130 лет психометрике:
от науки о тестах знаний к вычислительной науке о поведении

Дмитрий Аббакумов, *PhD*
ВШЭ и KU Leuven

Аннотация

Психометрика, наука о психологических и педагогических измерениях, отметила 130-летие. За это время из науки, основным подходом которой был подсчет правильных ответов на тестовые задания, психометрика превратилась в сложную междисциплинарную область знания на стыке психологии, педагогики, математической статистики, науки о данных и машинного обучения. На лекции мы познакомимся с историей развития психометрики, задачами, которые решала и решает эта наука, ее достижениями последних лет и перспективными областями, которыми будут заниматься будущие ученые.

Что такое психометрика?

Psychometrics is a scientific discipline concerned with the construction of assessment tools, measurement instruments, and formalized models that may serve to connect observable phenomena to theoretical attributes.

prof. dr. Denny Borsboom, UvA

Конструкт	**Наблюдения**
Знание иностранного языка	Количество ошибок в эссе
Числовой интеллект	Скорость решения задачи
Знание ПДД	Количество правильных ответов в тесте

Почему 130 лет?

'. . . it is intelligible to speak of the mean judgment of competent critics as the true judgment; and deviations from that mean as errors'

(Edgeworth, 1888, p. 622)

Francis I. Edgeworth (1845–1926)

Почему 130 лет?

James McKeen Cattell (1860–1944) и его психометрическая группа в Кембридже (1888/89)

- Научный метод в психологии
- Измерения интеллекта
- Президент American Psychological Association в 1895 г.
- Cattell, J. M. (1890). Mental tests and measurements. *Mind*, *15*, 373-381

- Measurement of Attitudes, 1929
- Primary Mental Abilities, 1938
- Factorial Studies of Intelligence, 1941
- Multiple Factor Analysis, 1947
- Основатель и первый президент **Психометрического общества** (1935/36)

Louis L. Thurstone (1887–1955)

SCIENCE

VOL. 85 FRIDAY, MARCH 5, 1937 No. 2201

Psychology as a Quantitative Rational Science: PROFESSOR L. L. THURSTONE 227

Obituary:
Edward Curtis Franklin: PROFESSOR C. A. KRAUS. *Recent Deaths and Memorials* 232

Scientific Events:
The Columbia University School of Medicine; The Electronics Institute at the University of Michigan; The Wisconsin Alumni Research Foundation; Meeting of Texas Geologists; The Foster Lecture Foundation at the University of Buffalo; The Installation of a Chapter of Sigma Xi at the George Washington University 234

Scientific Notes and News 237

Discussion:
Dissemination of Scientific Literature by Means of Microfilms: DR. ATHERTON SEIDELL. *Micro-photography or Photo-micrography:* PROFESSOR W. J. LUYTEN. *A Vasopressor Local Anesthetic:* PROFESSOR CHAUNCEY D. LEAKE. *The Action of Eserine and its Analogues on Skeleton Muscle:* Z. M. BACQ and G. L. BROWN 240

Scientific Books:
Recent Physics: PROFESSOR HENRY CREW. *Nuclear Physics:* PROFESSOR ALLAN C. G. MITCHELL 243

Special Articles:
Vitamin B_1 and the Growth of Excised Tomato Roots: DR. WILLIAM J. ROBBINS and MARY A. BARTLEY. *The Significance of the Adrenals for Adaptation:* PROFESSOR HANS SELYE. *The Synthesis of Substances Related to Lysergic Acid:* DR. WALTER A. JACOBS and DR. R. GORDON GOULD, JR. 246

Scientific Apparatus and Laboratory Methods:
An Automatic Dehydrating Device: JEAN PENNINGTON and DR. C. P. HICKMAN. *The Isolation of Muscle Nuclei:* GERMAIN CROSSMON 249

Science News 14

SCIENCE: A Weekly Journal devoted to the Advancement of Science, edited by J. MCKEEN CATTELL and published every Friday by

THE SCIENCE PRESS

New York City: Grand Central Terminal
Lancaster, Pa. Garrison, N. Y.

Annual Subscription, $6.00 Single Copies, 15 Cts.

SCIENCE is the official organ of the American Association for the Advancement of Science. Information regarding membership in the Association may be secured from the office of the permanent secretary, in the Smithsonian Institution Building, Washington, D. C.

PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE[1]

By Professor L. L. THURSTONE
THE UNIVERSITY OF CHICAGO

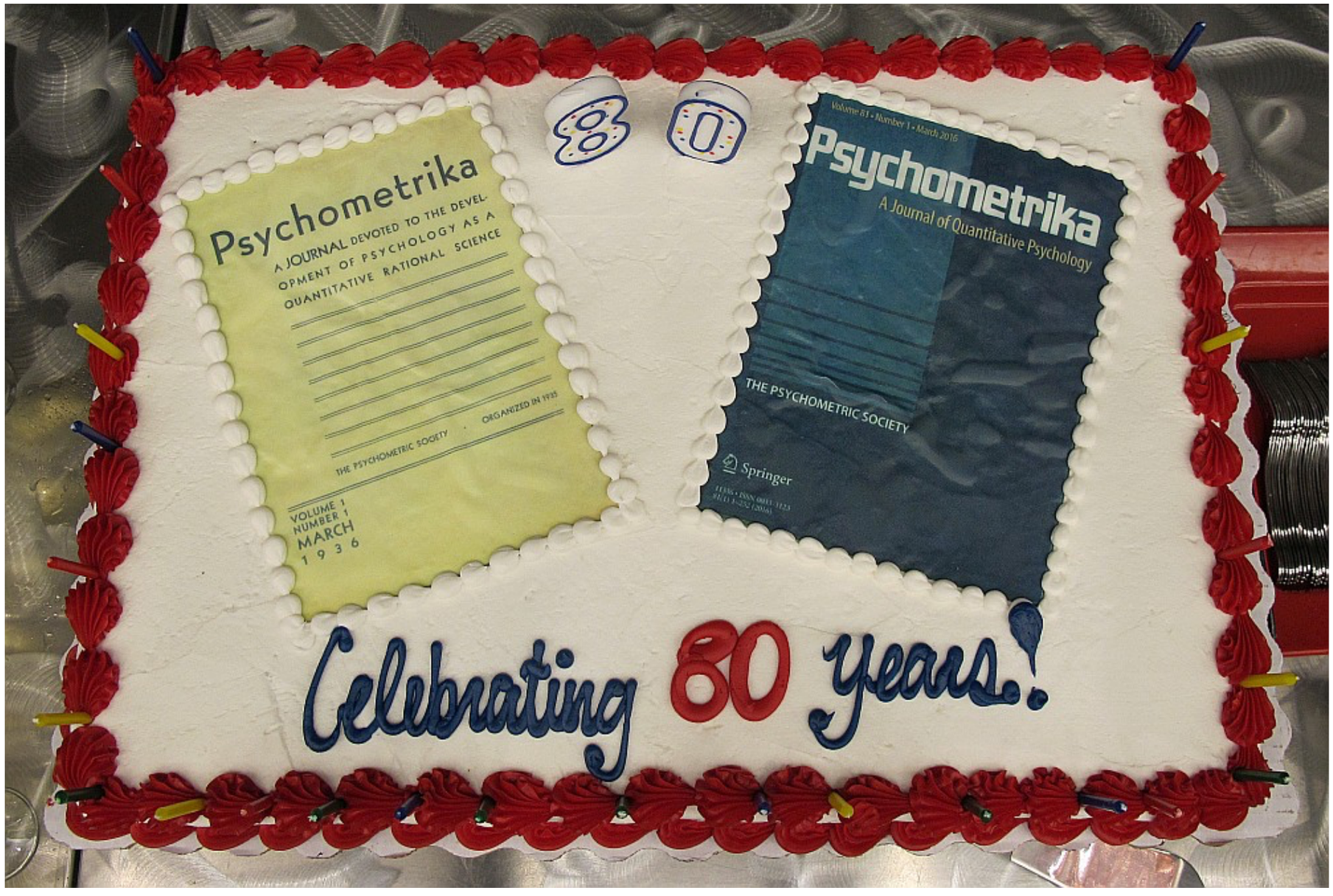
THE purposes of this society are not new, but they represent an emphasis and direction which have not hitherto received major consideration in psychological science. It seems proper that we should devote some share of our first program meeting to a consideration of our main objectives.

