

- уровней знаний в различных системах тестирования // Тезисы докладов III Всероссийской научно-методической конференции «Развитие системы тестирования в России» / Под ред. Л.С. Гребнева – М.: Центр тестирования Министерства образования РФ, 2001. – С. 205.
18. Маслак А.А., Бобрышев Е.А., Анисимова Т.С., Пушечкин Н.П. Исследование влияния числа дистракторов на точность оценивания уровня знаний // Тезисы докладов III Всероссийской научно-методической конференции «Развитие системы тестирования в России» / Под ред. Л.С. Гребнева – М.: Центр тестирования Министерства образования РФ, 2001. – С. 204.
 19. Отчет по НИР "Исследование эффективности системы тестирования на основе модели Раша" (руководитель проекта Маслак А.А.) - Славянск-на-Кубани, 2000. -177с.
 20. Отчет по проекту 4140 "Разработка системы комплексного мониторинга образовательного процесса в педагогическом вузе" (руководитель проекта Анисимова Т.С.) - Славянск-на-Кубани, 2000. -88с.
 21. Пашкова Л.М., Анисимова Т.С., Маслак А.А. Комплексная оценка здоровья первокурсников Славянского филиала Армавирского государственного педагогического института. /Материалы Второй всероссийской научно-практической конференции «Оценка эффективности образовательных инноваций и технологий». – Славянск-на-Кубани: Издательский центр Славянского филиала АГПИ, 2000. – С. 9-14.
 22. Пашкова Л.М., Анисимова Т.С., Маслак А.А. Мониторинг здоровья студентов в Славянском филиале Армавирского государственного педагогического института./ Здоровье школьников и студентов. Методы оценки, профилактические и реабилитационные технологии: Тезисы докладов. – Владимир, 2000. - С. 39-41.

ОЦЕНКА ЭФФЕКТИВНОСТИ МОДЕЛИ РАША ДЛЯ ЗАДАНИЙ С ВЫБОРОМ ПРАВИЛЬНОГО ОТВЕТА

*Маслак А.А., Осипов С.А. г. Курск, Славянск-на-Кубани,
КГПУ, Славянский филиал АГПИ*

Тема работы обусловлена следующим. Задания в закрытой форме (с выбором правильного ответа) являются наиболее технологическими и поэтому наиболее часто используются в тестах. Однако такая форма заданий предполагает и возможность угадывания правильного ответа. Вероятность угадывания правильного ответа зависит от числа дистракторов (числа возможных ответов в задании).

Цель работы состоит в исследовании влияния числа дистракторов и степени зашумленности данных на оценки параметров модели Раша.

Постановка задачи.

Точность оценок уровней знаний испытуемых является одним из наиболее важных показателей системы тестирования [3-5]. На основе имитационного моделирования исследуется точность оценивания уровней знаний испытуемых двух систем тестирования – традиционного тестирования (уровень знаний определяется по числу правильных ответов) и системы тестирования на основе модели Раша (уровень знаний определяется на основе параметров модели Раша). Варьируемыми факторами являются число дистракторов и степень зашумленности результатов тестирования.

Имитационное моделирование результатов тестирования осуществлялось следующим образом [6]. Выбранный интервал уровней знаний испытуемых в логитах составляет от -4 до $+4$. Всего используется 17 уровней знаний испытуемых, которые равномерно покрывают выбранный диапазон и отстоят друг от друга на расстоянии 0,5 логита. Уровни трудностей заданий выбраны аналогичным образом.

Для генерирования результатов тестирования используется следующая процедура. Прежде всего, для всех возможных комбинаций уровней знаний и трудности заданий вычисляется вероятность правильного ответа.

$$p_{ij} = c_k + (1 - c_k) \frac{e^{1,7(\theta_i - \beta_j)}}{1 + e^{1,7(\theta_i - \beta_j)}},$$

где p_{ij} - вероятность правильного ответа i -ого испытуемого с уровнем знаний θ_i на j -ое тестовое задание с трудностью τ_j , c_k – вероятность угадывания правильного ответа. Вероятность случайного угадывания c_k зависит от числа дистракторов в задании и числа правильных ответов. Коэффициент 1,7 означает, что дифференцирующая способность всех тестовых заданий при генерировании данных принята равной 1,7.

На основе вычисленных вероятностей генерируются результаты тестирования

$$x_{ij} = \text{Int} (p_{ij} + \text{Rnd}),$$

где $\text{Int} (Y)$ – целая часть числа Y , Rnd – случайное число, равномерно распределенное в интервале $(0; 1)$.

Каждый из 17 уровней знаний и трудностей заданий используется в трехкратной повторности, так что матрица результатов тестирования имеет размерность 51×51 .

Сгенерированные результаты тестирования использовались для вычисления трех видов оценок уровня знаний:

- доля правильных ответов, в соответствие с классической теорией тестирования;
- оценка θ_i модели Раша на основе алгоритма Cohen [3, 7];
- оценка θ_i модели Раша на основе алгоритма максимального правдоподобия.

В качестве примера эти три вида оценок для заданий открытой формы, в которой невозможно угадывание правильного ответа ($c_k = 0$), представлены в табл.1.

Таблица 1

Оценки уровня знаний на основе классической теории и модели Раша

Истин- тин- ный уро- вень знаний	Классическая теория			Модель Раша					
	Доля правиль- ных ответов			Алгоритм Cohen			Алгоритм макси- мального правдопо- добия		
	Повторения			Повторения			Повторения		
-4,0	0,06	0,06	0,12	-3,73	-3,73	-2,70	-4,37	-4,37	-3,64
-3,5	0,08	0,10	0,10	-3,31	-2,98	-2,98	-4,09	-3,86	-3,86
-3,0	0,14	0,24	0,20	-2,46	-1,57	-1,88	-3,43	-2,47	-2,84
-2,5	0,18	0,20	0,20	-2,06	-1,88	-1,88	-3,04	-2,84	-2,84
-2,0	0,27	0,27	0,29	-1,29	-1,29	-1,16	-2,12	-2,12	-1,92
-1,5	0,31	0,29	0,33	-1,03	-1,16	-0,91	-1,75	-1,92	-1,55
-1,0	0,37	0,35	0,39	-0,68	-0,79	-0,56	-1,16	-1,36	-0,95
-0,5	0,41	0,39	0,45	-0,45	-0,56	-0,24	-0,75	-0,95	-0,34
0,0	0,47	0,51	0,47	-0,13	0,08	-0,13	-0,15	0,24	-0,15
0,5	0,57	0,55	0,59	0,40	0,30	0,51	0,75	0,58	0,91
1,0	0,57	0,65	0,59	0,40	0,85	0,51	0,75	1,40	0,91
1,5	0,63	0,65	0,73	0,73	0,85	1,35	1,24	1,40	2,06
2,0	0,73	0,76	0,75	1,35	1,63	1,48	2,06	2,41	2,24
2,5	0,73	0,75	0,84	1,35	1,48	2,31	2,06	2,24	3,15
3,0	0,88	0,86	0,86	2,76	2,52	2,52	3,56	3,35	3,35
3,5	0,88	0,80	0,86	2,76	1,94	2,52	3,56	2,76	3,35
4,0	0,94	0,88	0,96	3,78	2,76	4,36	4,33	3,56	4,68

Точность трех видов оценок уровней знаний сравнивалась с помощью однофакторного дисперсионного анализа, в котором единственным исследуемым фактором являлся уровень знаний, который варьировался на 17 уровнях в трехкратной повторности. После этого проводилось попарное сравнение оценок уровней знаний – всего 136 пар. Чем больше пар, внутри которых уровни значимо отличаются друг от друга (по методу Шеффе), тем выше точность оценок. Для приведенных данных число значимых пар для трех методов равно 102, 92 и 104 соответственно.

Для увеличения общности вывода точность оценок уровня знаний исследуется в зависимости от следующих факторов.

