УДК 519.25

ИССЛЕДОВАНИЕ ВЛИЯНИЯ МНОГОМЕРНОСТИ ДАННЫХ НА ОЦЕНКУ КОМПЕТЕНЦИЙ С ИСПОЛЬЗОВАНИЕМ IRT-МОДЕЛЕЙ

Родионов А.В.

ФГБОУ ВПО «Байкальский государственный университет экономики и права», Иркутск, e-mail: avr-v@yandex.ru

Одним из перспективных способов оценки компетенций учащихся вузов является использование латентного анализа и в частности IRT-моделей, так как основные положения латентного анализа в наибольшей степени отвечают сущности понятия компетенций. Применение IRT-моделей требует выполнения ряда условий, в частности условия «одномерности». На практике выполнение этого условия крайне затруднительно, и оно зачастую опускается. Данное исследование ставит целью определить правомерность игнорирования этого условия в случае использования реальных экзаменационных оценок в качестве эмпирических данных. Результаты показали, что альтернативная модель, учитывающая влияние многомерности данных, в ряде случаев действительно обеспечивает лучшую точность измерения, однако увеличение количества параметров существенно усложняет расчеты и процедуру проведения оценки. Выигрыш в точности с практической точки зрения несущественен.

Ключевые слова: компетенция, оценка компетенции, IRT, латентная переменная, локальная независимость, многомерные данные

INVESTIGATION OF THE EFFECT OF MULTIDIMENSIONAL DATA TO ASSESSMENT COMPETENCE USING IRT-MODELS

Rodionov A.V.

Baikal State University of Economics and Law, Irkutsk, e-mail: avr-v@yandex.ru

One promising way to assess competencies of students of the university is the use of latent analysis – IRT-models, as the idea of latent analysis in the best essence of the concept of competence as a latent factor. Application of IRT-model requires a number of conditions, one of them is the condition of «one-dimensionality data». In practice, this condition is extremely difficult, and is often omitted. This Investigation aims to determine the validity of ignoring this condition in the case of using test scores as a real empirical data. The results showed that although the alternative model that takes into account the impact of local dependence data, in some cases, provides better measurement accuracy, more complex, and the gain in accuracy from the practical point of view, relatively unimportant.

Keywords: competencies, assessment of competence, IRT, latent variable, local independence, multidimensional data

Одним из условий применения IRT-моделей является необходимость соблюдения условия «одномерности» используемых эмпирических данных – ответы на задания конструкта должны зависеть только от одной латентной переменной (параметра). Но, несмотря на кажущуюся простоту, данное условие трудновыполнимо. Например, даже ответ на такой простой вопрос «сколько будет 2 + 2?» обусловлен не только умением складывать числа, но и рядом других латентных параметров: умением читать, знанием цифр и пр. Поэтому в настоящий момент считается, что если задание диагностирует исследуемый латентный параметр (главный), влияние других незначительно, задание удовлетворяет статистическим проверкам на соответствие [9], то его можно использовать для оценки латентного параметра. Однако некоторые теоретические исследования показывают, что игнорирование условия одномерности, а также локально-зависимых заданий [15] в ряде случаев приводит к необъективному оцениванию [14] либо к завышению надежности набора заданий (конструкта) [13].

В работе [2] рассматривается применение IRT-моделей для оценки компетенций студентов вузов. Базовая модель для оценки компетенций (латентная модель экзаменационных заданий, mRSM) формулируется в следующем виде [1]:

$$P\left(y_{ij}=k\mid\theta_i,\delta_j,\tau_{ju}\right)=\frac{e^{k\theta_i-k\delta_j-\sum_{u=0}^{k}\tau_{ju}}}{1+\sum_{v=1}^{K}e^{\sum_{u=1}^{v}(\theta_i-\delta_j-\tau_{ju})}},\quad(1)$$

где y_{ij} – ответ, который дал i-й студент на j-е задание, $i=\overline{1,N}$; $j=\overline{1,M}$; N – количество студентов; M – количество заданий; θ_i – латентный параметр (компетентность) i-го студента; τ_{ju} – параметр (сложность категории) «шага» шкалы оценки по каждому экзаменационному заданию j; δ_j – параметр задания; k – оценка за задание; K – максимальная оценка.

