УДК 528.8.04, 528.88

Д. А. Федерякин[1], Е. Ю. Карданова[2]

[1]e-mail: dafederiakin@hse.ru; [2]e-mail: ekardanova@hse.ru
Национальный исследовательский университет «Высшая школа экономики»,
Москва, Россия

ПСИХОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ СЛОЖНЫХ КОНСТРУКТОРОВ*

В этой работе описывается перечень исследований, которые необходимо провести для обоснования возможности одновременного использования как общего балла по тесту, так и баллов по субшкалам при измерении сложных конструктов. Эти исследования проиллюстрированы на примере компьютерного адаптивного инструмента PROGRESS-ML, измеряющего базовую математическую грамотность в 3-м классе.

Ключевые слова: сложные конструкты, композитные измерительные инструменты, многомерные Раш-модели, PROGRESS-ML.

Denis A. Federiakin[1], Elena Yu. Kardanova[2]

[1]e-mail: dafederiakin@hse.ru; [2]e-mail: ekardanova@hse.ru
National Research University Higher School of Economics, Moscow, Russia

PSYCHOMETRIC MODELLING OF COMPOSITE CONSTRUCTS

This paper describes the bunch of research necessary for the justified simultaneous use of both general test score and specific scores. We also provide an exemplary study using computerized adaptive test PROGRESS-ML which measures basic mathematical literacy in the third grade.

Keywords: complex constructs, composite measurement instruments, multidimensional Rasch-models, PROGRESS-ML.

В современном образовании и, шире, в социальных науках, наблюдается растущая потребность в композитных инструментах измерения – инструментах, имеющих сложную структуру, например, состоящих из субшкал, которые определенным образом вкладываются в итоговый тестовый балл. Это может являться следствием тренда на популярность измере-

* 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».

ния сложных конструктов, таких как, например, навыки XXI в., или новые компетенции. Такие конструкты состоят из различных компонентов и их сложно представить в виде классической одномерной однокомпонентной характеристики респондентов. В то же время, для практиков и лиц, принимающих решения на основе результатов измерений, является ценной информация как об уровне развития целостной характеристики, так и её составных частей. Это позволяет учитывать развитые и развивающиеся черты или способности респондентов, улучшая работу, например, системы образования или психологическую практику.

В стандартах образовательного и психологического тестирования [2] отмечается, что: 1) тестовые баллы не должны докладываться пользователям до тех пор, пока их валидность, сопоставимость и надежность не были установлены; 2) если инструмент выдает более, чем один тестовый балл, психометрическое качество всех докладываемых тестовых баллов должно быть установлено. Это важно, поскольку неточная информация на уровне общего тестового балла может привести к решениям с нежелательными социальными последствиями, а неточная информация на уровне балла по субшкалам – к неправильным решениям по исправлению или улучшению ситуации [18]. В академической среде использование низкокачественных баллов по субшкалам может привести к неверным выводам о природе изучаемого явления.

Психометрика композитных инструментов. На языке психометрики композитные инструменты являются многомерными, и задача состоит в том, чтобы оценить, по возможности, как общую способность респондентов, так и отдельные способности – ее составляющие. Психометрическое моделирование таких инструментов состоит из нескольких этапов. В первую очередь необходимо проверить, является ли тест существенно одномерным. Если да, то мы можем сообщать общий балл по тесту – разумеется, при условии, что будет доказана его валидность и психометрическая состоятельность. Если тест не является одномерным, то необходимо использовать многомерные модели, и в этом случае возможность сообщения общего балла требует дополнительного исследования с использованием иерархических моделей [17]. Особой популярностью пользуются два класса иерархических моделей – бифакторные модели [14] и модели с факторами высокого порядка [7]. Несмотря на алгебраические сходства [15] и то, что обе группы моделей предполагают использование тестового балла по всему тесту, их интерпретация сильно отличается [4, 11]. В то время как модели с факторами высокого порядка оценивают общий фактор, который проявляется в заданиях через баллы по субшкалам, бифакторные модели предполагают полное разделение эффектов общего фактора и баллов по субшкалам.

