**Программа дисциплины**

**«Углубленная психометрика»**

для образовательной программы «Образование»

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

Разработчик программы:

Карданова Елена Юрьевна,

канд.физ.-мат. наук

[ekardanova@hse.ru](mailto:ekardanova@hse.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 «Углубленная психометрика».

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 «Углубленная психометрика» является изучение продвинутых психометрических методов современной теории тестирования.

В частности, аспиранты познакомятся с моделями современной теории тестирования, особенностями и проблемами широкомасштабных программ оценивания, технологиями компьютерного тестирования, современными технологическими возможностями и сложностями процедур сбора данных.

Также аспиранты познакомятся с современной критикой теории валидности и подробно изучат различные модели, используемые для разработки измерительных шкал.

* 1. **Обзор содержания курса**

Данный курс начинается с изучения искажений в ответах испытуемых и дифференцированного функционирования тестовых заданий. При рассмотрении потенциальных причин искажений в тестовых заданиях разбираются вопросы, связанные со справедливостью оценивания и универсальным дизайном оценивания.

Затем изучаются модели IRT, используемые для изучения психометрических свойств шкал на политомических данных. Рассматриваются допущения каждой модели, математическое описание и процедуры оценивания параметров.

Далее аспиранты переходят к изучению тестовых заданий открытого типа, их особенностей и разновидностей. Изучаются правила разработки заданий с регламентированным ответом и со свободно конструируемым ответом, особенности подсчета баллов, правила разработки критериев и рубрик оценивания таких заданий. Особое внимание уделяется исследованию деятельности экспертов, оценивающих заданий, как в рамках традиционного подхода, так и с использованием многофасетных моделей IRT.

На следующем этапе аспиранты переходят к изучению специфических проблем и возможностей измерений в цифровом формате. Во-первых, изучаются техники и процедуры компьютерного адаптивного тестирования. Во-вторых, изучаются инновационные, технологически улучшенные формы тестовых заданий и цифровое моделирование. Аспиранты знакомятся с проблемами моделирования ответных данных, полученных с помощью технологически улучшенных тестовых заданий и цифровых симуляций, автоматического подсчета баллов за эссе. Рассматриваются тестовые задания сценарного типа, особенности их разработки и подходы к моделированию ответных данных, полученных с помощью этих заданий.

Далее рассматриваются вопросы, касающиеся крупномасштабных и лонгитюдных программ оценивания. Темы данного раздела включают сопоставимость баллов, эффекты режима администрирования, различные тестовые дизайны, процедуры взвешивания, создание банка заданий и безопасность тестов, а также связывание и выравнивание результатов тестирования. Кроме того, рассматриваются процедуры шкалирования и установления пороговых баллов. Наконец, исследуется валидизация программ оцениания, составление отчета о результатах и технического отчета.

В конце курса аспиранты пересматривают современную концепцию валидности и анализируют ее критику. Всестороннее переосмысливая модели каузального, рефлективного и формативного измерений, аспиранты углубляют свое понимание данных подходов к разработке измерительных шкал и доказательству валидности при каждом подходе.

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

После успешного завершения данного этапа обучения аспиранты будут способны:

* выполнять анализ дифференцированного функционирования тестовых заданий и корректировать инструменты для минимизации искажений;
* описывать математические модели и уметь применять политомические модели IRT;
* создавать задания открытого типа, в частности задания со свободно конструируемым ответом, разрабатывать критерии/рубрики и руководство по подсчету баллов;
* анализировать деятельность экспертов, оценивающих задания;
* разрабатывать дизайны адаптивного теста и процедуры администрирования;
* разрабатывать технические требования к системам оценивания в цифровом формате, включая соответствующие требования к функциональной совместимости и элементам дизайна интерфейса;
* конструировать технологически усовершенствованные тестовые задания и соответствующие правила подсчета баллов;
* создавать задания для диагностических тестов и использовать диагностические модели IRT для получения отчетов о баллах;
* исследовать сопоставимость результатов тестирования, проводить процедуры связывания и выравнивания;
* проектировать спиральное администрирование тестирования и рассчитывать выборку на основе спирального дизайна;
* разрабатывать план действия для крупномасштабной программы оценивания и составлять технический отчет, содержащий доказательства валидности;
* сравнивать и сопоставлять каузальные, рефлективные и формативные модели измерений и определять ключевые аспекты валидности, связанные с каждым классом модели.

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 описанные исследования, а также собственные планируемые и выполненные исследования.  Может идентифицировать проблемные места и предлагать способы их решения. | Анализ и обсуждения планируемых и выполненных исследований, своих и чужих.Написание эссе. |
| Способность генерировать оригинальные теоретические конструкции, гипотезы и исследовательские вопросы | УК-2 | Может генерировать собственные исследовательские вопросы и гипотезы,  описывать условия принятия/отвержения гипотез. | Рефлексия относительно происхождения гипотез и исследовательских вопросов в опубликованных работах. Практика постановки собственных вопросов. Групповое обсуждение сгенерированных гипотез и вопросов. |
| Способность выбирать и применять методы исследования, адекватные предмету и задачам исследования | УК-3 | Может разработать план собственного эмпирического исследования, подобрать адекватные методы исследования, обосновать их применение и применить. | Анализ и обсуждения различных методов исследований в психометрике. |
| 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 с описанием эмпирических исследований. Практика планирования, анализа и интерпретации результатов. |
| 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. | Выполнение индивидуальных проектных домашних заданий, связанных с проведением анализа данных и требующих подбора и обоснования метода исследования, выбора и применения соответствующего ПО. |
| 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. |
| Способность критически оценивать собственные результаты в контексте результатов современных педагогических, социально-психологических, социокультурных, социально-экономических исследований | ПК-6 | Интерпретирует собственные результаты исследования в контексте уже существующих результатов современных педагогических, социально-психологических, социокультурных, социально-экономических исследований, и связанных теорий. | Анализ части обсуждения в опубликованных работах. Практика построения собственных интерпретаций. Обсуждение в группе. |

# Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу дисциплин по выбору и изучается на 2-м году обучения.

