



Funded by
the European Union

КЫРГЫЗ РЕСПУБЛИКАСЫНЫН
БИЛИМ БЕРҮҮ ЖАНА ИЛИМ
МИНИСТРЛИГИ



МИНИСТЕРСТВО
ОБРАЗОВАНИЯ И НАУКИ
КЫРГЫЗСКОЙ РЕСПУБЛИКИ



THE WORLD BANK



Обеспечение качества инструментария исследования функциональной грамотности

Колачев Никита Игоревич

Иссык-Куль 2022

PISA

1. Психометрический анализ заданий.
2. Проверка надёжности измерений.
3. Проверка валидности результатов.

1. Психометрический анализ заданий.

Анализ заданий:

- Анализ трудности
- Анализ дискриминирующей способности
- Согласие заданий с моделью

Анализ категорий:

- Анализ наполненности категорий
- Анализ дискриминирующей способности категорий
- Согласие категорий с моделью

- Распределение трудности заданий в соответствии с теоретической рамкой (если возможно)
- Размах трудности заданий
- Наличие экстремальных значений

Значение	Интерпретация
Более 0,90	Очень легкое задание
0,61-0,89	Легкое задание
0,40-0,60	Задание оптимальной трудности
0,20-0,39	Трудное задание
$\leq 0,19$	Очень трудное задание



- Дискриминативность – мера связи выполнения отдельного задания с результатами выполнения теста в целом.
- Для исследования используются корреляционные методы оценки.



ИНТЕРПРЕТАЦИЯ ЗНАЧЕНИЙ КОЭФФИЦИЕНТА ДИСКРИМИНАТИВНОСТИ

- Если $D_j \geq 0,4$, то задание функционирует удовлетворительно
- Если $0,30 \leq D_j \leq 0,39$, то требуется небольшая коррекция задания
- Если $0,20 \leq D_j \leq 0,29$, то задание нуждается в пересмотре
- Если $D_j \leq 0,19$, то задание должно быть исключено из теста или полностью переделано



- Изучается только в рамках современной теории тестирования
- В рамках моделей Г. Раша используются два показателя:
 - Взвешенная статистика согласия (INFIT)
 - Невзвешенная статистика согласия (OUTFIT)
- Эти статистики представляют собой отклонения от ожидаемого значения, основаны на анализе остатков

Продуктивное измерение возможно, если статистики согласия не выходят за рамки интервала $[0.50; 1.50]$.



- **Взвешенная статистика** согласия связана с неожиданными ответами со стороны учащихся, для которых это задание по трудности является оптимальным.
- **Невзвешенная статистика**, в свою очередь, реагирует на неожиданные ответы со стороны либо слабо подготовленных, либо очень подготовленных учащихся.

Более подробная информация о том, как рассчитываются эти показатели: [Fit diagnosis: infit outfit mean-square standardized: Winsteps Help](#)



- Необходим для изучения функционирования шкалы
- Считается, что если категория набирает **менее 5%** наблюдений, то она не функционирует как необходимо.



- Дискриминирующая способность категорий задания – это мера связи между получением того или иного балла за задание и общим баллом по тесту.
- Категория «0 баллов» должна иметь отрицательную связь с итоговым баллом, «1 балл» и «2 балла» - положительную.
- Интерпретация аналогична дискриминативности задания.



- Изучается только в рамках современной теории тестирования
- В рамках моделей Г. Раша используется невзвешенная статистика согласия (OUTFIT)



2. Проверка надёжности измерений.

- Надёжность – один из ключевых показателей качества измерений.
- В теории измерений под надёжностью понимают долю ошибки в результатах измерения. Или иначе: доля истинных баллов в наблюдаемых.

$$r_{XT} = \frac{SD_T}{SD_X}$$



2. Проверка надёжности измерений.

- Существуют разные виды надёжности:
 - Надёжность-согласованность (проверить согласованность (гомогенность) поведения тестируемого между несколькими наборами тестовых заданий)
 - Надёжность-устойчивость (как правило, один и тот же инструмент; проверяем изменения во времени той или иной характеристики)
 - Ретестовая надёжность (как правило, разные формы (псевдопараллельные); доказываем, что формы теста хотя бы эквивалентны)



2. Проверка надёжности измерений.

- Наиболее распространённые показатели: альфа Кронбаха и надёжность Раш-измерений.
- Но есть и современные альтернативы: омега МакДональда, GLB (greater lower bound).
- См. [Frontiers | Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements \(frontiersin.org\)](https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01611/full)



3. Проверка валидности результатов.

- Проверка валидности результатов необходима, чтобы быть уверенным, что приписываемые респондентам баллы отражают интересующий нас конструкт.
- По сути все виды валидности – разновидности конструктивной (Messick, 1995).
- Содержательная валидность
- Структурная валидность
- Критериальная валидность
- Конвергентная валидность
- Дивергентная валидность

- Показывает, соответствует ли содержание теста заявленной теоретической рамке.
- Для проверки используется экспертная оценка (внешняя экспертиза).



- Показывает, получается ли на эмпирических данных заложенная в инструмент структура: одномерный или многомерный инструмент?
- Для проверки используются модели конфирматорного факторного анализа или современной теории тестирования.



Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --		Modeled
Total raw variance in observations	=	140.5	100.0%	100.0%
Raw variance explained by measures	=	47.5	33.8%	33.5%
Raw variance explained by persons	=	16.6	11.8%	11.7%
Raw Variance explained by items	=	30.9	22.0%	21.8%
Raw unexplained variance (total)	=	93.0	66.2%	66.5%
Unexplned variance in 1st contrast	=	2.2	1.6%	2.4%
Unexplned variance in 2nd contrast	=	2.0	1.5%	2.2%
Unexplned variance in 3rd contrast	=	1.9	1.3%	2.0%
Unexplned variance in 4th contrast	=	1.7	1.2%	1.8%
Unexplned variance in 5th contrast	=	1.6	1.1%	1.7%



- Показывает связь результатов измерений с похожими переменными, измеренными другим способом (обычно объективные меры, контрастные группы).
- **Например:** для тестов интеллекта критерием выступают оценки учащихся (годовые, семестровые, четвертные); в клинической психологии используют контрастные группы (нормативная/клиническая выборка).



- Показывает связь результатов измерений с результатами измерений методик измерения того же конструкта.
- **Например:** при разработке теста интеллекта можно проверить связь измерения этим инструментом с результатами теста Векслера или матриц Равена.



- Показывает отсутствие связи (слабые связи) с тем, с чем результаты наших измерений не должны быть связаны.
- **Например:** результаты тестов, измеряющих интеллект, обычно не связаны с весом взрослого человека.