КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ АБАЯ
Факультет Педагогики и психологии
Кафедра начального образования

Оценивание надежности и валидности педагогических тестов.

Дисциплина: «Технология оценивания в начальном образовании»

Алматы, 2025

Оценивание надежности и валидности педагогических тестов.

Вопросы для рассмотрения:

1. Оценивание надежности ретестовым методом (двукратное тестирование).
2. Метод параллельных форм.
3. Метод расщепления теста (однократное тестирование).
4. Метод Кьюдера— Ричардсона (для дихотомических оценок по заданиям теста).
5. Надежность и стандартная ошибка измерения .
6. Валидность гомогенных тестов.

1.1 Общие замечания о надежности и методах ее оценивания.

Методы оценки надежности нормативно-ориентированных тестов по способу осуществления

1 группа

базируется на двукратном тестировании, проводимом с помощью:

одного и того же теста

с помощью двух параллельных форм теста.

Предпочтение

2 группа

однократное тестирование

поскольку организация повторного тестирования и разработка параллельных форм всегда сопряжены с определенными трудностями и дополнительными затратами со стороны создателей тестов.

1.1Общие замечания о надежности и методах ее оценивания.

Вне зависимости от метода оценка надежности строится на подсчете корреляции между двумя наборами данных.

Чем выше  корреляция тем надежнее  тест

1.1 Общие замечания о надежности и методах ее оценивания.

Для маленькой выборки корреляцию можно оценить визуально:

Таблица 1

Результаты двукратного выполнения трех тестов

Номер ученика	Тест *A*		Тест *B*		Тест *C*	
	1-е тестирование	2-е тестирование	1-е тестирование	2-е тестирование	1-е тестирование	2-е тестирование
1	10	10	10	1	10	6
2	9	9	9	2	9	4
3	8	8	8	3	8	8
4	7	7	7	4	7	9
5	6	6	6	5	6	3
6	5	5	5	6	5	1
7	4	4	4	7	4	5
8	3	3	3	8	3	7
9	2	2	2	9	2	2
10	1	1	1	10	1	10

Тест А обладает оптимальной надежностью, так как результаты 10 учеников остались прежними: баллы и места учеников не изменились после повторного выполнения теста.

Тест В полностью ненадежен: тот, кто имел самые высокие баллы в первом тестировании, получает самые низкие баллы во втором тестировании после повторного применения этого же теста.

Тест С обеспечивает в целом хаотичное изменение результатов, хотя баллы отдельных учеников 3-го и 9-го) будут воспроизведены при повторном выполнении теста. Скорее всего, надежность третьего теста близка к нулю.

1.2 Подсчет коэффициента надежности.

Ретестовый метод оценки надежности (test-retest reliability)

- основан на подсчете корреляции индивидуальных баллов испытуемых, полученных в результате двукратного выполнения ими одного и того же теста.
- Обычно повторное тестирование проводится через 1 — 2 недели, когда испытуемые еще не успели забыть учебный материал и незначительно продвинулись в усвоении новых знаний.

Для подсчета коэффициента надежности по методу повторного тестирования используется формула 1:

Формула 1

$$(r_н)_{рет} = \frac{N\sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i\right)\left(\sum_{i=1}^{N} Y_i\right)}{\sqrt{N\sum_{i=1}^{N}(X_i)^2 - \left(\sum_{i=1}^{N} X_i\right)^2}\sqrt{N\sum_{i=1}^{N}(Y_i)^2 - \left(\sum_{i=1}^{N} Y_i\right)^2}},$$

где $(r_н)_{рет}$ — коэффициент надежности теста по ретестовому методу; X_i — индивидуальный балл i-го испытуемого в первом тестировании; Y_i — индивидуальный балл i-го испытуемого во втором тестировании ($i = 1, 2, \ldots, N$).

1.3 Пример подсчета.

Используя данные табл. 2 (первое тестирование) и добавляя к ним гипотетические данные второго тестирования, можно с помощью табл. 3 подсчитать коэффициент надежности ретестовым методом.

