**Программа дисциплины**

**«Разработка инструментов измерения»**

для образовательной программы «Образование»

по направлению подготовки научно-педагогических кадров в аспирантуре 44.06.01 Образование и педагогические науки

Разработчики программы:

Канонир Татьяна Николаевна, Dr.Psych.

[tkanonir@hse.ru](mailto:tkanonir@hse.ru)

Тюменева Юлия Алексеевна, к. психол. наук

тел.: +79104075140; email: jutu@yandex.ru

Утверждена Академическим советом Аспирантской школы по образованию

«11»октября 2018 г., протокол № 34

Академический директор

Е.А. Терентьев \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*(подпись)*

Москва,2018

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения-разработчика программы.*

1. **Область применения и нормативные ссылки**

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения аспиранта и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и аспирантов, обучающихся по образовательной программе «Образование» и изучающих дисциплину «Разработка инструментов измерения».

Программа учебной дисциплины разработана в соответствии с:

* образовательным стандартом НИУ ВШЭ по направлению подготовки кадров высшей квалификации 44.06.01 Образование и педагогические науки»;
* образовательной программой «Образование»;
* учебным планом образовательной программы «Образование».

1. **Цели и задачи освоения дисциплины**

Целью освоения дисциплины «Разработка инструментов измерения» является изучение фундаментальных основ разработки инструментов измерения. В частности, аспиранты познакомятся с историей развития области измерений и связанных с ней концепций. Дисциплина направлена на формирование концептуального понимания классического и современного подходов к валидности инструментария, а также понимания ключевых этапов и принципов разработки измерительных инструментов.

1. **Компетенции обучающегося, формируемые в результате освоения дисциплины**

В результате освоения дисциплины аспиранты должны будут:

* Знать основные этапы в развитии тестирования и шкалирования, а также подходов к валидности за прошедшее столетие и до наших дней;
* Знать основные этапы разработки и валидизации измерительных инструментов;
* Уметь разрабатывать тестовые задания с выбором ответа в соответствии с действующими принципами создания заданий;
* Знать принципы классической теории тестов, ее сильные и слабые стороны.
* Способны выбрать, расчитать и использовать соответствующую статистику тестовых заданий для выявления плохо функционирующих заданий;
* Способны оценить надежность тестового балла, ошибку измерения и одномерность с использованием классических психометрических методов;
* Способны внести изменения в инструменты с целью повысить одномерность и надежность и минимизировать ошибку измерения;
* Способны описать модель Раша, 1-, 2- и 3-параметрическую IRT-модели и объяснить, как они функционируют для моделирования дихотомических ответных данных;
* Способны применять модель Раша и 1-, 2- и 3-параметрические IRT-модели для разработки качественных измерительных шкал на основе дихотомических ответных данных.

В результате освоения дисциплины аспирант осваивает следующие компетенции:

| Компетенция | Код по ФГОС/ НИУ | Дескрипторы – основные признаки освоения (показатели достижения результата) | Формы и методы обучения, способствующие формированию и развитию компетенции |
| --- | --- | --- | --- |
| Способность к критическому анализу и оценке современных научных достижений, в том числе в междисциплинарных областях. | УК-1 | Критически анализирует и оценивает планируемые и выполненные исследования, в том числе собственные.  Может точно указывать и описывать проблемные места, и предлагать способы их решения. | Анализ и обсуждения опубликованных и планируемых исследований на лекциях, семинарах, самостоятельная работа с литературой. |
| Способность генерировать оригинальные теоретические конструкции, гипотезы и исследовательские вопросы. | УК-2 | Может генерировать собственные гипотезы и исследовательские вопросы.  Может оценивать их фальсифицируемость, описывать условия принятия/отвержения гипотез. | Рефлексия относительно происхождения гипотез и исследовательских вопросов в опубликованных работах, практика постановки собственных вопросов на лекциях, семинарах, самостоятельная работа с литературой. Разработка индивидуального проекта. |
| Способность выбирать и применять методы исследования, адекватные предмету и задачам исследования. | УК-3 | Может разработать дизайн собственного исследования, подобрать адекватные методы, обосновать их применение и применить. | Анализ и обсуждения различных методов исследований в области измерений. Разработка индивидуального проекта. |
| Способность собирать, анализировать, обрабатывать и хранить данные в соответствии с общепринятыми научными и этическими стандартами. | УК-4 | Знает правила формирования выборки, подходов и процедур сбора данных, их обработки и хранения.  Может выбрать для анализа данных соответствующий способ анализа данных и корректно представить результаты. | Анализ и обсуждение правил формирования выборки, подходов и процедур сбора данных, их обработки и хранения на лекциях и семинарах, практические задания на анализ данных в ходе семинаров, разработка индивидуального проекта, |
| Способность осуществлять комплексные исследования, в том числе междисциплинарные, на основе целостного системного научного мировоззрения. | УК-5 | Может осуществить планирование исследования, логически согласованное с текущим состоянием дел в области.  Может собрать и проанализировать данные в соответствии с планом.  Может интерпретировать результаты в рамках поставленных вопросов и существующих теорий в данной области. | Рефлексивный анализ существующих работ, практика планирования, анализа и интерпретации результатов в ходе лекций, семинаров, самостоятельной работы с литературой и разработки индивидуального проекта. |
| Способность самостоятельно осуществлять научно-исследовательскую деятельность в области образования с использованием современных методов исследования и информационно-коммуникационных технологий | ОПК-1 | Может самостоятельно подобрать метод исследования, обосновать его, применить, подобрать и применить программное обеспечение для анализа. | Разработка индивидуального проекта. |
| Способность следовать этическим нормам в профессиональной деятельности | ОПК-8 | Знает этические нормы своей профессиональной деятельности. Может описать типичные этические нарушения норм и ситуации, способствующие их возникновению. Может описать последствия нарушений для участников, самого исследователя и науки в целом. | Чтение соответствующей литературы, групповое обсуждение проблемных случаев на семинарах. |

# Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу дисциплин по выбору и изучается на 1-м году обучения. Освоению данной дисциплины должно предшествовать изучение дисциплин, посвященных методологии научного исследования в социальных науках.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении дисциплин:

* Углубленная психометрика,
* Углубленная статистика.

# Формы контроля знаний аспирантов:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | | | | Параметры \*\* |
| 1 | 2 | 3 | 4 |
| Текущий  (неделя) | Домашние задание |  |  | \* | \* | Ответы на вопросы на занятии |
|  |  |  | \* | Письменная работа |
| Домашнее задание |  |  | \* |  | Описание индивидуального проекта |
| Итоговый | Экзамен |  |  |  | \* | тест, презентация и защита индивидуального проекта |

* 1. **Критерии оценки знаний, навыков**

Для текущего контроля в течение курса используются опрос, групповая дискуссия, индивидуальные презентации, письменные и устные домашние работы. Примеры вопросов и заданий, а также критерии их оценки приведены в п. 9.1.

Итоговый контроль проводится в форме экзамена.Экзамен состоит из двух частей. В первой части студенты сдают тест. Вторая часть экзамена представляет собой защиту индивидуального проектного задания. Примеры заданий представлены в п.9.2.

# Содержание дисциплины

**Раздел 1. Введение в теорию измерений.**

Виды оценивания и их цели. Формирование современных представлений об измерениях. Краниология. Ииследования в области психофизологии: Вебер, Фехнер, Вундт. Ф.Гальтон и его исследования интеллекта. Первые стандартизированые тесты на интеллект. История измеренй в России до 1936 года.

Введение в шкалирование. Подходы к шкалированию в период создания первых тестов на интеллект. Эксперименты Терстоуна со шкалами. Шкалы Ликерта и Гутмана.

Факторная структура инструмента и идея факторного анализа. Тесты на интеллект как пример влияния использования факторного анализа на развитие теоретических представлений об интеллекте.

**Раздел 2. Развитие представлений о валидности**

Различные модели валидности в истории психометрики. Критериальная валидность как золотой стандарт в 1920-50 гг. Текущая и Прогностическая валидность. Проблема критерия. Факторы, влияющие на прогностическую валидность. Стандартная ошибка оценки. Фундаментальные ограничения критериальной модели.

Факторная структура – как база для конструктной валидности. Ограничения факторной модели.

Содержательная валидность как результат неудовлетворенности критериальной и факторными моделями. Выборка реального решения задач – как оценивание общего уровня навыков деятельности. Типовые области применения содержательной модели валидности. Преимущества и недостатки содержательной модели.

Происхождение идеи конструктной валидности. Теоретическая рамка конструктной валидности Кронбаха и Мила. Гипотетико-дедуктивная модель научной теории. Номологическая сеть конструктов и методы оценки конструктной валидности (внутренняя структура, эксперименты, групповые различия, MTMM матрицы и др.). Расширенная модель валидизации как научного приключения.

Непрямые результаты тестирования как основания обсуждать новый компонент валидности – социальные последствия тестирования. Неполная репрезентативность и конструктно-нерелевантная дисперсия в тестовых показателях. Невалидное использование теста.

Конструкная валидизация как база для Унифицированной модели валидности (Loevinger, Messick). Эффекты введения новой модели в стандарт APA. Подход к валидности как к аргументу (Kane, 1992).

**Раздел 3. Разработка и моделирование тестовых заданий с выбором ответа**

Разработка инструмента, рамка теста в парадигме ECD (Evidence-Centered Design). Будет рассмотрен весь процесс разработки от начала и до конца: анализа области исследования; определения цели тестирования; разработки 4-х процессной архитектуры предъявления теста (four- process delivery architecture for the assessment) респонденту в рамках инструмента оценивания.