Исходными данными являются оценки за ответы на выполненные экзаменационные задания. Экзаменационное задание по своей природе является собирательным, поэтому возникает вопрос об исследовании влияния множества компетенций на полученную экзаменационную оценку. Для этого можно использовать подход, рассмотренный в работах [6, 8], суть которого заключается в добавлении к исследуемой модели дополнительного параметра для учета влияния других латентных переменных – γ. Тогда модель (1) можно записать в следующем виде (TmRSM):

$$P\left(y_{ij}=k\,|\,\theta_i,\delta_j,\tau_{ju},\gamma_{d(j)i}\right)=$$

$$=\frac{e^{k\theta_i-k\delta_j-\sum_{u=0}^{k}\left(\tau_{ju}+\gamma_{d(j)i}\right)}}{1+\sum_{v=1}^{K}e^{\sum_{u=1}^{v}\left(\theta_i-\delta_j-\tau_{ju}-\gamma_{d(j)i}\right)}}. \qquad (2)$$

Однако при использовании модели (2) возникают три существенные сложности:

1) увеличение числа параметров приводит к увеличению вычислительной сложности алгоритма;

2) использование большого числа параметров в некоторых случаях может не только не привести к повышению согласия эмпирических данных и модели, но и ухудшить его;

3) возрастает «сложность» задания исходных данных и интерпретации результатов. Поэтому задача определения «границ» использования одномерной модели является важной и зачастую определяющей для успешного применения модели в практической деятельности.

План исследования

Для анализа используются эмпирические данные об оценках студентов направления подготовки «Прикладная информатика», профиль подготовки «Информационные системы и технологии в управлении», квалификация (степень) бакалавр.

Статистический анализ применимости моделей проводится с использованием fit-статистик Infit, OutFit, InFit(t), OutFit(t) – считается, что следует использовать задания, у которых значения Infit и OutFit находятся в диапазоне 0,4–1,6, а InFit(t) и OutFit(t): –2,5...+2,5 [9].

Для сравнения моделей между собой использован ряд критериев, связанных с понятием информационной энтропии и расстоянием Кульбака – Лейблера [7]. При применении критериев лучшей считается модель, в достаточной мере полно описывающая данные с наименьшим количеством параметров. В работе использованы критерии Акаике, Байесовский и их модификации: состоятельный критерий Акаике, скорректированный критерий Акаике и скорректированный Байесовский критерий [4, 12] (табл. 1).

Также для сравнения можно использовать информационную функцию, введенную А. Бирнбаумом. По одному из определений, количество информации, обеспеченное *j*-м заданием в конкретной точке θ, – это величина, обратно пропорциональная стандартной ошибке измерения данного значения θ с помощью задания *j* [3]. Информационная функция задания показывает соответствие количества информации, получаемой при оценивании параметра θ с помощью задания *j* и может быть выражена следующей формулой:

$$I_j(\theta)=-M\left[\frac{d^2}{d\theta^2}\ln L_j\left(\theta\,|\,y_j\right)\right], \qquad (3)$$

где $L_j\left(\theta\,|\,y_j\right)$ – есть функция правдоподобия задания *j*.

Таблица 1

Информационные критерии

Информационный критерий	Формула
Акаике (AIC)	$AIC=2p-2\ln(L)$
Скорректированный Акаике (AICc)	$AICc=AIC+\frac{2p(1-p)}{n-p-1}$
Байесовский (BIC)	$BIC=p\ln(n)-2\ln(L)$
Скорректированный Байесовский (aBIC)	$aBIC=\ln\left(\frac{n-2}{24}\right)p-2\ln(L)$
Состоятельный Акаике (CAIC)	$CAIC=(1+\ln(n))p-2\ln(L)$

где *p* – количество оцениваемых параметров модели; *L* – максимизированное значение функции правдоподобия, *n* – размер выборки.

Значения этой функции являются своеобразной характеристикой эффективности *j*-го задания: чем больше количество информации, тем лучше, образно говоря, работает задание на рассматриваемом интервале оси θ. Информация, полученная при измерении данного θ с помощью всего конструкта, складывается из отдельных значений ординат информационных функций, построенных для каждого задания [5]:

$$I(\theta) = \sum_{j=1}^{M} I_j(\theta). \qquad (4)$$

Оценка параметров моделей, расчеты статистик и значений критериев производились с помощью программы, написанной на языке R [11].

