СТАТИСТИЧЕСКИЕ ИССЛЕДОВАНИЯ ХАРАКТЕРИСТИК СРЕДСТВ ИЗМЕРЕНИЙ

д.т.н. Бояшова С.А.

Обобщенные рекомендации к разработке тестовых заданий и композиции целостного теста в соответствии с современной теорией и практикой педагогических измерений приводятся в работах В.С. Аванесова и А.Н. Майорова[1].

Основываясь на перечисленных выше авторских работах, опишем методику исследования качества средств измерений, исходя из особенностей построения АСИ.

Апробация средств измерений (тестов) начинается с исследований статистической трудности и дифференцирующей способности элементов теста (тестовых заданий) на основе данных тестирования различных выборок испытуемых.

В границах первоначального исследования структурный элемент системы теста рассматривается как одно тестовое задание, условно выполняемое всеми испытуемыми. В реальных условиях тестирования каждый испытуемый выполняет свое задание, относящееся к одному элементу системы теста, которое выбрано автоматически из базы АСИ в момент начала тестирования.

Если структурный элемент теста, рассматриваемый как одно тестовое задание, не соответствует статистическим требованиям, то пересматриваются все задания в базе АСИ, которые соответствуют данному элементу.

База заданий корректируется экспертами, и далее проводятся последующие исследования качества структурных элементов системы теста.

Каждое новое измерение позволяет автоматически улучшать качество тестов в АСИ.

Далее, исходя из вышесказанного, понятие «статистическая трудность задания» будем рассматривать как аналогичное понятию «статистическая трудность элемента теста» (СТЭТ), а понятие «дифференцирующая способность задания» как аналогичное понятию «дифференцирующая способность элемента теста» (ДСЭТ).

Основные статистические характеристики элементов теста (СТЭТ и ДСЭТ) рассчитываются по данным, полученным при апробации АСИ в режиме реального времени на двух различных по объему и качественному составу статистических выборках испытуемых.

Минимальная статистическая выборка отбирается по заявке образовательного учреждения в количестве тридцати испытуемых из общего количества тестируемых в следующем соотношении:

- 15 испытуемых – учащиеся, имеющие оценки в исследуемой предметной области «4» или «5» по пятибалльной шкале;
- 15 испытуемых – учащиеся, имеющие оценки в исследуемой предметной области «3» или «2» по пятибалльной шкале.

Полная статистическая выборка в качественном соотношении формируется произвольно по заявке образовательного учреждения, изъявившего желание пройти тестирование. Количественные характеристики выборок, на которых проходило

---

[1] Аванесов В.С. Теоретические основы разработки заданий в тестовой форме: учебное пособие для профессорско-преподавательского состава высшей школы. / В.С. Аванесов. – М.: МГТА, 1995. – 198 с. Майоров А.Н. Теория и практика создания тестов для системы образования. / А.Н. Майоров. – М.: Народное образование, 2000.

исследование основных статистических характеристик элементов моделей тестов по русскому языку и математике, представлены в табл. 1.

Таблица 1

Характеристики объемов выборок испытуемых

Наименование	2013 год	2014 год
Общая выборка испытуемых из числа школьников 4-х классов (русский язык, математика)	1665	1163
• выборка по русскому языку. Из них:	784	590
• СОШ	471	276
• гимназий	228	297
• лицеев	50	17
• коррекционных школ	35	0
• выборка по математике. Из них:	881	573
•СОШ	561	293
•гимназий	235	263
•лицеев	49	17
•коррекционных школ	36	0

СТЭТ рассчитывается по формуле статистической трудности тестового задания:

$$P = \frac{N_1}{N},$$

где: P – статистическая трудность задания;

N_1 – число испытуемых, правильно выполнивших задание;

N – общее число испытуемых, выполнявших задание.

Согласно тестовой теории, задания, имеющие статистическую трудность более 80% и менее 20%, не считаются тестовыми и выбраковываются.

Дифференцирующая способность элемента теста (ДСЭТ) оценивается аналогично дифференцирующей способности задания как мера соответствия между успешностью выполнения одного элемента теста в заданной выборке испытуемых и всего теста.

Количественной характеристикой ДСЭТ считают коэффициент дискриминации. Коэффициент дискриминации ДСЭТ равен:

$$\kappa_\partial = \frac{\overline{X}_i - \overline{X}}{\delta}\sqrt{\frac{N_n}{N - N_n}},$$

где: $\overline{X_i}$ – среднее арифметическое значение баллов, полученных испытуемыми, правильно выполнившими задание *I* (элемент теста);

$\overline{X}$ – среднее арифметическое значение баллов, полученных испытуемыми по всему тесту,

δ – среднеквадратическое отклонение баллов, полученных испытуемыми по всему тесту,

N_n – число испытуемых в данной выборке, правильно выполнивших задание (элемент теста) *i.*

Значения коэффициента дискриминации тестового задания могут находиться в интервале от – 1 до + 1.