Фактор А, – степень зашумления результатов тестирования, варьируется на 11 уровнях. Уровень a_1 соответствует “нулевому” зашумлению (т.е. результаты тестирования не зашумляются), уровень a_2 – зашумление равно 5%, т.е. 5% случайно выбранных результатов тестирования заменяются случайным образом 0 или 1, зашумление равно для уровня a_3 – 10%, уровня a_{11} – 50%.

Фактор В, – вероятность угадывания правильного ответа. При получении результатов тестирования на основе имитационного моделирования использовались следующие уровни фактора В: $b_1 = 0$, $b_2 = 1/2$, $b_3 = 1/3$, $b_4 = 1/4$, $b_5 = 1/5$, $b_6 = 1/6$, $b_7 = 1/7$, $b_8 = 1/8$, $b_9 = 1/9$, $b_{10} = 1/10$, $b_{11} = 1/14$, $b_{12} = 1/30$, $b_{13} = 1/62$, $b_{14} = 1/126$, $b_{15} = 1/254$. Уровень $b_1 = 0$ соответствует заданиям открытой формы, при использовании которых невозможно угадать правильный ответ. Строго говоря, для обработки именно таких результатов тестирования и предназначена модель Раша. Уровни $b_2 = 1/2$, $b_3 = 1/3$, ..., $b_{10} = 1/10$ соответствуют заданиям с выбором одного правильного ответа с числом дистракторов от 2 до 10 соответственно. Для уменьшения вероятности угадывания правильного ответа целесообразно увеличить число правильных ответов в заданиях, например от 1 до $n - 1$, где n – число дистракторов. Например, при $n = 3$, число возможных ответов N равно $2^3 - 2 = 6$, при $n = 4$ $N = 2^4 - 2 = 14$, при $n = 5$ $N = 2^5 - 2 = 30$, при $n = 6$ $N = 2^6 - 2 = 62$. Именно поэтому были выбраны уровни $b_{11} = 1/14$, $b_{12} = 1/30$, $b_{13} = 1/62$, ..., $b_{15} = 1/254$. Таким образом, фактор В варьируется на 15 уровнях.

Фактор С, – блок-фактор, который варьируется на двух уровнях, т.е. все результаты тестирования, полученные на основе моделирования, повторяются дважды.

Фактор D, – виды оценок уровня знаний испытуемых. Фактор варьируется на трех уровнях, уровень d_1 соответствует оценке уровня знаний по доле правильных ответов. Уровни d_2 и d_3 – соответствуют оценкам на основе модели Раша. Оценка d_2 вычисляется на основе алгоритма Cohen, оценка d_3 – на основе алгоритма максимального правдоподобия.

Откликом Y в данном исследовании является число значимых пар, внутри которых уровни отличаются друг от друга.

Выбор плана эксперимента. Для увеличения точности сравнений алгоритмов вычисления оценок для проведения эксперимента был выбран план с расщепленными делянками, в котором используются целые делянки, расщепленные делянки первого порядка и расщепленные делянки второго порядка [8, 9].

Блоки в этом плане образуют уровни фактора С. Целым делянкам назначаются уровни фактора В. Одна и та же сгенерированная целая делянка (матрица результатов тестирования) используется для всех одиннадцати уровней зашумления фактора А. В результате образуются одиннадцать рас-

щепленных делянок первого порядка. Фактически каждой полученной расщепленной делянке назначен свой уровень фактора А. Поэтому фактор А называется фактором расщепленных делянок. Далее одна и та же расщепленная делянка первого порядка использовалась для вычисления трех видов оценок уровней знаний. Таким образом, фактор D назначался расщепленным делянкам второго порядка. Альтернативой расщеплению является полная рандомизация, т.е. повторное генерирование результатов тестирования для каждой оценки.

На основе выбранного плана сгенерированы результаты тестирования, эти результаты были использованы для вычисления оценок уровней знаний, и затем с помощью однофакторного дисперсионного анализа для каждого алгоритма было определено число значимых пар, внутри которых уровни знаний значимо отличаются друг от друга.

Полученные таким образом экспериментальные данные представлены в табл. 2.

Таблица 2

Результаты попарных сравнений исследуемых уровней знаний

Вероятность угадывания (Фактор В)	Доля зашумленных данных (Фактор А)	Виды оценок уровня знаний (Фактор D)					
		Доля правильных ответов Р (d ₁)		Алгоритм Cohen (d ₂)		Алгоритм макс. правдоподобия (d ₃)	
		Блок-фактор		Блок-фактор		Блок-фактор	
		c ₁	c ₂	c ₁	c ₂	c ₁	c ₂
1	2	3	4	5	6	7	8
0 (b ₁)	0 % (a ₁)	112	104	81	61	107	95
	5 % (a ₂)	93	99	75	53	87	79
	10 % (a ₃)	90	93	70	82	77	88
	15 % (a ₄)	88	86	75	76	81	79
	20 % (a ₅)	78	80	76	65	76	69
	25 % (a ₆)	74	65	55	53	60	57
	30 % (a ₇)	66	53	62	45	63	49
	35 % (a ₈)	51	34	43	30	43	30
	40 % (a ₉)	54	36	48	30	48	31
	45 % (a ₁₀)	30	23	28	17	28	18
	50 % (a ₁₁)	36	22	33	17	33	18
1/2 (b ₂)	0 % (a ₁)	11	50	23	43	22	45
	5 % (a ₂)	3	30	15	17	15	17
	10 % (a ₃)	8	33	14	29	14	30
	15 % (a ₄)	7	31	22	28	22	28
	20 % (a ₅)	3	29	2	25	2	25

	2	3	4	5	6	7	8
	25 % (a ₆)	0	27	0	24	0	24
	30 % (a ₇)	1	3	3	2	3	2
	35 % (a ₈)	0	0	0	0	0	0
	40 % (a ₉)	0	5	0	6	0	6
	45 % (a ₁₀)	0	0	0	0	0	0
	50 % (a ₁₁)	0	3	0	2	0	2
1/3 (b ₃)	0 % (a ₁)	43	67	40	62	40	64
	5 % (a ₂)	36	62	44	62	42	64
	10 % (a ₃)	30	56	21	52	23	55
	15 % (a ₄)	40	41	34	23	35	23
	20 % (a ₅)	21	49	17	36	17	37
	25 % (a ₆)	10	29	14	26	14	26
	30 % (a ₇)	15	32	19	27	19	29
	35 % (a ₈)	2	24	3	19	3	19
	40 % (a ₉)	1	7	3	6	3	6
	45 % (a ₁₀)	15	17	9	15	10	15
	50 % (a ₁₁)	1	7	0	4	1	4
1/4 (b ₄)	0 % (a ₁)	62	77	49	66	51	72
	5 % (a ₂)	52	70	54	63	54	64
	10 % (a ₃)	47	64	42	57	43	60
	15 % (a ₄)	50	63	43	51	44	57
	20 % (a ₅)	44	54	35	48	35	49
	25 % (a ₆)	28	40	19	36	20	37
	30 % (a ₇)	18	39	18	39	18	39
	35 % (a ₈)	17	31	13	27	15	27
	40 % (a ₉)	20	31	19	28	19	29
	45 % (a ₁₀)	7	6	9	6	9	6
	50 % (a ₁₁)	2	6	3	4	3	4
1/5 (b ₅)	0 % (a ₁)	72	80	56	68	59	74
	5 % (a ₂)	66	66	59	57	60	62
	10 % (a ₃)	61	76	43	57	49	62
	15 % (a ₄)	52	53	40	50	45	50
	20 % (a ₅)	48	64	39	44	41	48
	25 % (a ₆)	56	53	45	46	48	47
	30 % (a ₇)	37	38	34	28	35	28
	35 % (a ₈)	48	23	45	17	46	18
	40 % (a ₉)	19	15	16	10	16	11
	45 % (a ₁₀)	8	17	7	10	7	10
	50 % (a ₁₁)	14	12	13	11	13	11

