ПАРАЛЛЕЛЬНЫЕ ВАРИАНТЫ ТЕСТА В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ ДИАГНОСТИКИ

Александр Колгатин
Харьковский национальный педагогический университет
имени Г.С. Сковороды
kolgatin@yahoo.com

Показано, что автоматизированное формирование вариантов теста приводит к появлению дополнительной погрешности педагогического измерения, связанной с недостаточной параллельностью выбираемых заданий. На основе модели G. Rash предложена формула для оценки указанной компоненты погрешности и определены её возможные значения. Исследуемая компонента погрешности сравнивается с погрешностью, вносимой угадыванием правильных ответов. Проведён анализ опыта применения автоматизированной системы педагогической диагностики с автоматически создаваемыми вариантами теста. Сопоставлены различные способы разработки параллельных заданий одинаковой трудности.

Введение

Автоматизированные системы педагогической диагностики позволяют генерировать варианты тестов индивидуально для каждого тестируемого, случайным образом выбирая задания из некоторой базы данных или используя фасеты. Такая возможность очень полезна при организации самостоятельной работы студентов по закреплению изученного материала, при подготовке к зачётам и экзаменам. Применение индивидуальных вариантов теста для каждого студента полезно и на зачётном тестировании, поскольку снижаются требования к «секретности» заданий. Однако случайный выбор заданий отрицательно сказывается на точности измерения. При формировании параллельных вариантов теста необходимо совершенствовать алгоритмы отбора заданий. А также

разрабатывать методы оценки той компоненты погрешности измерений, которая вызвана различием вариантов теста, предлагаемого разным студентам.

Традиционные подходы к оценке точности педагогических измерений

Точность педагогических измерений с помощью тестов всегда была в центре внимания исследователей как необходимый атрибут всякого измерения. Классическая теория надёжности тестовых результатов[1] предполагает оценку погрешности измерения на основе коэффициента надёжности, который определяется как отношение дисперсии истинных значений измеряемого признака к дисперсии результатов измерения (тестовых баллов) $r = s^2_{y\infty} / s_y$. Поскольку измеренные значения тестовых баллов y отличаются от истинных значений измеряемого признака y_∞ на величину погрешности E, которая статистически независима от y_∞, то коэффициент надёжности можно представить в виде $r` = 1 - \frac{s^2_E}{s^2_y}$, где s^2_E — дисперсия ошибки измерения. Отсюда выводится формула для вычисления стандартной ошибки измерения: $s_E = s_y\sqrt{1-r}$, где s_y — среднеквадратическое отклонение тестовых балов в группе тестированных.

В предположении о нормальном законе распределения тестовых баллов можно определить доверительный интервал для измеренного тестового балла с границами от $(y - \Delta y)$ до $(y + \Delta y)$, где $\Delta y = ts_E = ts_y\sqrt{1-r}$, множитель t определяется из уравнения $\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{x^2}{2}}dx = \beta_{над}$, где $b_{над}$ — доверительная вероятность[2]. В педагогике обычно принимают доверительную вероятность 95%, тогда $t \approx 1{,}96$.

Известно несколько подходов к определению коэффициента надёжности результатов:

• коэффициент стабильности («coefficient of stability») или надёжности[3], определяемый методом повторного тестирования. Вычисляется как коэффициент корреляции между результатами тестирований, проведённых одно за другим через определённое время в одной и той же группе, на основе одного и того же теста. Коэффициент стабильности учитывает погрешность, связанную с фактором времени, устойчивости уровня подготовленности, угадывания, невнимательности, пробелов в структуре учебных достижений[4];

1 *Brown F.G.* Principles of Educational and Psychological Testing. — Hinsdale: Dryden Press, 1970. 468 p.

2 *Жалдак М.І., Кузьміна М.Н., Берлінська С.Ю.* Теорія ймовірностей і математична статистика з елементами інформаційної технології. — К.: Вища шк., 1995. 351 с.

3 *Аванесов В.С.* Композиция тестовых заданий. — М.: Центр тестирования, 2002. 240 с.

4 *Колгатін О.Г.* Автоматизована педагогічна діагностика і точність вимірювання. // Вісник. Тестування і моніторинг в освіті. 2006. № 10–11, С. 29–33.