Если коэффициент дискриминации равен + 1, то элемент теста имеет высокую дифференцирующую способность и позволяет различать испытуемых с высоким и низким уровнем подготовки.

Если коэффициент дискриминации равен – 1, то элемент теста непригоден для теста.

Если коэффициент дискриминации равен 0, то элемент теста сформулирован некорректно.

Если тест содержит небольшое количество элементов или отдельные блоки, то дифференцирующая способность элементов теста оценивается по коэффициенту корреляции.

Задания, из которых формируются элементы теста в АСИ, исследуются на предмет оценки их СТЭТ и ДСЭТ с 2011 года, и на основании полученных результатов можно утверждать, что их следует использовать для формирования метрологически пригодных средств педагогического измерения.

Исследование системных свойств тестов, используемых в АСИ, проводится на основе исследования системных свойств средств измерений. К системным свойствам относится надежность шкалы средств измерений, которая оценивается по трем критериям[2]: обоснованность, устойчивость, точность.

Обоснованность (валидность) – это статистическая характеристика средств измерений, понимаемая как его способность измерять определенное заданное свойство или признак, не смешивая его с другими.

Контроль средств измерений на обоснованность его оценочной шкалы проводится в четыре шага.

Шаг первый – экспертиза кодификатора на его соответствие содержанию учебных программ и ФГОС в выделенной предметной области.

Шаг второй – экспертиза тестовых заданий на соответствие структурной единице кодификатора.

Шаг третий – проведение пробного тестирования минимальной статистической выборки испытуемых (30 человек из числа выпускников начальной школы) и сопоставление полученных данных с результатами итоговой аттестации за курс начальной школы в соответствующей предметной области.

Шаг четвертый – расчет коэффициента корреляции между двумя рядами случайных величин: X_i (ряд оценок, полученных в выборке испытуемых за тест) и Y_i (ряд итоговых оценок, полученных в той же выборке испытуемых).

Для определения корреляционной зависимости между переменными величинами в случае парной зависимости, используется коэффициент корреляции Пирсона[3]

$$r_{XY} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sqrt{\overline{X^2} - \left(\overline{X}\right)^2} \cdot \sqrt{\overline{Y^2} - \left(\overline{Y}\right)^2}}$$

Значение r всегда заключено в пределах $-1 < r < +1$.

Если $r = \mp 1$, то между рассматриваемыми величинами существует прямая связь в первом случае и обратная – во втором.

---

[2] Методы системного педагогического исследования. / Сб. ст. – М.: Народное образование, 2002

[3] Горелова Г.В., Кацко И.А. Теория вероятности и математическая статистика в примерах и задачах с применением Excel. / Г.В. Горелова, И.А. Кацко. – Ростов на/Д : Феникс, 2006. – 475 с.

Если $r = 0$, то это значит, что между рассматриваемыми величинами не существует ни функциональной, ни корреляционной обратной связи, но криволинейная корреляционная связь может существовать.

Чем ближе $r \to +1$ или $r \to -1$, тем точнее и теснее корреляционная связь между величинами X и Y. При: $r = 0 \div 0{,}2$ - связь незначительная, $r = \pm 0{,}2 \div \pm 0{,}40$ - низкая степень связи, $r = \pm 0{,}40 \div 0{,}70$ - ясно выраженная корреляция, $r = \pm 0{,}70 \div 1{,}0$ - высокая или очень высокая степень корреляции.

Если между выборочными значениями X_i и Y_i существует слабая взаимозависимость, то можно утверждать, что данные, полученные по шкалам экспертов, слабо сопоставляются с данными тестирования.

Вместе с тем, слабая взаимозависимость не является демонстрацией слабой валидности шкалы в АСИ по отношению к исследуемому признаку.

Если такой факт выявляется при апробации АСИ, то возникает новая задача исследования, связанная с выявлением влияния одного фактора (оценки за итоговый тест) на другой (итоговой экспертной оценки). Эта задача может быть решена методом дисперсионного анализа.

Устойчивость шкалы средств измерений определяется как однозначность данных, полученных при их использовании, со значительным временным промежутком (корреляция между первой и второй серией измерений должна быть высокой – 0,9).

Исследования средств измерений в АСИ на устойчивость шкалы проводятся их повторным испытанием на одной и той же выборке испытуемых: первое испытание – окончание обучения (4 класс); второе испытание – начало обучения на новой ступени (5 класс).

Точность средств измерений определяется из оптимального соотношения между чувствительностью измеряемого объекта и устойчивостью данных.

Оптимальность достигается экспериментально с помощью увеличения числа пунктов шкалы (числа элементов содержания в эталонном кодификаторе) и проверки шкалы на устойчивость.

Применение описанных выше статистических методов исследования позволяет обеспечить высокое качество тестов в АСИ.