Этапы разработки согласно ECD: анализ области компетенций (Что мы знаем об этой области? В каких ситуациях она важна? Как проявляются компетенции? Каковы ключевые компетенции?), моделирование области компетенций (Как мы концептуализируем ключевые аспекты области компетенций? Сколько субкомпетенций? Как они связаны между собой?), разработка концептуальной рамки тестирования (CAF) (Каковы основные элементы теста? Модель учащегося, модель тестового задания, модель свидетельства), воплощение теста (Наполнение основных элементов теста содержанием: создание конкретных заданий, статистические модели), применение теста (Взаимодействие участников тестирования с тестом, оценка компетенций, обратная связь).

Система предъявления теста респонденту состоит из: модуля выбора (выбор тестовых заданий из банка заданий и организация их в определенном порядке в соответствии с моделью тестирования); модуля предъявления задания респонденту; модуля ответа (предоставляет свидетельства об уровне тестируемого); модуля результата (собранные свидетельства используются для обновления данных о первоначальных представлениях об уровне трестируемого).

Анализ области компетенций. Латентные конструкты и их определение. Дискурс складывается из понимания самого латентного признака, его связи с другими концептами в социальном контексте, а также ситуации и способов измерения. SEMP – парадигма стандартных образовательных измерений. Что мы знаем об этой области? Какой подход к рассмотрению выбираем (когнитивный, бихевиористский, интеракционистский и т.п.). Ответ является свидетельством конструкта: как обобщать результаты тестирования и как полученные данные могут указывать на скрытые характеристики латентного конструкта.

Моделирование области компетенций. Представление предметной области и спецификации теста: как мы концептуализируем ключевые аспекты области компетенций, сколько суб-компетенций достаточно для измерения латентного конструкта, как они связаны между собой, как связаны элементы ситуаций, в которых компетенция может проявиться. Концептуальная рамка теста. Определения компетенций (student model). Определение уровней компетенций и их составляющих, и способов описания каждого из уровней. Модель сбора свидетельств (evidence model). Поиск наблюдаемых свидетельств наличия компетентности. Модель тестового задания (task model). Где измерять? Поиск конкретных индикаторов – результатов выполнения какой-либо тестовой задачи, то есть модель задания - это описание характеристик набора тестовых заданий (тип сценария, который требуется для получения явного свидетельства, тип стимульного материала, который может быть использован, то есть тип ситуации решения проблемы и т.д.).

Глубина знаний: тип и уровень мыслительных процессов тестируемых, необходимых для решения задачи и ответа на вопрос; особенности выполнения задания.

Согласование содержания и представления. Обзор разнообразных методик проведения согласования, которые варьируются по уровню сложности: низкая сложность - тестовые задания в тесте соотносятся с соответствующими содержательными стандартами; средняя сложность - тестовые задания соответствуют содержательным и когнитивным; высокая сложность - вводятся дополнительные критерии.

Разработка спецификации теста и разработка заданий с выбором ответа. Виды заданий, выбор ответных категорий.

Применение универсального дизайна при разработке инструмента. Как представлять свидетельства и тестовые задания таким образом, чтобы они отвечали особым потребностям некоторых студентов, как разработать вопросы, чтобы подходили как можно большему количеству разработчиков теста. Справедливость оценивания.

Полевое исследование и его цели. Использование когнитивных лабораторий, цели и общие процедуры проведения. Роль полевого исследования в ситуации адаптации инструмента и обосновании кроскультурной конструктной валидности. Выборка для полевого исследования.

**Раздел 4. Классическая теория тестирования**

4.1. Характеристики тестовых заданий: трудность и дискриминативность. Типы тестовых баллов: первичный (сырой) балл, шкалированный балл, процентилльный балл, z-балл. Классическая теория тестирования (КТТ): предположения, модели, ограничения. Концепция надежности в КТТ. Процедуры и методы вычисления надежности результатов тестирования (формула альфа Кронбаха; метод расщепления, тест-ретест). Концепция и ошибки измерения истинного тестового балла.

Проблема надежности шкалы. Способы оценки надежности: ретестовая, split-half, альфа Кронбаха. Связь надежности и ошибки измерения в классической тестовой теории. Ограничения альфа Кронбаха и альтернативные возможности (омега Макдональда и др.).

4.2. Анализ главных компонент и эксплораторный факторный анализ. Анализ главных компонент: цель и подход, общее представление о математических процедурах. Различия в задачах и результатах анализа главных компонентах и собственно факторного анализа. Вращения: ортогональные, косоугольные, бифакторный. Математическая эквивалентность разных видов вращения. Факторные оценки: проблема определённости факторных оценок и подходы к их расчёту.