• коэффициент эквивалентности («coefficient of equivalence») или надёжности параллельных вариантов теста[5]. Вычисляется как коэффициент корреляции между результатами тестирований, проведённых по параллельным вариантам теста. Этот коэффициент чувствителен к погрешности, вызванной угадыванием, невнимательностью, различиями в трудности заданий параллельных вариантов теста. Однако этот коэффициент не чувствителен к пробелам в структуре учебных достижений, если при формировании параллельных вариантов теста стараются сохранить содержательную направленность заданий;

• внутренняя согласованность результатов теста («internal consistency») — является мерой гомогенности теста[6] и показывает, действительно ли все задания теста измеряют один и тот же признак. Коэффициент надёжности оценивается коэффициентом α-Кронбаха:

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} s_j^2}{s_y^2}\right),$$

где s_j^2 – дисперсия баллов j-го задания, s_y^2 — дисперсия тестового балла, m — количество заданий. Этот коэффициент учитывает погрешности измерения, связанные с угадыванием, невнимательностью, пробелами в структуре учебных достижений;

• надёжность по частям теста также является показателем его гомогенности. Выбираются две эквивалентные по характеру и трудности группы заданий, которые рассматриваются как отдельные параллельные варианты теста. В случае, когда тест разделен на две равные части, коэффициент надёжности оценивается по формуле:

$$R = \frac{2r_{hh}}{1 + r_{hh}},$$

где r_{hh} — коэффициент корреляции между баллами, набранными по каждой части теста. Приведённая формула получается из общей формулы Спирмана–Брауна для случая разделения теста пополам;

• коэффициент структурированности знаний[7] в соответствии с теорией надёжности Л. Гуттмана определяется на основе анализа индивидуального профиля испытуемого по формуле:

$$r = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{m} e_{i,j}}{N_m},$$

где $e_{i,j}$ — ошибки в индивидуальном профиле тестируемого, N — количество тестируемых, m — количество заданий в тесте. Следует отметить, что ошибки в индивидуальном профиле отражают не только пробелы в структуре учебных

5 *Бурлачук Л.Ф., Морозов С.М.* Словарь-справочник по психологической диагностике. — К.: Наукова думка, 1989. 200 с.

6 *Дюк В.А.* Компьютерная психодиагностика. — СПб.: «Братство», 1994. С. 110.

7 *Аванесов В.С.* Конспект лекций.

достижений, но и влияние других факторов, таких, как погрешности в ранжировании заданий по трудности, угадывание, невнимательность.

Компоненты погрешности измерения

Все выше описанные способы оценки надёжности являются интегральными и не позволяют анализировать степень влияния различных источников на суммарную погрешность тестовых результатов, в частности, не позволяют определить вклад недостаточной эквивалентности параллельных вариантов теста в суммарную погрешность измерения. Целесообразно анализировать погрешность измерения, разложив её на компоненты в соответствии с источниками погрешности: угадывание, невнимательность, пробелы в структуре учебных достижений, недостаточная эквивалентность автоматически создаваемых компьютером параллельных вариантов теста. Поскольку все компоненты погрешности — независимые случайные величины, то можно представить дисперсию суммарной ошибки измерения в виде

$$s_E^2 = s_{угадывания}^2 + s_{невнимательности}^2 + s_{структуры}^2 + s_{вари\ антов}^2$$

и каждую компоненту погрешности рассматривать отдельно. Такой подход позволяет оценить погрешность тестовых результатов каждого генерируемого компьютером варианта теста по данным апробации входящих в него заданий ещё до того, как тестируемый выполнит этот тест. Для разработки оптимального алгоритма автоматической генерации теста особое значение имеет оценка компоненты погрешности, связанной со случайным выбором вариантов заданий, то есть $s_{вариантов}^2$.

Цель данной работы — создание метода предсказания возможного различия в результатах тестирования при замене некоторых заданий.

Вывод формулы для оценки возможного изменения тестового бала при замене некоторых заданий

Для проведения анализа будем полагать, что зависимость вероятности правильного ответа от подготовленности студента может быть описана двухпараметрической моделью G. Rash:

$$P_{i,j}\left\{X_{i,j}=1\middle|a_j,b_j\right\}= \left\}=\frac{\exp\left(a_j(\theta_i-b_j)\right)}{1+\exp\left(a_j(\theta_i-b_j)\right)}, \quad (1)$$

где $X_{i,j}$ =1, если ответ i-го испы-

туемого на j-е задание правильный; θ_i — логит знаний (уровень подготовленности); a_j — параметр, который даёт информацию о задании с точки зрения его дифференцирующей способности; b_j — уровень трудности j-го задания теста, выражённый в логитах.