Результаты исследования и их обсуждение

Рассмотрим компетенцию ПК-1. Определены 8 экзаменационных заданий, которые сформировали конструкт для оценки компетенции: правоведение, основы бизнеса, информационная безопасность, проектирование информационных систем, предметно-ориентированные экономические информационные системы, защита информации в банках, автоматизированные банковские системы, налогообложение. Дополнительно построена матрица (табл. 2), в которой цифрой 1 обозначены компетенции, влияющие на оценку соответствующего задания.

В табл. 3 приведены значения статистик параметров моделей, рассчитанные по экзаменационным оценкам.

Таблица 2

Матрица влияния компетенций на экзаменационные задания конструкта

Задание	ПК-1	ПК-2	ПК-4	ПК-5	ПК-6	ПК-8	ПК-11	ПК-13	ПК-14	ПК-15	ПК-18	ПК-19	ПК-22
1	1								1	1			
2	1												
3	1												
4	1			1	1	1	1						
5	1		1					1	1			1	
6	1		1								1		1
7	1	1	1			1					1		1
8	1												1

Таблица 3

Значения fit-статистик

Экзаменационное задание		Значения статистик							
		mRSM				TmRSM			
Наименование	Параметр	Outfit	Outfit (t)	Infit	Infit (t)	Outfit	Outfit (t)	Infit	Infit (t)
1	2	3	4	5	6	7	8	9	10
Правоведение	δ_1	0,99	–0,04	1,01	0,09	0,85	–0,92	0,93	–0,58
	τ_{11}	1,54	1,61	1,03	0,31	0,89	–0,05	0,90	–0,88
	τ_{12}	0,85	–1,24	0,91	–1,74	0,88	–0,90	0,96	–0,83
Основы бизнеса	δ_2	0,86	–0,89	0,90	–0,75	0,67	–2,17	0,76	–2,14
	τ_{21}	1,08	0,40	0,92	–0,57	0,84	–0,55	0,87	–0,92
	τ_{22}	0,82	–0,71	0,91	–1,18	1,02	0,15	0,94	–0,73
Информационная безопасность	δ_3	1,41	2,24	1,23	1,80	1,69	3,91	1,49	3,44
	τ_{31}	1,70	1,58	1,31	1,80	1,78	1,79	1,22	1,32
	τ_{32}	1,39	1,32	1,11	1,21	1,12	0,55	1,08	0,91
Проектирование информационных систем	δ_4	0,76	–1,10	0,77	–2,06	0,81	–1,67	0,82	–1,54
	τ_{41}	0,73	–1,21	0,80	–1,35	0,80	–0,93	0,82	–1,32
	τ_{42}	0,87	–0,54	0,90	–1,75	0,91	–0,25	0,97	–0,57

Окончание табл. 3

1	2	3	4	5	6	7	8	9	10
Предметно-ориентированные экономические информационные системы	δ_5	0,96	–0,18	1,00	0,06	0,89	–0,53	0,97	–0,18
	τ_{51}	0,92	–0,41	0,98	–0,22	0,83	–1,04	0,93	–0,69
	τ_{52}	0,87	–0,80	0,96	–0,69	0,95	–0,23	1,03	0,54
Защита информации в банках	δ_6	1,50	2,42	1,41	2,28	1,44	2,21	1,42	2,37
	τ_{61}	1,70	1,52	1,36	1,41	1,48	1,16	1,35	1,47
	τ_{62}	1,28	1,36	1,16	1,64	1,26	1,26	1,13	1,35
Автоматизированные банковские системы	δ_7	1,06	0,42	1,12	0,87	1,24	1,34	1,17	1,16
	τ_{71}	0,70	–1,87	0,74	–2,29	0,97	–0,14	0,95	–0,32
	τ_{72}	1,04	0,27	0,99	–0,05	0,97	–0,06	1,05	0,63
Налогообложение	δ_8	0,97	–0,17	1,00	0,03	0,98	–0,06	0,96	–0,25
	τ_{81}	0,76	–0,80	1,06	0,45	1,00	0,09	1,01	0,08
	τ_{82}	0,91	–0,29	0,95	–0,59	1,36	1,35	1,03	0,36

Анализируя таблицу, видно, что в целом обе модели достаточно хорошо описывают эмпирические данные, превышения значений статистик для некоторых параметров объясняются влиянием многомерности, к чему особенно чувствительна Outfit-статистика.