Таблица 2 (продолжение)

1	2	3	4	5	6	7	8
1/6 (b ₆)	0 % (a ₁)	87	89	64	68	74	81
	5 % (a ₂)	77	75	43	53	48	63
	10 % (a ₃)	71	77	41	68	47	72
	15 % (a ₄)	59	68	50	64	52	63
	20 % (a ₅)	55	72	43	62	44	63
	25 % (a ₆)	54	62	46	45	49	46
	30 % (a ₇)	39	33	39	27	39	28
	35 % (a ₈)	33	53	32	43	32	44
	40 % (a ₉)	37	37	33	29	35	29
	45 % (a ₁₀)	37	46	27	45	27	45
	50 % (a ₁₁)	14	20	14	18	14	18
1/7 (b ₇)	0 % (a ₁)	87	94	76	69	78	82
	5 % (a ₂)	84	88	63	74	67	80
	10 % (a ₃)	65	82	45	71	51	74
	15 % (a ₄)	66	74	45	71	52	72
	20 % (a ₅)	74	54	63	48	64	52
	25 % (a ₆)	43	39	31	32	31	32
	30 % (a ₇)	42	59	35	55	39	55
	35 % (a ₈)	48	50	46	49	46	49
	40 % (a ₉)	36	12	31	13	34	13
	45 % (a ₁₀)	14	13	10	13	10	13
	50 % (a ₁₁)	21	17	20	13	20	13
1/8 (b ₈)	0 % (a ₁)	92	93	71	67	81	79
	5 % (a ₂)	91	85	72	56	78	65
	10 % (a ₃)	72	81	71	60	71	68
	15 % (a ₄)	62	55	50	49	54	52
	20 % (a ₅)	65	53	64	40	64	46
	25 % (a ₆)	41	53	35	48	34	50
	30 % (a ₇)	64	59	56	55	59	50
	35 % (a ₈)	39	44	34	39	34	39
	40 % (a ₉)	48	24	43	22	43	22
	45 % (a ₁₀)	43	26	41	22	41	22
	50 % (a ₁₁)	21	25	22	21	22	21
1/9 (b ₉)	0 % (a ₁)	94	92	72	67	81	82
	5 % (a ₂)	84	84	70	57	75	70
	10 % (a ₃)	74	80	53	71	61	76

	2	3	4	5	6	7	8
	15 % (a ₄)	75	70	67	24	71	35
	20 % (a ₅)	54	54	44	45	46	47
	25 % (a ₆)	55	58	43	50	45	51
	30 % (a ₇)	42	44	37	38	38	37
	35 % (a ₈)	53	46	44	40	48	42
	40 % (a ₉)	26	39	26	32	26	34
	45 % (a ₁₀)	20	25	19	20	19	20
	50 % (a ₁₁)	25	18	26	17	26	17
1/10 (b ₁₀)	0 % (a ₁)	93	93	72	68	81	84
	5 % (a ₂)	84	79	72	65	78	70
	10 % (a ₃)	83	75	69	71	72	76
	15 % (a ₄)	82	73	79	60	80	62
	20 % (a ₅)	63	84	61	78	61	81
	25 % (a ₆)	54	65	49	57	49	60
	30 % (a ₇)	54	48	50	43	50	43
	35 % (a ₈)	42	55	42	51	43	53
	40 % (a ₉)	21	43	17	37	17	37
	45 % (a ₁₀)	30	21	29	20	30	20
	50 % (a ₁₁)	18	32	13	25	13	26
1/14 (b ₁₁)	0 % (a ₁)	99	91	70	69	83	82
	5 % (a ₂)	94	89	83	65	87	80
	10 % (a ₃)	71	76	47	62	58	67
	15 % (a ₄)	73	74	64	62	66	63
	20 % (a ₅)	71	60	67	52	68	54
	25 % (a ₆)	73	80	66	74	68	76
	30 % (a ₇)	54	62	42	48	45	49
	35 % (a ₈)	37	56	30	47	30	48
	40 % (a ₉)	43	44	41	40	41	41
	45 % (a ₁₀)	20	26	19	22	19	23
	50 % (a ₁₁)	36	33	36	30	36	30
1/30 (b ₁₂)	0 % (a ₁)	102	99	75	69	95	87
	5 % (a ₂)	88	84	73	56	79	70
	10 % (a ₃)	85	89	72	77	78	81
	15 % (a ₄)	67	71	41	60	50	62
	20 % (a ₅)	72	72	64	59	70	63
	25 % (a ₆)	74	42	58	35	63	36
	30 % (a ₇)	51	45	45	25	46	31
	35 % (a ₈)	40	50	33	42	34	44

	2	3	4	5	6	7	8
	40 % (a ₉)	29	39	21	37	22	37
	45 % (a ₁₀)	39	48	34	45	34	46
	50 % (a ₁₁)	25	45	26	37	26	37
1/62 (b ₁₃)	0 % (a ₁)	101	102	77	72	98	90
	5 % (a ₂)	94	90	65	79	77	84
	10 % (a ₃)	79	84	64	58	72	71
	15 % (a ₄)	87	80	79	64	82	70
	20 % (a ₅)	83	66	56	56	66	58
	25 % (a ₆)	79	82	68	72	71	76
	30 % (a ₇)	57	59	50	49	50	52
	35 % (a ₈)	60	52	47	46	50	47
	40 % (a ₉)	53	38	52	31	52	32
	45 % (a ₁₀)	23	43	20	37	20	37
	50 % (a ₁₁)	55	27	53	21	53	21
	50 % (a ₁₁)	21	45	18	35	18	35
1/254 (b ₁₅)	0 % (a ₁)	108	103	77	73	99	99
	5 % (a ₂)	98	95	81	63	89	78
	10 % (a ₃)	99	92	85	82	94	87
	15 % (a ₄)	85	90	58	78	70	81
	20 % (a ₅)	78	67	58	56	67	58
	25 % (a ₆)	50	55	40	46	42	48
	30 % (a ₇)	56	55	48	50	50	51
	35 % (a ₈)	46	61	40	55	41	57
	40 % (a ₉)	47	40	42	28	43	29
	45 % (a ₁₀)	39	33	30	28	31	28
	50 % (a ₁₁)	19	18	17	15	17	15

Выбор модели представления экспериментальных данных.

В соответствие с выбранным планом для анализа полученных результатов сравнений (табл. 2) используется модель с расщепленными делянками второго порядка.

$$y_{ijkl} = \mu + \beta_i + \gamma_j + \varepsilon_{\text{ош1}} + \alpha_k + (\alpha\gamma)_{kj} + \varepsilon_{\text{ош2}} + \\ + \delta_l + (\alpha\delta)_{kl} + (\gamma\delta)_{jl} + (\alpha\gamma\delta)_{kjl} + \varepsilon_{\text{ош3}}$$

где y_{ijkl} - отклик, полученный при i -ом уровне фактора В ($i = 1, 15$), j -ом уровне фактора С ($j = 1, 2$), k -ом уровне фактора А ($k = 1, 11$), l -ом уровне фактора D ($l = 1, 3$);

μ - генеральное среднее;

β_i – эффект i -ого уровня фактора В;
 γ_j – эффект j -ого уровня фактора С;
 $\varepsilon_{\text{ош1}}$ – ошибка для уровней факторов, назначаемых целым делянкам;
 α_k – эффект k -ого уровня фактора А;
 $(\alpha\beta)_{ki}$ – эффект взаимодействия k -ого уровня фактора А с i -ым уровнем фактора В;
 $\varepsilon_{\text{ош2}}$ – ошибка для уровней факторов, назначаемых расщепленным делянкам первого порядка;
 δ_l – эффект l -ого уровня фактора D;
 $(\alpha\delta)_{kl}$ – эффект взаимодействия k -ого уровня фактора А с l -ым уровнем фактора D;
 $(\beta\delta)_{il}$ – эффект взаимодействия i -ого уровня фактора В с l -ым уровнем фактора D;
 $(\alpha\beta\delta)_{kil}$ – эффект взаимодействия k -ого уровня фактора А, i -ого уровня фактора В и l -ого уровня фактора D;
 $\varepsilon_{\text{ош3}}$ – ошибка для уровней факторов, назначаемых расщепленным делянкам второго порядка.

