частичная эквивалентность конструктов. Из построения общей шкалы были исключены ещё 2 задания.

Установить методную эквивалентность заданий не представляется возможным, поскольку все возможные способы, рассмотренные в первой главе данной работы, должны быть заложены при разработке инструмента.

Эквивалентность заданий устанавливалась с помощью DIF-анализа. DIF-анализ проводился с помощью 4-ёх методов: Мантель-Ханцель, Стандартизации, Логистической регрессии и t-статистики. В результате мы установили, что 11 из 28 заданий демонстрируют DIF. Мы показали, что DIF могут вызывать не только изменения в формулировке задания при переводе, но также и представление, оформление задания. Сюда входит расположение вариантов ответа и масштаб и чёткость рисунков.

Построение общей шкалы для двух стран осуществлялось методом вертикального выравнивания. Было выделено 9 заданий, которые показали близкое функционирование (как в рамках КТТ, так и в IRT), не имеют никаких искажений при переводе и в визуальном представлении, а также свободны от DIF. На основе этих 9-ти заданий была построена общая шкала. В результате подготовленность испытуемых из Новгородской области и Таджикистана оценена на одной шкале и возможно проведение сравнений по баллам на 1000-балльной шкале. Однако, пороги для распределения учеников по уровням были рассчитаны для каждой страны отдельно.

Таким образом, можно говорить о том, что достигнута частичная эквивалентность результатов тестирования SAM в Новгородской области и Таджикистане.

# Заключение

Данная работа продолжает обширный ряд исследований, посвящённых вопросу сопоставимости результатов кросс-культурных исследований. В отличие от большинства исследований, мы рассматривали сопоставимость результатов тестирования, которое изначально не планировалось как кросс-культурное. Эта особенность вносит существенные коррективы в те меры установления сопоставимости, которые можно применить к тесту. Так, все меры, которые должны быть проведены до начала сбора на данных, на этапе разработки, оказываются недоступными в нашем случае. Фокусом нашего исследования была возможность установления сопоставимости результатов исследования SAM, полученных в Новгородской области и Таджикистане.

Для достижения этой цели мы в первую очередь обратились к анализу мировой практики сопоставимости результатов тестирования. В мировой практике сопоставимость результатов исследования обозначают понятием эквивалентность. Выделяется 3 вида эквивалентности: конструктная, методная и эквивалентность заданий, достижение которых позволяет говорить о сопоставимости результатов тестирования в разных странах. С помощью обзора литературы нам удалось выделить ряд мер, направленных на установление сопоставимости результатов кросс-культурных исследований.

Эмпирическая часть данного исследования была направлена на то, чтобы показать сопоставимость результатов тестирования SAM в Новгородской области и Таджикистане, используя ряд мер, сформулированных в теоретической части данной работы. Эмпирическая часть состоит из двух этапов: подготовительного и основного. В рамках подготовительного этапа было проведено исследование баз данных Новгородской области и Таджикистана, сделана выборка для Новгородской области с тем, чтобы уравновесить число испытуемых в Новгородской области и Таджикистане. Также, было показано, что в Таджикистане 3-ий уровень освоения материала не сформирован. В рамках содержательного этапа мы исследовали конструктную эквивалентность и эквивалентность заданий.

Особенностью нашей работы является то, каким образом устанавливалась конструктная эквивалентность. Мы не использовали эксплораторный и конфирматорный факторные анализы. Для установления эквивалентности использовались «карты конструкта», а также анализ перевода заданий. Нам удалось показать, что между российской и таджикской версиями существует частичная эквивалентность. В результате проведения DIF-анализа мы установили, что задания могут несправедливо оценивать испытуемых не только из-за изменения в формулировке, но также из-за разницы в представлении задания. В предыдущих исследованиях на тему эквивалентности мы не нашли указаний на то, что следует обращать внимание на представление заданий. В результате было выделено несколько заданий, на которых была построена общая шкала для Таджикистана и Новгородской области методом вертикального выравнивания.

