УДК 528.8.04, 528.88

Д. А. Федерякин
e-mail: dafederiakin@hse.ru
Национальный исследовательский университет «Высшая школа экономики»,
Москва, Россия

ВРЕМЯ ОТВЕТА В КОМПЬЮТЕРНОМ АДАПТИВНОМ ТЕСТИРОВАНИИ*

В данной работе изучаются возможности разработки компьютерных адаптивных тестов с использованием времени ответа на задания в качестве коллатеральной информации. Показано, что при включении в измерительную модель времени ответа происходит такой же прирост надежности, как и в случае линейного теста. Тем не менее наличие пропущенных ответов может привести к смещению оценок способности.

Ключевые слова: компьютерное адаптивное тестирование, коллатеральная информация, время ответа, современная теория тестирования.

Denis A. Federiakin
e-mail: dafederiakin@hse.ru
National Research University Higher School of Economics, Moscow, Russia

ITEM RESPONSE TIMES IN COMPUTERIZED ADAPTIVE TESTING

The paper describes the ways to develop a computerized adaptive test using item response times as collateral information. The paper shows that introducing item response times in the measurement model has the same effect on the reliability of computerized adaptive tests as on the reliability of linear tests. Nonetheless, the presence of missing responses may bias the estimates of the ability.

Keywords: computerized adaptive testing, collateral information, item response times, item response theory.

Психометрические исследования за последние 40 лет связаны в основном с разработкой современной теории тестирования (Item Response Theory, IRT) [23]. Эта теория является гибким статистическим подходом, предполагающим, что разделение ненаблюдаемых параметров респондентов и заданий позволяет оценить вероятность наблюдения правильного от-

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-14110 «Использование контекстной информации и информации из цифровой среды оценивания при измерении индивидуального прогресса учащихся начальной школы с помощью цифровых технологий».

вета индивидуально для каждого респондента и каждого задания. Однако для применения математико-статистического аппарата IRT основным требованием остается теоретическая проработка изучаемого конструкта [5]. Для того чтобы применение IRT-моделей было осмысленным, требуется разработка теоретической рамки инструмента измерения, которая способна детально определить как природу изучаемой черты респондентов, так и её поведенческие проявления – в ответах на какие задания проявляется конструкт. Теоретически обоснованную информацию о респондентах и заданиях называют целевой.

Дальнейшее развитие IRT привело к формированию направления исследований, связанного с компьютерным адаптивным тестированием (Computerized Adaptive Testing, CAT) [9]. Развитие этого направления основано на том, что респондент может ответить не на весь набор заданий, а лишь на какую-то его часть, тем не менее на основе его ответов можно получить оценки вероятностей решения остальных непредъявленных заданий. Более того, используя аппарат IRT, можно получить на одной шкале оценки способностей респондентов, которые отвечали на два не пересекающихся набора заданий.

Другим направлением развития IRT является изучение возможностей использования коллатеральной информации о респондентах, заданиях и их взаимодействии. Коллатеральной информацией называют любую информацию о заданиях, респондентах или их взаимодействии, включение которой в измерительную модель не меняет интерпретацию параметров, но которая позволяет уменьшить неопределенность в оценках этих параметров [24]. Чаще всего в рамках такой информации используют контекстную информацию о респондентах (например, социально-демографическую информацию) [4] или время ответа на задания [3]. Однако в качестве этой информации могут использоваться такие данные, как информация о повторных посещениях веб-страницы с заданием [1], передвижениях глаз респондента по экрану [14] и многое другое. Тем не менее, на сегодняшний день существует относительно мало исследований, которые синтезируют эти два направления развития IRT. Целью данной работы является разведывательный анализ возможностей разработки CAT с использованием времени ответа на задания в качестве коллатеральной информации.

Компьютерное адаптивное тестирование. Основной идеей CAT является постепенное (с каждым следующим предъявленным заданием) приближение оценки способности респондента к значению его истинной способности [11]. При этом начинается тестирование, как правило, с нескольких заданий, которые одинаковы для всех респондентов, часто – заданий средней трудности [7].