В этой модели ошибка для уровней факторов, называемых целыми делянками, формируется на основе взаимодействия ВС, ошибка для расщепленных делянок первого порядка – на основе взаимодействий FC и ABC, и, наконец, ошибка для расщепленных делянок второго порядка – на основе взаимодействий CD, ACD, BCD и ABCD [8].

Статистический анализ результатов парных сравнений

Поскольку наиболее важный исследуемый фактор является качественным в качестве метода статистической обработки выбран дисперсионный анализ.

Результаты дисперсионного анализа в соответствие с выбранной моделью представлены в табл. 3.

Таблица 3

Результаты дисперсионного анализа числа значимых пар
по плану с расщепленными делянками

Источник дисперсии	Знаменатель $F_{\text{эксп}}$	Сумма квадратов	Степени свободы	Средний квадрат	$F_{\text{эксп}}$	$F_{\text{табл}} \alpha=0,05$	$\alpha_{\text{эксп}}$
С		938,7	1	938,7	1,33	4,60	0,00
В	Ош1	191460,0	14	13676,0	19,45*	2,47	0,00
Ош1	Ош2	9844,6	14	703,2	3,68*	1,73	0,00

A	Ош2	339150,0	10	33915,0	177,58*	1,90	0,00
AB	Ош2	37001,0	140	264,3	1,38*	1,20	0,00
Ош2	Ош3	28648,1	150	191,0	18,53*	1,10	0,00
D	Ош3	13869,0	2	6934,6	672,9*	3,07	0,00
AD	Ош3	5609,8	20	280,5	27,22*	1,60	0,00
BD	Ош3	2577,3	28	92,0	8,93*	1,51	0,00
ABD	Ош3	3527,4	280	12,6	1,22*	1,15	0,04
Ош3		3401,0	330	10,3			
Всего		636030,0	989				

Символом * в табл. 3 отмечены значимые источники дисперсии. Второй столбец в таблице дисперсионного анализа введен для того, чтобы указывать, на фоне какой ошибки (среднего квадрата) проверяется тот или иной источник дисперсии.

Интерпретация результатов статистического анализа

Проинтерпретируем полученные результаты.

Таблица 4

Результаты парных сравнений уровней фактора D

Уровень фактора	Среднее значение	Парный контраст
d ₁	52,7	1 2* 1 3*
d ₂	43,6	2 3*
d ₃	46,9	

1. Прежде всего, отметим, что, как и следовало ожидать, средние квадраты ошибок уменьшаются с увеличением расщепления $MS_{\text{ош1}} = 703,2 > MS_{\text{ош2}} = 191,0 > MS_{\text{ош3}} = 10,3$. Следовательно, с наибольшей точностью сравниваются уровни наиболее важного для нас фактора – фактора D, т.е. алгоритмы вычисления оценок уровней знаний.

2. Фактор D (сравниваемые алгоритмы оценки уровня знаний) значим, поскольку $F_{\text{эксп}} = 672,9 > F_{\text{табл}} = 3,07$. Это означает, что среди уровней фактора D существуют такие, которые значимо отличаются друг от друга. Для того чтобы определить, какие уровни значимо отличаются друг от друга, ниже приведены результаты парных сравнений уровней этого фактора.

Полуширина 95%-ого доверительного интервала для уровней равна 0,35, а для пар уровней (по методу множественных сравнений Шеффе) 0,62. Из приведенных сравнений следует, что в среднем для всех уровней исследованных факторов все виды оценок значимо отличаются друг от друга. Наибольшее число значимых пар обеспечивает оценка уровней знаний по доле правильных ответов (52,7), наименьшее – алгоритм Кохена (43,6), алгоритм максимального правдоподобия занимает промежуточное положение (46,9).

3. Взаимодействие AD значимо ($F_{\text{эксп}} = 27,22 > F_{\text{табл}} = 1,60$). Это означает, что эффект фактора A зависит от того, на каком уровне находится фактор C в среднем для всех уровней факторов B и C. Таким образом, число значимых пар зависит от степени зашумления данных тестирования и вида алгоритма получения оценки уровня знаний. Соответствующие средние приведены в табл. 5.

Таблица 5

Число значимых пар в зависимости от степени зашумления результатов тестирования и вида оценки уровня знаний

Вид оценки	Степень зашумления результатов тестирования										
	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
Доля правильных ответов Р	86,87	77,90	72,13	66,60	60,77	52,57	45,23	40,43	30,70	25,03	21,27
Алгоритм Cohen	65,77	61,77	58,93	55,27	51,37	44,20	39,60	35,43	27,10	21,73	18,80
Алгоритм макс. правдоподобия	78,60	68,83	64,23	58,77	53,90	45,97	40,67	36,27	27,57	21,97	18,90

Эта зависимость представлена на рис. 1.

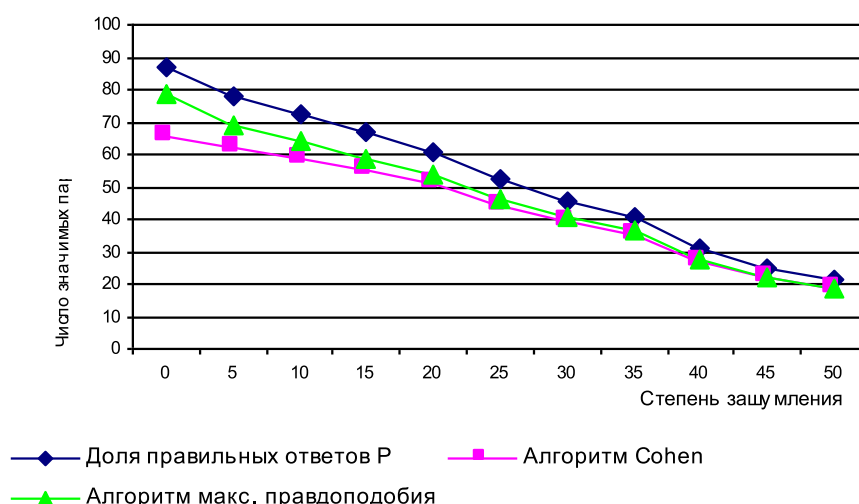


Рис. 1 Число значимых пар в зависимости от степени зашумления результатов тестирования и вида оценки уровня знаний

Как и следовало ожидать, с увеличением степени зашумления число значимых пар уменьшается, кроме того, уменьшается различие между исследуемыми алгоритмами.

4. Взаимодействие BD значимо ($F_{\text{эксп}} = 8,93 > F_{\text{табл}} = 1,51$). Следовательно, эффект фактора B зависит от того, на каком уровне находится фактор D (в среднем для всех уровней факторов A и C). Таким образом, число значимых пар зависит от вероятности угадывания правильного ответа и вида

алгоритма получения оценки уровня знаний. Соответствующие средние приведены в табл. 6.

Таблица 6

Число значимых пар в зависимости от вероятности угадывания правильного ответа и алгоритма получения оценки уровня знаний

Вероятность угадывания ответа	Вид оценки		
	Доля правильных ответов Р	Алгоритм Cohen	Алгоритм макс. правдоподобия
0	66,68	53,41	59,82
1/2	11,09	11,59	11,68
1/3	27,50	24,36	24,96
1/4	37,64	33,14	34,32
1/5	44,46	36,14	38,18
1/6	54,32	43,36	46,05
1/7	52,82	44,23	46,68
1/8	56,18	47,18	49,77
1/9	55,09	43,73	47,59
1/10	58,73	51,27	53,91
1/14	61,91	51,64	55,18
1/30	61,64	49,27	54,14
1/62	67,91	55,27	60,41
1/126	69,09	57,64	62,59
1/254	65,18	52,27	57,91

Эта зависимость представлена на рис. 2.