Таблица 2

Матрица результатов с индивидуальными баллами испытуемых и количеством правильных ответов на задания теста

Номер испытуемого i	Номер заданий j										Индивидуальные баллы (множество X_i)
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	1	1	0	0	0	0	6
2	1	1	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	1	0	0	1
4	1	1	0	1	1	1	1	1	1	1	9
5	1	0	1	0	1	1	0	0	0	0	4
6	1	1	1	0	0	0	0	1	0	0	4
7	1	1	1	1	0	1	0	0	0	0	5
8	1	1	1	1	0	0	0	0	0	0	4
9	1	1	1	1	1	1	1	1	1	0	9
10	1	1	1	1	1	0	1	0	0	0	6
Число правильных ответов (множество R_j)	9	8	7	6	5	5	3	4	2	1	50

Таблица 3

Пример данных для оценки надежности

Номер ученика i	Балл при первом тестировании X_i	Балл при втором тестировании Y_i	X_iY_i	$(X_i)^2$	$(Y_i)^2$
1	6	5	30	36	25
2	2	4	8	4	16
3	1	2	2	1	4
4	9	7	63	81	49
5	4	6	24	16	36
6	4	3	12	16	9
7	5	7	35	25	49
8	4	6	24	16	36
9	9	7	63	81	49
10	6	8	48	36	64
	$\Sigma X_i = 50$	$\Sigma Y_i = 55$	$\Sigma X_iY_i = 309$	$\Sigma(X_i)^2 = 312$	$\Sigma(Y_i)^2 = 337$

После подстановки чисел из нижней строчки таблицы в формулу 1 коэффициент надежности будет равен:

$$(r_{\text{н}})_{\text{рет}} = \frac{10 \cdot 309 - 50 \cdot 55}{\sqrt{10 \cdot 312 - 50^2}\sqrt{10 \cdot 337 - 55^2}} = \frac{340}{\sqrt{620}\sqrt{345}} \approx 0,78.$$

Значение $r_{\text{н}} = 0,78$ указывает на невысокую надежность теста.

2. Метод параллельных форм.

Метод параллельных форм (parallel-form reliability)

- малоэффективен в тех случаях, когда при тестировании используется один вариант теста.
- Необходима тщательная ротация вариантов в группе испытуемых для обеспечения сходных выборок учащихся на параллельных вариантах теста.
- Даже при стратификации выборки испытуемых и ротации вариантов достоверность оценок надежности снижается из-за того, что, как правило, обнаруживаются статистически значимые различия в характеристиках параллельных вариантов.

3. Метод расщепления теста (однократное тестирование)

3.1 Описание метода.

Наиболее распространен из-за своего удобства.

Позволяет вычислить коэффициент надежности при однократном выполнении учениками теста.

Для оценки надежности результаты тестирования делят на две части: в одну включают данные испытуемых по четным, а в другую — по нечетным заданиям, считая при этом, что получены сходные по содержанию части теста.

3.2 Подсчет коэффициента надежности.

Для оценивания надежности методом расщепления результаты учеников заносят в табл. 4

табл. 4 **Сводная таблица для оценки надежности (метод расщепления)**

Номер ученика i	Баллы по четным заданиям X_i	Баллы по нечетным заданиям Y_i	$X_i Y_i$	$(X_i)^2$	$(Y_i)^2$
1	X_1	Y_1	$X_1 Y_1$	$(X_1)^2$	$(Y_1)^2$
2	X_2	Y_2	$X_2 Y_2$	$(X_2)^2$	$(Y_2)^2$
...	...	...	...	...	...
N	X_N	Y_N	$X_N Y_N$	$(X_N)^2$	$(Y_N)^2$
	$\sum_{i=1}^{N} X_i$	$\sum_{i=1}^{N} Y_i$	$\sum_{i=1}^{N} X_i Y_i$	$\sum_{i=1}^{N} (X_i)^2$	$\sum_{i=1}^{N} (Y_i)^2$

Далее для таблицы данных используют формулу 1, в которой роль результатов в первом тестировании выполняют данные по четным, а во втором — по нечетным заданиям.

.

3.3 Коррекция коэффициента надежности.

Для коррекции оценки надежности в соответствии с длиной исходного теста используется:

Формула Спирмена— Брауна

$$r_н = \frac{2(r_н)_{расш}}{1+(r_н)_{расш}}$$

Формула 2

- где в числителе и знаменателе дроби стоит коэффициент надежности для половины заданий теста
- а слева — **скорректированный** коэффициент надежности с учетом всех заданий теста.

4. Метод Кьюдера— Ричардсона (для дихотомических оценок по заданиям теста).

4.1 Описание метода.

Как и метод расщепления теста, основан на однократном тестировании

но в отличие от него не зависит от искусственных допущений о полной параллельности двух частей теста.

Однако сфера его применения ограниченна, так как он годится лишь при использовании дихотомических оценок по результатам выполнения заданий гомогенных тестов.

4.2 Формула Кьюдера—Ричардсона.