Подводя итог, хотелось бы отметить, что установление эквивалентности между версиями теста, который изначально не задумывался как кросс-культурный, требует особой изобретательности от исследователей. Наша работа имеет несколько аспектов, которыми могут воспользоваться другие исследователи в данной области. Во-первых, мы приводим подробный анализ перевода заданий и обращаем внимание на необходимость проверки визуального представления задания, который включает в себя такие черты как шрифт, масштаб, расположение вариантов ответов и рисунки. Во-вторых, мы описываем анализ «карты конструкта», что также мало освещено в литературе на сегодняшний день. В-третьих, мы используем метод вертикального выравнивания, который не является очень распространённым, но позволяет создать одну шкалу для данных, имеющих только частичную эквивалентность. Однако, не стоит забывать, что наше исследование использовало не все потенциально доступные средства для установления сопоставимости результатов тестирования. Мы не проводили эксплораторный и конфирматорный факторные анализы, в силу специфики имеющихся данных. В дальнейшем мы планируем преодолеть это ограничение с помощью привлечения других данных: результатов тестирования SAM в Киргизии и Казахстане.

# Список использованных источников:

1. AERA, APA & NCME, 1999. Standards for Educational and Psychological Testing.
URL: << http://www.apa.org/science/programs/testing/standards.aspx>>
2. Barbara B. Ellis (1989). Differential Item Functioning: Implications for Test Translations. Journal of Applied Psychology, Vol. 74, No. 6,912-921
3. Bracken, B. A., & Barona, A. (1991). State-of-the-art Procedures for Translating, Validating, and Using Psychoeducational Tests for Cross-Cultural Assessment. School Psychology International, 12, 119-132.
4. Chiu P.C. The effect of English proficiency on mathematics performance: a comparison of the Item Response Theory-based Area and Mantel-Haenszel methods.
5. Clauser, E. B., & Mazor, M. K. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practice. 17 , 31-44.
6. difR package manual:
URL: <<http://cran.r-project.org/web/packages/difR/difR.pdf>>
7. Dif-R Package официальный сайт программы:
URL:<< http://cran.r-project.org/web/packages/difR/index.html >>
8. Dorans N.J., Kulick E (1986). Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. Journal of Educational Measurement, Vol. 23, No. 4, pp. 355-368.
9. Fons J. R. van de Vijver, Ype H. Poortinga (1997).Towards an Integrated Analysis of Bias in Cross-Cultural Assessment. European Journal of Psychological Assessment, Vol. 13, Issue 1, pp. 29–37
10. Fons van de Vijver, Hambleton, R. K (1996). Translating Tests: Some Practical Guidelines. European Psychologist, Vol. 1, No. 2, pp. 89-99.
11. Geisinger K.F. (1994). Cross-Cultural Normative Assessment: Translation and Adaptation Issues Influencing the Normative Interpretation of Assessment Instruments. Psychological Assessment, Vol. 6, No. 4, 304-312.
12. Gómez-Benito J., Hidalgo D., Padilla J.-L (2009). Efficacy of Effect Size Measures in Logistic Regression. An Application for Detecting DIF. Methodology; Vol. 5(1):18–25.
13. Hambleton, R. K. (Chair). (1993, August). Technical standards for translating tests and establishing test score equivalence. In Symposium conducted at the 101st Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.
14. International Test Commission Guidelines for Translating and Adapting Tests, version 2010.
URL: <<http://www.intestcom.org/upload/sitefiles/40.pdf>>
15. Karami H. (2012). An introduction to Differential Item Functioning. The International Journal of Educational and Psychological Assessment, Vol. 11(2), pp. 59-76.
16. Magis D., Beland S., Tuerlinckx F., De Boeck P (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 42 (3), 847-862.Vijver, Hambleton 1996
17. Magis D., De Boeck P. (2012). A Robust Outlier Approach to Prevent Type I Error Inflation in Differential Item Functioning. Educational and Psychological Measurement, 72(2), 291–311.
18. Martha L. Stocking, Frederic M. Lord (1983). Developing a Common Metric in Item Response Theory. APPLIED PSYCHOLOGICAL MEASUREMENT Vol. 7, No. 2, pp. 201-210.
19. McNamara, T., & C. Roever (2006) Language testing: The social dimension .Malden, MA & Oxford: Blackwell.
20. Millsap R.E., Everson H.T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. Applied Psychological Measurement, vol. 17 no. 4, pp. 297-334.
21. Mylonas K., Furnham A.(2013). Bias in Terms of Culture and a Method for Reducing It: An Eight-Country ''Explanations of Unemployment Scale'' Study. Educational and Psychological Measurement.
22. Nezhnov, P. (2011). SAM – toolkit to assess primary school students’ academic achievements. CADMO. Innovations in assessment to meet changing needs. ANNO XIX, 1, pp.85-98.
23. Greenfield P. (1997). YOU CAN’T TAKE IT WITH YOU: Why Ability Assessments Don’t Cross Cultures. American Psychologist, Vol. 52, No. 10, 1115-1124.
24. Rogers H.J., Swaminathan H. (1993). A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. Applied Psychological Measurement, 17: 105.
25. Hambleton R.K., John H.A.L. de Jong (2003). Advances in translating and adapting educational and psychological tests. Language Testing, 20 (2) 127–134
26. SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // под ред. Нежнова П.Г., Кардановой Е.Ю., 2011, 104 с.
27. Scheuneman J.D. Bleistein С.А. (1989). A Consumer's Guide to Statistics for Identifying Differential Item Functioning. Applied Measurement in Education, Volume 2, Issue 3, рр.255-275.
28. Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361–370.
29. Uttaro T., Millsap R.E. (1994). Factors influencing M-H procedure in the detection of DIF. Applied Psychological Measurement, vol.18, no. 1, pp.15-25.
30. Vale C.D (1986). Linking Item Parameters Onto a Common Scale. Applied Psychological Measurement, 10: 333
31. Van deVijver F., Tanzer N.K (2004). Bias and equivalence in cross-cultural assessment: an overview. Revue européenne de psychologie appliquée 54, pp. 119–135.
32. Wang W.-C., Su Y.-H. (2004). Factors Influencing the Mantel and Generalized Mantel-Haenszel Methods for the Assessment of Differential Item Functioning in Polytomous Items. Applied Psychological Measurement, 28: 450.
33. Winsteps manual:
URL:<< http://www.winsteps.com/a/winsteps-manual.pdf>>
34. Winsteps официальный сайт:
URL: <<http://www.winsteps.com/>>
35. Poortinga Y.P. (1989) Equivalence of Cross-Cultural Data: An Overview of Basic Issues, International Journal of Psychology, 24:6, 737-756
36. Zumbo B.D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. LANGUAGE ASSESSMENT QUARTERLY, 4(2), 223–233.
37. Заключительный отчет по проекту: Локализация, адаптация и пилотирование инструмента оценки школьных достижений (SAM) учащихся начальной школы Республики Таджикистан (2013).
38. Карданова Е.Ю., Нейман Ю.М (2003). Проблема выравнивания в современной теории тестирования // Вопросы тестирования в образовании, № 8.
39. Нежнов П.Г., Карданова Е.Ю., Эльконин Б.Д. Оценка результатов школьного образования: структурный подход // Вопросы образования, 2011, № 1, стр. 26-43.