Our main purpose is briefly stated in the subtitle of the new journal, *Psychometrika*, namely, to encourage the development of psychology as a quantitative rational science. More briefly, this may be called mathematical psychology. We should justify our emphasis upon quantification and upon rationalization in science, as well as our conception of the fundamental nature of science.

I assume that we are in complete agreement that we can not suddenly quantify our comprehension of psychological phenomena over their entire range. As psychologists, we are as interested as ever in making exploratory studies of new psychological effects and in discovering hitherto unknown effects. At present, the range of psychological phenomena that can be profitably reduced to mathematical formulation is limited, and it is likely that every man who works on a problem of mathematical psychology will also concern himself with exploratory studies of other problems that are as yet too new for detailed rationalization.

After the discovery of a psychological effect, we naturally turn to the second phase of scientific inquiry, namely, to relate the new effect in a simple descriptive manner to what is already known. In this stage theories are devised to explain the experimentally known effects, and we try, of course, to make psychological theories less complicated than the effects that are to be explained. In this phase the descriptions of psycho-

[1] Abstract of address by the retiring president of the Psychometric Society at Hanover, N. H., September 4, 1936.


Psychometrika
A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE
THE PSYCHOMETRIC SOCIETY
ORGANIZED IN 1935
VOLUME 1
NUMBER 1
MARCH
1936
Volume 81 • Number 1 • March 2016
Psychometrika
A Journal of Quantitative Psychology
THE PSYCHOMETRIC SOCIETY
Springer
80
Celebrating 80 years!

Dorothy C. Adkins (1912–1975)
Президент Психометрического общества (1949/50)

- Adkins, D. C. (1947). Construction and analysis of achievement tests.

- Adkins, D. C. (1958). **Measurement in relation to the educational process**. *Educational and Psychological Measurement*, *18*, 221-240.

- Flanagan, J. C., Adkins, D., & Cadwell, D. H. B. (1950). Major developments in examining methods.

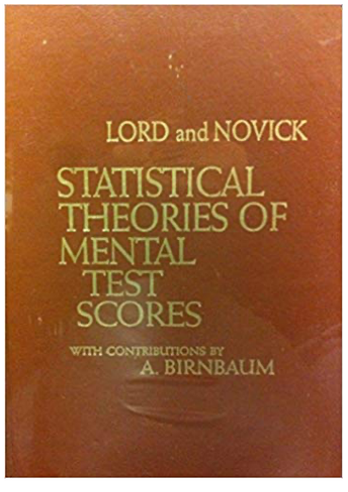
Классическая психометрическая теория

$$X = T + E$$

X – наблюдаемый балл
T – истинная подготовленность
E – ошибка измерения

Melvin R. Novick (1932–1986)
През. ПО (1979/80)

Frederik M. Lord (1912–2000)
През. ПО (1958/59)



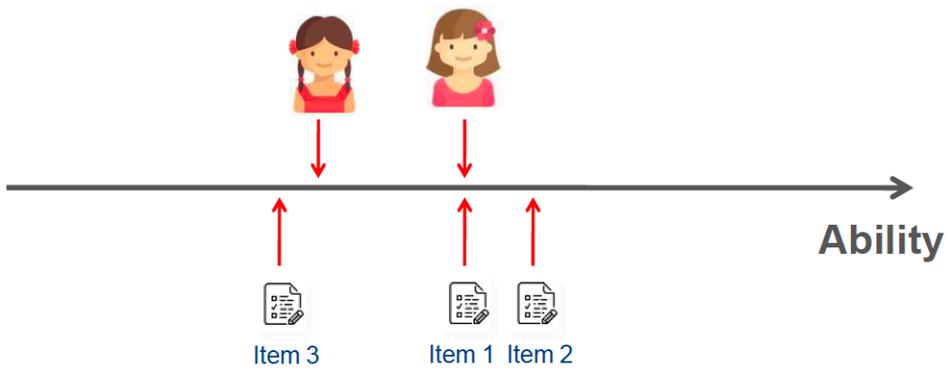
Lord & Novick (1968)