Критерии определения количества факторов: график собственных значений, Minimum Average Partial, параллельный анализ. Содержательные различия между структурами с разным количеством измерений (в методы главных компонент и факторном анализе). Иерархические структуры данных и факторный анализ. Сравнение факторных структур: проблема инвариантности. Оценка метрической инвариантности с использованием прокрустова вращения и коэффициентов сходства факторов (коэффициент пропорциональности фи Такера и др.).

Различия между моделью эксплораторного и конфирматорного факторного анализа. Индексы соответствия. Вложенные модели: EFA, ESEM, CFA. Решение проблемы инвариантности измерений с помощью мультигруппового КФА: этапы и процедуры.

Анализ качества заданий по итогам апробации.Рекомендации по удалению и переработке заданий. Модификация заданий и наборов заданий на основе эмпирических свидетельств.

**Раздел 5. Введение в современные методы**

5.1. Модели Раша. Обзор основных моделей Раша. Примеры. Карты переменных. Принципы измерений при построении тестов и шкал.

Понятие характеристической кривой задания и характеристической кривой категории задания. Информационные функции. Оценивание параметров методами максимального правдоподобия. Отделение параметров. Процедура Ньютона-Рафсана.

Исследование согласия экспериментальных данных тестирования с используемой моделью измерения в рамках моделей Раша. Статистики согласия, основанные на стандартизованных уклонениях. Свойства статистик согласия и их распределений. Исследование согласия: параллельный анализ, анализ принципиальных компонент остатков.

Инвариантность параметров: d-plots, z-tests, карты переменных и их устойчивость. Применение инвариантности для практических проблем тестирования.

Компьютерные программы анализа данных в рамках моделей Раша: Winsteps, Conquest. Интерпретация аутпутов компьютерных программ анализа данных в рамках моделей Раша.

**5.2. Современная теория тестирования (IRT).** Однопараметрическая модель IRT – математическая репрезентация, предположения, свойства параметров. Двухпараметрическая модель IRT – математическая репрезентация, предположения, свойства параметров.Трехпараметрическая модель IRT – математическая репрезентация, предположения, свойства параметров. Оценивание параметров 1PL, 2PL и 3PL IRT моделей.

Исследование согласия экспериментальных данных тестирования с используемой моделью измерения для IRT моделей.

Компьютерные программы анализа данных в рамках моделей IRT. R- пакет как универсальное ПО для анализа данных в рамках IRT.

# Образовательные технологии

Лекции, семинары, самостоятельная работы, письменные самостоятельные работы, практические занятия, работа по группам, общегрупповая дискуссия, презентации и их защита.

# Оценочные средства для текущего контроля и аттестации аспиранта

Для текущего контроля используются опрос, групповая дискуссия, индивидуальные презентации, письменные и устные домашние работы.

## Тематика заданий текущего контроля

***Примерные вопросы для текущего контроля***:

1. Пример вопроса для обсуждения: Каким образом использование факторного анализа способствовало формированию представлений об интеллекте?
2. Пример задания для письменного домашнего задания: разработать задания с выбором ответа по заданной тематике.

***Задание для индивидуального проекта:***

Индивидуальный проект "Разработка и валидизация инструмента оценки".

Ожидается, что в качестве демонстрации усвоения материала дисциплина аспиранты готовят и защищают индивидуальный проект по разработке и валидизации инструмента оценки учебных достижений. С учетом временных ограничений для такой работы предполагается, что результатом проекта станет не полностью готовый инструмент, но детальный план его создания и валидизации, включающий спецификацию, формат теста, примеры заданий и ответов, с обоснованием их типа, описание процедуры скоринга и шкалирования. Должна быть описана процедура администрирования, и докладывание результатов тестирования. Валидизация должна быть описана как подробное руководство для ее проведения.

Вместо реальных данных аспиранты будут использовать симулированные данные для анализа и интерпретации.

В качестве предмета оценки аспиранты могут выбрать учебную программу 1-6 класса школы по любому предмету, и основывать свой инструмент на содержании соответствующих методических материалов (учебных комплектов), а также на ФГОС и имеющихся оценочных инструментах (внешняя оценка, итоговые контрольные и т.д.).

Критерии оценки индивидуального проекта включают умения:

1. строить стратегию разработки и валидизации инструмента с учетом цели его дальнейшего использования, особенностей всех участников оценивания, и всего широкого социального контекста;

2. планировать и обосновывать дизайн инструмента (форма, типы заданий и ответов, скоринг, шкалирование);

3. планировать, проводить психометрический анализ и давать интерпретацию результатам;

4. указывать на ограничения используемых психометрических теорий и моделей; предлагать варианты их преодоления;

5. адекватно применять или обосновывать слабости разных дизайнов валидизации (содержательной, критериальной, конструктной);

6. предвидеть угрозы для валидности инструмента и интерпретации результатов (например, специальная подготовка детей к тесту, или резкие различия в социально-экономическом статусе учащихся);

7. указывать на ограничения в применении теста, предвидеть и указывать на способы превенции негативных социальных последствий;

8. степень детализации и конкретизации проекта.