# Приложение 1. Наиболее ожидаемые типы ошибок

|  |  |
| --- | --- |
| **Тип ошибки** | **Источник** |
| Конструкт | Неполное совпадение конструктов в странах |
| Различная допустимость формулировок |
| Метод | Различный уровень социальной желательности |
| Различные стили ответов в разных странах (Н: тенденция к среднему в одной стране, к крайним ответам в другой) |
| Знакомство с форматом тестирования |
| Сопоставимость выборок |
| Различия в процедуре проведения тестирования |
| Эффекты исследователя |
| Задания | Неадекватный перевод заданий |
| Измерение дополнительного конструкта заданиями |
| Различная допустимость формулировок |

# Приложение 2. Разделы содержания целевого конструкта SAM

|  |  |
| --- | --- |
| **Разделы содержания** | **Средства ориентировки математических действий** |
| *Числа и вычисления* | * последовательность натуральных чисел;
* числовая прямая;
* позиционный принцип;
* свойства арифметических действий;
* порядок действий
 |
| *Измерение величин* | * отношение между числом, величиной и единицей;
* отношение «целого и частей»;
* формула площади прямоугольника
 |
| *Закономерности* | * «индукционный шаг»;
* повторяемость (периодичность)
 |
| *Зависимости* | * отношения между однородными величинами (равенство, неравенство, кратности, разностное, «целого и частей»);
* прямая пропорциональная зависимость между величинами;
* производные величины: скорость, производительность труда и др.;
* соотношения между единицами
 |
| *Элементы геометрии* | * форма и другие свойства фигур (основные виды геометрических фигур);
* пространственные отношения между фигурами;
* симметрия
 |

(Карданова, Нежнов, 2011)

|  |  |  |
| --- | --- | --- |
| **Разделы** | **Количество блоков** | **Количество задач** |
| *Числа и вычисления* | 4 | 12 |
| *Измерение величин*  | 5 | 15 |
| *Закономерности*  | 2 | 6 |
| *Зависимости между величинами* | 2 | 6 |
| *Элементы геометрии* | 2 | 6 |
| *ВСЕГО* | 15 | 45 |

# Приложение 3. Карта заданий и испытуемых

