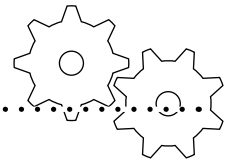
Технология и практика обучения



О.Г. Деменчёнок, *заведующий кафедрой математики и информатики Восточно-Сибирского института МВД России*

СОВЕРШЕНСТВОВАНИЕ КОМПЬЮТЕРНОГО КОНТРОЛЯ ЗНАНИЙ НА ОСНОВЕ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Проблема случайного угадывания правильных ответов при компьютерном тестировании весьма актуальна для сегодняшней российской системы образования. Автор попробовал рассмотреть её при помощи средств теории вероятности.

Возможно ли снизить процент угадывания до допустимого уровня? Какова случайная погрешность результата тестирования? Можно ли найти алгоритм перевода доли правильных ответов в педагогическую оценку? Может ли накопительный статистический анализ результатов тестирования служить основой для автоматизированной коррекции значений весовых коэффициентов тестовых заданий? Проведя соответствующие исследования, автор ответил на все эти вопросы, а полученные результаты оформил в виде компьютерной программы: разработанная им автоматизированная система Assistent позволяет создавать тесты и проводить тестирование (сайт программы www.asksystem.narod.ru).

Устраняя субъективизм оценивания, тесты не гарантируют объективной оценки знаний. Среди факторов, снижающих точность педагогических измерений, достаточно значимыми представляются:

— случайное угадывание правильных ответов;

— жёсткость алгоритмов оценивания: несущественное различие результатов выполнения теста может привести к существенно разным оценкам. Например, 59 баллов — «неудовлетворительно», а 60 баллов — «удовлетворительно;

— необоснованность значений весовых коэффициентов заданий: обычно вес зада-

ний или принимается равным (за каждое задание — одинаковое количество баллов), или назначается разработчиком теста на основе интуиции. При этом значения весовых коэффициентов могут быть далеки от оптимальных.

Анализ влияния случайного угадывания правильных ответов

Тестирование часто критикуют за возможность случайного угадывания испытуемыми правильных ответов. Ответ засчитывается как верный, независимо от того, был ли он угадан или выбран на основе знаний. Такая практика искажает тестовый балл, снижает точность педагогического измерения.

Действительно, для задания с выбором одного правильного ответа вероятность случайного угадывания обратно пропорциональна числу предложенных вариантов k

$$P_1 = \frac{1}{k} \qquad (1)$$

Насколько существенно влияние случайного угадывания на результат теста? Для ответа на этот вопрос рассмотрим тест из m заданий с выбором одного правильного ответа. Предположим, что студент на все задания выбирает ответы случайным образом. Тогда по формуле Бернулли[1] вероятность угадывания a правильных ответов:

$$P_m(a) = C_m^a p_1^a (1 - p_1)^{m-a}, \qquad (2)$$

где $C_m^a = \frac{m!}{a!(m-a)} = \frac{m(m-1)...(m-(a-1))}{a!}$

— число сочетаний.

[1] *Вентцель Е.С.* Теория вероятностей. М.: Высшая школа, 2001. 576 с.

Результаты расчётов при $m = 10$ и $p_1 = 0{,}25$ (рис. 1) показывают, что с вероятностью 0,056 не будет угадано ни одного ответа, вероятность угадывания 2–3 ответов равна 0,25–0,28. Получить положительную оценку, для которой обычно требуется набрать более половины правильных ответов, исключительно за счёт угадывания маловероятно — вероятность угадать 6 или более ответов менее 0,02 (2%). Вместе с тем вероятность завышения оценки высока — 0,944; а тестовый балл будет «улучшен» за счёт угадывания на 25%.

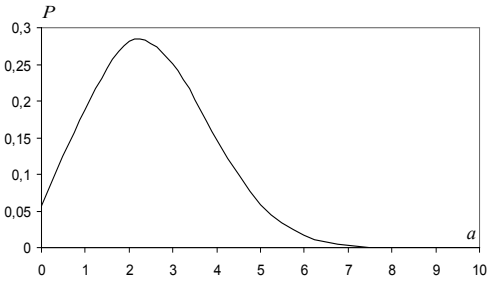


Рис. 1. Вероятность случайного угадывания a правильных ответов в 10 заданиях с выбором одного правильного ответа из 4-х вариантов

Однако в реальных ситуациях студент способен решить часть заданий (обозначим эту часть заданий w), а ответы на остальные пытается угадать (результаты расчётов приведены на рис. 2).