## Проведение итогового контроля

Итоговый контроль проводится в форме экзамена.Экзамен состоит из двух частей. В первой части студенты сдают тест. Вторая часть экзамена представляет собой защиту индивидуального проекта "Разработка и валидизация инструмента оценки".

Критерии оценки защиты индивидуального проекта:

- логичность и аргументированость изложения;

- отражена вся необходимая информация (см.п.8.1.);

- корректное использование терминологии.

# Порядок формирования оценок по дисциплине

Оценивается работа на занятиях по обязательной литературе и домашние задания. На оценку влияет: правильность выполнения домашних работ (и их своевременная сдача), демонстрация и знания заданного материала, активность на занятиях.

Преподаватель оценивает работу аспирантов на семинарских и практических занятиях: - воспроизведение прочитанного материала, демонстрация способности видеть связи между основными понятиями, задавание вопросов, отвеы на вопросы преподавателя, участие в дисскуссиях и в групповой работе. Оценки за работу на семинарских занятиях преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале за работу на семинарских и практических занятиях определяется перед промежуточным или итоговым контролем - *Оаудиторная*.

Преподаватель оценивает самостоятельную работу аспирантов: оценивается правильность выполнения заданий домашних работ и индивидуального проекта и их своевременная сдача, полнота выполнения, ответы на вопросы аудитории при докладах. Оценки за самостоятельную работу аспиранта преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале за самостоятельную работу определяется перед промежуточным или итоговым контролем – *Осам. работа*.

Накопленная оценка за текущий контроль учитывает результаты аспиранта по текущему контролю следующим образом:

*Отекущий* = *0.3·Оауд. + 0.2·Одз + 0.5·Оиндив.проект* ;

Способ округления накопленной оценки текущего контроля: арифметический (например, оценка 4,4 округляется до 4, а оценка 4,5 до 5.

Результирующая оценка за итоговый контроль в форме теста выставляется в 10бальной шкале и формируется из оценки за тест и оценки за защиту проекта следующим образом:

*Орезульт* = *0.5·Отест + 0.5·Озащита*

Итоговая оценка складывается из суммы ткущей оценки и итогового теста и преводится в 10-бальную шкалу.

Аспирант не имеет возможности пересдать низкие результаты за текущий контроль или работу на занятиях, самостоятельную работу, если за любой из этих видов работы была выставлена оценка.

В случае если аспирант пропустил срок сдачи отчетности по текущему контролю или самостоятельной работе по уважительной причине (к уважительным причинам может быть отнесена болезнь в указанный период, подтвержденная справкой из медицинского учреждения) допускает сдача форм контроля не позднее чем через две недели с момента окончания действия справки, если другие сроки не установлены деканатом. В случае, если аспирант не сдает работу в указанный срок, ему выставляется оценка «0».

За каждый день просрочки сдачи домашней работы по неуважительной причине итоговая оценка за эту домашнюю работу снижается на 1 балл (по 10-бальной шкале)

***Условия пересдач***

На пересдаче аспиранту не предоставляется возможность получить дополнительный балл для компенсации оценки за текущий контроль. Первая пересдача проводится преподавателем, отвечающим за чтение дисциплины на факультете. При выставлении результирующей оценки учитывается накопленная оценка *Отекущий* за текущий контроль.

Вторая пересдача проводится в присутствие комиссии, включающей не менее трех преподавателей. При выставлении результирующей оценки учитывается накопленная оценка *Отекущий* за текущий контроль. В исключительных случаях комиссия может принять решение об отказе от учета накопленной аспирантом текущей оценки. В данном случае комиссией оценивается выполнение теста и защита проектного задания.

В исключительных случаях комиссия может принять решение об отказе от учета накопленной аспирантом оценки. В данном случае комиссией используются следующие критерии оценки ответа аспиранта: полнота ответа на вопрос, самостоятельность при ответе на вопрос, точность в ответе на поставленные вопросы.

В диплом выставляет результирующая оценка по учебной дисциплине, которая формируется по следующей формуле:

*Одисциплина = 0,5· Отекущий* + *0,5*·*Оитоговый*

Способ округления результирующей оценки по учебной дисциплине: арифметический (например, оценка 4,4 округляется до 4, а оценка 4,5 до 5).