Следствием использования IRT-моделей для разработки CAT является глубокая интеграция в CAT идеи информации Фишера [2]. Для CAT

важно предъявлять не просто случайный набор заданий, а именно те задания, которые дают наибольшее количество информации о текущей оценке способности респондента. Соответственно, оптимизация предъявления заданий происходит за счет того, что конкретному респонденту не предъявляются слишком трудные *для него* задания (которые он, скорее всего, решит неправильно) и слишком легкие *для него* задания (которые он, скорее всего, решит правильно). Важной особенностью выбора следующего задания является частая необходимость соблюдения баланса содержания теста [25]. Большие банки заданий часто являются неодномерными – они состоят нескольких из тематических секций измеряемого конструкта, между которыми требуется соблюсти баланс присутствия в предъявляемом наборе заданий.

Другим ключевым вопросом для CAT является правило остановки [18]. Существуют три наиболее популярных (но не исчерпывающих) способа определения, когда следует останавливать адаптивное тестирование [8]: достижение определенной точности оценки способности/классификации; предъявление какого-либо числа заданий (направлен на то, чтобы дать одинаковые возможности проявить свои способности всем респондентам); исчерпывание банка заданий для конкретного респондента (когда заданий, дающих много информации о способности респондента не остается в банке).

Можно перечислить пять фундаментальных вопросов, на которые пытаются ответить исследователи, занимающиеся разработкой CAT:

1) с каких (ого) задний (я) начать тестирование? (правило начала тестирования, Starting Rule);

2) какие (ое) задания (е) предъявлять следующими? (правило выбора следующего задания, Next Item Rule);

3) как начислять баллы за задания? Как переоценивать способность после наблюдения новых ответов? (правило начисления баллов, Scoring Rule);

4) когда закончить тестирование? (правило окончания тестирования, Stopping Rule);

5) как защитить банк заданий от утечки в открытый доступ?

В то время как первые четыре вопроса являются необходимым ядром CAT, пятый вопрос набирает всё большую популярность в последние годы. Это направление исследований фокусируется на развитии статистических индексов, оценивающих «сверхпредъявление» заданий (Item Overexposure) [19]. Главной целью этого направления является обеспечение равномерного предъявления всех заданий в целях избегания ситуации, когда какое-то задание предъявляется почти всем респондентам.

Время ответа на задания как коллатеральная информация. Исследователи в психологии уже долгое время интересуются вопросами времени реакции на стимулы. В своей книге на тему измерения интеллекта

Торндайк с соавторами писал, что «при прочих равных, если интеллект А может сделать на каждом уровне (трудности) сколько же заданий, сколько и интеллект Б, но в меньшее время, интеллект А выше» [20, с. 33]. Такой подход привел к развитию учения о компромиссе между скоростью и точностью (Speed-Accuracy Trade-off, SAT) [17]. Этот подход к изучению времени реакции подразумевает, что у каждого респондента есть своя функция наибольшей продуктивности от времени. При уменьшении времени, доступного на решение задачи, снижается точность её решения.

Прямой реализацией идеи SAT в психометрике является иерархический байесовский фреймворк [22]. Этот фреймворк подразумевает, что параметры респондентов (способность и быстрота) и параметры заданий (минимум – трудность и трудозатратность) следуют своим многомерным латентным распределениям. В этом случае целью IRT модели является описание параметров этих распределений. Этот фреймворк является концептуальной основой для включения в IRT-модели любой другой информации о взаимодействии респондентов и заданий.

Позднее было показано, как этот фреймворк анализа времени ответа может быть реализован в терминах фриквентистского факторного анализа [12]. Для этого требуется специальная репараметризация модели, основанная на сходствах IRT-моделей и факторного анализа. Это приводит к формулированию B-GLIRT (Bivariate Generalized Linear IRT) фреймворка для анализа времени и точности ответов [13]. Показано, что такая репараметризация позволяет описать множество других моделей IRT для времени реакции как частные случаи этого фреймворка. При этом именно этот фреймворк ориентирован на использование времени в качестве коллатеральной информации и направлен на измерение способности в оригинальном IRT-понимании. Таким образом, в этом исследовании используется именно этот фреймворк.