Для определённости допустим также, что расчёт тестового балла осуществляется по методике, предложенной на основе работы В.В. Кромера[8]. Чтобы исключить систематическое влияние правильных ответов, полученных в результате случайного угадывания, тестовый балл i-го испытуемого рассчитывается как доля осознанно правильных ответов[9]:

$$y_i = \frac{\sum_{j=1}^{m} X_{i,j}}{m}, \qquad (2)$$

где $X_{i,j} = \begin{cases} 1, \text{ правильный ответ} \\ 0, \text{ отказ от ответа} \\ \dfrac{-c_j}{1-c_j}, \text{ неправильный ответ} \end{cases}$,

j — номер задания; i — номер испытуемого; m — количество заданий в тесте; c_j — вероятность случайного предоставления правильного ответа для j-го задания.

Величину v_i можно рассматривать как среднюю вероятность того, что студент способен правильно выполнить взятое наугад задание теста или как долю правильных ответов. Таким образом, можно прогнозировать изменение тестового балла v_i при замене j-го задания другим при условии, что параметры a и b модели уже определены по результатам апробации заданий теста на некоторой репрезентативной случайной выборке студентов:

$$(\Delta y_i)_j = \frac{1}{m}\left(P_{i,j_1}\left\{X_{i,j_1}=1 \middle| a_{j_1}, b_{j_1}\right\} - P_{i,j_2}\left\{X_{i,j_2}=1 \middle| a_{j_2}, b_{j_2}\right\}\right) =$$

$$= \frac{1}{m}\left(\frac{\exp\left(a_{j_1}(\theta_i - b_{j_1})\right)}{1+\exp\left(a_{j_1}(\theta_i - b_{j_1})\right)} - \frac{\exp\left(a_{j_2}(\theta_i - b_{j_2})\right)}{1+\exp\left(a_{j_2}(\theta_i - b_{j_2})\right)}\right),$$

где приближенное значения уровня подготовленности может быть оценено через тестовый балл:

$$\theta_i = \ln\left(\frac{y_i}{1-y_i}\right).$$

Безусловно, приведённая оценка является вероятностной и верна только как средняя при большом числе опытов. Истинное изменение тестового балла при одном замененном задании дискретно и может принимать значения

$$0,\ \pm\frac{c_j}{m(1-c_j)},\ \pm\frac{1}{m} \text{ или}$$

$$\text{или } \pm\frac{1}{m}\left(1+\frac{c_j}{1-c_j}\right).$$

8 *Кромер В.В.* О некоторых вопросах тестовых технологий. // Тез. докл. Второй Всеросс. конфер. «Развитие системы тестирования в России», г. Москва 23–24 ноября 2000 г. Ч. 4. — М: Прометей, 2000. С. 59–61.

9 *Колгатін О.Г.* Статистичний аналіз тесту з різними за формою завданнями. //Засоби навчальної та науково-дослідної роботи./ За заг. ред. В.І. Євдокимова і О.М. Микитюка. — ХДПУ ім. Г.С. Сковороди. — Харків: ХДПУ, 2003. Вип. 20. С. 50–54.

10

Колгатін Олександр. Вплив вгадування на надійність тестових результатів у комп'ютерних системах педагогічної діагностики. // Математика в школі. 2008. № 2 (78) С. 36–41.

Для достаточно большого числа заменённых заданий в тесте получаем оценку изменения тестового балла:

$$\Delta y_i = \sum_{j=1}^{m} (\Delta y_i)_j .$$

Предположим, что автоматизированная система формирует индивидуальные варианты теста путем формирования каждого j-го задания на основе фасета или путем выбора из некоторой j-й совокупности однотипных заданий в базе данных. Для краткости будем называть множество заданий, из которого осуществляется случайный выбор j-го задания для теста, j-м блоком заданий. В этом случае изменение тестового балла есть сумма взаимно независимых случайных величин. Следовательно, компонента дисперсии тестового балла, определяемая неэквивалентностью заданий, есть сумма дисперсий величин $(\Delta y_i)_j$:

$$s^2_{вариантов,i} = \sum_{j=1}^{m} s^2_{(\Delta y_i)_j},$$

где, $$s^2_{(\Delta y_i)_j} = \frac{1}{m^2} \sum_{k=1}^{z_j} \frac{\left(P_{i,j,k} - \overline{P_{i,j}}\right)^2}{z_j}. \quad (3)$$

$P_{i,j,k}$ — вероятность события, заключающегося в том, что i-й студент способен дать правильный ответ на k-ое задание из j-го блока — оценивается на основе модели (1); $\overline{P}_{i,j}$ — средняя для j-го блока вероятность события, заключающегося в том, что i-й студент способен дать правильный ответ на задание; k — номер задания в блоке; z_j — количество заданий в блоке.

