Дизайн теста FORtitude

Тест FORtitude измеряет вербальные и числовые способности. Каждый респондент получает свой собственный набор заданий. Банк заданий делится на сектора по следующим параметрам:

1. Измеряемый конструкт — вербальные и числовые способности.
2. Измеряемый компонент каждой способности (по 3 компонента в каждой способности).
3. Трудность каждого задания: лёгкое / среднее / сложное.

Всего каждый респондент получает 20 заданий (10 из которых направлены на измерение вербальных способностей, другие 10 — на измерение числовых). На выполнение каждого задания теста отводится 2 минуты, таким образом, максимальное время тестирования — 40 минут. Среднее время тестирования — 21 мин. Тестирование проводится онлайн на платформе Formatta Assessment https://assessment.formatta.pro

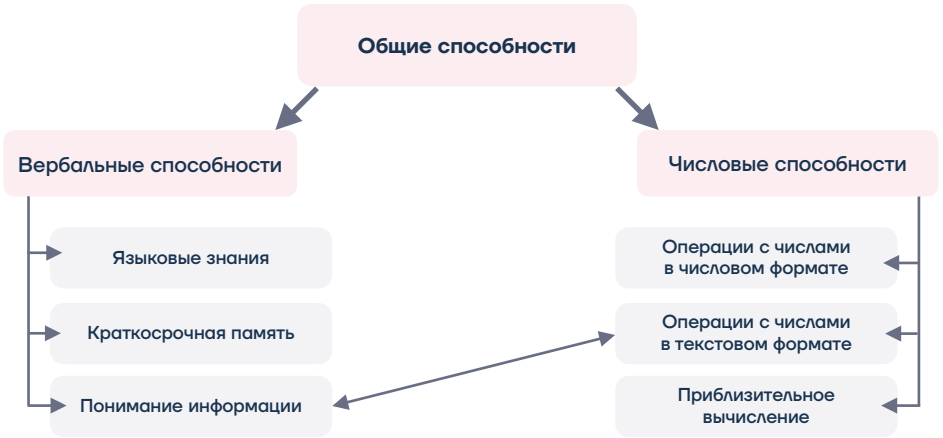
Сырые баллы (верно и неверно выполненные задания) тестируемого отправляются на сервер, где автоматически обрабатываются в рамках современной теории тестирования IRT (https://files.eric.ed.gov/fulltext/EJ1133065.pdf). Моделируются конструкты вербальных и числовых и общие способностей в логитах. Оценка способностей производится на основе параметров заданий, полученных в процессе количественных апробаций.

Отчёт с результатами тестирования заказчик выгружает с платформы Formatta Assessment после тестирования в форматах .xls (индивидуальный отчёт тестируемого и общий по группе тестируемых) и pdf (индивидуальный отчёт тестируемого).

Как разрабатывался тест

1. Определение измеряемых конструктов и их структуры, подготовка спецификации теста, в частности — обоснование измеряемых конструктов и их операциональных определений.
2. Разработка заданий, направленных на измерение конструктов.
3. Экспертиза и качественная апробация разработанных заданий.
4. Количественная апробация разработанных заданий 1.0.
5. Моделирование конструктов и установление психометрических характеристик заданий и теста в целом.
6. Разработка заданий, направленных на измерение конструктов (наполнение банка заданий).
7. Количественная апробация разработанных заданий 2.0.
8. Подготовка технического отчета по тесту.
9. Проведение дополнительных валидизационных исследований.

Вербальные и числовые способности определялись на основе анализа литературы: теорий, концепций и инструментов, накопленных наукой за десятилетия исследования. Мы выделили следующую структуру конструктов:



Задания теста разрабатывались действующими сотрудниками коммерческих компаний и психометриками по следующей процедуре:

1. Опрос потенциальных разработчиков заданий — чтобы убедиться в их принадлежности целевой аудитории тестирования и соответствия роли разработчика.
2. Индивидуальный тренинг по разработке заданий для каждого разработчика: объясняем суть измеряемых конструктов, правила разработки стема и вариантов ответа, прорабатываем типичные ошибки при разработке заданий.
3. Каждый разработчик индивидуально разрабатывает 10 заданий закрытого типа, самостоятельно выбирая субконструкт и контекст, формулируя проблему стема и ответные опции.
4. После разработки заданий дизайнер отрисовывает картинки в тех заданиях, где они необходимы.