Формула Кьюдера — Ричардсона (KR-20) имеет следующий вид:

$$(r_{\text{н}})_{\text{KR-20}} = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_j q_j}{S_X^2}\right),$$

Формула 3

где p_j — доля правильных ответов на j-е задание; q_j — доля неправильных ответов, $q_j = 1 - p_j$; S_X^2 — дисперсия по распределению наблюдаемых баллов; n — число заданий теста [87].

4.3 Общие рекомендации по применению метода Кьюдера —Ричардсона.

В целом при оценке надежности нельзя полагаться лишь на один показатель.

поскольку каждый из них имеет свои ограничения, сметающие оценки надежности теста в сторону завышения или занижения.

Для достоверной проверки качества теста следует учитывать несколько показателей надежности, подсчитанных по разным формулам.

4.3 Общие рекомендации по применению метода Кьюдера —Ричардсона.

В качестве нижнего предела допустимых значений надежности обычно выбирают **0,7**

При более низком значении использование теста вряд ли целесообразно в силу большой погрешности измерения.

Если тест разрабатывается профессионалами, то к нему предъявляют более жесткие требования.

тесты с надежностью менее **0,8** считаются непригодными в организованных службах и центрах тестирования.

значения коэффициента надежности, превышающие **0,9** свидетельствуют о высоком качестве теста.

Они желательны, но встречаются редко.

Обычно в тестологической практике надежность тестов колеблется в интервале 0,8; 0,9

5. Надежность и стандартная ошибка измерения.

Для установления связи между стандартной ошибкой измерения и надежностью теста необходимо преобразовать формулу (4) в формулу (5)

Формула 4

$$r_{11} = 1 - \frac{S_E^2}{S_X^2}.$$

Формула 5

$$S_E = S_X \sqrt{1 - r_{11}},$$

где S_X — стандартное отклонение по распределению индивидуальных баллов; r_{11} — коэффициент надежности теста; S_E — стандартная ошибка измерения. Это выражение обычно используется для вычисления S_E по известным величинам r_{11} и S_X.

5.2 Построение доверительного интервала.

Общераспространен подход, когда доверительный интервал выстраивается вокруг наблюдаемого показателя ученика как две симметричные окрестности (левая и правая)

хотя это не совсем верно, поскольку речь должна идти об окрестностях, расположенных слева и справа от истинного балла

Тем не менее этот факт вынужденно игнорируется в прикладных исследованиях в силу отсутствия истинного балла, и доверительный интервал при заданном риске допустить ошибку t= 0,05 (в пяти случаях из ста) принимается равным:

$(X_i - 1{,}96S_E; X_i + 1{,}96S_E)$, где X_i — наблюдаемый балл i-го испытуемого; 1,96 — константа, табличное число, используемое при $t = 0{,}05$.

5.3 Численный пример.

табл. 5

Номер испытуемого i	Номер заданий j										Индивидуальные баллы (множество X_i)
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	1	1	0	0	0	0	6
2	1	1	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	1	0	0	1
4	1	1	0	1	1	1	1	1	1	1	9
5	1	0	1	0	1	1	0	0	0	0	4
6	1	1	1	0	0	0	0	1	0	0	4
7	1	1	1	1	0	1	0	0	0	0	5
8	1	1	1	1	0	0	0	0	0	0	4
9	1	1	1	1	1	1	1	1	1	0	9
10	1	1	1	1	1	0	1	0	0	0	6
Число правильных ответов (множество R_j)	9	8	7	6	5	5	3	4	2	1	50

Для рассматриваемого ранее примера матрицы тестовых результатов (см. табл. 5), коэффициента надежности г„= 0,78 и стандартного отклонения Sx= 2,62, вычисленного ранее для матрицы, 5E-будет равно

$$S_E = 2,62\sqrt{1-0,78} \approx 1,23.$$

В данном случае доверительный интервал для истинного балла первого ученика со значением X1 = 6 будет составлять 6 - 1,23; 6 + 1,23) или 4,77; 7,23).

Истинный балл первого ученика может находиться в любой точке этого интервала.

5.4 Предсказание истинных баллов на основе регрессионной модели.

Методы регрессионного анализа позволяют прогнозировать оценки истинных баллов испытуемых по распределению наблюдаемых баллов и коэффициенту надежности теста.

Прогноз получается путем подстановки в регрессионное уравнение

$$T_i = \bar{X} + r_{н}\left(X_i - \bar{X}\right),$$

где T_i — истинный балл; X_i — индивидуальный балл i-го испытуемого; $\bar{X}$ — среднее значение баллов испытуемых [60].

6. Валидность гомогенных тестов.

6.1 Критерии для оценки валидности.

