

Лекция №10

Тема: Шкалирование результатов тестирования

Цель: Раскрыть сущность шкалирования результатов тестирования в образовательной оценке. Определить специфику методов шкалирования, их применение и влияние на интерпретацию данных в педагогической практике.

Вопросы для рассмотрения:

1. Шкалирование результатов тестирования: понятие, цели и принципы.
2. Методы шкалирования (линейное, эквиперцентильное, на основе теории IRT).
3. Преимущества и недостатки различных подходов к шкалированию в образовании.
4. Этапы шкалирования и его роль в анализе образовательных достижений.
5. Применение шкалированных результатов в мониторинге качества образования и принятии решений.

Краткое содержание лекции: Под шкалированием результатов тестирования следует понимать статистический процесс преобразования сырых баллов (raw scores), полученных за выполнение тестовых заданий, в стандартизированные шкалированные баллы (scaled scores), которые позволяют сравнивать результаты разных форм тестов, учитывая их различия в сложности и обеспечивая объективность оценки. Это необходимо для создания единой метрики, где, например, балл 500 может соответствовать среднему уровню знаний независимо от версии теста. Шкалирование решает проблему эквивалентности: разные тесты (формы) имеют разную сложность, и без корректировки сравнение результатов будет несправедливым. В образовательной практике шкалирование применяется в крупных системах оценки, таких как NAEP (США), SAT, ACT или ЕНТ (Казахстан), где результаты выражаются на шкалах от 0 до 500 или 0 до 140 баллов, что облегчает интерпретацию и прогнозирование. Основная цель — повышение валидности и надежности оценки, позволяя учителям, администраторам и политикам анализировать тенденции, выявлять пробелы и корректировать программы. Шкалирование не создает интервальные шкалы (как отметки), но обеспечивает ординальную или равномерную метрику для агрегации данных.

К составляющим шкалирования относятся: сырые баллы как исходные данные, выбор метода (линейный для простых корректировок, эквиперцентильный для сохранения распределения, IRT — для моделирования способностей), эталоны (нормы по возрастным группам или стандартам), программное обеспечение (например, на базе R или SPSS) и валидация результатов. Ведущими функциями шкалирования в образовании являются: сравнительная (для межгруппового анализа), диагностическая (выявление уровней освоения), прогностическая (предсказание успеха) и информативная (для отчетов). Основная сфера применения — отчетность в системах мониторинга, где шкалированные баллы используются для ранжирования школ, учителей и учеников, а также для грантов и аттестации.

Совокупность показателей шкалирования всегда согласована с образовательными стандартами (например, ГОСО РК), где учитываются демографические факторы (возраст, регион). Сбор данных о сырых баллах и их преобразование имеет не только педагогический, но и управленческий смысл: статистика по шкалированным результатам демонстрирует эффективность реформ, выявляет неравенство и обосновывает бюджетные решения. Некоторые критики отмечают риски: манипуляция эталонами может исказить данные в

пользу политических целей, усиливая централизованный контроль. Однако использование стандартных критериев (перцентили, z-оценки) и статистической коррекции (учет сложности заданий) минимизирует это, обеспечивая репрезентативность.

Шкалирование позволяет решать задачи интерпретации: данные помогают выявить систематические ошибки в тестах, оценить влияние инноваций (например, онлайн-форм), диагностировать причины низких результатов (связь с социально-экономическими факторами). В целом, оно обеспечивает обратную связь для улучшения обучения, делая результаты сопоставимыми во времени и пространстве.

Условия эффективного шкалирования. Специалисты по психометрии выделяют ряд условий его реализации, среди которых:

- использование валидных и надежных тестов как базы данных;
- сочетание количественных методов (статистическая эквивалентность) с качественными (экспертная валидация);
- репрезентативная выборка (учет разнообразия учеников);
- корректная интерпретация с учетом контекста (возраст, мотивация);
- репрезентативность эталонных групп для норм;
- привлечение квалифицированных психометристов и IT-специалистов;
- использование современного ПО (IRT-модели в BILOG или IRTPRO) для обработки;
- методическая поддержка от министерств образования.

Наиболее простая модель шкалирования — линейная (простая пропорциональная корректировка), нацеленная на базовое выравнивание форм тестов по среднему и стандартному отклонению. Достоинство — простота и скорость; недостатки — игнорирование нелинейных различий в сложности, что ограничивает точность для продвинутых анализов.

Динамическая модель шкалирования. Большинство образовательных систем используют вертикальное шкалирование для отслеживания прогресса (например, от 4-го к 8-му классу), основанное на якорных заданиях (общих для смежных тестов). Это позволяет строить непрерывную шкалу, учитывая начальные способности и семейные факторы. Преимущество — учет динамики роста; динамическая модель рекомендуется для мониторинга, где измеряется скорость прироста достижений. Благодаря теории откликов на задания (IRT), субтесты адаптируются без потери сопоставимости, размещая результаты на единой шкале для долгосрочного анализа.

При проведении шкалирования, как и при любом психометрическом анализе, выделяют три основных этапа:

- подготовка (сбор сырых данных, выбор метода и эталонов);
- обработка (статистическая корректировка, моделирование);
- анализ и интерпретация (формирование отчетов, валидация).

Эти этапы конкретизируются в зависимости от уровня (школьный — простая нормировка; национальный — IRT-модели), расширяя функции до комплексного прогнозирования.

Вопросы для самоконтроля:

1. Сущность и цели шкалирования результатов тестирования.
2. Методы шкалирования и их сравнение.
3. Преимущества и недостатки шкалирования в образовании.
4. Этапы шкалирования и условия его эффективности.
5. Роль IRT в динамическом шкалировании для регионального уровня.
6. Выбор эталонов и интерпретация шкалированных баллов.

Литература:

1. Mayorov, A.N. Theory and practice of creating tests for the education system. M.: Intellect-Center, 2001. – 296 p. [https://www.ets.org/Media/Research/pdf/RD_Connections16.pdf].
2. Chelyshkova, M.B. Theory and practice of constructing pedagogical tests: Textbook. M.: Logos, 2002. – 432 p. [<https://www.thoughtco.com/understanding-scaled-scores-4161300>].
3. Kolen, M.J., Brennan, R.L. Test Equating, Scaling, and Linking: Methods and Practices. Springer, 2014. – 574 p. [https://nces.ed.gov/nationsreportcard/guides/scores_achv.aspx].
4. von Davier, A.A. Statistical Models for Test Equating, Scaling, and Linking. Springer, 2011. – 271 p. [<https://scalingcommunityofpractice.com/scaling-approaches-for-education-assessments/>].
5. Educational Testing Service (ETS). Why Do Standardized Testing Programs Report Scaled Scores? [https://www.ets.org/Media/Research/pdf/RD_Connections16.pdf].