Frederik M. Lord (1912–2000)
Item Response Theory

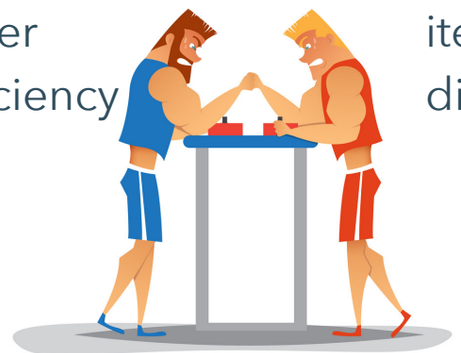
Georg Rasch (1901–1980)
The Rasch Model

Современная психометрическая теория

Вероятность правильного ответа на задание описывается функцией разности уровня подготовленности студента и уровня трудности задания



learner proficiency



item difficulty

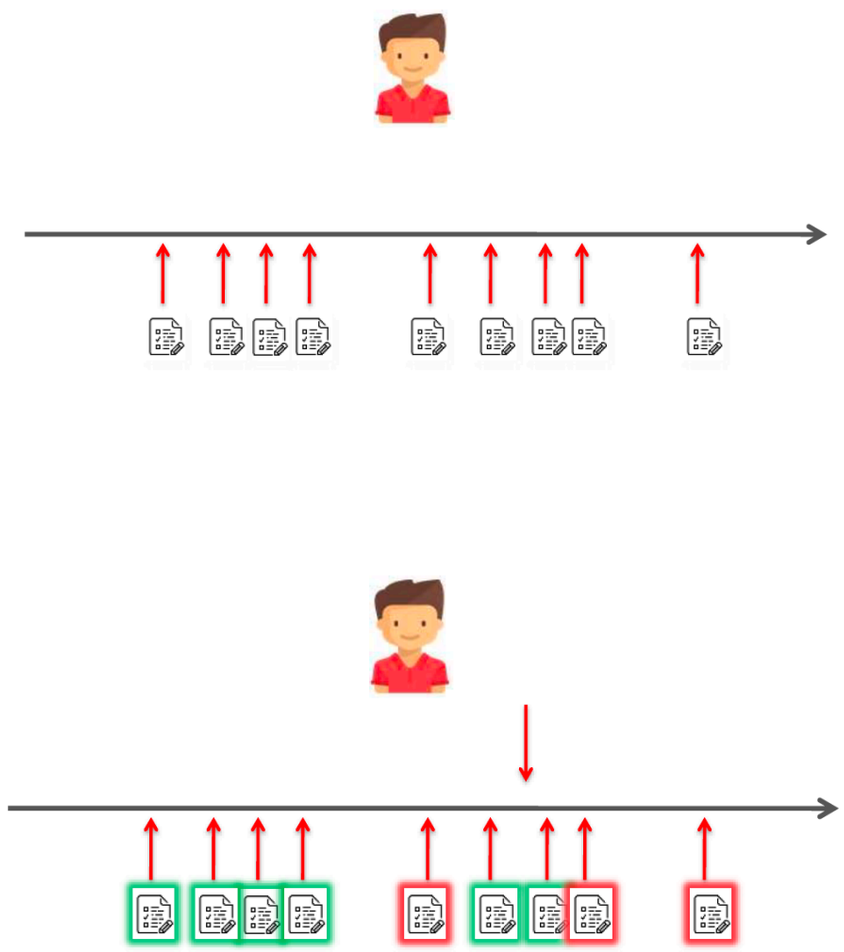
Rasch simple logistic model (SLM)

$$P_{ni} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

One-parameter logistic IRT model

$$P_{ni} = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

Современная психометрическая теория



Bloxom, B. (1985). Considerations in psychometric modeling of response time. *Psychometrika*, *50*(4), 383-397

Bruce Bloxom
През. ПО (1984/85)

John M. Linacre

Linacre, J. (1992). Many-Facet Rasch Measurement

'. . . it is intelligible to speak of the mean judgment of competent critics as the true judgment; and deviations from that mean as errors'

(Edgeworth, 1888, p. 622)

Susan Embretson
През. ПО (1998/99)

Embretson, S. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407-433

van Der Linden, W., & Glas, C. (2000). Computerized Adaptive Testing: Theory and Practice.

Wim Van der Linden
През. ПО (1999/00)

Cees Glas
През. ПО (2017/18)

Нидерландское психометрическое чудо

Van der Heijden, P. & Sijtsma, K. (1996). Fifty years of measurement and scaling in the Dutch social sciences. *Statistica Neerlandica*, *50*, 111-135

- от национальной к интернациональной ориентации, от изоляции к экспансии
- от нескольких публикаций в национальных журналах к осознанному увеличению публикационной активности на международном уровне
- от поддерживающей статистики к продвинутым разработкам в области методологии
- от малых наборов данных к большим данным

Сегодня Нидерланды и Фландрия глобальный лидер (инфлюенсер, простите) в области психометрики, например

Пример 1: Elo Rating System

Arpad Elo

Computers & Education 57 (2011) 1813–1824

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation

S. Klinkenberg*, M. Straatemeier, H.L.J. van der Maas

Department of Psychology, University of Amsterdam, Roeterstraat 15, Room A522, 1018 WB Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 26 November 2010
Received in revised form
8 February 2011
Accepted 8 February 2011

Keywords:
IRT
CAT
CAP
Computer adaptive practice
Serious gaming
Progress monitoring
Item calibration

ABSTRACT

In this paper we present a model for computerized adaptive practice and monitoring. This model is used in the Maths Garden, a web-based monitoring system, which includes a challenging web environment for children to practice arithmetic. Using a new item response model based on the Elo (1978) rating system and an explicit scoring rule, estimates of the ability of persons and the difficulty of items are updated with every answered item, allowing for on the fly item calibration. In the scoring rule both accuracy and response time are accounted for. Items are sampled with a mean success probability of .75, making the tasks challenging yet not too difficult. In a period of ten months our sample of 3648 children completed over 3.5 million arithmetic problems. The children completed about 33% of these problems outside school hours. Results show better measurement precision, high validity and reliability, high pupil satisfaction, and many interesting options for monitoring progress, diagnosing errors and analyzing development.