В случае намерения сообщать дополнительно баллы по отдельным субшкалам (например, когнитивным операциям или содержательным об-

ластям), можно применить несколько подходов. Первый состоит в том, чтобы одномерную модель применять к каждой субшкале отдельно [6]. Этот подход наименее привлекателен, так как число заданий в каждой субшкале, как правило, невелико, и, соответственно, надежность измерения будет недостаточно высокой, а ошибка измерения слишком велика. Это приводит к тому, что баллы по отдельным субшкалам сообщать будет неправомерно [2]. Второй подход предполагает использование бифакторных моделей. Эти модели, гипотетически, допускают одновременное сообщение общего балла с баллами по субшкалам в качестве дополнительной независимой информации. Однако, как показывают исследования, баллы по субшкалам, оцененные в рамках бифакторных моделей, редко обладают удовлетворительной надежностью, потому что они описывают информацию, которая не была извлечена с помощью балла по всему тесту, а она часто подавлена случайными шумами [8].

Третий подход предполагает применение не-компенсаторных многомерных моделей [13]. Такие модели, фактически, являются несколькими одномерными моделями, записанными в одном уравнении правдоподобия. Каждая латентная характеристика переменных вычисляется на основе ответов респондентов только на соответствующие задания и с учетом оцененных корреляций между самими латентными переменными. Таким образом, многомерные модели используют информацию о каждой размерности и моделируют вероятность выполнения заданий как функцию не одной латентной переменной, а нескольких, с учетом связей между ними. В итоге надежность измерений будет выше, чем при первом подходе, и более вероятно, что можно будет сообщать баллы по отдельным субшкалам. Применение этих моделей может рассматриваться в контексте анализа коллатеральной информации – любой информации об инструменте, респондентах или их взаимодействии, включение которой в измерительную модель не меняет интерпретацию параметров, но которая позволяет уменьшить неопределенность в оценках этих параметров [20]. В этом случае для каждой из субшкал ответы по всем остальным субшкалам (вместе с матрицей корреляций ненаблюдаемых размерностей) являются коллатеральной информацией [23]. Это тот подход, который мы используем в данной работе.

Таким образом, для использования результатов композитных инструментов измерения необходимо проведение психометрических исследований с целью принятия решения о том, являются ли результаты по всему тесту и его субшкалам надежными и могут ли они сообщаться пользователям. В этой работе различные подходы к моделированию композитных инструментов демонстрируются на примере теста базовой математической грамотности PROGRESS-ML.

Тест математической грамотности PROGRESS-ML. Тест PROGRESS-ML оценивает, насколько хорошо учащийся ориентируется в мате-

матике после прохождения двух лет обучения в школе. При разработке теста мы опирались на следующее определение математической базовой грамотности [24]: «Базовая математическая грамотность (включая работу с данными) – способность применять математические инструменты, аргументацию, моделирование в повседневной жизни, в том числе в цифровой среде».

Тест базовой математической грамотности PROGRESS-ML состоит из 30 заданий, оцениваемых дихотомически. Оценивание проводится в формате компьютерного адаптивного тестирования.

Содержание теста отбиралось таким образом, чтобы, с одной стороны, отвечать определению базовой математической грамотности, а с другой стороны, учитывать содержание программы начального общего образования. В результате было выделено пять тематических областей: Пространственные представления, Измерения величин, Закономерности, Моделирование, Работа с информацией. Задания в тесте сгруппированы в блоки в соответствии с тематической областью.

Дополнительно, тест PROGRESS-ML оценивает когнитивные процессы учащихся, необходимые для выполнения заданий. При разработке теста использовалась теоретическая рамка международного исследования TIMSS для 4-го класса [12], в котором, помимо оценки владения предметным содержанием, измеряются три группы когнитивных операций – знание, применение, рассуждения.