Гипотетический эксперимент и обсуждение результатов

Из формулы (3) видно, что рассматриваемая компонента погрешности уменьшается при увеличении числа заданий в тесте. Погрешность, связанная со случайным формированием варианта теста не может превышать максимального значения, которое определяется асимптотическим случаем, когда комплект заданий формируется путем равновероятного их выбора из множества предельно трудных ($P = 0$) и предельно лёгких заданий ($P = 1$). В этом случае

$$\overline{P} = 0{,}5, \; s^2_{вариантов,i} = \sum_{j=1}^{m} \frac{1}{4m^2} = \frac{1}{4m}.$$

Из сравнения полученного результата с нашими оценками компоненты погрешности, связанной с угадыванием правильных ответов[10],

$$s^2_{угадывания} = \frac{c(1-y)}{(m-1)(1-c)},$$

видно что погрешность, вносимая случайным формированием варианта теста не может превышать погрешности, вносимой угадыванием при выбо-

ре одного правильного ответа из 5 возможных.

Чтобы понять, какую величину может иметь на практике погрешность измерения, связанная со случайным выбором заданий, рассмотрим гипотетический пример теста. Пусть тест состоит из 41 задания с равномерно распределёнными от −1 до +1 значениями параметра трудности b, пусть параметр разделяющей способности для всех заданий. Пусть логит знаний испытуемых $\theta = 0$, поскольку именно в этом случае погрешность максимальна. Пусть внутри каждого j-го блока параметр трудности заданий равновероятно принимает значения $b_j - \Delta b$ и $b_j + \Delta b$. Тогда

$$\overline{P_{i,j}} = \frac{1}{2}\left(\frac{\exp\left(\theta_i - b_j - \Delta b\right)}{1+\exp\left(\theta_i - b_j - \Delta b\right)} + \frac{\exp\left(\theta_i - b_j + \Delta b\right)}{1+\exp\left(\theta_i - b_j + \Delta b\right)}\right),$$

$$s^2_{(\Delta y_i)_j} = \frac{1}{2m^2}\left(\left(\frac{\exp\left(\theta_i - b_j - \Delta b\right)}{1+\exp\left(\theta_i - b_j - \Delta b\right)} - \overline{P_{i,j}}\right)^2 + \left(\frac{\exp\left(\theta_i - b_j + \Delta b\right)}{1+\exp\left(\theta_i - b_j + \Delta b\right)} - \overline{P_{i,j}}\right)^2\right) =$$

$$= \frac{1}{4m^2}\left(\frac{\exp\left(\theta_i - b_j - \Delta b\right)}{1+\exp\left(\theta_i - b_j - \Delta b\right)} - \frac{\exp\left(\theta_i - b_j + \Delta b\right)}{1+\exp\left(\theta_i - b_j + \Delta b\right)}\right)^2$$

$$s^2_{вариантов,i} = \sum_{j=1}^{m} s^2_{(\Delta y_i)_j}.$$

На рис. 1 представлены графики зависимости $s^2_{вариантов}$ от вариации параметров трудности заданий в блоке Δb. На рис. 2 представлена соответствующая оценка погрешности измерения $\Delta y = 1{,}96\sqrt{s^2_{вариантов,i}}$ для доверительной вероятности 95%, в предположении, что вклад других источников погрешности пренебрежимо мал по сравнению с влиянием различия трудности вариантов теста.

Величина погрешности, связанной с формированием вариантов в реальных тестах

Для обеспечения заданного коэффициента надёжности необходимо $s^2_{вариантов} \le s^2_y(1-r)$. В предположении равномерного распределения тестового бала $s^2_y \le 0{,}083$, получаем грубую оценку сверху $s^2_{вариантов} \le 0{,}083(1-r)$. В реальных тестах эта величина меньше. Так, например, дисперсия тестового бала по результатам внешнего оценивания по математике 2006 года[11] составляла 95,1 при максимально возможном тестовом балле 62, что в пересчёте на нашу систему обозначений даёт $s^2_y \approx 0{,}025$ и $s^2_{вариантов} \le 0{,}025(1-r)$. Например, чтобы обеспечить надёжность 0,95 для реального теста

11

Дворецька Л.П. Результати зовнішнього оцінювання з математики 2006 року. // Вісник. Тестування і моніторинг в освіті. 2006. № 9, С. 18–27.