Обучение в данном блоке требует от аспирантов глубокого понимания модели Раша, одно-, двух- и трехпараметрических моделей IRT и процедур оценки, используемых этими моделями. Аспиранты также должны хорошо знать принципы и процедуры конструирования инструментов измерения и создания тестовых заданий. Наконец, аспиранты должны быть хорошо знакомы с современной концепцией валидности теста.

Данный блок опирается на темы курса «Введение в разработку и валидизацию инструментов измерения», углубляя понимание более сложных моделей IRT, используемых для политомически оцениваемых данных и инструментов оценивания. Кроме того, при рассмотрении сложностей и возможностей, которые возникают в процессе разработки инструментов, презентуемых на цифровых платформах, данный блок исходит из того, что аспиранты уже знакомы с процедурами разработки инструментов измерения в традиционном виде. Наконец, исследуя критику современной концепции валидности теста и альтернативные теории, данный блок основывается на уже имеющемся у аспирантов понимании валидности теста.

# Формы контроля знаний аспирантов:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля | 2 год / модули | | | | Параметры |
| 1 | 2 | 3 | 4 |
| Текущий  (модуль) | Контрольная работа |  |  | \* | \* | письменная работа на занятии |
| Домашнее задание |  |  | \* | \* | письменная работа 3-4 стр. (эссе по статье, отчет о проведенном анализе данных) |
| Итоговый | Экзамен |  |  |  | \* | тест, презентация и защита результатов анализа |

## Критерии оценки знаний, навыков

Для текущего контроля в течение курса используются опрос, групповая дискуссия, индивидуальные презентации, письменные домашние и контрольные 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 аспиранты сдают тест. Вторая часть экзамена представляет собой защиту индивидуального проектного задания. Примеры заданий представлены в п.9.3.

# Содержание дисциплины

**Раздел 1. Справедливость оценивания. DIF анализ.**

**Тема 1.1.** Справедливость оценивания. Универсальный дизайн оценивания.

**Тема 1.2.** Анализ различного функционирования заданий по отношению к различным группам испытуемых (DIF анализ). Инструкции по удалению или переработке заданий на основе DIF анализа.

**Раздел 2. Политомические модели IRT**

**Тема 2.1.** Политомические модели Раша:Partial Credit Model, Rating Scale Model. Описание моделей, параметры моделей и их оценивание и интерпретация. Области применения различных моделей.

**Тема 2.2.** Политомические модели IRT: Graded Response Model, Modified Graded Response Model, Generalized Partial Credit Model. Описание моделей, параметры моделей и их оценивание и интерпретация. Области применения различных моделей. Сравнение моделей и их свойств. Выбор модели измерения.

**Тема 2.3.** Особенности анализа теста при наличии комбинации дихотомических и политомических заданий. Анализ размерности теста. Анализ функционирования политомических заданий. Анализ совместного функционирования дихотомических и политомических заданий различных форм.

**Раздел 3. Задания открытого типа. Оценка деятельности экспертов.**

**Тема 3.1.** Различные формы заданий открытого типа: задания с коротким ответом (на дополнение), задания с регламентированным ответом, задания со свободно конструируемым ответом, их особенности, плюсы и минусы. Правила разработки заданий открытого типа.

**Тема 3.2.** Разработка рубрик и критериев оценивания заданий открытого типа. Различные виды рубрик и их особенности. Способы оценивания (скоринга) заданий открытого типа.

**Тема 3.3.** Надежность экспертного оценивания. Обучение экспертов. Контроль качества оценок экспертов и различные эффекты экспертного оценивания.

**Тема 3.4.** Многофасетные модели IRT и их применение для оценки деятельности экспертов.

**Раздел 4. Использование технологий в образовательной оценке**

**Тема 4.1.** Введение в модели компьютерного тестирования: Компьютерное Адаптивное Тестирование и Многоступенчатое Тестирование: различия, процедуры разработки, ограничения, достоинства и возможности.

**Тема 4.2.** Компьютерное адаптивное тестирование (КАТ) – различные виды, процедуры, области применения. Различные алгоритмы КАТ: правила выбора первого задания (блока заданий), правила выбора следующего задания, правила завершения тестирования, процедуры оценивания и начисления баллов.

**Тема 4.3.** Технологически-усовершенствованные задания. Способы использования технологий в модификации заданий: классификации технологически-усовершенствованных заданий. Технологически-усовершенствованные задания и ECD. Сопоставимость результатов между классическими и новыми форматами заданий. Калибровка технологически-усовершенствованных заданий.

**Тема 4.4.** Справедливость измерений: доступность тестов для лиц с ОВЗ. Стандарт доступности заданий для лиц с ОВЗ. Универсальный дизайн тестов и технологически-усовершенствованные задания. Сопоставимость результатов измерений между группами лиц с ОВЗ и без ОВЗ. Аккомодации. Стандарт доступности заданий для лиц с ОВЗ APIP.

Платформы доставки тестов до респондентов на примере TAO. Стандарты хранения данных: QTI-формат заданий, форматы сбора и хранения данных о процессе ответа на задания. Возможности и ограничения TAO. Особенности хранения и сбора данных в QTI-формате.

**Тема 4.5.** Методы автоматической проверки ответов на открытые вопросы. Тренировка и валидация сети.

**Тема 4.6.** Сценарные типы заданий и симуляции. Разработка сценариев в многофасетных теориях Г.Раша и Л.Гуттмана: принципы, ограничения, анализ и примеры.

**Раздел 5. Оценивание многомерных конструктов**

**Тема 5.1.** Одномерный, последовательный и многомерный подход к моделированию многомерных конструктов. Многомерные IRT модели и их характеристики.

**Тема 5.2.** Не-компенсаторные (Between-Item Multidimensionality) и компенсаторные (Within-Item Multidimensionality) модели современной теории тестирования. Моделирование локальной зависимости заданий.