Нетрудно заметить: с увеличением w возрастает вероятность того, что не будет угадано ни одного ответа (т.е. угадывание никак не повлияет на результат тестирования). Однако влияние случайного угадывания остаётся существенным: для w = 0,2... 0,8 вероятность случайного угадывания одного ответа составляет 0,26...0,42, а двух ответов — 0,06... 0,31.

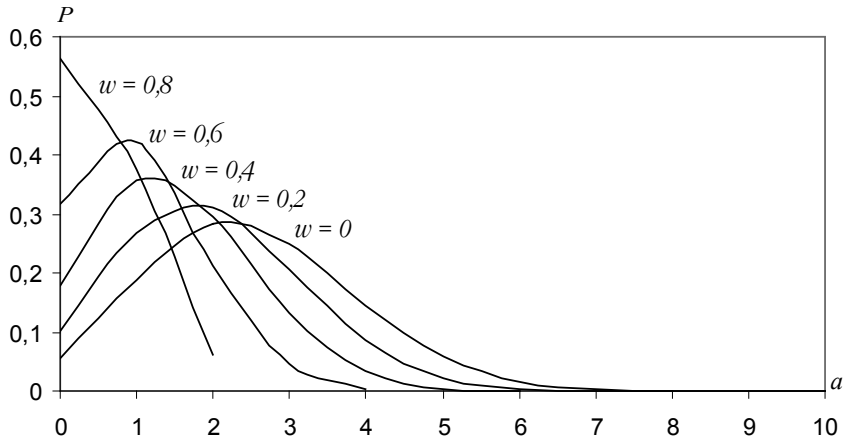


Рис. 2. Вероятность случайного угадывания ***a*** правильных ответов в 10 заданиях с выбором одного правильного ответа для различных значений ***w***

Разумеется, на угадывание влияет и количество дистракторов (приводимых в задании неправильных ответов). На рис. 3 представлены результаты расчётов для различных значений ***k***.

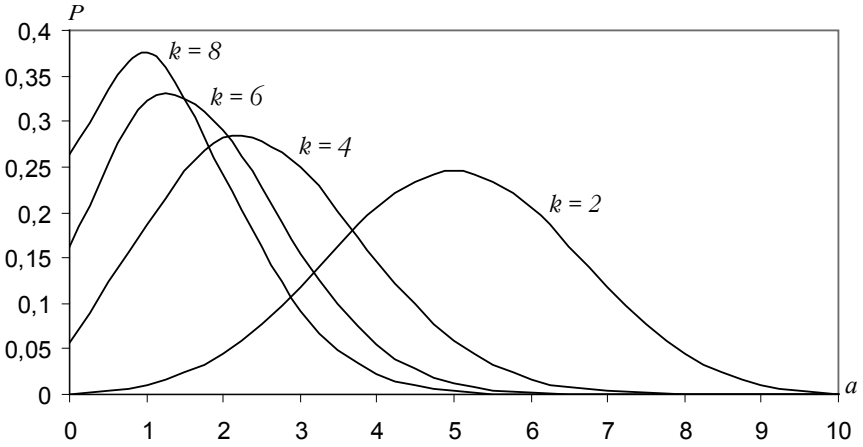


Рис. 3. Вероятность случайного угадывания ***a*** правильных ответов в 10 заданиях с выбором одного правильного ответа для различных значений ***k***

Анализ свидетельствует, что увеличение числа дистракторов несколько снижает актуальность проблемы угадывания, но устранить её полностью не в состоянии. Так, для

k = 6... 8 вероятность угадать не менее одного правильного ответа выше 0,73; не менее двух ответов — 0,36; не менее трёх — 0,11. Дальнейшее увеличение числа дистракторов нецелесообразно: кроме увеличения трудоёмкости составления теста, оно приведёт к нарушению известного из эргономики правила, гласящего, что человек может удерживать в кратковременной памяти 7 ± 2 элемента.

Если учесть, что часть заданий студент способен решить, то при ***k*** = 8 и ***w*** = 0,2... 0,8 получим вероятность случайного угадывания не менее одного ответа 0,23...0,65, а двух и более ответов — 0,02...0,26.

Осталось проверить воздействие на угадывание количества заданий в тесте. На рис. 4 даны графики вероятности случайного угадывания ***a*** правильных ответов в 50 заданиях с выбором одного правильного ответа для различных значений ***k***.

Вероятность не угадать ни одного ответа исчезающе мала — она равна 0,001; для ***k*** = = 4...8 вероятность случайного угадывания не менее 5 ответов составляет 0,76...0,99, а 10 и более ответов — 0,09...0,84.