# Учебно-методическое и информационное обеспечение дисциплины

## Основная литература

1. Adams, R. J., Wu, M. L., & Wilson, M. R. (2012). ACER ConQuest 3.0.[computer program]. Melbourne: ACER.
2. American Educational Research Association et al. Standards for educational and psychological testing. – American Educational Research Association, 1999.
3. Ames, A. J., & Penfield, R. D. (2015). An NCME Instructional Module on Item Fit Statistics for Item Response Theory Models. Educational Measurement: Issues and Practice, 34(3), 39-48.
4. Badia X, Prieto L, Linacre J.M. (2002). Differential Item and Test Functioning (DIF & DTF). Rasch Measurement Transactions, 16:3 p.889
5. Brown, T.A. (2015). Confirmatory Factor Analysis for Applied Research. 2nd Ed. N.Y.: The Guilford Press.
6. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.
7. Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1-29.
8. Cronbach, L. J. & Meehl, P. E. (1955).  Construct validity in psychological tests.  Psychological Bulletin, 52, 281-302.
9. Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. Educational and psychological measurement, 64(3), 391-418.
10. DeVellis, R. F. (2006). Classical test theory. Medical care, S50-S59.
11. DiStefano, C., Zhu, M., &Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, Research & Evaluation, 14(20), 1-11.
12. Ebel, R. L. (1961). Must all tests be valid? American Psychologist, 16(10), 640.
13. Embretson S.E., Reise S.P. (2000). Item Response Theory for Psychologists. New Jersey: Lawrence Erlbaum Associates, Publishers. Ch. 4, 7-13.
14. Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. Psychometrika, 46, 59–77.
15. Fischer, G. H., & Molenaar, I. W. (Eds.). (2012). Rasch models: Foundations, recent developments, and applications. Springer Science & Business Media.
16. Gregory, J.R. (2011). Psychological Testing. History, Principles, and Applications. 6th edition, Allyn & Bacon: Boston. Chapter 2; pp.132-138 (про шкалирование); Chapter 5 (про факторный анализ).
17. Hambleton R.K., Swaminathan H., Rogers H.J. Fundamentals of Item Response Theory (Measurement Methods for the Social Science). Sage Publications, Inc., 1991
18. International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). [www.InTestCom.org]
19. ITC Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores. URL: https://www.intestcom.org/files/guideline\_quality\_control.pdf
20. Kane, M. T. (1992). An argument-based approach to validity. Psychological bulletin, 112(3), 527.
21. Kane, M. T. (2016). Explicating validity. Assessment in Education: Principles, Policy & Practice, 23(2), 198-211.
22. Kiefer, T., Robitzsch, A., & Wu, M. (2018). TAM: Test analysis modules. R package.
23. Kirkpatrick, R., Way, W.D. (2008). Field Testing and Equating Designs for State Educational Assessments. Paper presented at the annual meeting of the American Educational Research Association, New York, NY, March 2008
24. Lewis, C. (2006). Selected Topics in Classical Test Theory. Handbook of statistics, 26, 29-43.
25. Linacre J.M. (2010) When to stop removing items and persons in Rasch misfit analysis? Rasch Measurement Transactions, 2010, 23:4, 1241
26. Linacre, J. M. (2011). Winsteps®(Version 3.72. 3). Computer Software]. Beaverton, Oregon: Winsteps. com. Available from http://www. winsteps. com.
27. Linacre, J.M. (2010). Predicting Responses from Rasch Measures. Journal of Applied Measurement. Vol. 11(1).
28. Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological reports, 3(3), 635-694.
29. Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. Methodology, 2(2), 57-64.
30. Mair, P., and Hatzinger, R. 2007. Extended Rasch modeling: The eRm package for the application of IRT models in R. Journal of Statistical Software, 20, 1–20.
31. Messick, S. (1987). Validity. ETS Research Report Series, 1987(2).
32. Messick, S. (1990). Validity of Test Interpretation and Use. Report. ETS. http://files.eric.ed.gov/fulltext/ED395031.pdf
33. Miller P., Linn R. L., Gronlund N. E. (2008). Validity: Measurement and assessment in teaching. Chapter 7, 8, 10.
34. Mislevy, R., & Almond, R., and Lukas, J. (2004). A Brief Introduction to Evidence-Centered Design, CSE Report 632, Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; US Department of Education.
35. Murphy, K. R., & Davidshofer, C. O. (1988). Psychological testing. Principles, and Applications, Englewood Cliffs. Chapter 11.
36. Partchev I. A visual guide to item response theory. (2004). Friedrich-Schiller-Universit¨at Jena. https://www.metheval.uni-jena.de/irt/VisualIRT.pdf
37. Reckase, M. D. (1997). The past and future of multidimensional item response theory. Applied Psychological Measurement, 21(1), 25-36.
38. Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. Educational and psychological measurement, 48(3), 657-667.
39. Stage, C. (2003). Classical test theory or item response theory: the Swedish experience (Vol. 42). Univ..
40. Stevens, S.S. (1946). On the theory of scales of measurement. Science, 103,667-680.
41. Tabachnick B.L., Fidell, L.S. (2007). Using Multivariate Statistics. 5th Ed. Pearson Education. Chapter 13. Principal Components and Factor Analysis.
42. Traub RE (1997). Classical test theory in historical perspective. Educational Measurement: Issues and Practice, 8-14.
43. Wilson, M. (2004). Constructing measures: An item response modeling approach. Routledge.
44. Wright, B., Stone, M. (1999). Measurement Essentials. WIDE RANGE, INC. Wilmington, Delaware.
45. Wright, B.D., Linacre, J.M. (2001) Glossary of Rasch Measurement Terminology. Rasch Measurement Transactions, 15:2 p. 824-5.
46. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов. Москва: Издательская группа "Логос", 2010.