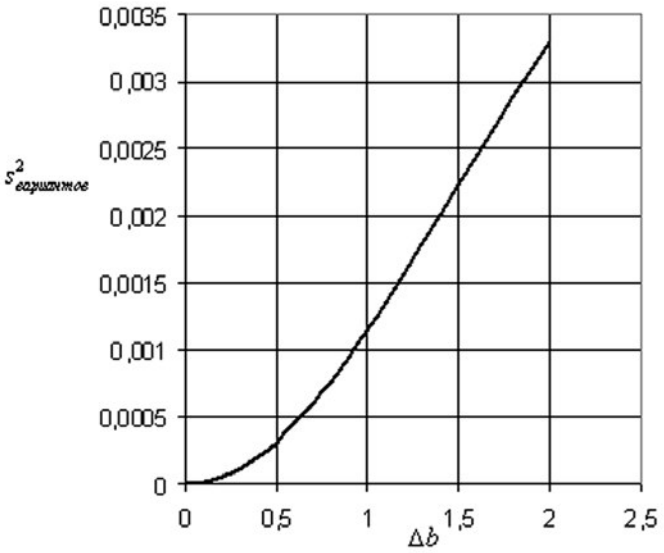

Рис. 1. Компонента дисперсии тестового балла, связанная с вариацией трудности вариантов теста при $m = 41$ и $\theta = 0$

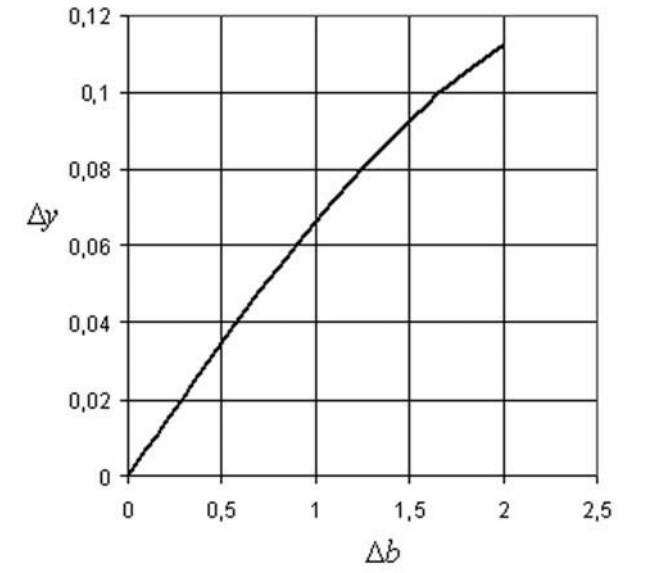

Рис. 2. Погрешность тестового балла для 95% доверительной вероятности, определяемая разбросом трудности вариантов теста при отсутствии других источников погрешности ($m = 41$, $\theta = 0$)

при отсутствии других источников погрешности, кроме случайного выбора заданий, необходимо выполнение условия $s^2_{варіантів} \le 0{,}025 \cdot (1 - 0{,}95) \approx 0{,}0012$, что достигается при $\Delta b < 1$, как это видно из рис. 1.

Из вышесказанного следует, что погрешность измерения, определяемая случайным формированием вариантов теста, достаточно существенна. Поэтому важно изучить возможную вариацию трудности заданий в фасете, выяснить причины, вызывающие различие трудности внешне эквивалентных заданий, выработать педагогические рекомендации по созданию блоков параллельных заданий.

Рассмотрим экспериментальные результаты, полученные при использовании автоматизированной системы педагогической диагностики «Эксперт». Она была создана для организации самостоятельной работы студентов и проведения модульного контроля по курсам «Математические методы в психологии» и «Теоретические основы информатики» в Харьковском национальном педагогическом университете имени Г.С. Сковороды. В табл. 1 представлены статистические характеристики для некоторых блоков заданий. Оценка параметра трудности b производилась по однопараметрической модели G. Rash (a = 1), без выполнения итераций, то есть логит подготовленности студента вычислялся на основе тестового балла по формуле

$$\theta_i = \ln\left(\frac{y_i}{1 - y_i}\right).$$

Такой подход оправдывается тем, что модель G. Rash применяется в данной работе только для оценки погрешности; цель работы — определение погрешности тестового балла, вычисленного по формуле (2).