Проведённый анализ показывает, что влияние случайного угадывания правильного ответа уменьшается с увеличением числа дистракторов и доли тестовых заданий, которые студент выполняет, не прибегая к угадыванию. Влияние различается количественно в зависимости от параметров рассматриваемой ситуации и не проявляется только в одном случае — когда студент самостоятельно решает все задания. Во всех остальных случаях влияние угадывания на результат выполнения теста с заданиями на выбор одного правильного ответа не может быть признано пренебрежимо малым.

Снижение влияния угадывания путём рационального применения различных форм тестовых заданий

Устранению проблемы угадывания поможет изменение формы тестовых заданий. В.С. Аванесов рекомендует переходить от зада-

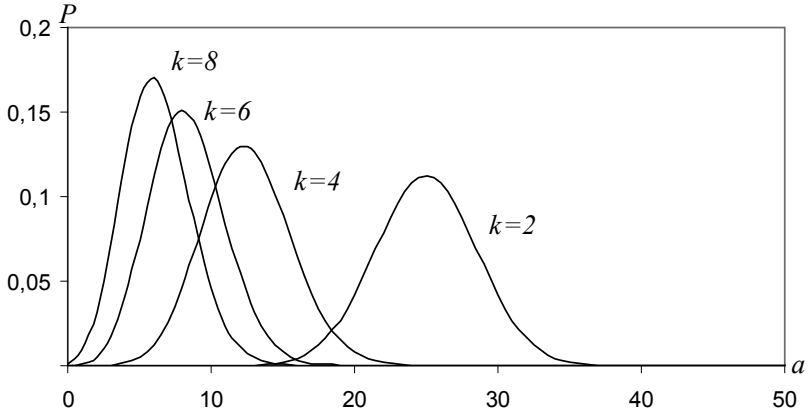


Рис. 4. Вероятность случайного угадывания ***a*** правильных ответов в 50 заданиях с выбором одного правильного ответа для различных значений ***k***

ний с выбором одного правильного ответа к заданиям с выбором нескольких правильных ответов[2], которые благодаря своей форме устойчивы к угадыванию правильного ответа.

Рассмотрим формы тестовых заданий с точки зрения снижения вероятности случайного угадывания ответа. Начнём с заданий с выбором одного правильного ответа. Устранить угадывание невозможно в принципе; снизить вероятность угадывания можно только увеличением числа дистракторов. Рекомендуемое число дистракторов — от четырёх до семи. Если дистракторов меньше четырёх, то слишком высока вероятность угадывания; если больше семи, то существенно усложняется анализ вариантов ответа при тестировании, а также повышается трудоёмкость разработки задания. При соблюдении этой рекомендации вероятность угадывания составит 0,125...0,2.

Следующая форма тестовых заданий — задания с выбором нескольких правильных ответов. Пример:

1. КАРЛ И КЛАРА УКРАЛИ ДРУГ У ДРУГА

1) крекер

2) кораллы

3) крем-брюле

4) кредитную карту

5) кларнет

В этом случае каждый из элементов выбирается независимо от остальных. Вероятность случайно сделать правильный выбор для любого из элементов равна 0,5, так как нужно угадать, какой из двух возможных вариантов правильный: «выбрано» или «не выбрано». По теореме умножения вероятностей независимых событий вероятность угадывания правильного ответа задания определяется произведением вероятностей угадывания для всех k элементов:

$$P_2 = \frac{1}{2}\cdot\frac{1}{2}\cdots\frac{1}{2} = \frac{1}{2^k}. \qquad (3)$$

При $k = 4$ вероятность угадывания равна 0,06, при $k = 6$ вероятность 0,016, при $k = 10$ вероятность менее 0,001. Случайно полностью правильно угадать ответ при $k > 4$ нереально, однако иногда учитываются частично правильные ответы, что ослабляет стойкость к угадыванию. В этом случае могут использоваться различные индикаторы меры близости ответа тестируемого и полностью правильного ответа, например коэффициент Джекарда:

$$S = \frac{d}{b + c}, \qquad (4)$$

где d — количество выбранных тестируемым правильных ответов; b — число правильных ответов в задании; c — количество несовпадений (число невыбранных правильных ответов плюс количество выбранных дистракторов).

К сожалению, практика показала неприемлемость подобных подходов для целей педагогической диагностики. Тестируемый, не зная ответа, выбирает все варианты и гарантированно получает весомую прибавку к тестовому баллу. Так, при $k = 6$ и двух дистракторах в случае выбора всех вариантов $d = b = 4$, $c = 2$:

$$S = \frac{4}{4 + 2} \approx 0,67$$

Более приемлемым представляется предложение В.С. Аванесова при двухбалльной

[2] *Аванесов В.С.* Применение тестовых форм в Rasch Measurement // Педагогические измерения, 2005, № 4. С. 3–20.