Han van der Maas

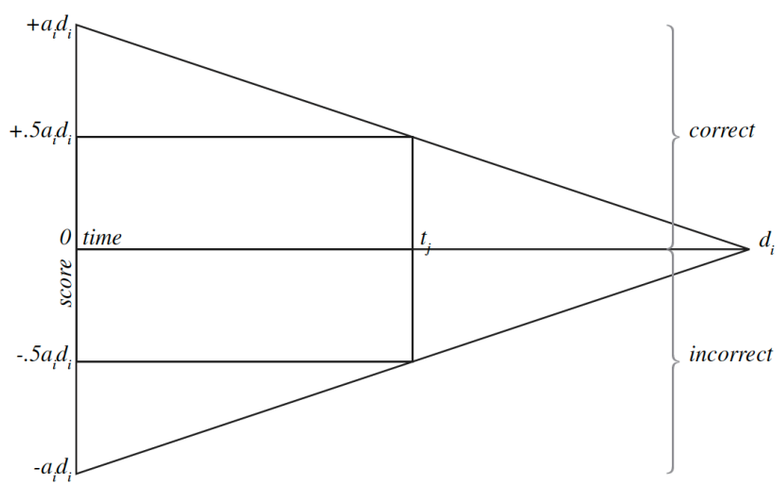
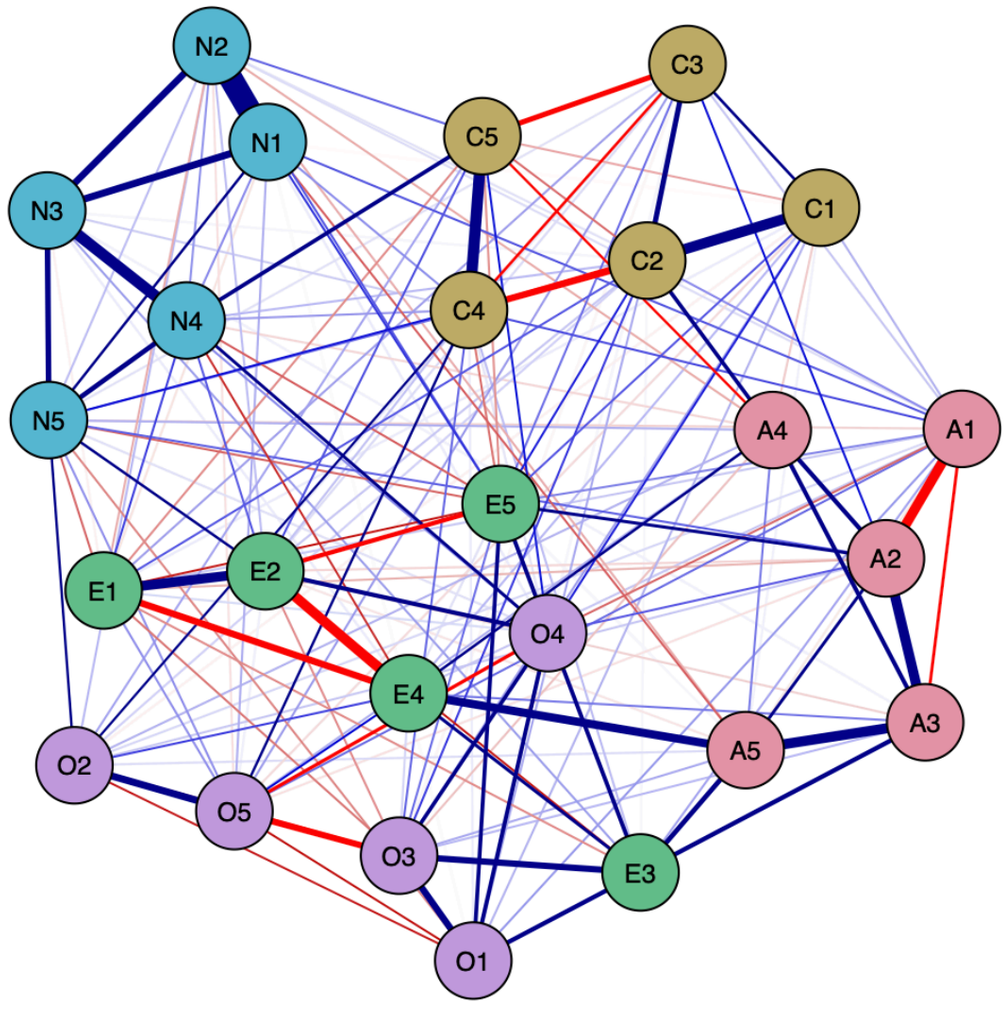


Fig. 2. High speed, high stakes scoring rule.

Пример 2: Network Psychometrics



http://sachaepskamp.com/Dissertation

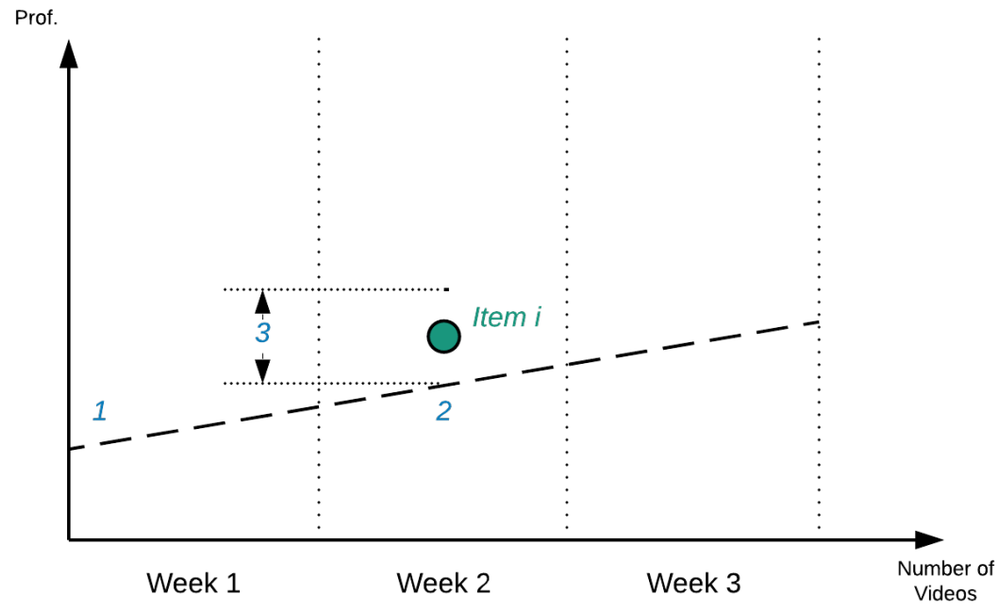
Примеры из практики: трекинг обучения

Abbakumov, D., Desmet, P., Van Den Noortgate, W. (2019). Measuring growth in students' proficiency in MOOCs: Two component dynamic extensions for the Rasch model. *Behavior Research Methods*, *51*(1), 332-341.
Abbakumov, D., Desmet, P., Van Den Noortgate, W. (2018). Measuring student's proficiency in MOOCs: multiple attempts extensions for the Rasch model. *Heliyon*, *4*(12), 1-15.