## Дополнительная литература

1. Abdi, H., & Williams, L. J. (2010). Principal component analysis. WIREs Computational Statistics, 2, 433-459.
2. Allalouf, A. (2007). Quality Control Procedures in the Scoring, Equating, and Reporting of Test Scores. Educational Measurement: Issues and Practice, 26: 36-43.
3. Andersen, E. B. 1973. A goodness of fit test for the Rasch model. Psychometrika, 38, 123–140.
4. Baker, F.B. (2001), The Basics of Item Response Theory. – ERIC Clearinghouse on Assessment and Evaluation
5. Bazaldua, D. A. L., Lee, Y. S., Keller, B., & Fellers, L. (2017). Assessing the performance of classical test theory item discrimination estimators in Monte Carlo simulations. Asia Pacific Education Review, 18(4), 585-598.
6. Bond Tr.G., Fox C.M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers, 2008.
7. Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement, 38(4), 295-317.
8. Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. American Psychologist. Advance online publication. doi: 10.1037/a0022389.
9. Chou, Y. T., & Wang, W. C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. Educational and Psychological Measurement, 70(5), 717-731.
10. Cultures: A Review of the ITC Guidelines for Test Adaptations. European Journal of Psychological Assessment, 15 (3), 258 – 269.
11. de Leeuw, J. (1997). Review of Rasch models—Foundations, recent developments, and applications. Statistics in Medicine, 16, 1431–1434.
12. DeVellis RF (2012) (3rd Ed.). Scale development: Theory and applications. Ch 7: An overview of item response theory. Sage.
13. Drasgow F & Lissak RI (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68(3), 363-373.
14. Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399-412.
15. Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research, 16(1), 5.
16. Eggen, T. J. H. M. 2000. On the loss of information in conditional maximum likelihood estimation of item parameters. Psychometrika, 65(3), 337–362.
17. Etz, A. (2018). Introduction to the Concept of Likelihood and Its Applications. Advances in Methods and Practices in Psychological Science, 1(1), 60–69.
18. Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. Review of Educational Research, 87(6), 1082-1116.
19. Gorin, J. S. (2006). Test design with cognition in mind. Educational Measurement: Issues and Practice, 25(4), 21-35.
20. Graham, J.M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them? Educational and Psychological Measurement, 66(6), 930-944.
21. Green, SB, Redell N, Thompson MS & Levy R. (2016). Accuracy of revised and traditional parallel Analyses for assessing dimensionality with binary data. Educational and Psychological Measurement, 76(1), 5-21.
22. Guilford, J. P. (1946). New standards for test evaluation. Educational and psychological measurement, 6(4), 427-438.
23. Guilford, J.P. (1985). A Sixty-Year Perspective on Psychological Measurement. Applied Psychological Measurement, Vol. 9 (4), pp.341-349.
24. Gulliksen, H. (1950). Intrinsic validity. American Psychologist, 5(10), 511
25. Guttman, 1947 The Cornell technique for scale and intensity analysis. Educational and psychological measurement, 7, 247-280
26. Guttman, L. (1944) A Basis for Scaling Qualitative Data. American Sociological Review, 9, 139-150.
27. Haladyna, T. M., & Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15, 309-334.
28. Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground. International Journal of Testing, 16(2), 142–163. https://doi.org/10.1080/15305058.2015.1108322.
29. Levy, R. (2013). Psychometric and Evidentiary Advances, Opportunities, and Challenges for Simulation-Based Assessment. Educational Assessment, 18(3), 182–207. https://doi.org/10.1080/10627197.2013.814517
30. Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? Journal of outcome measurement, 2, 266-283.
31. Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14(2), 117-138.
32. Ludlow LH, Matz-Costa C, Johnson C, Brown M, Besen E, & James JB (2014). Measuring engagement in later life activities: Rasch-based scenario scales for work, caregiving, informal helping, and volunteering. Measurement and Evaluation in Counseling and Development, 47(2), 127-149.
33. Ludlow, L. H. (2002). Residuals: Trash or treasures. Popular Measurement, 4, 1-7.
34. Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E. J. (2017). A tutorial on Fisher information. Journal of Mathematical Psychology, 80, 40-55.
35. Massof, R.W. (2001). Understanding Rasch and Item Response Theory Models: Applications to the Estimation and Validation of Interval Latent Trait Measures from Responses to Rating Scale Questionnaires. Ophthalmic Epidemiology, 18(1): 1-19.
36. Maydeu-Olivares, A. (2013). Why Should We Assess the Goodness-of-Fit of IRT Models?. Measurement: Interdisciplinary Research and Perspectives, 11(3), 127-137.