- всегда проводится путем соотнесения характеристик результатов измерения с внешними критериями
  - содержательная валидность;
  - конструктная валидность;
  - прогностическая валидность.
- Высокая корреляция между анализируемыми результатами испытуемых и внешними критериями подтверждает высокую валидность теста.
- Основная трудность при такой валидизации носит не практический, а методологический характер
  - поскольку она состоит в выборе значимого внешнего критерия.

6.2 Связь надежности и валидности.

Для повышения полноты охвата содержания и роста содержательной валидности теста

желательно отбирать задания с малыми коэффициентами интеркорреляции

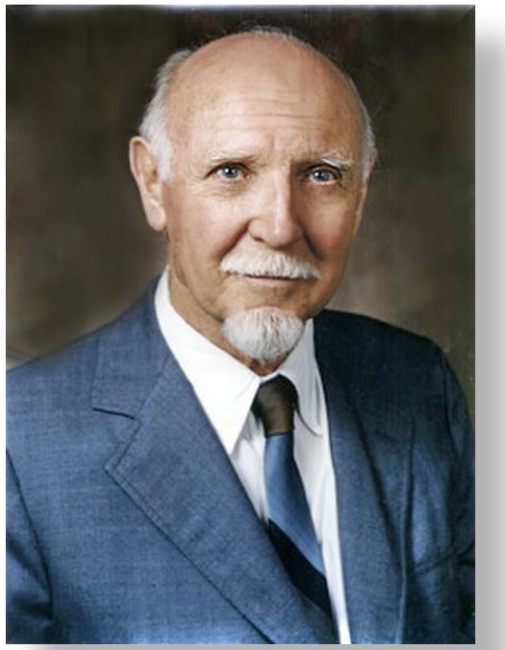
Отбирая задания с большими коэффициентами интеркорреляции

можно обеспечить высокую однородность содержания и надежность теста.

противоречие, получившее название «парадокс Ф.Лорда», приводит к возникновению серьезных проблем при конструировании теста.

Таким образом, при конструировании гомогенного теста следует стремиться к повышению в разумных пределах его надежности, чтобы не снизить существенным образом содержательную валидность теста.

6.2 Связь надежности и валидности.



Раймонд Кеттелл

Максимум валидности может быть получен тогда, когда все задания слабо, но положительно коррелируют друг с другом, однако каждое из них имеет высокую корреляцию с критерием по тесту.

Внутренняя согласованность теста — непременное условие его высокой содержательной валидности, и потому высокая надежность является предпосылкой оптимальной валидности теста.

Джой Пол Гилфорд

6.3 Количественные оценки валидности.

При количественных оценках валидности для педагогических тестов в качестве критерия обычно берутся оценки экспертов, выставленные ими при традиционной проверке знаний учеников без использования тестов.

Процесс валидизации осложняется необходимостью установления меры согласованности оценок экспертов, которых обычно бывает не менее трех человек.

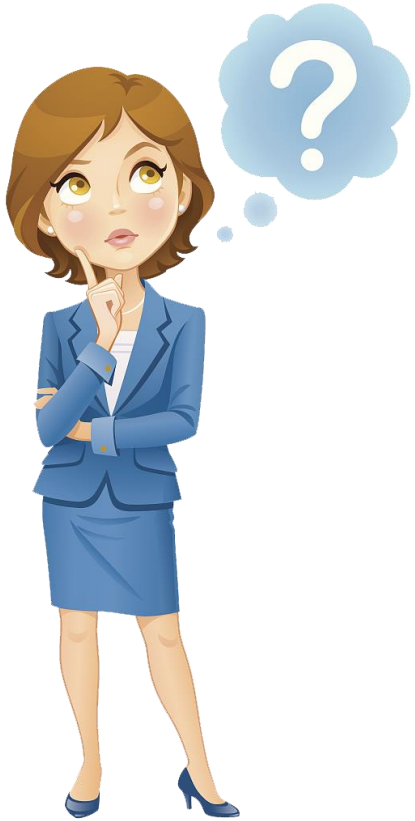
Если мера согласованности достаточно высока, то для оценки валидности используется формула:

$$r_{в} = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(X_{m_i} - \bar{X}_э\right)}{N_m \sqrt{S_X^2 \cdot S_{m_э}^2}},$$

где $X_i - \bar{X}$ — отклонение тестового балла i-го ученика от среднего балла по тесту; $X_{m_i} - \bar{X}_э$ — отклонение балла i-го ученика у экспертов от $\bar{X}_э$ — среднего арифметического экспертных оценок; S_X^2 — дисперсия баллов учеников по тесту; $S_{m_э}^2$ — дисперсия баллов m-го эксперта; m — число экспертов.