- Подготовленность
  - Item Response Theory (G. Rasch, F. Lord, 1960-е)
    - подготовленность – константа
- Трекинг обучения
  - Bayes Knowledge Tracing (Corbett & Anderson, 1985)
    - прирост одинаков для всех
- … долгое время всех все устраивало

Новый взгляд на моделирование прироста

- Видео – центральный элемент контента MOOK
  - рост внутри курса
- Попытка – новый вид взаимодействия с заданием в MOOK
  - локальный рост (внутри единицы содержания)



Результаты

Table 2 Parameters of the extension for Course 1

				Rasch model (Eq. 5)	Extension with fixed growth effects (Eq. 6)	Extension with random growth effects and univariate distribution (Eq. 7)	Extension with correlated random growth effects and multivariate distribution (Eq. 7)		
Fixed		Intercept	b_0	0.96 (0.09)	– 0.52 (0.16)	– 0.38 (0.17)	– 0.32 (0.17)		
		Video	b_1		4.45 (0.45)	3.71 (0.50)	3.72 (0.46)		
		Attempt	b_2		0.43 (0.01)	0.80 (0.02)	0.82 (0.02)		
Random	Student	Intercept	σ_{u1}	0.72	0.79	0.80	0.95	Corr.	
		Video	σ_{b1}			1.52	2.07	– .67	
		Attempt	σ_{b2}			0.52	0.51	.14	.02
	Item	Intercept	σ_{u2}	1.03	1.09	1.09	1.08		
AIC				146,979	143,024	140,367	140,221		

In the table, Course 1 is “Economics for Non-Economists” (Higher School of Economics, n.d.-a). For fixed effects, standard errors are presented in parentheses. For random effects, standard deviations are presented

Результаты

Table 2 Parameters of the extension for Course 1

				Rasch model (Eq. 5)	Extension with fixed growth effects (Eq. 6)	Extension with random growth effects and univariate distribution (Eq. 7)	Extension with correlated random growth effects and multivariate distribution (Eq. 7)		
Fixed		Intercept	b_0	0.96 (0.09)	− 0.52 (0.16)	− 0.38 (0.17)	− 0.32 (0.17)		
		Video	b_1		4.45 (0.45)	3.71 (0.50)	3.72 (0.46)		
		Attempt	b_2		0.43 (0.01)	0.80 (0.02)	0.82 (0.02)		
Random	Student	Intercept	σ_{u1}	0.72	0.79	0.80	0.95	Corr.	
		Video	σ_{b1}			1.52	2.07	− .67	
		Attempt	σ_{b2}			0.52	0.51	.14	.02
	Item	Intercept	σ_{u2}	1.03	1.09	1.09	1.08		
AIC				146,979	143,024	140,367	140,221		

In the table, Course 1 is “Economics for Non-Economists” (Higher School of Economics, n.d.-a). For fixed effects, standard errors are presented in parentheses. For random effects, standard deviations are presented

Результаты

Table 3 Dynamics of antilogits throughout Course 1

	Start	Week 1			Week 5		
	Avg.	Avg.	– *SD*	+ *SD*	Avg.	– *SD*	+ *SD*
Rasch model (Eq. 5)	.72	.72			.72		
Extension with fixed growth effects (Eq. 6)	.37	.47			.83		
Extension with random growth effects and univariate distribution (Eq. 7)	.41	.49	.45	.52	.80	.66	.89
Extension with correlated random growth effects and multivariate distribution (Eq. 7)	.42	.50	.46	.55	.81	.62	.92

In the table, antilogits for a student with average (initial) ability and for an item with average difficulty are presented. The dynamics of antilogits for the start, the end of the first week, and the end of the fifth week of the course are presented. For the model with random growth effects, the average, one standard deviation lower than the average, and one standard deviation higher than the average continuous growth effect are presented. Course 1 is "Economics for Non-Economists" (Higher School of Economics, n.d.-a)

Результаты

Table 3 Dynamics of antilogits throughout Course 1

	Start	Week 1			Week 5		
	Avg.	Avg.	– *SD*	+ *SD*	Avg.	– *SD*	+ *SD*
Rasch model (Eq. 5)	.72	.72			.72		
Extension with fixed growth effects (Eq. 6)	.37	.47			.83		
Extension with random growth effects and univariate distribution (Eq. 7)	.41	.49	.45	.52	.80	.66	.89
Extension with correlated random growth effects and multivariate distribution (Eq. 7)	.42	.50	.46	.55	.81	.62	.92

In the table, antilogits for a student with average (initial) ability and for an item with average difficulty are presented. The dynamics of antilogits for the start, the end of the first week, and the end of the fifth week of the course are presented. For the model with random growth effects, the average, one standard deviation lower than the average, and one standard deviation higher than the average continuous growth effect are presented. Course 1 is "Economics for Non-Economists" (Higher School of Economics, n.d.-a)

Результаты

Table 4 Dynamics of antilogits with attempts in Course 1

	Att. 1	Attempt 2			Attempt 3		
	Avg.	Avg.	– *SD*	+ *SD*	Avg.	– *SD*	+ *SD*
Rasch model (Eq. 5)	.72	.72			.72		
Extension with fixed growth effects(Eq. 6)	.37	.48			.58		
Extension with random growth effects and univariate distribution (Eq. 7)	.41	.60	.48	.72	.77	.54	.91
Extension with correlated random growth effects and multivariate distribution (Eq. 7)	.42	.62	.50	.73	.79	.57	.91

In the table, antilogits for a student with average ability and for an item with average difficulty at the start of the course are presented. The dynamics of antilogits for the first, second, and third attempts are presented. For the model with random growth effects, the average, one standard deviation lower than the average, and one standard deviation higher than the average local growth effect are presented. Course 1 is “Economics for Non-Economists” (Higher School of Economics, n.d.-a)

Результаты

Table 4 Dynamics of antilogits with attempts in Course 1

	Att. 1	Attempt 2			Attempt 3		
	Avg.	Avg.	– *SD*	+ *SD*	Avg.	– *SD*	+ *SD*
Rasch model (Eq. 5)	.72	.72			.72		
Extension with fixed growth effects(Eq. 6)	.37	.48			.58		
Extension with random growth effects and univariate distribution (Eq. 7)	.41	.60	.48	.72	.77	.54	.91
Extension with correlated random growth effects and multivariate distribution (Eq. 7)	.42	.62	.50	.73	.79	.57	.91

In the table, antilogits for a student with average ability and for an item with average difficulty at the start of the course are presented. The dynamics of antilogits for the first, second, and third attempts are presented. For the model with random growth effects, the average, one standard deviation lower than the average, and one standard deviation higher than the average local growth effect are presented. Course 1 is “Economics for Non-Economists” (Higher School of Economics, n.d.-a)

Результаты

Table 7 Accuracy in predicting correctness

	Overall		Course 1		Course 2		Course 3	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Rasch model	.743	.047	.724	.002	.699	.002	.806	.000
Extension with fixed growth effects	.760	.038	.737	.002	.732	.002	.812	.001
Extension with random growth effects and univariate distribution	.766	.034	.747	.002	.740	.001	.813	.001
Extension with random growth effects and multivariate distribution	.766	.034	.747	.002	.739	.001	.812	.001

In the table, Course 1 is “Economics for Non-Economists” (Higher School of Economics, n.d.-a), Course 2 is “Game Theory” (Higher School of Economics, n.d.-b), and Course 3 is “Introduction to Neuroeconomics: How the Brain Makes Decisions” (Higher School of Economics, n.d.-c)

Где применяем?