**Раздел 6. Психометрические аспекты широкомасштабных программ тестирования.**

**Тема 6.1.** Исследование сопоставимости результатов тестирования различных тестовых форм.

Проблема сопоставимости результатов тестирования различных тестовых форм. Проблема сопоставимости результатов тестирования в бланковой и компьютерной формах. Возможность построения единой шкалы для двух тестовых форм. Факторы, влияющие на сравнимость результатов тестирования. Дизайны исследования эквивалентности тестовых форм.

**Тема 6.2.** Связывание и выравнивание: принципы, подходы, процедуры, применение. Практические аспекты выравнивания. Горизонтальное и вертикальное выравнивание.

**Тема 6.3.** Международные бенчмарки. Дизайн широкомасштабных опросов и тестирований. Создание единой шкалы при спиралевидном дизайне.

**Тема 6.4.** Проблема миссингов (пропущенных данных) и методы интерполяции.

**Тема 6.5.** Оценивание испытуемых - plausible values, взвешивание

**Раздел 7. Проблемы лонгитюдных исследований**

**Тема 7.1.** Оценивание роста (прогресса). Вертикальное выравнивание. Различные методы оценивания роста (изменения тестового балла на eдиной шкале, регрессионные методы, процентилльные методы), их достоинства и недостатки. Банк заданий для лонгитюдного исследования. Поддержка банка.

**Тема 7.2.** Вопросы информационной безопасности тестирования. Разработка и использование мер по обеспечению информационной безопасности. Проблема нечестной сдачи теста (списывания) и методы его обнаружения.

**Раздел 8. Уровни достижений: названия и дескрипторы. Методы установления пороговых баллов. Работа с экспертами.**

**Тема 8.1.** Нормативно-ориентированное и критериально-ориентированное тестирование. Описание уровней достижений. Классификация методов установления пороговых баллов. Общий фреймворк процедуры установления пороговых баллов: разработка описания уровней достижений, выбор метода установления пороговых баллов, отбор экспертов, обучение экспертов, модерация сессий установления пороговых баллов, документация, оценка валидности полученных результатов, принятие итогового варианта пороговых баллов. Методы установления пороговых баллов. Особенности работы с экспертами. Раунды установления пороговых баллов. Виды обратной связи для экспертов (нормативная, про характеристики заданий, о последствиях).

**Тема 8.2.** Корректировка и утверждение итоговых пороговых баллов, вертикальное выравнивание пороговых баллов, общественные и политические вопросы. Анализ последствий принятия тех или иных пороговых баллов, цена ложноположительных и ложноотрицательных решений, ошибки измерения (тестовых заданий, согласованность экспертов), сопоставление результатов различных методов установления пороговых баллов.

**Раздел 9. Валидность программы оценивания. Сообщение результатов тестирования.**

**Тема 9.1.** Исследование валидности программы оценивания. Правильное использование инструмента оценивания, обоснованность принятых решений. Рекомендации по анализу программы оценивания. Технический отчет, обоснование качества инструментария и валидности измерений.

**Тема 9.2.** Дизайн отчета с результатами измерений для разных групп пользователей. Руководство к тесту. Преобразование тестовых баллов для сообщения пользователям, различные шкалы.

**Раздел 10. Диагностические оценивание**

**Тема 10.1.** Диагностические оценивание – определение, различные формы, цели. Дизайны диагностирующего оценивания. Обзор различных аспектов формирующего оценивания. Итоговое оценивание.

**Тема 10.2.** Диагностические IRT модели.

**Раздел 11. Критика современной теории тестирования**

**Тема 11.1.** Позиция Мичелла относительно психологического тестирования. Дефицит правдоподобных самостоятельных теорий, способных управлять измерениями ментальных атрибутов. Три структуры атрибутов, измеряющихся в науке: классификация, гетерогенный порядок и количественная структура. Измеримость атрибутов и неколичественно-упорядоченные атрибуты. Психометрические заблуждения.

**Тема 11.2.** Критика Мичелом современной теории тестирования и теория соединенного измерения(Conjoint measurement).Операциональные и репрезентационные концепции измерения. Теория соединенного измерения (conjoint measurement): необходимые и достаточные условия для квантификации атрибута. Допущения модели Раша. Роль ошибки в количественном атрибуте. Природа парадокса Раша. Релевантность теории для психометрики. Отношения между заданиями теста.

**Тема 11. 3.** Критика современной концепции валидности: Позиция Борбума и Лизитца. Возвращение к теориям валидности: тесты как измерение конструктов vs. тесты как выборка поведения. Номологический подход Кронбаха и Мила. Валидность как интерпретация (Мессик). Введение причинности как главной причины дисперсии в результатах измерения. Статус психологического конструкта: онтология vs. эпистемология, референция vs. значение, причинность vs. корреляция.

**Раздел 12. Возвращение к теориям валидности**

**Тема 12.1.** Возвращение к теориям валидности: каузальные и рефлективные модели. Рефлективные модели измерения: эксплицитное использование идеи причинности между конструктами и тестовыми показателями. Путевые диаграммы. Теории регулярности и рефлективные модели. Теории от противного как особый тип вероятностной принности. Предсказание и объяснения. Процессуальные теории. Модели когнитивной переработки информации. Выводы, основанные на теории.

Возвращение к теориям валидности: оценка формативных измерительных моделей. Валидность как оценивание. Наблюдаемые и латентные переменные. Каузальные индикаторы и понимание, основанное на фактах. Формативные модели vs. модели регулярности vs. доказательства от противного. Процессуальная интерпретация наблюдаемых индикаторов и латентные переменные.

**Тема 12.2.** Возвращение к теориям валидности: открытые вопросы о значении тестовых показателей. Интегративный взгляд на валидность (позиция Борсбума и Маркуса). Доказательства валидности, связанные с перспективами рефлективных, формативных и каузальных моделей. Сетевые модели измерения.

# Образовательные технологии

Проблемные лекции, семинары, самостоятельная работа, письменные самостоятельные работы, практические занятия, работа по группам, общегрупповая дискуссия, презентации и их защита, мастер-классы.