Симуляционное исследование. Для того чтобы изучить эффекты включения времени ответа согласно B-GLIRT фреймворку в CAT, в симуляционном примере используется дизайн инструмента START-PROGRSS, направленного на измерение базовой математической грамотности для первоклассников. Тест состоит из 7 тематических блоков, каждый из которых содержит от 5 до 18 дихотомических заданий (всего 57 заданий). Результаты тестирования обрабатываются с использованием Раш-моделирования. Благодаря этому задания образуют однозначную иерархию по трудности. Тестирование основано на постепенном увеличении трудности заданий – от самых простых к самым трудным, соответственно, порядок предъявления заданий одинаков для всех респондентов. Однако внутри тематического блока используется правило остановки – если респондент совершил 3 ошибки подряд или 4 ошибки всего, ему предъявляется первое задание следующего блока.

Как и любая модель из B-GLIRT фреймворка, модель, которая используется в этой работе, имеет три компонента: субмодель для правильности ответов, субмодель для времени ответов, функция их отношений (cross-relation function). В данной работе для описания правильности ответов используется Раш-модель, для описания быстроты ответов используется факторно-аналитическая субмодель без ограничений. Благодаря этому используемая модель способна описывать как вопросы, демонстрирующие положительную корреляцию между временем, затраченным на решение и правильностью ответа (вопросы на точность), так и вопросы, демонстрирующие отрицательную корреляцию между ними (вопросы на скорость). Однако описанная модель накладывает ограничение на корреляцию между параметрами быстроты и способности респондентов – она должна быть зафиксирована равной нулю. Несмотря на то, что это теоретически нереалистичное допущение, подобная модель способна предоставить расширенную информацию на уровне отдельных заданий, а смещения в оценках способности, которые вызваны таким допущением, незначительны [6].

В рамках имитационного моделирования симулируются ответы на вопросы и наблюдаемое время ответа, аналогичные реальным данным. Объем выборки составил 2 000 человек. После оценки параметров модели проводится сравнение двух статистик, оцененных с использованием информации о времени ответа и без нее: корреляция оценённой способности и той, которая использовалась для симуляции, а также надежность оценок способности. Симуляция была реплицирована 50 раз.

Весь анализ проводился с помощью пакетов TAM v. 3.5-19 [15] и lavaan v. 0.6-7 [16] для программного обеспечения R v. 3.6.2. Модели IRT без учета времени ответа были оценены методом максимального маргинального правдоподобия, модели со временем ответа были оценены с использованием диагонально-взвешенного метода наименьших квадратов. Надежность оценивалась с помощью коэффициента ω для факторно-аналитических моделей со скоррелированными ошибками [10]. Латентные переменные с учетом времени были оценены с помощью метода регрессии (Regression Factor Scores) [21]. Результаты анализа симуляций приведены в таблице.

Таблица

Результаты симуляционного исследования

Параметр	Среднее	SD
Надежность EAP-оценок способности без времени	0.892	0.014
Корреляция EAP-оценок способности без времени с истинными значениями	0.951	0.023
Надежность оценок способности со временем	0.999	0.001
Корреляция оценок способности со временем с истинными значениями	0.949	0.008

Из приведенных результатов видно, что включение времени как коллатеральной информации в измерительную модель приводит к увеличению надежности. Небольшой прирост надежности в абсолютных значениях объясняется использованием достаточно длинного компьютерного адаптивного теста, который сам по себе (без времени) обладает высокой надежностью. На тестах меньшей длины наблюдается больший прирост надежности оценок способности [6]. Однако также результаты показывают, что корреляция оценок способности с истинными значениями несколько снижается при введении времени в модель. Подобный результат объясняется наличием большого количество пропущенных ответов, которые взаимодействуют с методом вычисления значений латентной переменной. На линейных тестах этого эффекта не наблюдается [6].