Экспертиза и качественная апробация разработанных заданий

После разработки заданий проходила тестологическая экспертиза заданий. Проверялись следующие параметры заданий:

- Наличие проблемы в стеме
- Наличие избыточных слов в стеме
- Верность ключа (правильного ответа)
- Наличие частично верных ответов
- Однородность вариантов ответа (отражение ими одного аспекта проблемы в стеме)
- Наличие принципа разработки вариантов ответа (например, сочетаемости и др.)

Соответствие разработанного задания измеряемому конструкту

После тестологической экспертизы были изменены формулировки 60% заданий (основная причина — слишком длинный стем), 10% заданий были удалены из банка как не проверяющие нужные конструкты.

После тестологической экспертизы заданий проводится качественная апробация заданий: проведение когнитивных лабораторий с помощью протокола «мысли вслух» и доработка формулировок заданий. Суть этого метода заключается в том, что респонденты проговаривают всё, о чем они думают в процессе чтения задания и во время и после его решения. Перед интервью респонденты получают инструкцию о том, как мыслить вслух. Цель этого этапа — установить, как понимаются задания потенциальными тестируемыми и какие когнитивные действия они совершают при решении тестовых заданий. Эти данные сопоставляются с данными о заданиях в спецификации теста. При обнаружении расхождений задания дорабатываются или удаляются из банка заданий совсем (если не подлежат исправлению).

Для каждого задания теста проводится 3–5 когнитивных интервью.

По результатам когнитивных интервью были внесены изменения в 35% заданий. Основная причина изменений — неясная формулировка, которая может быть понята по-разному, а также орфографические или пунктуационные ошибки.

Количественная апробация разработанных заданий 1.0

Первая количественная апробация заданий теста FORtitude проводилась с мая до сентябрь 2021 года. Апробировались 40 заданий из банка. В апробации участвовали 5 компаний: PWC, Трансгелеком, Ростелеком, Билайн, 3М. Общее количество респондентов составило 422 человека, при этом на каждое конкретное задание ответило от 250 до 422 человек. Апробация заданий

проводилась в условиях, приближенных к условиям тестирования: респонденты проходили задания в компьютерном формате, с таймером.

На основе результатов IRT моделирования мы проанализировали все тестовые задания и оценили их трудность и дискриминативность. По итогам этого анализа из банка заданий были удалены 6 заданий, которые имели экстремально низкую трудность и недостаточно высокую дискриминативность. Кроме того, на основе этого анализа были выделены задания с наиболее хорошими показателями трудности и дискриминативности, которые составили эквивалентный блок заданий в апробации 2.0.

Количественная апробация разработанных заданий 2.0

Вторая количественная апробация заданий теста FORtitude проводилась с ноября 2021 года до февраля 2022 года. Апробировались 70 заданий из банка.

Цель этой апробации заключалась в исследовании свидетельств валидности результатов теста FORtitude, поэтому основные результаты апробации будут представлены в разделе «Дизайн валидизационных исследований и важные свидетельства валидности».

На март 2022 в количественной апробации FORtitude приняли участие более 4000 человек, которые в общей сложности решали 70 заданий (35 — на вербальные способности, 35 — на числовые). 9,9% (403 человека) выборки были отфильтрованы при чистке данных: тренировочные прохождения и данные респондентов, которые потратили на выполнение теста менее 16 минут (это время требуется только на прочтение всех заданий). Всего в апробации приняли участие сотрудники и кандидаты 19 компаний.