Проблемная лекция предполагает проблемное изложение учебного материала на лекции: новый теоретический материал представляется в форме проблемной задачи, которую студентам предстоит решить совместно с преподавателем. Предполагается, что проблемная лекция реализуется в следующем виде: преподаватель обозначает проблемную ситуацию, побуждает студентов к поискам решения проблемы, шаг за шагом подводя их к искомой цели. При подготовке к проблемной лекции преподаватель должен специальным образом обработать лекционный материал, представив его в виде одной или нескольких проблем, продумать принцип развертывания его на лекции в форме диалогического общения со студентами.

# Оценочные средства для текущего контроля и аттестации аспиранта

Для текущего контроля используются опрос, групповая дискуссия, индивидуальные презентации, письменные контрольные работы.

## Тематика заданий текущего контроля

***Примеры вопросов и заданий контрольных работ***

1. Для темы «Критика современной теории тестирования. Возвращение к теориям валидности»:

а) Предложите пример применения Раш моделирования к измерению достижений и прокомментируйте результаты в терминах Мичела, например, с точки зрения доказательства непрерывности и одномерности изучаемого конструкта.

6) Предложите пример валидазационных исследований, вовлекающих близкие конструкты, но выполненных в рамках рефлективной и формативной модели. Аргументируйте свою интерпретацию. Предложите план валидизации этих инструментов в сетевой модели.

1. Для темы «Задания открытого типа» - пример задания контрольной работы:

Перед вами задание открытого типа (эссе) и дополнительная информация к нему из спецификации к тесту. Это задание составлено корректно. Вам нужно составить аналитическую рубрику для оценки этого задания. Свой ответ представьте в виде таблицы.

*Задание*:

"На основании результатов проведенного анализа Петя заключил, что дополнительные занятия математикой снижают балл в PISA. Корректен ли вывод Пети и почему? Приведите два аргумента в защиту позиции.

*Сведения из спецификации:*

Цель задания: Формулирует ограничения интерпретации результатов применения методов

Тема: Дисперсионный анализ

Уровень по таксономии Блума: оценка

Максимальный балл: 4

Время выполнения задания: 8 минут

***Письменные домашние задания***

1. Эссе по статье

Студенту необходимо написать эссе по статье – либо предложенной преподавателем, либо выбранной самостоятельно по указанной преподавателем теме. Домашнее задание по написанию эссе состоит из двух частей:

а) Перевод аннотации к статье и составление плана эссе.

б) Написание эссе.

Эссе представляет собой собственное рассуждение автора по проблеме, которая поднимается в статье.

Эссе содержит следующие элементы:

* Вступление. Например, какой основной вопрос статьи?
* Позиция автора. Например, главные идеи и результаты исследования, область их применения. Достоинства и недостатки используемых методов.
* Заключение. Например, вопросы, которое ставит это исследование.

При написании эссе важно также учитывать следующие моменты:

1. Вступление и заключение должны фокусировать внимание на проблеме (во вступлении она ставится, в заключении - резюмируется мнение автора).
2. Необходимо выделение абзацев, красных строк, установление логической связи абзацев: так достигается целостность работы.
3. Стиль изложения: четкий, ясный, лаконичный, но отражает особенности личности.

Объем эссе - 3 страницы.

Критерии оценивания эссе:

* Наличие всех элементов
* Полнота ответов - отмечены ли все ключевые идеи, правильно ли они изложены
* Ясность стиля изложения
* Наличие собственной позиции автора

Эссе оценивается в 10 баллов.

Критерии оценивания эссе:

10 баллов – очень хороший текст. Все основные аспекты учтены. Все выводы обоснованы. Текст хорошо организован, мысли следуют четкой логике. Текст структурно оформлен как академический. Приведены обоснования для каждого утверждения. Используются профессиональные термины. Грамматически и стилистически без ошибок.

9 баллов – то же самое, что и 10, но страдает оформление текста, допущены незначительные грамматические или стилистические ошибки.

8 баллов – текст включает все основные аспекты. Удерживается логика изложения. Основные мысли аргументируются.

7 баллов – то же самое, что и 8, но страдает логика и язык изложения.

6 баллов – учтены лишь некоторые из основных аспектов. Аргументация иногда недостаточная или нерелевантная.

5 баллов – то же что и 6, но язык изложения очень скуден, страдает оформление. Много грамматических ошибок.

4 балла – очень поверхностное изложение некоторых из нужных аспектов. Небрежный текст. Не развернута логика, отсутствует основная аргументация.

1-3 балла – не отражен ни один из необходимых аспектов, остальное как в 4.

0 – работа не сдана

1. Письменное домашнее задание по анализу данных тестирования

Домашнее задание по заданию данных тестирования представляет собой полный анализ данных тестирования с выбором метода исследования, программы обработки данных, написанием отчета с результатами и его презентацией. Работа выполняется в группах по 2 человека.

Объем отчета не лимитируется.

Каждое домашнее задание оценивается в 10 баллов.

Критерии оценивания домашнего задания:

* Полнота и правильность проведенного анализа (максимум 3 балла)
* Качество написания отчета (максимум 3 балла)
* Представление результатов - качество презентации (максимум 2 балла)
* Умение вести дискуссию, отстаивать свою точку зрения (максимум 1 балл)
* Умение работать в группе (максимум 1 балл)

## Проведение итогового контроля

Итоговый контроль проводится в форме экзамена.Экзамен состоит из двух частей. В первой части аспиранты сдают тест. Вторая часть экзамена представляет собой защиту индивидуального проектного задания. Проектное задание предполагает решение поставленной проблемы: выбор и обоснование метода решения, модели измерения и компьютерной программы, а также выполнение всех необходимых действий, получение результатов, их интерпретацию и представление.

Примеры заданий представлены в п.9.3.