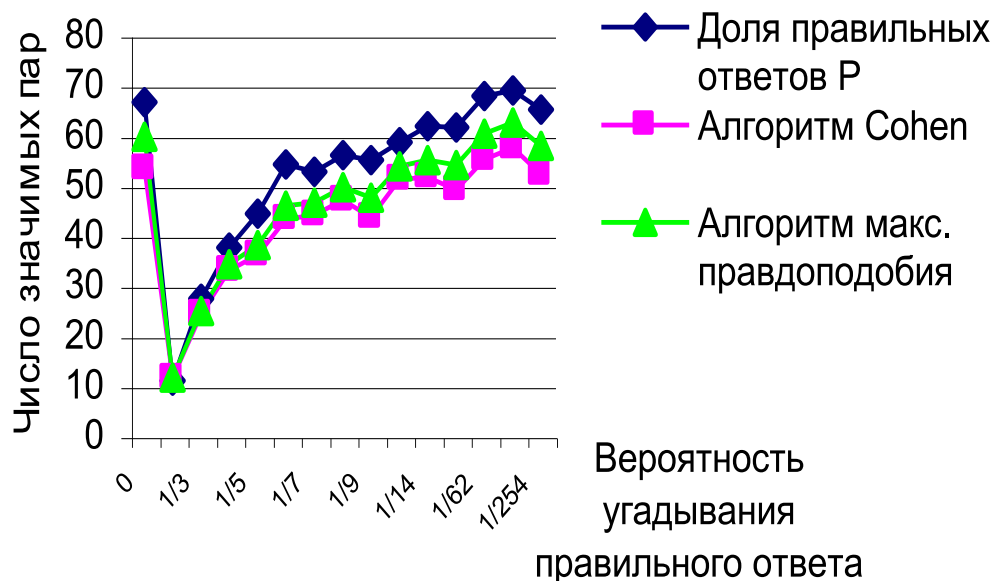


Рис. 2 Число значимых пар в зависимости от вероятности угадывания правильного ответа и вида оценки уровня знаний

Как и следовало ожидать, дифференцирующая способность тестовой системы (число значимых пар) сильно зависит от вероятности правильного ответа. И только при малой вероятности угадывания правильного ответа ($P < 1/30$) дифференцирующая способность исследуемых алгоритмов такая же, как и при отсутствии угадывания.

5. Взаимодействие ABD значимо ($F_{\text{эксп}} = 1,22 > F_{\text{табл}} = 1,15$). Это означает, что эффект фактора D зависит от того, на каких уровнях находятся факторы A и B в среднем для всех уровней фактора C. Таким образом, число значимых пар зависит от степени зашумления результатов тестирования, вероятности угадывания правильного ответа и вида алгоритма получения оценки уровня знаний. Соответствующие средние приведены в табл. 7.

Таблица 7

Число значимых пар в зависимости от степени зашумления результатов тестирования, вероятности угадывания правильного ответа и алгоритма получения оценки уровня знаний

Вид оценки	Вероятность угадывания ответа	Степень зашумления результатов тестирования										
		0	5	10	15	20	25	30	35	40	45	50
Доля правильных ответов Р	0	108,0	96,0	91,5	87,0	79,0	69,5	59,5	42,5	45,0	26,5	29,0
	1/2	30,5	16,5	20,5	19,0	16,0	13,5	2,0	0,0	2,5	0,0	1,5
	1/3	55,0	49,0	43,0	40,5	35,0	19,5	23,5	13,0	4,0	16,0	4,0
	1/4	69,5	61,0	55,5	56,5	49,0	34,0	28,5	24,0	25,5	6,5	4,0
	1/5	76,0	66,0	68,5	52,5	56,0	54,5	37,5	35,5	17,0	12,5	13,0
	1/6	88,0	76,0	74,0	63,5	63,5	58,0	36,0	43,0	37,0	41,5	17,0
	1/7	90,5	86,0	73,5	70,0	64,0	41,0	50,5	49,0	24,0	13,5	19,0
	1/8	92,5	88,0	76,5	58,5	59,0	47,0	61,5	41,5	36,0	34,5	23,0
	1/9	93,0	84,0	77,0	72,5	54,0	56,5	43,0	49,5	32,5	22,5	21,5
	1/10	93,0	81,5	79,0	77,5	73,5	59,5	51,0	48,5	32,0	25,5	25,0
	1/14	95,0	91,5	73,5	73,5	65,5	76,5	58,0	46,5	43,5	23,0	34,5
	1/30	100,5	86,0	87,0	69,0	72,0	58,0	48,0	45,0	34,0	43,5	35,0
	1/62	101,5	92,0	81,5	83,5	74,5	80,5	58,0	56,0	45,5	33,0	41,0
	1/126	104,5	98,5	85,5	88,0	78,0	68,0	66,0	59,0	38,5	41,0	33,0
	1/254	105,5	96,5	95,5	87,5	72,5	52,5	55,5	53,5	43,5	36,0	18,5
Алгоритм Cohen	0	71,0	64,0	76,0	75,5	70,5	54,0	53,5	36,5	39,0	22,5	25,0
	1/2	33,0	16,0	21,5	25,0	13,5	12,0	2,5	0,0	3,0	0,0	1,0
	1/3	51,0	53,0	36,5	28,5	26,5	20,0	23,0	11,0	4,5	12,0	2,0
	1/4	57,5	58,5	49,5	47,0	41,5	27,5	28,5	20,0	23,5	7,5	3,5
	1/5	62,0	58,0	50,0	45,0	41,5	45,5	31,0	31,0	13,0	8,5	12,0
	1/6	66,0	48,0	54,5	57,0	52,5	45,5	33,0	37,5	31,0	36,0	16,0
	1/7	72,5	68,5	58,0	58,0	55,5	31,5	45,0	47,5	22,0	11,5	16,5
	1/8	69,0	64,0	65,5	49,5	52,0	41,5	55,5	36,5	32,5	31,5	21,5
	1/9	69,5	63,5	62,0	45,5	44,5	46,5	37,5	42,0	29,0	19,5	21,5
	1/10	70,0	68,5	70,0	69,5	69,5	53,0	46,5	46,5	27,0	24,5	19,0
	1/14	69,5	74,0	54,5	63,0	59,5	70,0	45,0	38,5	40,5	20,5	33,0
	1/30	72,0	64,5	74,5	50,5	61,5	46,5	35,0	37,5	29,0	39,5	31,5
	1/62	74,5	72,0	61,0	71,5	56,0	70,0	49,5	46,5	41,5	28,5	37,0
	1/126	74,0	82,0	67,0	75,5	69,0	56,5	59,5	53,0	36,0	35,0	26,5
	1/254	75,0	72,0	83,5	68,0	57,0	43,0	49,0	47,5	35,0	29,0	16,0

Таблица 7 (продолжение)