Стоит отметить, что дизайн апробации отличается от дизайна теста: в апробационном исследовании задания предъявлялись тестируемым вариантами (по 36 заданий в каждом варианте, 18 — на вербальные, 18 — на числовые), в каждом варианте был эквивалентный блок заданий — одинаковый для всех вариантов. Эквивалентный блок заданий использовался для того, чтобы была возможность оценить параметры заданий из разных вариантов на одной шкале (для этого мы использовали метод параллельного выравнивания).

Психометрические характеристики заданий

Цель этого анализа состоит в том, чтобы оценить, насколько разнообразны задания теста FORtitude по трудности и дискриминативности и удалить из теста задания, которые функционируют плохо. Параметры заданий позволяют судить о том, насколько хорошо тестовые задания измеряют способности респондентов.

Для оценки параметров заданий мы использовали двухпараметрическую IRT модель Бирнбаума, в которой оцениваются параметры трудности и дискриминативности. Анализ проводился в среде R, пакеты ltm и mirt. На первом этапе анализа все задания, прошедшие апробацию, были объединены методом параллельного выравнивания для того, чтобы у нас была возможность установить общий 0 на шкале и таким образом положить все показатели трудности и дискриминативности на одну шкалу.

На втором этапе анализа все задания, участвовавшие в апробации, были оценены с помощью двухпараметрической IRT модели, чтобы выявить плохо функционирующиеся задания. По результатам психометрического анализа были удалены 22 задания. Часть из них не согласовалась с моделью (моделирование описано в разделе «Свидетельство: соответствие внутренней структуры теста концептуальной рамке (структурный аспект валидности») — 4 задания, другая часть демонстрировала DIF — 6 заданий, часть заданий демонстрировала локальную зависимость (3 задания), остальные задания имели слишком низкие (ниже 0,5 логита) показатели дискриминативности или экстремальные показатели трудности.

При нормальном распределении результатов 99% респондентов имеют способности, находящиеся от -3 до 3 логитов. Показатели трудности заданий должны колебаться в таких же пределах, чтобы оценивать способности максимально точно. Мы получили список хорошо функционирующих заданий, которые покрывают большой разброс способностей.

Структура конструктов и определения вербальных и числовых способностей и их компонентов основаны на анализе следующих источников:

1. Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. Journal of Occupational and Organizational psychology, 78(3), 387-409.
2. Carretta, T. R., & Ree, M. J. (2000). General and specific cognitive and psychomotor abilities in personnel selection: The prediction of training and job performance. International Journal of Selection and Assessment, 8(4), 227-236.
3. Dehaene, S. (1992). Varieties of numerical abilities. Cognition, 44(1-2), 1-42.
4. Fuhs, M. W., & Day, J. D. (2011). Verbal ability and executive functioning development in preschoolers at head start. Developmental psychology, 47(2), 404.
5. Grudnik, J. L., & Kranzler, J. H. (2001). Meta-analysis of the relationship between intelligence and inspection time. Intelligence, 29(6), 523-535
6. Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs.
7. Hunt, E. (1978). Mechanics of verbal ability. Psychological Review, 85(2), 109–130. doi:10.1037/0033-295x.85.2.109
8. Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological bulletin, 96(1), 72.
9. Salgado, J. F., & Moscoso, S. (2019). Meta-analysis of the validity of general mental ability for five performance criteria: Hunter and Hunter (1984) revisited. Frontiers in psychology, 10, 2227.
10. Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European Community meta-analysis. Personnel Psychology, 56(3), 573-605.
11. Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. Human performance, 15(1-2), 187-210.
12. Schmidt, F. L., & Hunter, J. E. (2000). Select on intelligence. Handbook of principles of organizational behavior, 3-14.
13. Аббакумов, Д. Ф. (2011). Сравнительный анализ эффективности числового и вербального тестов при прогнозировании результатов работы сотрудников. Организационная психология, 1(2).
14. Орел, Е. А. (2007). Диагностика особенностей мыслительной деятельности специалистов в области информационных технологий (программистов) (Doctoral dissertation, Московский государственный университет им. МВ Ломоносова (МГУ). Факультет психологии).