В табл. 4 приведены значения информационных критериев – мер относительно качества моделей, учитывающих степень «подгонки» модели под данные с корректировкой (штрафом) на используемое количество оцениваемых параметров.

Даже если исключить из рассмотрения AICc, который накладывает очень большой штраф на количество параметров, все равно увеличение количества параметров является очень существенным и сказывается на значениях всех прочих информационных критериев, согласно которым следует выбрать mRSM.

На рис. 1 в графическом виде представлены оценки компетенций студентов в логитах, рассчитанные по двум моделям.

Таблица 4

Значения информационных критериев

Информационный критерий	TmRSM	mRSM
Акаике (AIC)	1964,5	1775,37
Скорректированный Акаике (AICc)	4187,83	1788,12
Байесовский (BIC)	2292,48	1846,67
Скорректированный Байесовский (aBIC)	1925,19	1766,83
Состоятельный Акаике (CAIC)	2407,48	1871,67

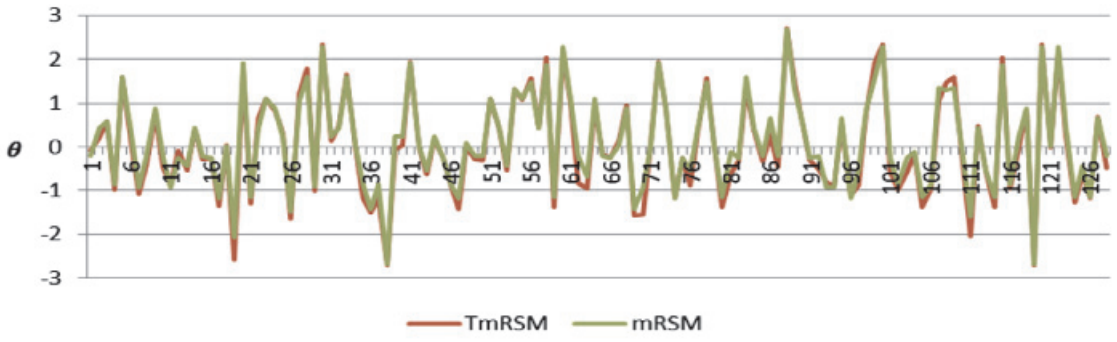

Рис. 1. График оценок компетенций студентов

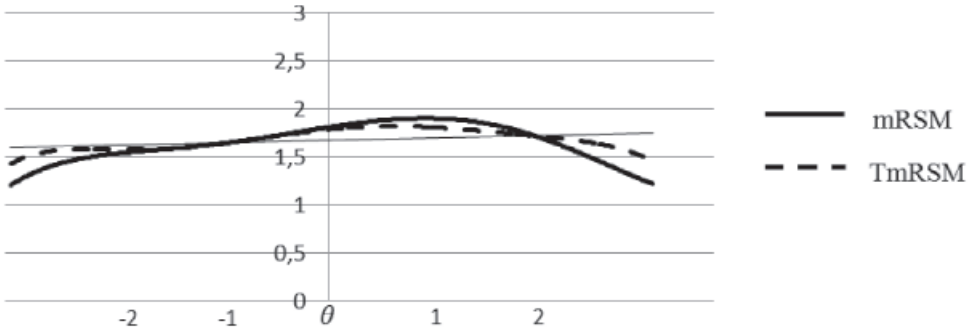

Рис. 2. Графики информационных функций конструкта

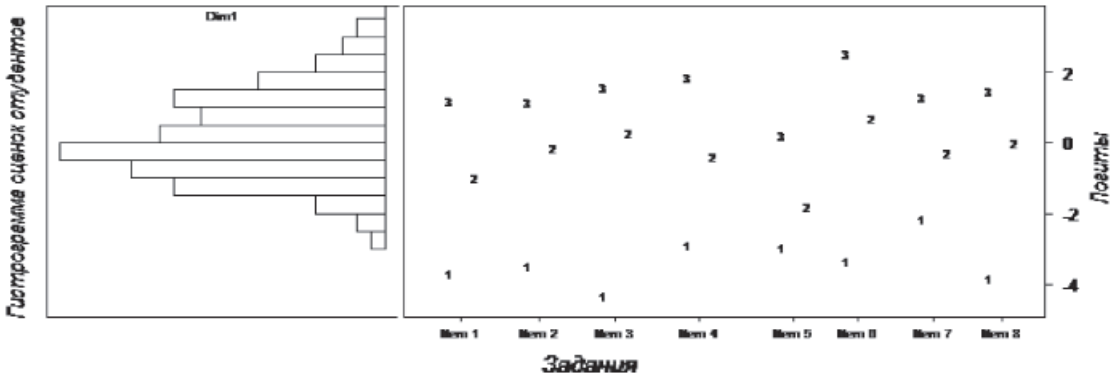

Рис. 3. Карта Райта

Абсолютная максимальная разница в оценках составила 0,6 логита, средняя разница – 0,13 логита, наибольшая разница проявляется на концах шкалы, при оценке θ в диапазоне (–1,5; +1,5) логита разница не превышает среднюю. Корреляция оценок – 0,99. В подавляющем большинстве случаев разница в оценке не превышает среднюю ошибку вычисления. На рис. 2 представлены графики информационных функций конструкта.

Для обеих моделей информационные функции имеют один выраженный максимум, что говорит о хорошем качестве конструкта. Сравнивая точности измерения компетенций, можно отметить, что у mRSM наибольшая точность достигается в диапазоне (примерно) $\theta \in (-0,5; +2)$, в диапазоне $\theta \in (-1,2; +0,5)$ точности равны, а по краям шкалы TmRSM показывает большую точность измерения. Однако данные различия не являются значительными, и в общем случае нельзя сделать однозначный вывод о том, какой модели следует отдать предпочтение. На рис. 3 представлена карта Райта для набора заданий оценки компетенции ПК-1. Оценки большинства студентов находятся в диапазоне (–2; +2) логита, где по точности mRSM не уступает TmRSM.

Результаты расчётов и анализа по остальным компетенциям ФГОС «Прикладная информатика» аналогичны приведённым для компетенции ПК-1.

Выводы