Вид оценки	Вероятность угадывания ответа	Степень зашумления результатов тестирования										
		0	5	10	15	20	25	30	35	40	45	50
Алгоритм максимального правдоподобия	0	101,0	83,0	82,5	80,0	72,5	58,5	56,0	36,5	39,5	23,0	25,5
	1/2	33,5	16,0	22,0	25,0	13,5	12,0	2,5	0,0	3,0	0,0	1,0
	1/3	52,0	53,0	39,0	29,0	27,0	20,0	24,0	11,0	4,5	12,5	2,5
	1/4	61,5	59,0	51,5	50,5	42,0	28,5	28,5	21,0	24,0	7,5	3,5
	1/5	66,5	61,0	55,5	47,5	44,5	47,5	31,5	32,0	13,5	8,5	12,0
	1/6	77,5	55,5	59,5	57,5	53,5	47,5	33,5	38,0	32,0	36,0	16,0
	1/7	80,0	73,5	62,5	62,0	58,0	31,5	47,0	47,5	23,5	11,5	16,5
	1/8	80,0	71,5	69,5	53,0	55,0	42,0	54,5	36,5	32,5	31,5	21,5
	1/9	81,5	72,5	68,5	53,0	46,5	48,0	37,5	45,0	30,0	19,5	21,5
	1/10	82,5	74,0	74,0	71,0	71,0	54,5	46,5	48,0	27,0	25,0	19,5
	1/14	82,5	83,5	62,5	64,5	61,0	72,0	47,0	39,0	41,0	21,0	33,0
	1/30	91,0	74,5	79,5	56,0	66,5	49,5	38,5	39,0	29,5	40,0	31,5
	1/62	94,0	80,5	71,5	76,0	62,0	73,5	51,0	48,5	42,0	28,5	37,0
	1/126	96,5	91,5	75,0	81,0	73,0	59,5	61,5	53,0	35,5	35,5	26,5
	1/254	99,0	83,5	90,5	75,5	62,5	45,0	50,5	49,0	36,0	29,5	16,0

Эта зависимость для четырех значений вероятности угадывания 0, 1/2, 1/5 и 1/30 представлена на рис. 3 – 5.

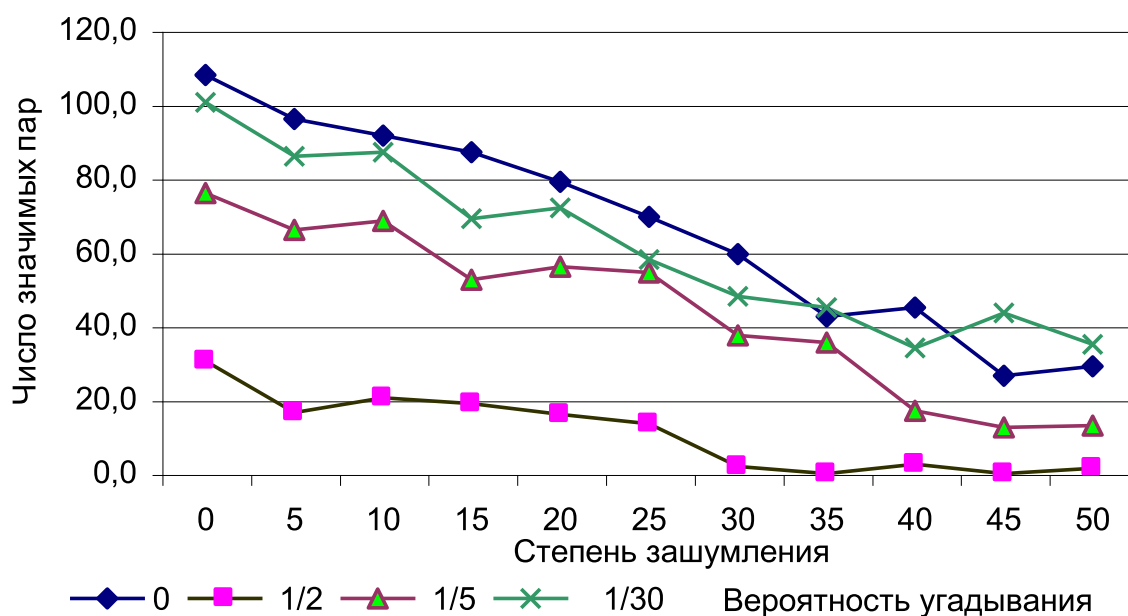


Рис. 3 Число значимых пар в зависимости от степени зашумления результатов тестирования и вероятности угадывания правильного ответа для доли правильных ответов p .

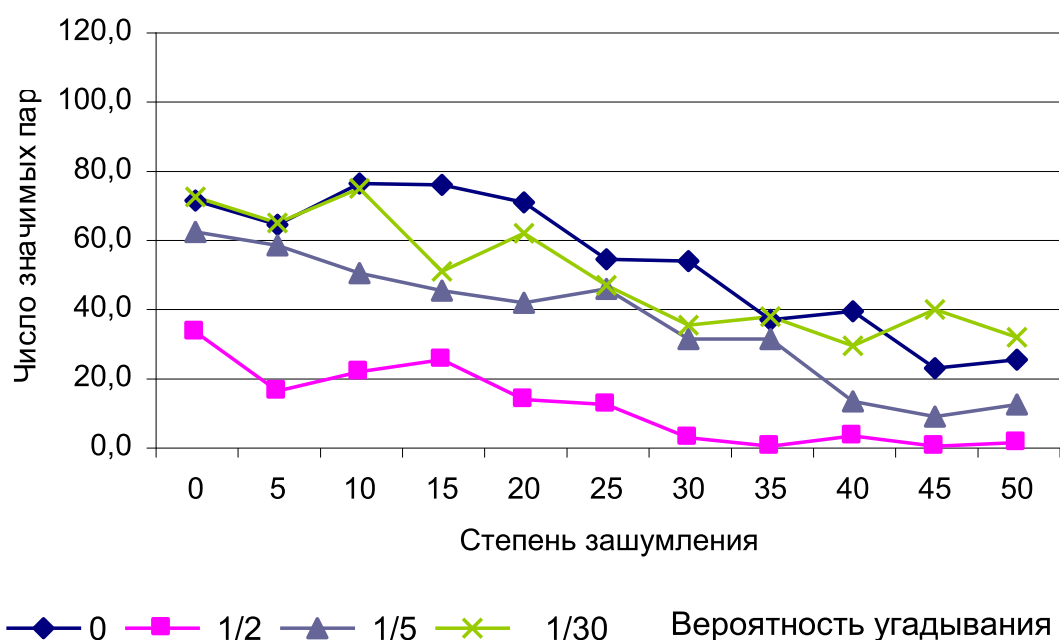


Рис. 4 Число значимых пар в зависимости от степени зашумления результатов тестирования и вероятности угадывания правильного ответа для алгоритма Cohen

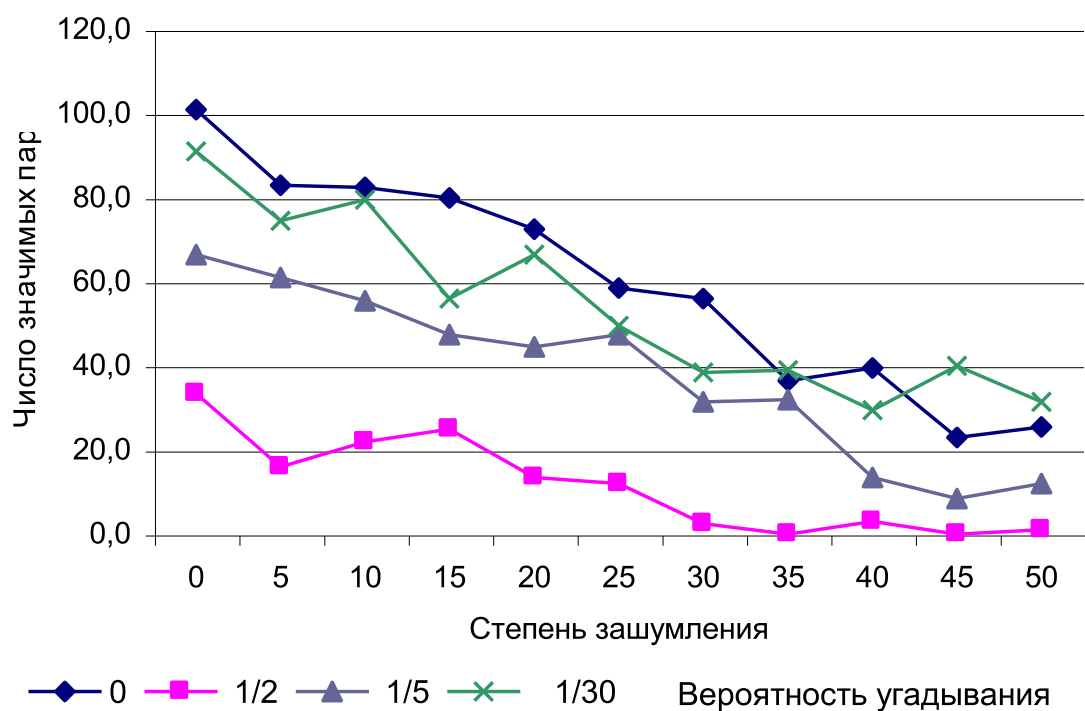


Рис. 5 Число значимых пар в зависимости от степени зашумления результатов тестирования и вероятности угадывания правильного ответа для алгоритма максимального правдоподобия