## Примеры заданий итогового контроля

1. Пример задания теста

В таблице представлены профили ответов нескольких испытуемых с указанием неожиданных ответов.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Балл | Ряд ответов  с пометками существенных отклонений  (\* — существенное отрицательное,  + — существенное положительное)  Часть А | Часть В | Часть С | INFIT/ OUTFIT  mnsq | Возможная причина |
| 1 | 68 | 11100 11010 11111 11011 11111 11110 10010 10001 10001 22213  \*\* \* \* \* ++ + |  |  |
| 2 | 82 | 11110 11011 11110 11011 11101 11111 01020 11111 11111 22213  \* \* \* \* \* \* \* \* + |  |  |
| 3 | 73 | 10001 10011 11111 11111 11101 11111 11020 11121 00111 12112  \*\*\* \*\* |  |  |
| 4 | 24 | 01000 00000 10000 00001 00010 00000 00001 00120 00012 21100  + + + + ++ + + |  |  |
| 5 | 56 | 11111 11111 11011 11101 11111 01111 10010 00000 01001 01000 |  |  |

a)Укажите в таблице подходящие значения статистик согласия INFIT и OUTFIT для каждого испытуемого.

б) Укажите в таблице наиболее возможную причину искажений для каждого профиля.

1. Пример проектного задания

**Тест по математике, файл math-1 (предоставляется файл с данными)**

**Описание ситуации**

К вам обратился руководитель команды разработчиков теста по биологии для школьников выпускного класса с просьбой помочь провести анализ. Тест состоит из трех частей – А, В и С, различающихся типом заданий и, предположительно, их трудностью. До вас в группе работал другой психометрик, который сейчас уволился. Работал он на своем компьютере и использовал какую-то свою программу. Он прислал по почте файл, который вам передали. Также вам передали спецификацию теста и сам тест. Руководитель говорит, что у них нет лицензии на какое-либо специализированное программное обеспечение. Руководитель что-то слышал о каких-то моделях IRT и просит вас ответить на следующие вопросы:

1. Стоит ли им использовать IRT? Какие преимущества для их анализа даст эта теория?
2. Вроде как есть много моделей IRT. А чем они отличаются? Зачем их так много? Как выбрать подходящую модель?
3. Задания части А – это задания с выбором одного правильного ответа из пяти предложенных. Есть предположение, что могло иметь место гадание. Можно ли это проверить?
4. В тесте есть задания части С, которые намеренно создавались очень трудными. Что вы можете сказать об этих заданиях? Полезны ли они в тесте? Какую дополнительную информацию они дают?
5. Ну и наконец, от вас ждут заключение – хороший ли они разработали тест.

# Порядок формирования оценок по дисциплине

Оценивается работа на занятиях(*Оауд.*), домашние задания (*Осам. работа*) и контрольные работы (*Ок/р*). На оценку влияет своевременная сдача и правильность выполнения заданий контрольных и домашних работ, демонстрация знания заданного материала, активность на занятиях.

На семинарских занятиях оценивается знание заданного к изучению материала, способность видеть связи между основными понятиями, участие в дискуссии и в групповой работе, презентация результатов проведенного анализа. Накопленная оценка по 10-ти балльной шкале за работу на семинарских занятиях определяется перед промежуточным или итоговым контролем - *Оауд.*.

Преподаватель оценивает самостоятельную работу аспирантов: оценивается своевременная сдача и правильность выполнения заданий домашних работ, полнота выполнения, ответы на вопросы аудитории при докладах. 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* = *0,4·Ок/р + 0,3· Оауд. + 0,3· Осам. работа*.

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

Аспирант не имеет возможности пересдать низкие результаты за текущий контроль -работу на занятиях, самостоятельную работу, контрольную 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), допускается сдача форм контроля не позднее, чем через две недели с момента окончания действия справки, если другие сроки не установлены департаментом образования. В случае, если аспирант не сдает работу в указанный срок, му 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-балльной шкале).

Результирующая оценка за итоговый контроль выставляется в 10-балльной шкале. Форма итогового контроля: экзамен. Экзамен состоит из двух частей. В первой части аспиранты сдают тест (вес в оценке за экзамен 0,5). Вторая часть экзамена представляет собой защиту индивидуального проектного задания (вес в оценке за экзамен 0.5). Проектное задание предполагает решение поставленной проблемы: выбор и обоснование метода решения, модели измерения и компьютерной программы, а также выполнение всех необходимых действий, получение результатов, их интерпретацию и представление.

Итоговая оценка складывается из суммы текущей оценки и итоговой оценки и преводится в 10-балльную шкалу.