6.3 Количественные оценки валидности.

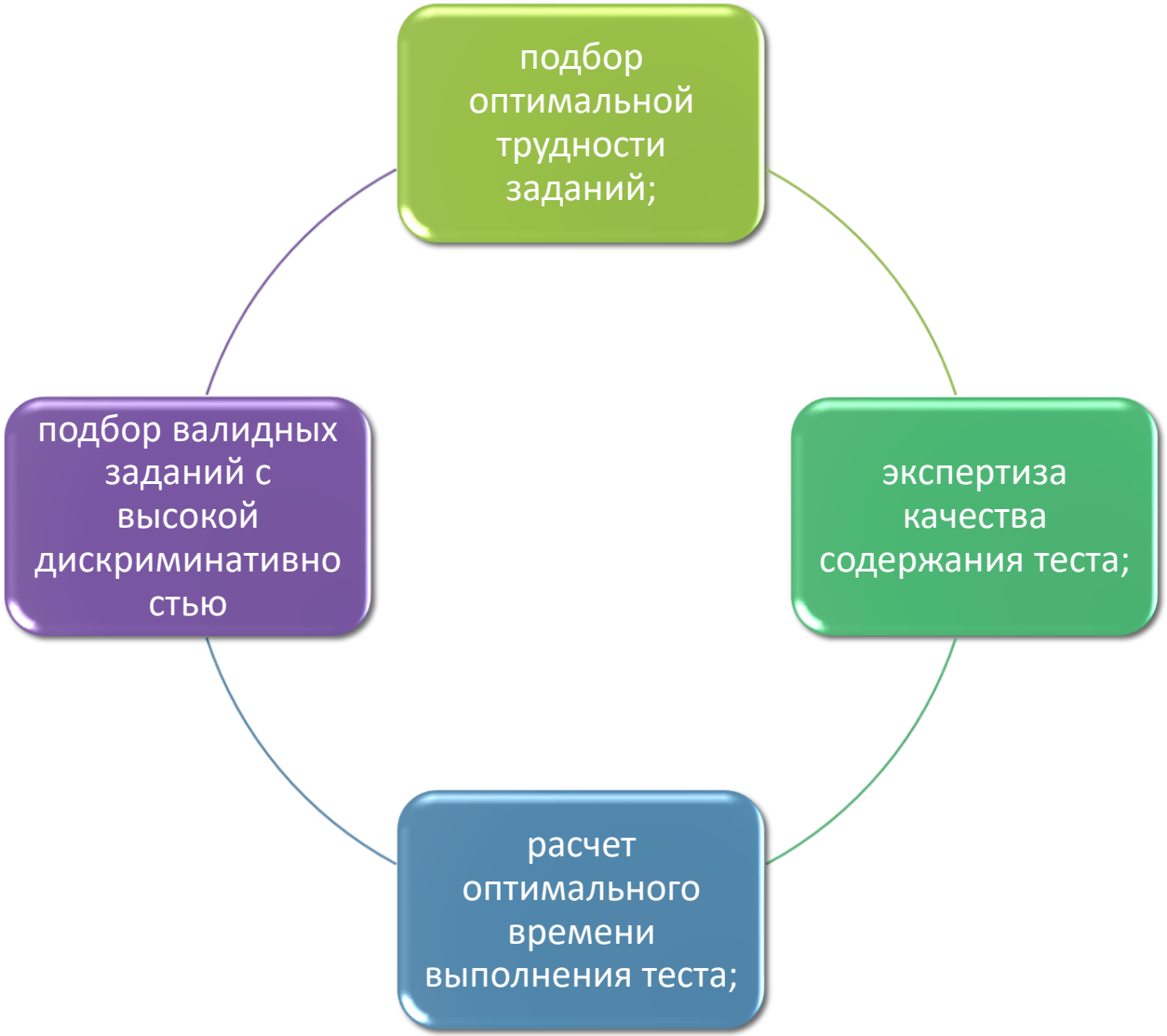
Бывают случаи, когда **педагог** заинтересован в оценке прогностической валидности, указывающей меру вероятности прогноза успешности дальнейшего обучения по результатам выполнения теста.



- В этом случае результаты по тесту коррелируют с результата- 164 ми поступивших абитуриентов после окончания первого года обучения в вузе.

- Высокая корреляция означает, что разработанные тесты для отбора абитуриентов в вуз пропюстичны.

6.4 Источники повышения валидности теста.



Спасибо за внимание!