37. McGrew, K.S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. Intelligence, 37, 1-10.
38. McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9(1), 49-57.
39. McNeish, D. (2018). Thanks coefficient alpha, we’ll take it from here. Psychological Methods, 23(3), 412-433.
40. Michell, J. (2004) Measurement in Psychology: A Critical History of a Methodological Concept. Cambridge University Press.
41. Millett, C., Payne, D., Dwyer, C., Stickler, L., Alexiou, J. (2008). A Culture of Evidence: An evidence-centered approach to accountability for student learning outcomes. Educational Testing Service, Princeton, NJ.
42. Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and abilities. In R. Linn (Ed.), Educational measurement.
43. Myung, J. (2003). Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology, 47, 90–100.
44. Partchev I. A visual guide to item response theory. (2004). Friedrich-Schiller-Universit¨at Jena. https://www.metheval.uni-jena.de/irt/VisualIRT.pdf
45. Reese, T.W. (1943). The application of the Theory of Physical Measurement to the Measurement of Psychological Magnitudes, with Three Experimental Examples. Psychological Monographs. Vol. 55(3), 251, pp. 6-20.
46. Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement, 14(2), 127-137.
47. Reise, S.P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47(5), 667-696.
48. Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice, 24(2), 3-13.
49. Schultz, D.P., Schultz, S.E. (2011). A History of Modern Psychology. 10th edition. Wadsworth, Cengage Learning.
50. Smith, Jr. E. V. (2002). Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals. Journal of Applied Measurement, 3 (2), 205-231.
51. Spearman, C. (1904). General Intelligence Objectively Determined and Measured. The American Journal of Psychology, Vol. 15(2), pp. 201–292. www.jstor.org/stable/1412107.
52. Steyer, R., Smelser, N. J., & Jena, D. (2001). Classical (psychometric) test theory. International encyclopedia of the social and behavioral sciences. Logic of inquiry and research design, 1955-1962.
53. Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied psychological measurement, 7(2), 201-210.
54. Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. Journal of Educational Measurement, 40(4), 331-352.
55. Tanzer, N.K., & Sim, C.Q.E. (1999). Adapting Instruments for Use in Multiple Languages and
56. Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2015). PerFit: An R package for person-fit analysis in IRT. Journal of Statistical Software, 74(5), 1-27.
57. Thomson, B., Daniel, L.G. (1996). Factor Analytic Evidence for the Construct Validity of Scores: A Historical Overview and Some Guidelines. Educational and Psychological Measurement, Vol 56 (2), pp. 197 – 208. https://doi.org/10.1177/0013164496056002001
58. Thurstone, L.L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16(7), 433-451.
59. Turilova-Miščenko, T., & Raščevska, M. (2011). Scientific Concepts Test in Latvian and Russian Language: Evidence for reliability and validity. Baltic Journal of Psychology, 12 (1, 2), 73- 82.
60. Von Mayrhauser, R.T. (1989) Making intelligence functional: Walter Dill Scott and applied psychological testing in Word War I. Journal of the History of the Behavioral Science, Vol.25, pp.60-72.
61. Wang, W.-C., Chen, C.-T. (2005), Item Parameter Recovery, Standard Error Estimates, and Fit Statistics of the Winsteps Program for the Family of Rasch Models, Educational and Psychological Measurement, Vol.65, No.3, p.376-404
62. Wells, C. S., & Hambleton, R. K. (2016). Model fit with residual analyses. Handbook of item response theory. Volume two, Statistical tools, 395-413.
63. Wilson, M., Zheng, X., McGuire, L. (2012). Formulating Latent Growth Using an Explanatory Item Response Model Approach. Journal of Applied Measurement, 13(1).
64. Zegers, F. E., & ten Berge, J. M. (1985). A family of association coefficients for metric scales. Psychometrika, 50(1), 17-24.
65. Zwick, W. R., &Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological bulletin, 99(3), 432-442.
66. Карданова Е.Ю., Карданов Р.С. О некоторых свойствах характеристической и информационной функций политомического тестового задания // Вестник НовГУ. – 2010, №55, с.19-24
67. Нейман Ю.М., Хлебников В.А. (2000). Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей. Гл. 2-3.
68. Челышкова М.Б. Теория и практика конструирования педагогических тестов. М.: Логос, 2002.

## Справочники, словари, энциклопедии

Нет

## Программные средства

Аспиранты должны заниматься в компьютерном классе Института образования, в котором на компьютерах установлена программа Winsteps и пакет R.

# Материально-техническое обеспечение дисциплины

Проектор (для лекций или семинаров) и компьютеры (ноутбуки).