- В аналитических дэшбордах
- В рекомендательных и навигационных движках

Пример из практики: измерение учебной активности

Abbakumov, D., Desmet, P., Van Den Noortgate, W. (2019). Measuring student’s activity in MOOCs using extensions of the Rasch model. Presented at the International Meeting of the Psychometric Society, Santiago, Chile, 15 Jul 2019-19 Jul 2019.

Проблема

Активность студентов МООК обычно описывают, используя пропорции (просмотренных видео, решенных заданий). Такие меры активности просты и интуитивны. Но! Обобщая данные по студенту, мы упускаем важную информацию о том, как отдельный студент взаимодействовал с отдельной единицей содержания. О персонализации учебного опыта можно даже не думать…

Цель

Смоделировать и измерить активность студентов на микро-уровне:

1 студент * 1 единица содержания

Фреймворк

Item Response Theory

- Есть скрытая переменная, объясняющая наблюдение
- Функциональная связь между наблюдаемым и латентным – нелинейная

Ответ студента на задание объясняется подготовленностью студента и трудностью задания

Просмотр студентом видео-лекции объясняется активностью студента и '___' видео-лекции

Данные

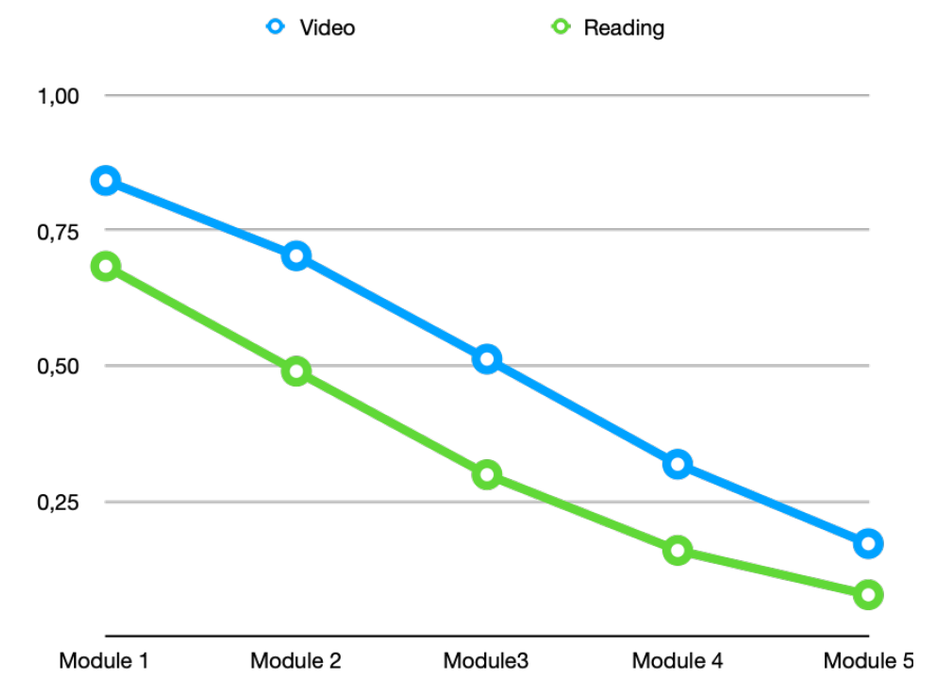
> **70,000** студентов онлайн-курсов ВШЭ на Coursera

> 2,500,000 единичных взаимодействий студентов с контентом

3 курса в кросс-валидизации

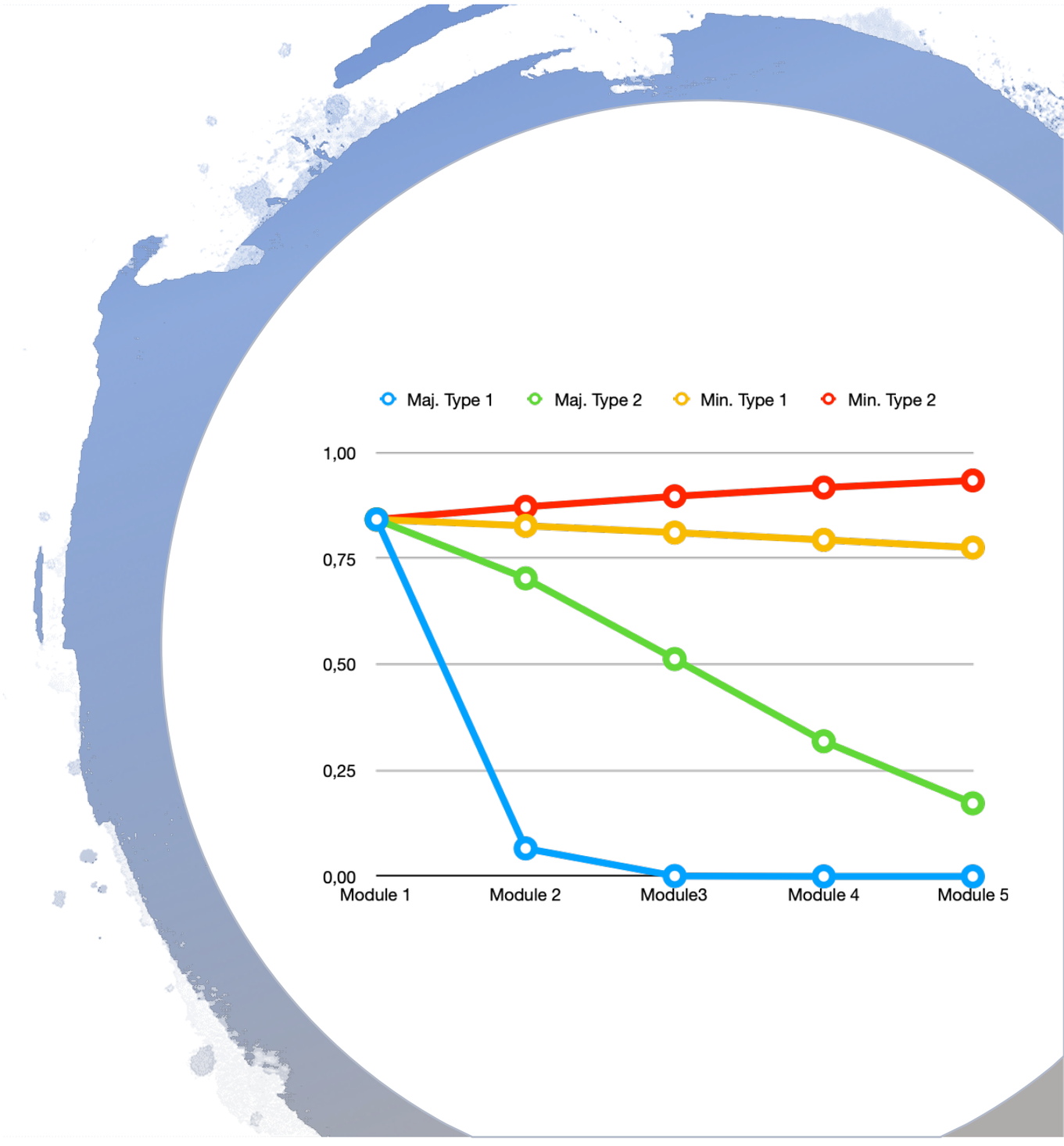
Студенты активнее смотрят видео-лекции, чем читают.