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 диплом выставляется 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. Accessibility and Accommodations: http://www.smarterbalanced.org/assessments/accessibility-and-accommodations/
2. Adams, R.J., Wilson, M. and Wang, W. (1997), The MultidimensionalRandom Coefficients Multinomial Logit Model, Applied Psychological Measurement,21(1), 1-23.
3. Allen, D.D., Wilson, M. (2006). Introducing multidimensional item response modelingin health behavior and health education research. Health education research:Theory & Practice, 21(1),73-84.
4. Almond, R. G., Steinberg, L., &Mislevy, R. J. (2001). A sample assessment using the four process framework. CSE Report 543. Center for study of evaluation. Los Angeles, University of California.
5. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). AERA, APA, & NCME. Standards for educational and psychological testing. Washington, DC: American Educational Research Association. Chapter 4, Test Design and Development, 75-95.
6. Blood, I. A. (2011). Automated Essay Scoring: A Literature Review. Working Papers in TESOL & Applied Linguistics, 11(2), 1-25.
7. Bond, Tr.G., Fox, C.M. (2008), Applying the Rasch Model: Fundamental Measurement in the Human Sciences. – New Jersey.: Lawrence Erlbaum Associates, Inc., Publishers.
8. Bouhlila, D. S., & Sellaouti, F. (2013). Multiple imputation using chained equations for missing data in TIMSS: a case study. Large-Scale Assessments in Education, 1(1), 4.
9. Briggs, D. C., Domingue B. (2013). The Gains From Vertical Scaling. Journal of Educational and Behavioral Statistics, 38 (6), pp. 551–576
10. Chang, W.-C.C. (2017). Measuring the complexity of teachers' enactment of practice for equity: A Rasch model and facet theory-based approach. Published PhD Dissertation, Boston College.
11. Che-Yu Kuo, Hsin-Kai Wu (2013) Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science, Computers & Education, Volume 68, Pages 388–403
12. Davier, M. von, Gonzales, E. J., &Mislevy, R. J. (2009). What are Plausible Values and Why are They Useful? ERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2, 9–36.
13. Dolan, B., Strain-Seymour, E., Deokar, A., &Ostler, W. (2010). Next-generation assessment interoperability standards: A white paper from Pearson.
14. Duval, E. (2004). Learning technology standardization: making sense of it all. Computer Science and Information Systems, 1(1), 33-43.
15. Embretson, S. E., Reise, S. P. (2000). Item Response Theory for Psychologists. New Jersey.: Lawrence Erlbaum Associates, Publishers. Ch.5.
16. ETS (2006) Test and Score Data Summary for TOEFL Computer-Based and Paper-Based Tests
17. Gonzalez, E., &Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. IEA-ETS Research Institute Monograph, 3, 125–156.
18. Haladyna, T. M., & Downing, S. M. (2011). Handbook of test development. Routledge. Taylor & Francis Group. Ch. 13.
19. Hambleton R.K., Swaminathan H., Rogers H.J. Fundamentals of Item Response Theory (Measurement Methods for the Social Science). Sage Publications, Inc, 1991.
20. IMS Global Learning Consortium. Better learning from better learning technology. (2014). https://www.imsglobal.org/apip/apipv1p0/APIP\_BPI\_v1p0.html
21. International Test Commission (2001). International Guidelines for Test Use, International Journal of Testing, 1(2), 93-114.
22. Ip, E.H. (2010). Empirically indistinguishable multidimensionalIRT and locally dependent unidimensional itemresponse models. British Journal of Mathematical and Statistical Psychology, 63, 395–416.
23. ITC Guidelines on the Security of Tests, Examinations, and Other Assessments. URL: https://www.intestcom.org/files/guideline\_test\_security.pdf
24. Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. Applied Measurement in Education, 16(4), 277-298.
25. Linacre, J.M., Wright, B.D. (2002), Construction of Measures from Many-facet Data, Journal of Applied Measurement, Vol. 3, No. 4, p. 486-512.
26. Linn, R. L., Gronlund, N. E., & Miller, M. D. (2013). Measurement and assessment in teaching. Boston: Pearson. Ch. 9, 10, 11
27. Lissitz, R. W. & Huynh H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. Practical Assessment, Research & Evaluation, 8(10). Retrieved February 16, 2006 from http://PAREonline.net/getvn.asp?v=8&n=10.
28. Ludlow, L. H., Matz-Costa, C., Johnson, C., Brown, M., Besen, E., & James, J. B. (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.
29. Magis, D., Yan, D., & Von Davier, A. A. (2017). Computerized Adaptive and Multistage Testing with R: Using Packages CatR and MstR. New York, NY: Springer.
30. Markus, K. A., & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. Routledge.
31. McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing, 23, 35-59.
32. Michael D. Kurtz (2018) Value-Added and Student Growth Percentile Models: What Drives Differences in Estimated Classroom Effects?, Statistics and Public Policy, 5:1, 1-8
33. Michell, J. (2000). Normal science, pathological science and psychometrics. Theory & Psychology, 10(5), 639-667.
34. Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. Theory & Psychology, 14(1), 121-129. 10.1177/0959354304040201
35. Michell, J. (2008). Conjoint measurement and the Rasch paradox: A response to Kyngdon. Theory & Psychology, 18(1), 119-124. 10.1177/0959354307086926
36. Michell, J. (2009). The psychometricians' fallacy: Too clever by half? British Journal of Mathematical and Statistical Psychology, 62(1), 41-55. 10.1348/000711007X243582
37. Michell, J. (2013). Constructs, inferences, and mental measurement. New Ideas in Psychology, 31(1), 13-21.
38. Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma. Theory & Psychology, 24(1), 111-123. 10.1177/0959354313517524
39. OECD. (2009). PISA Data Analysis Manual: SPSS. Paris: OECD Publishing.
40. Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. Review of Educational Research, 74(4), 525–556.
41. PISA 2015 Technical report www.oecd.org/pisa/data/2015-technical-report/
42. Question And Test Interoperability (QTI) Specification: https://www.imsglobal.org/question/index.html
43. Ramalingam, V. V., Pandian, A., Chetry, P., & Nigam, H. (2018, April). Automated Essay Grading using Machine Learning Algorithm. In Journal of Physics: Conference Series (Vol. 1000, No. 1, p. 012030). IOP Publishing.
44. Rose, D. H., Meyer, A., Strangman, N., & Rappolt, G.(2002). Teaching every student in the digital age: Universal design for learning..[Электронныйресурс]. URL: https://www.bcpss.org/bbcswebdav/institution/Resources/Summer%20TIMS/Section%205/A%20Teaching%20Every%20Student%20in%20the%20Digital%20Age.pdf
45. Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. The Journal of Technology, Learning and Assessment, 1(2).
46. Russell M., Hoffmann T., and Higgins J. Meeting the Needs of All Students: A Universal Design Approach to Computer-Based Testing: http://www.innovateonline.info/index.php?view=article&id=676
47. Russell, M., Hoffmann, T., & Higgins, J. (2009). Meeting the needs of all students: A universal design approach to computer-based testing. Innovate: Journal of online education, 5(4).
48. Schafer, W. D. (2006). Growth Scales as an Alternative to Vertical Scales. Practical Assessment, Research & Evaluation. 11 (4),
49. Smith, Ev.V., Smith, R.M. (2004). Introduction to Rasch Measurement. Maple Grove, Minnesota: JAM Press.
50. Smith, R.M. (2004). Detecting Item Bias with the Rasch Model. Journal of Applied Measurement, 5(4), 430-449.
51. TAO User guide 3.2: https://userguide.taotesting.com/3.2/introduction/what-is-tao.html
52. Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7(3), 215-231.
53. Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. Practical Assessment, Research & Evaluation, Ch. 16.
54. Van de Vijver, F. J., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 47, 263-279.
55. van der Linden, W.J., Glas, C.A.W. (2010). Elements of adaptive testing. New York, NY: Springer. pp. 3-30).
56. Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008) Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: a metaanalysis of testing mode effects. Educational and Psychological Measurement, 68(1), 5-24.
57. Wiberg, M. (2012). Can a multidimensional test be evaluated with unidimensional item response theory?. Educational Research and Evaluation, 18(4), 307-320.
58. Wright, B.D., Masters, G.N. (1982), Rating Scale Analysis. Rasch Measurement. – Chicago: Mesa Press.
59. Wu, M. (2005).The role of plausible values in large-scale surveys. Studies in Educational Evaluation, 31(2–3), 114–128.
60. Zumbo, B.D. 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.
61. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов. Москва: Издательская группа "Логос", 2010.