Использование коллатеральной информации в компьютерных адаптивных тестах до сих пор остается достаточно неразработанной темой. Это связано как со сложностью используемого моделирования, так и с дороговизной разработки подобных продуктов. Тем не менее эта тема обладает большим психометрическим потенциалом и способна привести к разработке высоко оптимальных тестовых инструментов.

В данном докладе была проиллюстрирована возможность использования времени ответа в качестве коллатеральной информации для компьютерного адаптивного тестирования. С применением имитационного моделирования показано, что включение этой информации в измерительную модель позволяет повысить надежность измерения. Тем не менее это может снизить корреляцию оценок латентной переменной с её истинными значениями в ситуации наличия большого количества пропущенных ответов. Разработка других методов вычисления значений латентных переменных, которые были бы робастны в присутствии пропущенных ответов, является многообещающим направлением исследований для развития CAT с коллатеральной информацией.

Список литературы

1. Bezirhan U., von Davier M., Grabovsky I. (2020). Multiple group hierarchical speed accuracy revisit model. Paper presented at the International Meeting of Psychometric Society – 2020.

2. Chang H. H., Ying Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20(3), 213–229.

3. de Boeck P., Jeon M. (2019). An overview of models for response times and processes in cognitive tests. Frontiers in psychology, 10, 102.

4. de Boeck P., Wilson M. (2004). Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. NY: Springer-Verlag.

5. Embretson S. E. (1996). The new rules of measurement. Psychological assessment, 8(4), 341.

6. Federiakin D. (2020). Rasch model with time parameters for tests with speediness-items and achievement-items. Paper presented at the International Meeting of Psychometric Society – 2020.

7. Lunz M. E., Bergstrom B. A. (1994). An empirical study of computerized adaptive test administration conditions. Journal of Educational Measurement, 31(3), 251–263.

8. Magis D., Barrada J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. Journal of Statistical Software, 76(1), 1–19.

9. Magis D., Yan D., Von Davier A. A. (2017). Computerized adaptive and multistage testing with R: Using packages catr and mstr. Springer.

10. McDonald R. P. (1999). Test theory: A unified approach.

11. Meijer R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction.

12. Molenaar D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. British Journal of Mathematical and Statistical Psychology, 68(2), 197–219.

13. Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. Multivariate Behavioral Research, 50(1), 56–74.

14. Moustaki I. (2020). Keynote: Psychometric analysis of eye movements during search and choice. Paper presented at the International Meeting of Psychometric Society - 2020.

15. Robitzsch A., Kiefer T., Wu M. (2020). Package 'TAM'. Test Analysis Modules – Version: 3.5-19.

16. Rosseel Y., Jorgensen T. D., Rockwood N., Oberski D., Byrnes J., Vanbrabant L., Savalei V., Hallquist M., Rhemtulla M., Katsikatsou M., Barendse M., Scharf S. (2020). Package 'lavaan'. Latent Variable Analysis. – Version: 0.6-7.

17. Schouten J. F., Bekker J. A. M. (1967). Reaction time and accuracy. Acta psychologica, 27, 143–153.

18. Stafford R. E., Runyon C. R., Casabianca J. M., Dodd B. G. (2019). Comparing computer adaptive testing stopping rules under the generalized partial-credit model. Behavior research methods, 51(3), 1305–1320.

19. Stocking M. L., Lewis C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. Journal of Educational and Behavioral Statistics, 23(1), 57–75.

20. Thorndike E. L., Bregman E. O., Cobb M. V., Woodyard E. (1926). The measurement of intelligence. New York, NY: Teachers College Bureau of Publications.

21. Thurstone L. L. (1935). The vectors of mind: Multiple-factor analysis for the isolation of primary traits.

22. van der Linden W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. Psychometrika, 72(3), 287.

23. van der Linden W. J. (Ed.). (2016). Handbook of Item Response Theory: Volume 1: Models. CRC Press.

24. Wang W., Chen P., Cheng Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. Psychological Methods, 9(1), 116-136.

25. Yi Q., Chang H. H. (2003). A-stratified cat design with content blocking. British Journal of Mathematical and Statistical Psychology, 56(2), 359–378.