Как этим воспользоваться на практике?

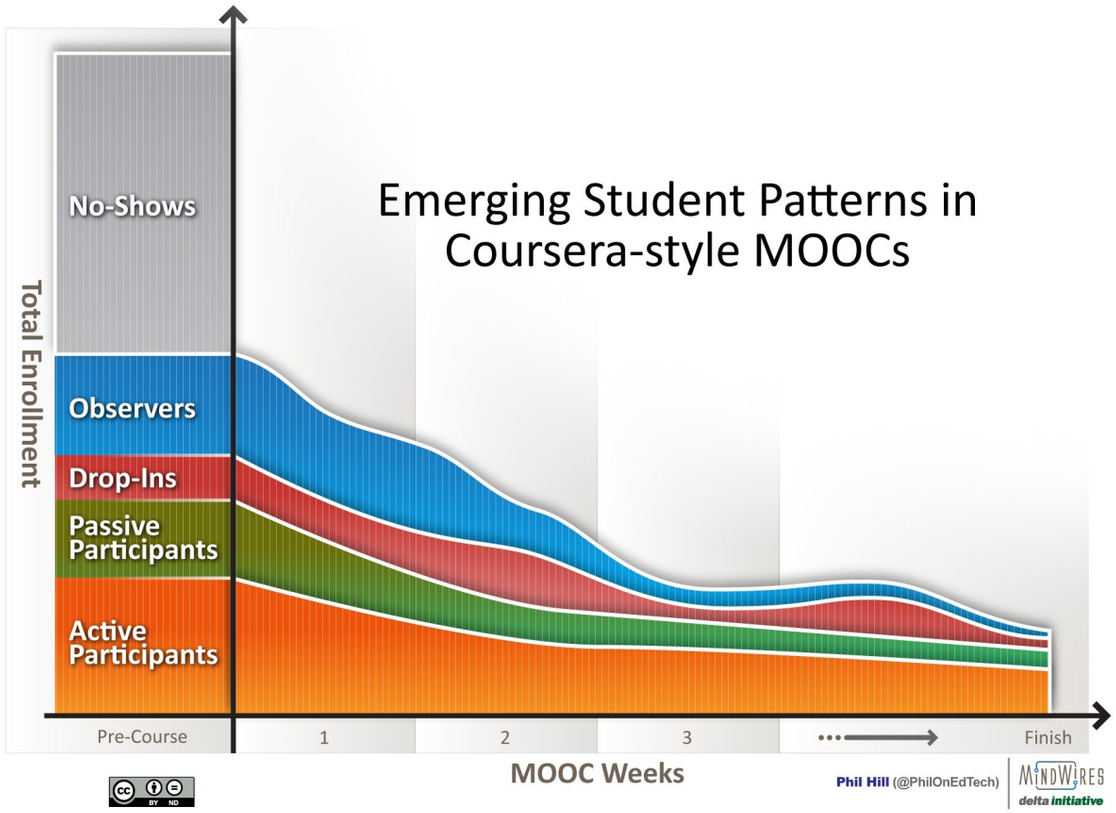


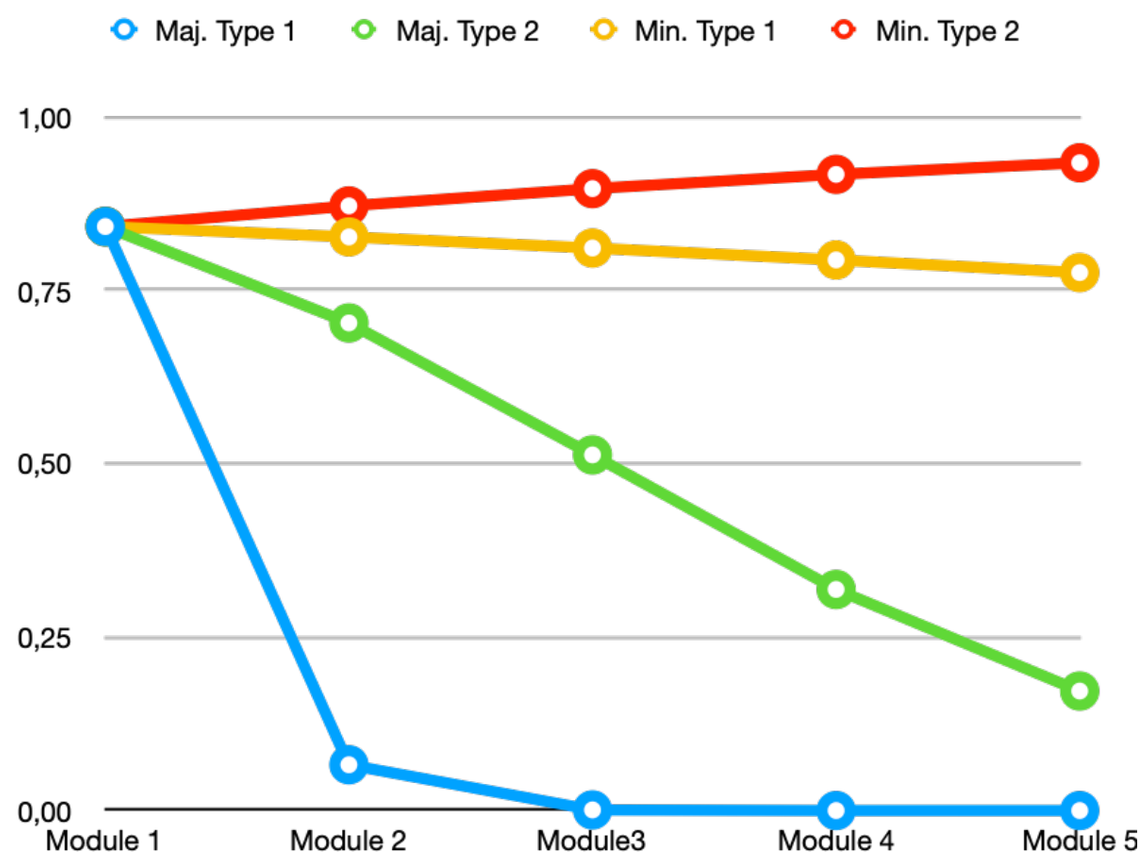
У большинства студентов активность в течение курса снижается.