В статье с практической точки зрения рассмотрена проблема игнорирования условия одномерности при использовании одномерной IRT-модели оценки компетенции. Проведено сравнение одномерной модели mRSM с многомерной TmRSM, учитывающей влияние нескольких латентных переменных – компетенций. Проведенные расчеты показывают, что TmRSM в некоторых случаях более точно позволяет оценить компетенции, особенно у студентов с низким и/или высоким уровнем освоения компетенции, однако для студентов со средним уровнем освоения компетенции (а таких большинство) более предпочтительной является одномерная модель. Оценки компетенций студентов, полученные по mRSM и TmRSM, обладают очень высоким коэффициентом корреляции – минимальное значение 0,985 (по компетенции ПК-15), а абсолютная разница оценок в 96% случаев не превышает ошибку измерения. Разумеется, нельзя не списывать со счетов и размер

выборки – расчеты, проведенные на имитационных данных, например в работе [10], показывают, что с увеличением объема выборки расхождение в оценках параметров моделей растет. Но при небольшом и среднем размере выборки применение одномерной модели не приводит к существенному ухудшению точности оценки. В совокупности с усложнением процедуры оценки и алгоритмов расчетов TmRSM использование одномерной модели для практического применения следует считать допустимым.

Список литературы

1. Родионов А.В. Модификация рейтинговой параметрической модели оценки латентных факторов для измерения уровня сформированности компетенций // Известия Иркутской государственной экономической академии. – 2014. – № 6. – С. 168–174.

2. Родионов А.В., Братищенко В.В. Применение IRT-моделей для анализа результатов обучения в рамках компетентностного подхода // Современные проблемы науки и образования. – 2014. – № 4. – URL: www.science-education.ru/118-13858 (дата обращения: 20.08.2015).

3. Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. – 432 с.

4. Akaike H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. – 1974. – Vol. 19(6). – P. 716–723.

5. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability // Reading MA: Addison-Wesley. itle. – 1968. – P. 397–472.

6. Bradlow E.T., Wainer H., & Wang X. A Bayesian random effects model for testlets // Psychometrika. –1999. – Vol. 64. – P. 153–168.

7. Kullback S., Leibler R.A. On information and sufficiency // The Annals of Mathematical Statistics. – 1951. – Vol. 22. – № 1. – P. 79–86.

8. Li Y., Bolt D. M., & Fu J. A comparison of alternative models for testlets // Applied Psychological Measurement. – 2006. – Vol. 30. – P. 3–21.