Таким образом, тест PROGRESS-ML является композитным инструментом: он включает пять тематических областей и отражает три группы когнитивных операций. Предполагается, что по итогам тестирования будет сообщаться общий тестовый балл учащегося (в данном случае уровень его базовой математической грамотности), а также баллы по субшкалам (в данном случае это могут быть тематические области и/или когнитивные операции).

Методология. Выборку составили 6 078 учеников 3-го класса двух регионов РФ. Относительно регионов выборки были репрезентативными. Средний возраст – 9,06 лет (SD = 0.46), количество девочек = 52,36 %.

Анализ проводился в рамках моделей Раша из парадигмы современной теории тестирования (IRT). Для проверки гипотезы о возможности сообщения общего балла по тесту мы использовали одномерную модель [22]. Мы исследовали размерность теста с помощью метода главных компонент, примененного к стандартизированным модельным остаткам [10, 19]. Для проверки возможности сообщения баллов по отдельным субшкалам мы использовали многомерные модели без кросс-нагрузок (between-item multidimensional models; [1]). Все использованные модели могут быть рассмотрены как частные случаи Многомерной мультиномиальной логит-модели смешанных эффектов (Multidimensional Random Coefficients Multi-

nomial Logit Model [1]). Для оценки всех моделей использовался квази-Монте-Карло алгоритм, внедренный в пакет TAM v. 3.5-19 [16] для программного обеспечения R v. 3.6.2.

Результаты. В результате анализа стандартизированных остатков методом главных компонент было обнаружено, что собственное значение первого компонента равняется 1.45, что соответствует 4,2 % дисперсии остатков. Собственные значения следующих четырех компонент находятся в промежутке (1.15, 1.2), и распределение объясненной дисперсии остатков среди компонент практически равномерное – около 4 % на компоненту. Следовательно, одномерная модель в достаточной мере описывает распределение вероятностей ответа, и тест можно считать одномерным.

Модельная Expected-a-Posteriori надежность [3] всего тестового балла из одномерной модели составила 0.76. Для сравнения мы вычислили надежность с помощью методов классической теории тестирования (КТТ): Greatest Lower Bound (GLB [9] надежность составила 0.86, индекс α Кронбаха [5] составил 0.81. Однако важно заметить, что дизайн тестирования (компьютерное адаптивное) подразумевает, что не все респонденты выполняли все задания, а параметры КТТ становятся нестабильны в условиях наличия пропущенных ответов. Поэтому даже несмотря на то, что в нашем примере надежность, оцененная в рамках КТТ (GLB и α Кронбаха), несколько превосходит надежность баллов, оценённых в IRT, этим индексам не следует доверять.

Таким образом, результаты анализа предполагают, что тест может рассматриваться как одномерный, даже несмотря на различные способы группирования заданий. Это подразумевает, что по результатам тестирования представляется возможным докладывать один общий тестовый балл математической грамотности, который будет обладать хорошими психометрическими характеристиками.

Результаты анализа надежности согласно группировке по содержательным областям приведены в табл. 1, по когнитивным операциям – в табл. 2.

Из анализа таблиц можно заключить, что все размерности обладают надежностью, достаточной для мониторингового использования инструмента. Несмотря на малое количество заданий (такое малое количество заданий, фактически, делает сырые тестовые баллы по отдельным размерностям неиспользуемыми), это возможно благодаря используемому подходу к моделированию инструмента. Дополнительно были оценены корреляции между размерностями: все тематические области коррелируют друг с другом приблизительно одинаково – на уровне 0.8–0.9, то же самое справедливо для когнитивных операций. Это может рассматриваться как дополнительный аргумент в пользу одномерности теста даже несмотря на то, что многомерные модели статистически лучше подходят данным, чем одномерная.