оценке за правильное выполнение задание снимать один балл за одну допущенную ошибку и снимать два балла за вторую допущенную ошибку. Используя формулу Бернулли, несложно получить выражение для вероятности угадывания с одной ошибкой:

$$P_2^1 = C_k^{k-1} p^{k-1}(1-p)^1 = \frac{k!}{1!(k-1)} \cdot \left(\frac{1}{2}\right)^{k-1} \cdot \left(1-\frac{1}{2}\right) = \frac{k}{2^k}, \quad (5)$$

где $\boldsymbol{p}$ = 0,5 — вероятность угадывания при выборе одного из элементов ответа.

Вероятность случайно угадать ответ, допустив не более одной ошибки

$$P_2 = \frac{1}{2^k} + 0,5\frac{k}{2^k} = \frac{0,5k+1}{2^k}. \quad (6)$$

Формулу (5) легко модифицировать для расчёта вероятности угадывания с $\boldsymbol{a}$ ошибками ($0 \le \boldsymbol{a} \le \boldsymbol{k}$):

$$P_2^a = C_k^{k-a} p^{k-a}(1-p)^a = \frac{k!}{a!(k-a)!} \cdot \left(\frac{1}{2}\right)^{k-a} \cdot \left(1-\frac{1}{2}\right)^a =$$

$$= \frac{k!}{a!(k-a)!\,2^k} \quad (7)$$

Рассмотрим задание на установление правильной последовательности. Студенту предоставляется набор готовых элементов (например, технологических операций). В его задачу входит расстановка этих элементов в правильной последовательности. Задания такой формы результативны в тех предметных областях, где требуется чёткое знание последовательности операций, порядка действий или взаимного расположения объектов. Пример:

Установить правильную последовательность:

2. ЦВЕТА ПОЛОТЕН ФЛАГА РОССИИ, НАЧИНАЯ С НИЖНЕГО

— белый

— красный

— синий

Если все $\boldsymbol{k}$ элементов входят в ответ, то вероятность угадывания обратно пропорциональна числу перестановок:

$$P_3 = \frac{1}{k!} = \frac{1}{1 \cdot 2 \ldots (k-1)k}. \quad (8)$$

Так, вероятность случайно расставить в правильном порядке три цвета 0,17. С увеличением числа элементов вероятность угадывания быстро снижается. Так, при $\boldsymbol{k}$ = 5 вероятность угадывания равна 0,008.

Ещё одна форма заданий предлагает восстановить соответствия между элементами двух списков. Например:

Установите соответствие:

3. ПИСАТЕЛИ ПРОИЗВЕДЕНИЯ

1. Л.Н. Толстой	А) Евгений Онегин
2. А.С. Пушкин	Б) Герой нашего времени
3. М.Ю. Лермонтов	В) Война и мир
	Г) Дубровский
	Д) Анна Каренина

Ответы: 1__ 2__ 3__

Так как каждому из элементов одного списка может соответствовать один или несколько элементов другого списка, то вероятность угадывания

$$P_4 = \frac{1}{2^{k_1 k_2}}, \quad (9)$$

где $\boldsymbol{k_1}$ и $\boldsymbol{k_2}$ — количество элементов первого и второго списка.

Если за одну ошибку снижать балл наполовину

$$P_4^1 = C_{k_1 k_2}^{k_1 k_2 - 1} p^{k_1 \cdot k_2 - 1}(1-p)^1 = \frac{k_1 \cdot k_2}{2^{k_1 k_2}},$$

где $\boldsymbol{p}$ = 0,5 — вероятность угадывания

при восстановлении одного из возможных соответствий.

Тогда вероятность случайно угадать ответ, допустив не более одной ошибки

$$P_4 = \frac{1}{2^{k_1 k_2}} + 0,5\frac{k_1 \cdot k_2}{2^{k_1 k_2}} = \frac{0,5k_1 \cdot k_2 + 1}{2^{k_1 k_2}}. \quad (10)$$

Вероятность угадывания очень низкая: при $k_1 = k_2 = 3$ вероятность безошибочного угадывания 0,002, а с одной ошибкой — 0,011.

Нельзя обойти вниманием и задания открытой формы, где ответ испытуемый дописывает сам. Например:

4. КУЛИКОВСКАЯ БИТВА СОСТОЯЛАСЬ В _____ ГОДУ.