Наиболее сбалансированными по трудности заданий оказались блоки 241 и 240 (табл. 1). Примеры заданий из блока 240 (всего в блоке 6 заданий) представлены на рис. 3.

Задания, близкие по трудности

Блок 290 содержит 4 задания, отличающиеся только взаимным расположением дистракторов (рис. 4). Трудность этих заданий действительно очень близка, однако наблюдаемые различия эмпирической трудности являются статистически значимыми. Оценка параметра b модели G. Rash проведена нами приближённо. Поэтому для доказательства значимости различий лучше использовать классическую характеристику эмпирической трудности задания, определяемую как долю правильных ответов. Здесь используется эта характеристика с поправкой на угадывание.

Таблица 1. **Статистические характеристики блоков заданий**

№ блока	Способ создания вариантов задания	Коэффициент корреляции задания с оценкой по 12-бальной шкале[12]	Среднее по блоку значение параметра трудности задания b	Стандартное отклонение значений b среди заданий блока, s_b	Объём выборки по всем заданиям блока
241	Фасет	0,42	–0,69	0,14	6230
240	Фасет	0,43	–0,70	0,14	6230
290	Перестановка дистракторов	0,33	–0,75	0,24	2797
390	перестановка дистракторов	0,42	0,54	0,27	1389
248	Фасет	0,45	0,00	0,33	6230
217	Фасет	0,33	–1,07	0,36	110
254	Ситуации	0,32	–0,17	0,41	6230
38	Числовой фасет	0,32	–0,79	0,46	697
214	Числовой фасет	0,30	–0,58	0,46	3004
35	Ситуации	0,35	0,29	0,47	703
37	Числовой фасет	0,34	–0,51	0,47	463
88	Два числовых фасета	0,34	–0,07	0,48	47
31	Числовой фасет	0,47	–0,84	0,54	3007
213	Числовой фасет	0,33	–0,13	0,55	3003
40	Два фасета	0,43	0,63	0,57	3014
30	Числовой фасет	0,48	–0,79	0,61	3007
89	Два числовых фасета	0,44	0,66	0,67	62
42	Два фасета	0,39	0,19	0,74	657
32	Числовой фасет	0,45	–0,39	0,76	3007
41	Два фасета	0,40	–0,43	0,81	657
36	Числовой фасет	0,40	0,87	0,82	700
327	Три фасета	0,33	–1,08	0,87	318
252	Три фасета	0,33	–0,09	0,88	1135
249	Фасет	0,36	–0,78	0,92	6230

91	Разные наборы дистракторов и разное задание	0,67	1,58	0,98	643
269	Фасет	0,71	–0,32	1,05	1580
259	Ситуации	0,36	–0,35	1,17	1135
86	Фасет	0,39	–0,13	1,23	643
218	Три числовых фасета	0,77	–1,21	1,43	102

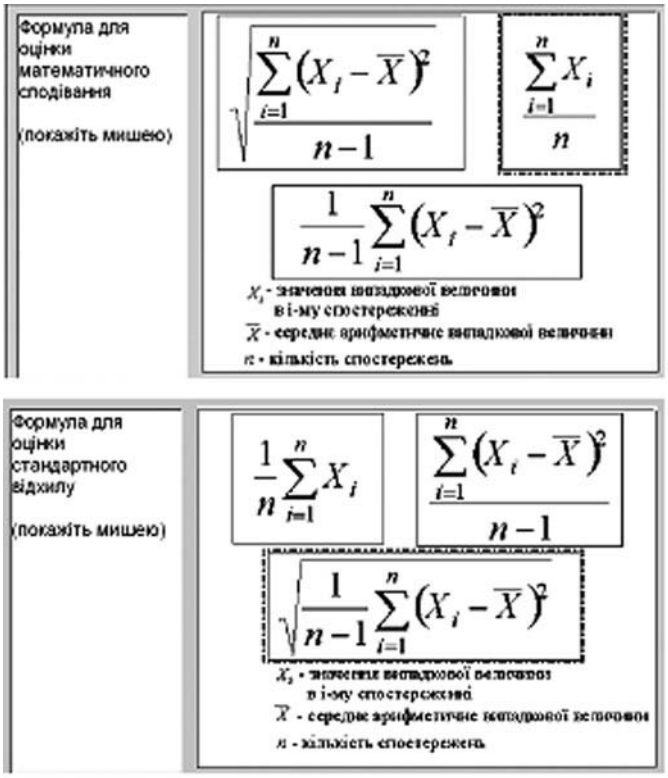

Рис. 3. Задания близкие по трудности

12

Білоусова Л.І., Колгатін О.Г. Методика обробки та інтерпретації результатів педагогічної діагностики //Комп'ютер у школі та сім'ї. 2003. № 8, С. 28–31.