9. Linacre J.M. What do Infit and Outfit, Mean-square and Standardized mean? – Режим доступа: http://www.rasch.org/rmt/rmt162f.htm (дата обращения: 20.08.2015).

10. Ou Zhang. Polytomous irt or testlet model: an evaluation of scoring models in small testlet size situations. – 2010. – URL: http://ufdc.ufl.edu/UFE0042638/00001 (дата обращения: 20.08.2015).

11. R Core Team. R: A language and environment for statistical computing // R Foundation for Statistical Computing, Vienna, Austria. – 2014. – URL: http://www.R-project.org (дата обращения: 20.08.2015).

12. Schwarz G. Estimating the dimension of a model // Annals of Statistics. – 1978. – Vol. 6. – P. 461–464.

13. Sireci S.G., Thissen D., & Wainer H. On the reliability of testlet-based tests // Journal of Educational Measurement. – 1991. – Vol. 28. – P. 237–247.

14. Wainer H., & Wang X. Using a new statistical model for testlets to score TOEFL // Journal of Educational Measurement. – 2000. – Vol. 37. – P. 203–220.

15. Yen W.M. Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model // Applied Psychological Measurement. – 1984. – Vol. 2. – P. 125–145.

References

1. Rodionov A.V. Modifikacija rejtingovoj parametricheskoj modeli ocenki latentnyh faktorov dlja izmerenija urovnja sformirovannosti kompetencij, *Izvestija Irkutskoj gosudarstvennoj jekonomicheskoj akademii*, 2014, no. 6, pp. 168–174.

2. Rodionov A.V., Bratishhenko V.V. Primenenie IRT-modelej dlja analiza rezultatov obuchenija v ramkah kompetentnostnogo podhoda, *Sovremennye problemy nauki i obrazovanija*, 2014, no. 4, Available at: www.science-education.ru/118-13858 (accessed 20 August 2015).

3. Chelyshkova M.B. Teorija i praktika konstruirovanija pedagogicheskih testov: Uchebnoe posobie, *M.: Logos*, 2002, 432 p.

4. Akaike H. A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 1974, Vol. 19(6), pp. 716–723.

5. Birnbaum A. Some latent trait models and their use in inferring an examinees ability, *Reading MA: Addison-Wesleyitle*, 1968, pp. 397–472.

6. Bradlow E.T., Wainer, H., & Wang, X. A Bayesian random effects model for testlets, *Psychometrika*, 1999, Vol. 64, pp. 153–168.

7. Kullback S., Leibler R.A. On information and sufficiency, *The Annals of Mathematical Statistics*, 1951, Vol. 22, no. 1, pp. 79-86.

8. Li Y., Bolt D.M., & Fu J. A comparison of alternative models for testlets, *Applied Psychological Measurement*, 2006, Vol. 30, pp. 3–21.

9. Linacre J.M. What do Infit and Outfit, Mean-square and Standardized mean? Available at: http://www.rasch.org/rmt/rmt162f.htm (accessed 20 August 2015).

10. Ou Zhang. Polytomous irt or testlet model: an evaluation of scoring models in small testlet size situations (2010), Available at: http://ufdc.ufl.edu/UFE0042638/00001.

11. R Core Team. R: A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria (2014), Available at: http://www.R-project.org (accessed 20 August 2015).

12. Schwarz G. Estimating the dimension of a model, *Annals of Statistics*, 1978, Vol. 6, pp. 461–464.

13. Sireci S.G., Thissen D., & Wainer H. On the reliability of testlet-based tests, *Journal of Educational Measurement*, 1991, Vol. 28, pp. 237–247.

14. Wainer H., & Wang X. Using a new statistical model for testlets to score TOEFL, *Journal of Educational Measurement*, 2000, Vol. 37, pp. 203–220.

15. Yen W.M. Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model, *Applied Psychological Measurement*, 1984, Vol. 2. pp. 125–145.

Рецензенты:

Пархомов В.А., д.ф.-м.н., профессор кафедры информатики и кибернетики, ФГБОУ ВПО «Байкальский государственный университет экономики и права», г. Иркутск;

Боровский А.В., д.ф.-м.н., профессор кафедры информатики и кибернетики, ФГБОУ ВПО «Байкальский государственный университет экономики и права», г. Иркутск.