Таблица 1

Надежности для модели с тематическими областями

Размерность (Тематическая область)	Пространственные представления	Измерения	Закономерности	Моделирование	Работа с информацией
Надежность	0.68	0.71	0.67	0.68	0.63
Число заданий	7	6	6	6	5

Таблица 2

Надежности для модели с группами когнитивных операций

Размерность (Группа когнитивных операций)	Знание	Применение	Рассуждение
Надёжность	0.75	0.74	0.61
Число заданий	12	14	4

Таким образом, применение многомерных моделей современной теории тестирования позволило получить достаточно надежные баллы по отдельным субшкалам (тематическим областям и когнитивным операциям) и тем самым сделало возможным их сообщение пользователям. Более того, приведенные оценки надежности приведены из IRT моделей, в которые не были введены никакие контекстные переменные, а введение этих переменных в модель приведет к вычислению еще более надежных баллов по субшкалам.

Современная литература отмечает всё более растущую популярность композитных инструментов измерения, которые призваны выдавать как единый тестовый балл, так и баллы по субшкалам. Одной из стратегий использования результатов по подобным инструментам может являться использование сырых тестовых баллов [21]. Однако в большинстве случаев сырые тестовые баллы невозможно использовать, в частности, из-за низкой их надежности [8].

В этой работе на примере теста базовой математической грамотности PROGRESS-ML продемонстрировано, что использование моделей современной теории тестирования позволяет проверить гипотезы о возможности сообщения общего балла респондента по тесту и баллов по отдельным субтестам. Главным результатом тестирования является общий балл респондента. Однако повторные рекалибровки данных с учетом различных тематических областей и разных групп когнитивных операций, требуемых для решения заданий, позволяют докладывать и баллы по субтестам. Сутью такого использования результатов является декомпозиция общего тестового балла на компоненты, которые его составляют.

Список литературы

1. Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. Applied psychological measurement, 21(1), 1–23.

2. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 2014.

3. Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied psychological measurement, 6(4), 431–444.

4. Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. Journal of personality, 80(4), 796–846.

5. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297–334.

6. Davey, T. Hirsch, T.M. (1991). Concurrent and Consecutive estimates of examinee ability profiles. Paper presented at the Annual Meeting of the Psychometric Society, New Brunswick, NJ.

7. Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor?. Psychology Science, 50(1), 21.

8. Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. Psychometrika, 75(2), 209–227.

9. Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. Psychometrika, 42(4), 567–578.

10. Linacre, J. M. (1998). Structure in Rasch residuals: why principal components analysis. Rasch measurement transactions, 12(2), 636.

11. Mansolf, M., & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. Intelligence, 61, 120–129.

12. Mullis, I. V., & Martin, M. O. (2017). TIMSS 2019 Assessment Frameworks. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

13. Reckase, M. D. (2009). Multidimensional item response theory models. In Multidimensional item response theory (pp. 79-112). Springer, New York, NY.

14. Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate behavioral research, 47(5), 667–696.

15. Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. Journal of Educational Measurement, 47(3), 361–372.

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

17. Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. Psychometrika, 22(1), 53–61.

18. Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. Educational Measurement: Issues and Practice, 30(3), 29–40.

19. Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement, 3, 205–231.

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

21. Wilson, M., & Gochyyev, P. (2020). Having your cake and eating it too: Multiple dimensions and a composite. Measurement, 151, 107247.

22. Wright, B. D., & Stone, M. H. (1979). Best test design.

23. Wu, M., Tam, H. P., & Jen, T. H. (2016). Multidimensional IRT Models in Book: Educational measurement for applied researchers. Theory into practice.

24. Фрумин, И. Д., Добрякова, М. С., Баранников, К. А., & Реморенко, И. М. (2018). Универсальные компетентности и новая грамотность: чему учить сегодня для успеха завтра. Предварительные выводы международного доклада о тенденциях трансформации школьного образования (2(19); Современная Аналитика Образования). https://ioe.hse.ru/data/2018/07/25/1152380855/CAO_2(19)_электронная_версия.pdf.