6. Фактор В значим ($F_{\text{эксп}} = 19,45 > F_{\text{табл}} = 2,47$). Следовательно, в среднем для всех уровней факторов А, С и D (всех значениях зашумления результатов тестирования, для трех видов оценок уровня знаний) уровни фактора В (вероятности угадывания правильного ответа) значимо отличаются друг от друга по показателю Y . Причем, эмпирический уровень значимости менее 0,001.

Для определения того, какие уровни фактора В отличаются друг от друга, в табл. 7 приведены результаты парных сравнений.

Таблица 7

Результаты парных сравнений уровней фактора В

Уровень фактора	Среднее значение	Парный контраст
b_1	60,0	1 2* 1 3* 1 4* 1 5* 1 6* 1 7* 1 8* 1 9* 1 10 1 11 1 12 1 13 1 14 1 15
b_2	11,5	2 3* 2 4* 2 5* 2 6* 2 7* 2 8* 2 9* 2 10* 2 11* 2 12* 2 13* 2 14* 2 15*
b_3	25,6	3 4* 3 5* 3 6* 3 7* 3 8* 3 9* 3 10* 3 11* 3 12* 3 13* 3 14* 3 15*
b_4	35,0	4 5 4 6* 4 7* 4 8* 4 9* 4 10* 4 11* 4 12* 4 13* 4 14* 4 15*
b_5	39,6	5 6* 5 7* 5 8* 5 9* 5 10* 5 11* 5 12* 5 13* 5 14* 5 15*
b_6	47,9	6 7 6 8 6 9 6 10 6 11* 6 12 6 13* 6 14* 6 15*
b_7	47,9	7 8 7 9 7 10 7 11* 7 12 7 13* 7 14* 7 15*
b_8	51,0	8 9 8 10 8 11 8 12 8 13* 8 14* 8 15
b_9	48,8	9 10 9 11 9 12 9 13* 9 14* 9 15*
b_{10}	54,6	10 11 10 12 10 13 10 14* 10 15
b_{11}	56,2	11 12 11 13 11 14 11 15
b_{12}	55,0	12 13 12 14* 12 15
b_{13}	61,2	13 14 13 15
b_{14}	63,1	14 15
b_{15}	58,5	

Полуширина 95%-го доверительного интервала для уровней равна 2,2, а для пар уровней (по методу множественных сравнений Шеффе) 7,8.

7. Фактор А значим ($F_{\text{эксп}} = 177,58 > F_{\text{табл}} = 1,90$). Это означает, что в среднем для всех уровней факторов В, С и D (всех вероятностях угадывания правильного ответа, для трех видов оценок уровня знаний) уровни фактора А (степень зашумления результатов тестирования) значимо отличаются друг от друга по показателю Y . Причем, эмпирический уровень значимости менее 0,001.

Для определения того, какие уровни фактора В отличаются друг от друга, в табл. 8 приведены результаты парных сравнений.

Таблица 8

Результаты парных сравнений уровней фактора А

Уровень фактора	Среднее значение	Парный контраст
a ₁	77,1	1 2* 1 3* 1 4* 1 5* 1 6* 1 7* 1 8* 1 9* 1 10* 1 11*
a ₂	69,5	2 3 2 4* 2 5* 2 6* 2 7* 2 8* 2 9* 2 10* 2 11*
a ₃	65,1	3 4 3 5* 3 6* 3 7* 3 8* 3 9* 3 10* 3 11*
a ₄	60,2	4 5 4 6* 4 7* 4 8* 4 9* 4 10* 4 11*
a ₅	55,3	5 6* 5 7* 5 8* 5 9* 5 10* 5 11*
a ₆	47,6	6 7 6 8* 6 9* 6 10* 6 11*
a ₇	41,8	7 8 7 9* 7 10* 7 11*
a ₈	37,4	8 9* 8 10* 8 11*
a ₉	28,5	9 10 9 11*
a ₁₀	22,9	10 11
a ₁₁	19,7	

Полуширина 95%-го доверительного интервала для уровней равна 1,9, а для пар уровней (по методу множественных сравнений Шеффе) 5,9.

8. Взаимодействие АВ значимо ($F_{\text{эксп}} = 1,38 > F_{\text{табл}} = 1,20$). Это означает, что эффект фактора А зависит от того, на каком уровне находится фактор В в среднем для всех уровней факторов С и D. Таким образом, число значимых пар зависит от степени зашумления результатов тестирования и вероятности угадывания правильного ответа. Соответствующие средние приведены в табл. 9.

Таблица 9

Число значимых пар в зависимости от степени зашумления результатов тестирования и вероятности угадывания правильного ответа

Вероятность угадывания	Степень зашумления результатов тестирования										
	0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
0	93,33	81,00	83,33	80,83	74,00	60,67	56,33	38,50	41,17	24,00	26,50
1/2	32,33	16,17	21,33	23,00	14,33	12,50	2,33	0,00	2,83	0,00	1,17
1/3	52,67	51,67	39,50	32,67	29,50	19,83	23,50	11,67	4,33	13,50	2,83
1/4	62,83	59,50	52,17	51,33	44,17	30,00	28,50	21,67	24,33	7,17	3,67
1/5	68,17	61,67	58,00	48,33	47,33	49,17	33,33	32,83	14,50	9,83	12,33
1/6	77,17	59,83	62,67	59,33	56,50	50,33	34,17	39,50	33,33	37,83	16,33
1/7	81,00	76,00	64,67	63,33	59,17	34,67	47,50	48,00	23,17	12,17	17,33
1/8	80,50	74,50	70,50	53,67	55,33	43,50	57,17	38,17	33,67	32,50	22,00
1/9	81,33	73,33	69,17	57,00	48,33	50,33	39,33	45,50	30,50	20,50	21,50
1/10	81,83	74,67	74,33	72,67	71,33	55,67	48,00	47,67	28,67	25,00	21,17
1/14	82,33	83,00	63,50	67,00	62,00	72,83	50,00	41,33	41,67	21,50	33,50
1/30	87,83	75,00	80,33	58,50	66,67	51,33	40,50	40,50	30,83	41,00	32,67
1/62	90,00	81,50	71,33	77,00	64,17	74,67	52,83	50,33	43,00	30,00	38,33
1/126	91,67	90,67	75,83	81,50	73,33	61,33	62,33	55,00	36,67	37,17	28,67
1/254	93,17	84,00	89,83	77,00	64,00	46,83	51,67	50,00	38,17	31,50	16,83

Для наглядности на рис.3 показана дифференцирующая способность системы тестирования (в среднем для всех алгоритмов) только для четырех значений вероятности угадывания 0, 1/2, 1/5 и 1/30.