## Дополнительная литература

1. Albarak Sari, A., KELECİOĞLU, H. (2016). Assessment of Achievement and Growth by Vertical Scaling: Comparison of Vertical Scaling Methods. Journal of Educational Sciences Research, 6(2).
2. Andrich D., Hagquist C. (2011). Real and Artificial Differential Item Functioning. Journal of Educational and Behavioral Statistics, 1076998611411913.
3. Arne Weigold, Ingrid K. Weigold, Naomi M. Drakeford, Stephanie A. Dykema, Charity A. Smith (2016) Equivalence of paper-and-pencil and computerized self-report surveys in older adults, Computers in Human Behavior, Volume 54, Pages 407–413
4. Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. Applied Psychological Measurement, 34(6), 438-452.
5. Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. Science education, 85(5), 536-553.
6. Betebenner, D. W. (2011). A Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections/Trajectories. The National Center for the Improvement of Educational Assessment, Dover, New Hampshire, https://www.nj.gov/education/njsmart/performance/SGP\_Technical\_Overview.pdf
7. Beuk, C. H. (1984). A Method for Reaching a Compromise between Absolute and Relative Standards in Examinations. Journal of Educational Measurement, 21(2), 147–152.
8. Black, P., &Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), 7-74.
9. Black, P., Harrison, C., Hodgen, J., Marshall, B., &Serret, N. (2010). Validity in teachers’ summative assessments. Assessment in Education: Principles, Policy & Practice, 17(2), 215-232.
10. Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. Psychological Methods, 22(3), 581.
11. Bolt, S. E., &Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. Journal of Psychoeducational Assessment, 26(2), 121-138.
12. Boo J. & Vispoel W.. (2012). Computer versus Paper-and-Pencil Assessment of Educational Development: A Comparison of Psychometric Features and Examinee Preferences. Psychological Reports: Mental & Physical Health, 111 (2), 443-460
13. Borsboom, D. (2016). Zen and the art of validity theory. Assessment in Education: Principles, Policy & Practice, 23(3), 415-421.
14. Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. Journal of Educational Measurement, 50(1), 110-114.
15. Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. Intelligence, 30(6), 505-514.
16. Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological review, 111(4), 1061.
17. Brandt, S. (2017). Concurrent Unidimensional and Multidimensional Calibration within Item Response Theory. Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana, 54(2), 1-18.
18. Briggs, D. C., Weeks, J. P.,Wiley E. (2008). Vertical Scaling in Value-Added Models for Student Learning University of Colorado. Paper presented at the National Conference on Value-Added Modeling, April 22-24, 2008, Madison, WI.
19. Brown, M., Dehoney, J., & Millichap, N. (2015). The next generation digital learning environment. A Report on Research. ELI Paper. Louisville, CO: Educause April.
20. Bryant, W. (2017). Developing a Strategy for Using Technology-Enhanced Items in Large-Scale Standardized Tests. Practical Assessment, Research & Evaluation, 22(1).
21. Calhoun, M.B., Fuchs, L., & Hamlett, C. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. Learning Disability Quarterly, 23, 271-282.
22. Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children–Fifth Edition: Confirmatory factor analyses with the 16 primary and secondary subtests. Psychological Assessment, 29(4), 458-472.
23. Chang W-C, Ludlow LH, Grudnoff A, Ell F, Haigh M, Hill M & Cochran-Smith M (under review). Measuring the complexity of teaching practice for equity: Development of a scenario-format scale. Teaching and Teacher Education.
24. Chen, T.-A. (2010). Random or Fixed Testlet Effects: A ComparisonOf Two Multilevel Testlet Models. Published PhD Dissertation at The University of Texas at Austin.
25. Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. Psychometrika, 74(4), 619.
26. Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. Educational and Psychological Measurement, 71(1), 37-53.
27. Cizek, G. J. (2005). Standard-Setting Guidelines. Educational Measurement: Issues and Practice, 15(1), 13–21.
28. Crawford, L., & Tindal, G. (2004). Effects of a read-aloud modification on a standardized reading test.  Exceptionality, 12(2), 89-106.
29. Cuevas M., Cervantes V. H. (2012). Differential Item Functioning detection with Logistic Regression. Mathematics and Social Sciences, 3, 45–59.
30. DeMars C. E., Jurich D. P. (2015). The Interaction of Ability Differences and Guessing When Modeling Differential Item Functioning With the Rasch Model Conventional and Tailored Calibration. Educational and Psychological Measurement, 75(4), 610-633.
31. Geer, J. G. (1988). What do open-ended questions measure? Public Opinion Quarterly, 52(3), 365-367.
32. Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. Educational Measurement: Issues and Practice, 29(1), 38–44
33. Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. Computers & education, 57(4), 2333-2351.
34. Glas, C. A., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. Applied Psychological Measurement, 27(4), 247-261.
35. Green, D. R., Yen, W. M., Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2(4), p. 297-312.
36. Greene, P. (2018). Automated Essay Scoring Remains an Empty Dream. Forbes.
37. Harlen, W. (2007). Criteria for evaluating systems for student assessment. Studies in Educational Evaluation, 33(1),15-28
38. Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? Journal of Educational Measurement, 38(3), 249-266.
39. Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. Psychological methods, 12(2), 205.
40. Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425–461
41. Keller, L. (2005), Item response theory (IRT) models for polytomous response data, in B. S. Everitt & D. C. Howell. (Eds.). Encyclopedia of statistics in behavioral sciences. West Sussex, England: John Wiley & Sons, p. 990-1003.
42. Kristjansson, E., Aylesworth, R., McDowell, I., Zumbo, B.D. (2005). A comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items. Educational and Psychological Measurement, 65(6), 935-953
43. Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. Communications in Statistics - Theory and Methods, 46(22), 11341–11357.
44. Laukaityte, I., &Wiberg, M. (2017a). Importance of sampling weights in multilevel modeling of international large-scale assessment data.Communications in Statistics - Theory and Methods, 0(0), 1–22.
45. Leung, S. O. (2003). A practical use of vertical equating by combining IRT equating and linear equating. Practical Assessment, Research & Evaluation, 8(23). Retrieved February 16, 2006 from http://PAREonline.net/getvn.asp?v=8&n=23.
46. Linacre, JM. (2002). Facets, factors, elements and levels. Rasch Measurement Transactions, 16:2 p.880
47. Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational researcher, 36(8), 437-448.
48. Liu, O.L., Wilson, M., Paek, I. A Multidimensional Rasch Analysis of Gender Differences in PISA Mathematics, Journal of Applied Measurement, Vol. 9, No.1, p. 18-35.
49. Livingston, S. A., & Zieky, M. J. (1982). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton: Educational Testing Service.
50. Lord F.M. Application of Item Response Theory to practical testing problems. Hillsdale, NJ: Erlbaum.
51. Ludlow LH, Matz-Costa C & Klein K (In press). Enhancement and validation of the Productive Engagement Portfolio-Scenario (PEP-S8) Scales. Measurement and Evaluation in Counseling and Development.
52. Markus, K. A., & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. Routledge.
53. Meinck, S. (2015).Computing sampling weights in large-scale assessments in education.Survey Insights: Methods from the Field, Weighting: Practical Issues and ‘How to’ Approach.
54. Meinck, S., Cortes, D., & Tieck, S. (2017). Evaluating the risk of nonresponse bias in educational large-scale assessments with school nonresponse questionnaires: a theoretical study. Large-Scale Assessments in Education, 5(1), 3.
55. Meyer, R. H., Dokumaci, E. (2009). Value-Added Models and the Next Generation of Assessments. https://www.ets.org/Media/Research/pdf/MeyerDokumaciPresenterSession4.pdf
56. Michell, J. (2003). Epistemology of measurement: The relevance of its history for quantification in the social sciences. Social Science Information, 42(4), 515-534. 10.1177/0539018403424004
57. Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. Frontiers in Psychology, 3(AUG)10.3389/fpsyg.2012.00261
58. Michell, J. (2017). On substandard substantive theory and axing axioms of measurement: A response to Humphry. Theory & Psychology, 27(3), 419-425.
59. Mislevy R.J., Haertel G., Riconscente M., Rutstein D.W., Ziker C. (2017). Evidence-Centered Assessment Design. In: Assessing Model-Based Reasoning using Evidence- Centered Design. Springer Briefs in Statistics. Springer, Cham.
60. Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O’Sullivan, C. Y., &Preuschoff, C. (2009). TIMSS 2011 Assessment Frameworks. Boston: International Association for the Evaluation of Educational Achievement. Chapter 4.
61. Myers, N. D., Wolfe, E.W., Feltz, D.L., Penfield, R.D. (2006). Identifying Differential Item Functioning of Rating Scale Items with the Rasch Model: An Introduction and an Application. Measurement in Physical education and exercise science, 10(4), 215–240
62. Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3–19.
63. Perie, M. (2008). A guide to understanding and developing performance-level descriptors. Educational Measurement: Issues and Practice, 27(4), 15–29.
64. Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. Applied Psychological Measurement, 19(1), 23-37.
65. Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using Response Time to Detect Item Preknowledge in Computer Based Licensure Examinations. Educational Measurement: Issues and Practice, 35(1), 38-47.
66. Reckase, M.D.(2009). Multidimensional Item Response Theory. NY: Springer.
67. Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. Applied Psychological Measurement, 15(3), 217-226.
68. Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. Developments in applied statistics, 19(1), 160-117.
69. Rhemtulla, M., van Bork, R., & Borsboom, D. (2015). Calling models with causal indicators “measurement models” implies more than they can deliver. Measurement: Interdisciplinary Research & Perspectives, 13(1), 59-62.
70. Rubin, D. B. (1987).Multiple imputation for nonresponse in surveys. New York, NY: Wiley.
71. Santos, P., Hernández-Leo, D., Pérez-Sanagustín, M., & Blat, J. (2012). Modeling the computing based testing domain extending IMS QTI: Framework, models and exemplary implementations. Computers in Human Behavior, 28(5), 1648-1662.
72. Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. New ideas in psychology, 31(1), 43-53.
73. Swaminathan H., Rogers H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, Vol. 27, pp. 361–370.
74. Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal Design Applied to Large Scale Assessments. Synthesis Report.
75. Tindal, G., Heath, B., Hollenbeck, K., Almond, P., &Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study. Exceptional Children, 64(4), 439-450.
76. van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. Psychometrika, 68(2), 251-265.
77. Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting Aberrant Behavior and Item Preknowledge: A Comparison of Mixture Modeling Method and Residual Method. Journal of Educational and Behavioral Statistics, 1076998618767123.
78. Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. Contemporary Issues in Technology and Teacher Education, 8(4), 310-325.
79. Wiliam, D., & Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? British Educational Research Journal, 22(5), 537-548.
80. Wilson, M., Zheng X. (2012). Formulating Latent Growth Using an Explanatory Item Response Model Approach. Journal of applied measurement,13(1).
81. Wise, S. L. (2018). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. Education Inquiry, 1-13.
82. Wright B.D., Stone M.N. (1979). Best Test Design

## 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).