Статистический анализ проведён на основе критерия Пирсона. Нулевая гипотеза «трудность всех четырёх заданий одинакова» должна быть отклонена на уровне значимости 0,05 в пользу альтернативной гипотезы «задания различаются по трудности», поскольку рассчитанное значение вероятности ошибки первого рода равно 0,02 и меньше, чем 0,05. Задание, в котором правильный вариант ответа расположен в конце списка, оказалось труднее других. Повторение статистического анализа без этого задания доказывает параллельность по трудности оставшихся трёх заданий (вероятность ошибки первого рода равна 0,83, что не позволяет отклонить нулевую гипотезу).

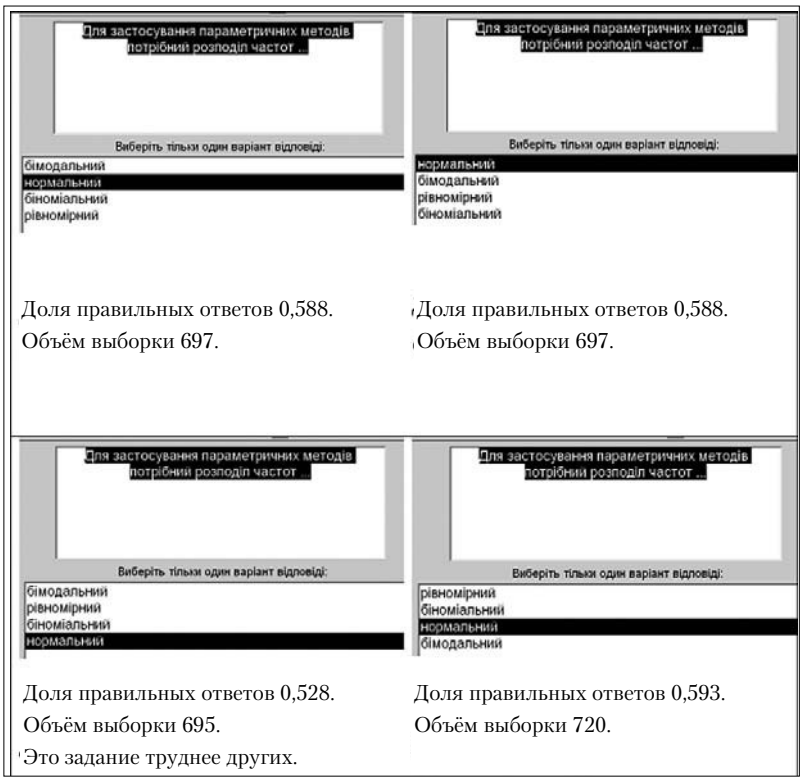

Рис. 4. Задания, отличающиеся только взаимным расположением дистракторов (так выглядит задание после того, как студент выберет правильный ответ)

Близкие по трудности задания блока 390 также отличаются только перестановкой дистракторов. Наиболее трудными оказались 2 задания из 7, одно из них имеет расположение правильного ответа в конце списка дистракторов. Однако объём выборки недостаточен для доказательства значимости различий с доверительной вероятностью 95%.

Задания блока 248 являются обратными по отношению к заданиям блоков 241 и 240, в них требуется выбрать из списка, что вычисляет формула. Различие эмпирической трудности внутри блока определяется содержанием заданий, но удовлетворяет описанным выше требованиям.

Задания блока 217 формируются на основе фасета с изменяющимся числовым параметром в задании и фиксированным набором дистракторов. Задание, для которого правильный ответ оказался расположенным в конце списка дистракторов, характеризуется долей правильных ответов 0,65 против 0,74 у другого задания, однако объём выборки недостаточен для доказательства значимости различий.