Но есть те, у которых активность растет, но их меньшинство. Все разные.

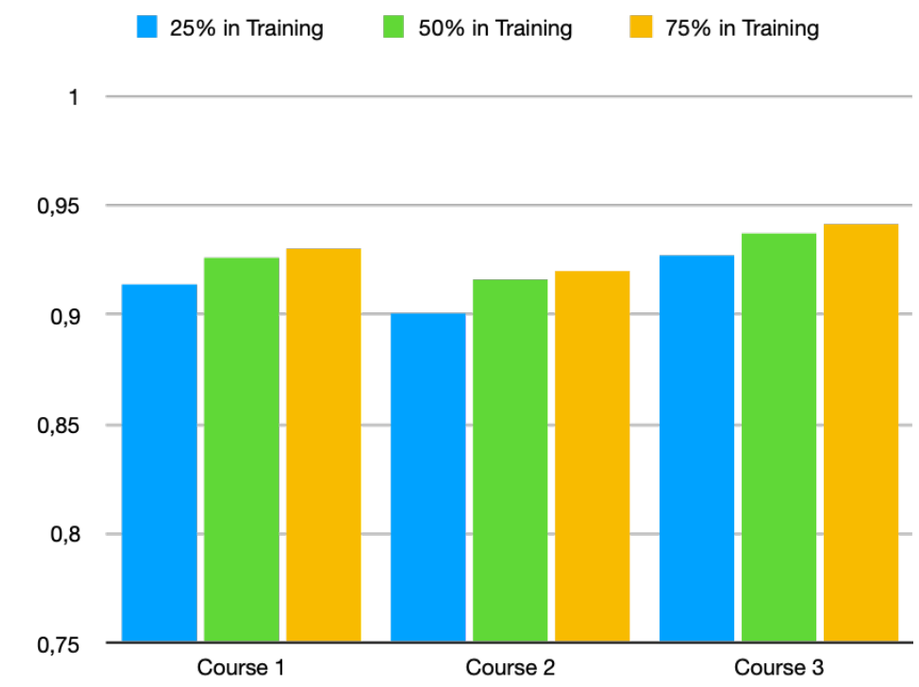


Emerging Student Patterns in Coursera-style MOOCs





Результаты кросс-валидизации



Что дальше?

объясним латентные переменные

проведем расширенную валидизацию

… подумаем о переходе от аналитики к практике

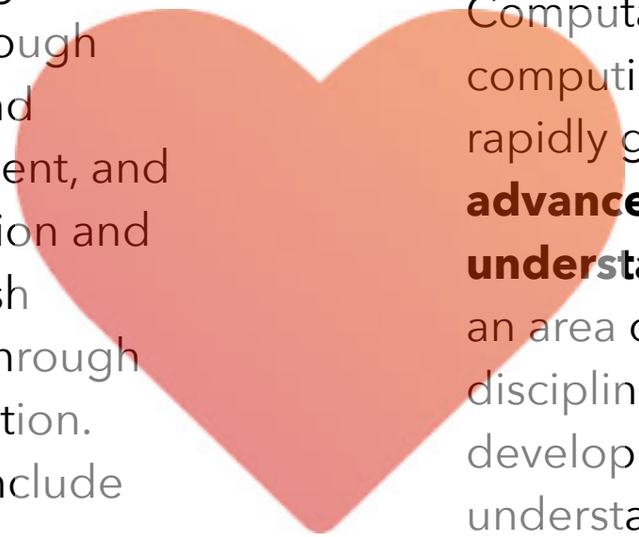
Computational Behavioral Science

Behavioral sciences **explore the cognitive processes** within organisms and **the behavioral interactions** between organisms in the natural world. It involves the systematic analysis and investigation of human and animal behavior through the study of the past, controlled and naturalistic observation of the present, and disciplined scientific experimentation and modeling. It attempts to accomplish legitimate, objective conclusions through rigorous formulations and observation. Examples of behavioral sciences include **psychology, psychobiology, anthropology, and cognitive science**. Generally, behavior science deals primarily with human action and often seeks to generalize about human behavior as it relates to society.

Computational science (also scientific computing or scientific computation (SC)) is a rapidly growing multidisciplinary field that uses **advanced computing capabilities to understand and solve complex problems**. It is an area of science which spans many disciplines, but at its core it involves the development of models and simulations to understand natural systems.

Computational Behavioral Science

Behavioral sciences **explore the cognitive processes** within organisms and **the behavioral interactions** between organisms in the natural world. It involves the systematic analysis and investigation of human and animal behavior through the study of the past, controlled and naturalistic observation of the present, and disciplined scientific experimentation and modeling. It attempts to accomplish legitimate, objective conclusions through rigorous formulations and observation. Examples of behavioral sciences include **psychology, psychobiology, anthropology, and cognitive science**. Generally, behavior science deals primarily with human action and often seeks to generalize about human behavior as it relates to society.



Computational science (also scientific computing or scientific computation (SC)) is a rapidly growing multidisciplinary field that uses **advanced computing capabilities to understand and solve complex problems**. It is an area of science which spans many disciplines, but at its core it involves the development of models and simulations to understand natural systems.

Спасибо!

www.abbakumov.com