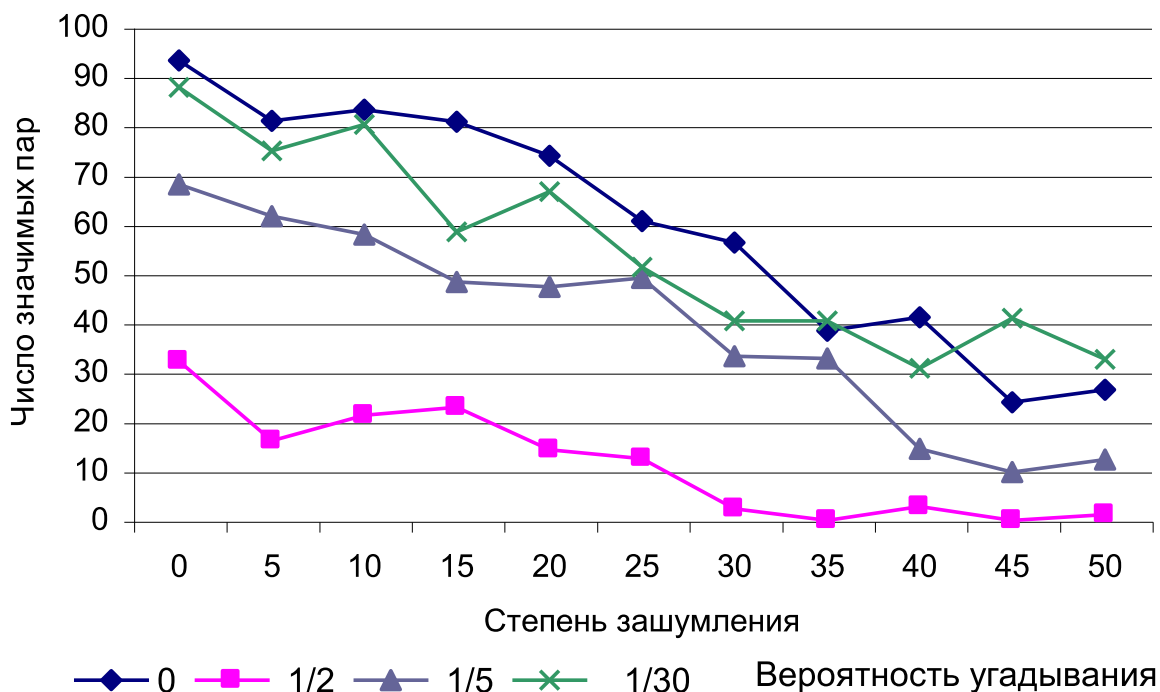


Рис. 6 Число значимых пар в зависимости от степени зашумления результатов тестирования и вероятности угадывания правильного ответа

7. Ошибка для расщепленных делянок второго порядка значимо меньше ошибки для расщепленных делянок первого порядка $F_{0ш2} = 18,53 > F_{0ш3} = 1,10$. Это свидетельствует о том, что уровни фактора D и взаимодействия этого фактора с факторами A и B оцениваются с большей точностью, чем уровни фактора A и взаимодействие AB.

Оценка эффекта расщепления экспериментальных делянок

Особенности планирования на расщепленных делянках подробно рассмотрены в [8]. Одним из наиболее важных преимуществ этих планов является то, что они позволяют с большей точностью, чем в полностью рандомизированных или блочных планах оценить уровни основных факторов.

Оценим выигрыш в точности оценивания для уровней основного в данном исследовании фактора D – алгоритмов вычисления оценок уровней знаний. Прежде всего, на основе результатов дисперсионного анализа (табл. 3) вычислим приведенную ошибку для проверки значимости фактора D в полностью рандомизированном плане.

Приведенная ошибка MS_{прив} для полноблочного плана включает следующие источники дисперсии [8]: Ош1 с 14 степенями свободы, Ош2 с 150 степенями свободы и Ош3 с $\nu_{\text{ош3}} = 330$ степенями свободы. В результате получаем, что MS_{прив} = 84,8 с $\nu_{\text{прив}} = 494$ степенями свободы.

Относительная эффективность RE (relative efficiency) реализованного плана по сравнению с полностью рандомизированным равна

$$RE = \frac{(\nu_{\text{ош3}} + 1)}{(\nu_{\text{ош3}} + 3)MS_{\text{ош3}}} : \frac{(\nu_{\text{прив}} + 1)}{(\nu_{\text{прив}} + 3)MS_{\text{прив}}} =$$

$$(331 \cdot 497 \cdot 84,8) / (333 \cdot 10,3 \cdot 495) \approx 8,2.$$

Таким образом, реализованный план с расщепленными делянками обеспечивает в 8,2 раза большую точность сравнения оценок алгоритмов вычисления уровней знаний по сравнению с полностью рандомизированным планом.

Выводы

На основе проведенного исследования можно сделать следующие выводы.

1. Наибольшее число значимых пар в среднем для уровней всех остальных факторов обеспечивает оценка уровней знаний по доле правильных ответов (52,7), наименьшее – алгоритм Кохена (43,6), алгоритм максимального правдоподобия занимает промежуточное положение (46,9). Следовательно, традиционная система тестирования обладает наибольшей разрешающей способностью. Этот несколько неожиданный результат можно объяснить тем, что при использовании модели Раша результаты тестирования подвергаются двойному нелинейному преобразованию – логарифмирование и взятие обратной величины.

2. Зашумление результатов тестирования снижает дифференцирующую способность всех трех алгоритмов. Причем, чем больше зашумление результатов тестирования, тем меньше различие между алгоритмами.

3. Вероятность угадывания правильного ответа значительно влияет на дифференцирующую способность всех алгоритмов. При малом числе дистракторов, особенно при двух и трех, дифференцирующая способность тестов очень мала. Поэтому не рекомендуется использовать менее четырех дистракторов. При большом числе дистракторов (не менее 14) дифференцирующая способность алгоритмов такая же, как для тестов открытой формы, для которой вероятность угадывания правильного ответа равна 0.

4. Показана эффективность планов с расщепленными делянками в многофакторных сравнительных исследованиях на основе имитационного моделирования.

В целом получена предварительная оценка алгоритмов определения уровней знаний в традиционной системе тестирования и в системе тестирования на основе модели Раша. Для получения более полной оценки систем тестирования объем исследования необходимо увеличить и рассмотреть факторы, влияющие на адекватность модели Раша результатам тестирования.

Для получения более объективной оценки и увеличения общности выводов целесообразно проводить многофакторные сравнительные исследования. Многофакторный сравнительный эксперимент позволяет увеличить общность выводов. Получаемые выводы относительно каждого источника дисперсии (фактора или взаимодействия) распространяются на все уровни учитываемых в эксперименте факторов. Кроме того, в многофакторном эксперименте становится возможным получить информацию о взаимодействии факторов, т.е. определить, зависит ли эффект одного фактора от того, на каком уровне находится другой фактор (или другие факторы).

Литература

1. Овчинников В.В. Оценка латентных параметров теста // Материалы конференции “Анализ качества образования и тестирование”. – М.: МЭСИ, 2001. – с. 215 - 223.
2. Колпаков А.В., Захаров А.В. Анализ модели G. Rasch методом численного эксперимента // Актуальные вопросы совершенствования оценочной деятельности в образовании и культуре. - М.: Исследовательский центр проблем качества подготовки специалистов, 2000. - с. 70 - 72.
3. Wright B.D. Solving measurement problem with the Rasch model // Journal of Educational Measurements.- 1977.- V.14.- N 2.- P. 97-116.
4. Аванесов В.С. Композиция тестовых заданий. – М.: Адепт, 1998. – 217с.
5. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. – М.: Прометей, 2000. – 168с.
6. Анисимова Т.С., Маслак А.А., Осипов С.А. Сравнительный анализ модели Раша и классической модели по точности оценивания // Материалы конференции “Анализ качества образования и тестирование”. – Москва: МЭСИ, 2001. – с. 38 - 42.
7. Cohen I. A modified logistic response model for item analysis. Unpublished manuscript, 1976.
8. Маркова Е.В., Маслак А.А. Рандомизация и статистический вывод. –М.: Финансы и статистика, 1986. – 208 с.
9. Маслак А.А. Основы планирования и анализа эксперимента в педагогике и психологии. - Курск: РОСИ, 1998. - 167 с.