Блок 214 формируется на основе фасета заданий открытой формы (ввод числового ответа):

ДЕСЯТИЧНОЕ ЧИСЛО { 2 3 4 5 6 7 } В ДВОИЧНОЙ СИСТЕМЕ СЧИСЛЕНИЯ ИМЕЕТ ВИД __

Задание приведено в переводе с украинского языка. Различие трудности заданий, получаемых на основе этого фасета доказано с помощью критерия Пирсона с вероятностью ошибки 10^{-12}. Доля правильных ответов на задания, предполагающие перевод в двоичную систему счисления чётных чисел, находится в пределах 0,6...0,7, в то время как эта характеристика для заданий по переводу нечётных чисел составляет 0,75...0,78. Более высокая трудность перевода чётных чисел объясняется особенностями алгоритма выполнения действий. Данный факт показывает, что сам по себе фасетный подход к формированию вариантов заданий не обеспечивает параллельности по трудности. Даже незначительные различия числовых параметров в формулировках заданий могут приводить к существенным изменениям умственных действий, которые необходимо выполнить испытуемому для правильного ответа

на задание. Следует с осторожностью подходить к автоматизации формирования заданий и обязательно предусматривать экспертную оценку качества каждого задания, которое может быть сформировано на основе фасета, с последующим наблюдением за статистическими характеристиками задания в процессе эксплуатации теста.

Как пример блока, построенного на основе фасета с внешне однородными вариантами, но разной трудностью, целесообразно рассмотреть блок 86. Ниже приводятся задания этого блока в переводе с украинского языка с указанием доли правильных ответов:

1. ЧИСЛО БИТ В 1 БАЙТЕ __
(доля правильных ответов 0,76)

2. ЧИСЛО БАЙТ
В 1 КИЛОБАЙТЕ _____
(доля правильных ответов 0,52)

3. ЧИСЛО КИЛОБАЙТ
В 1 МАГАБАЙТЕ _____
(доля правильных ответов 0,31)

Это задание соответствует начальному уровню учебных достижений, оно предлагалось только тем студентам, которые получили неудовлетворительные оценки по результатам предварительного тестирования[12], поэтому доля правильных ответов невелика. Из приведённых данных видно, что различия трудности заданий блока 86 существенны, статистический анализ на основе критерия Пирсона подтверждает этот вывод с вероятностью ошибки 10^{-14}. Единицы измерения количества информации по-разному усваиваются студентами, недоучившими материал.

Из проведённого анализа экспериментального материала видно, что наименьшая вариация по трудности заданий блока обеспечивается, когда варианты заданий отличаются только взаимным расположением дистракторов. В этом случае стандартное отклонение параметра трудности b в блоке заданий может достигать s_b = 0,27, доля правильных ответов на задания может отличаться на 0,07...0,09. Формирование заданий на основе фасета в нашем эксперименте приводит к стандартному отклонению параметра трудности заданий в блоке от 0,14 до 1,23. В случаях, когда фасет содержит только один числовой параметр, s_b = 0,46...0,82 задания с одинаковыми наборами дистракторов, предполагающие оценку сходных ситуаций обеспечивают s_b от 0,41 до 1,17. Наибольшая вариация параметров трудности заданий в блоке имеет место, когда задания блока отличаются по формулировке и набору дистракторов,

хотя и направлены на проверку одного учебного содержания (s_b = 0,98), и в случаях, когда одно задание содержит несколько фасетов (s_b = 0,48...1,43). Понятно, что все приведённые цифры отражают только опыт конкретного эксперимента и не могут быть обобщены для других тестов без дополнительного анализа.

Выводы

1. Предложена формула для оценки компоненты погрешности измерения учебных достижений с помощью автоматизированных систем диагностики, которая вызвана случайным формированием вариантов теста.

2. Проведённые оценки показывают, что погрешность измерения, связанная со случайным формированием вариантов теста, существенна, однако она не превышает погрешности угадывания при случайном выборе одного правильного ответа из 5 предложенных.

3. Исследуемая компонента погрешности уменьшается при увеличении числа заданий в тесте.

4. На основе анализа экспериментальных данных, полученных при проведении диагностики учебных достижений студентов в реальном учебном процессе, определены возможные пределы изменения параметров, характеризующих вариацию по трудности заданий.

5. Обнаружено, что перестановка местами дистракторов может приводить к статистически значимым изменениям трудности заданий. В нашем примере доля правильных ответов на 0,07...0,09 меньше для заданий, в которых правильный вариант ответа расположен в конце списка дистракторов.

6. Показано, что получение параллельных заданий на основе фасета эффективно. Однако для каждого из заданий, которые могут быть получены на основе фасета, необходим предварительный экспертный анализ их трудности с последующим анализом доли правильных ответов. Часто задания, отличающиеся только числовым параметром, имеют существенно различную трудность.

В перспективе данного исследования предполагается разработать методические рекомендации авторам тестов для автоматизированных систем педагогической диагностики, а также создать системы, способные оценивать точность результатов непосредственно в процессе тестирования на основе предложенных формул.