Вероятность угадывания минимальна, в первом приближении равна нулю. Целесообразно ограничиться заданиями с кратким свободным ответом, на которые тестируемый должен записать ответ словом, словосочетанием или числом. В отличие от заданий открытой формы с развёрнутым ответом, задания с кратким свободным ответом относительно технологичны.

Для неоднородного по числу дистракторов или форме заданий теста средняя вероятность угадывания определяется как средняя арифметическая:

$$\overline{P} = \frac{\Sigma P_i}{N}, \quad (11)$$

где P_i _ вероятность угадывания правильного ответа для i-того задания.

Очевидно, что для уменьшения влияния угадывания следует увеличивать количество дистракторов и снижать долю заданий с выбором одного правильного ответа.

В ходе пробных расчётов установлено, что влияние угадывания снижается до допустимого уровня при значениях средней вероятности угадывания меньших 0,1. Результаты расчётов для $\overline{P}$ = 0,096 и N = 50 приведены на рис. 5.

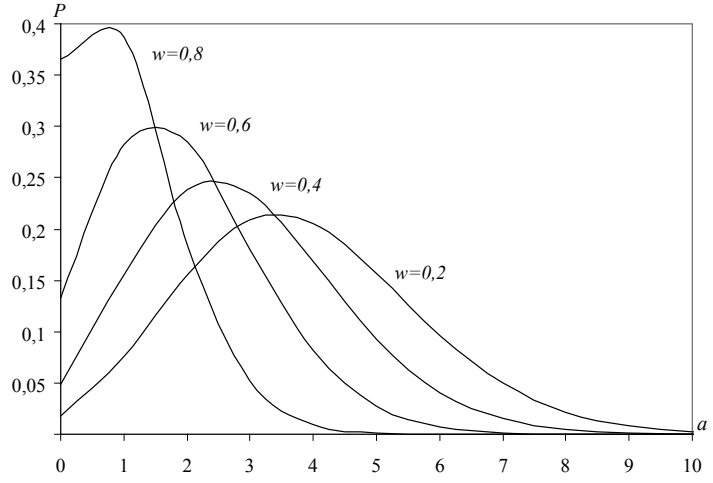


Рис. 5. Вероятность угадывания a правильных ответов в 50 заданиях при $\overline{P}$ = 0,096

Анализ показывает, что в этом случае влияние угадывания на тестовый балл пренебрежимо мало:

• очень слабому студенту ($\boldsymbol{w}$ = 0,2) угадывание не поможет. Вероятность угадать 15 и более ответов (и в сумме с честно решёнными $50 \cdot \boldsymbol{w} = 10$ заданиями набрать хотя бы половину правильных ответов) не превышает 10^{-6};

• слабому студенту ($\boldsymbol{w}$ = 0,4) угадывание также не поможет. Вероятность угадать 5 и более ответов (и в сумме с решёнными $50 \cdot \boldsymbol{w} = 20$ заданиями набрать половину правильных ответов) равна 0,15. Однако если установить порог для положительной оценки в 60% правильных ответов, то вероятность его преодоления всего 0,003;

• средний студент ($\boldsymbol{w}$ = 0,6) к честно набранным 30 ответам с вероятностью 0,57 угадает 1–2 ответа, с вероятностью 0,26–3–4, с вероятностью 0,036 — более 4 ответов. Это позволит ему улучшить тестовый балл на 3–13%;

• сильный студент ($\boldsymbol{w}$ = 0,8) мало выиграет за счёт угадывания: с вероятностью 0,39 он угадает 1 ответ, с вероятностью 0,18–2, с вероятностью 0,06 — более 3 ответов. Увеличение тестового балла — до 5%.

Число дистракторов и доля заданий каждой формы и могут варьироваться в широких пределах. Главное — добиться того, чтобы средняя величина вероятности угадывания была менее 0,1. Тогда влияние угадывания на тестовый балл будет сведено до приемлемого уровня.

Полученные теоретические результаты реализованы в рамках системы автоматизированного обучения и контроля знаний Assistent:

• увеличено с 6 до 12 количество вариантов ответа, что позволяет создавать тестовые задания, более защищённые от случайного угадывания правильного ответа;

• в среду разработки тестов добавлена возможность анализа влияния угадывания.

Редактор тестов постоянно рассчитывает среднюю вероятность случайного угадывания (её значение индицируется в строке состояния).

Так, для теста на рис. 6 средняя вероятность угадывания равна 0,0858. Очевидно, что при стремлении этой величины к нулю влияние случайного угадывания на результат тестирования также снижается до нулевой отметки. Поэтому желательно, чтобы эта величина была возможно более низкой. Для получения более полной информации и рекомендаций можно щёлкнуть поле со значением средней вероятности (рис. 7).

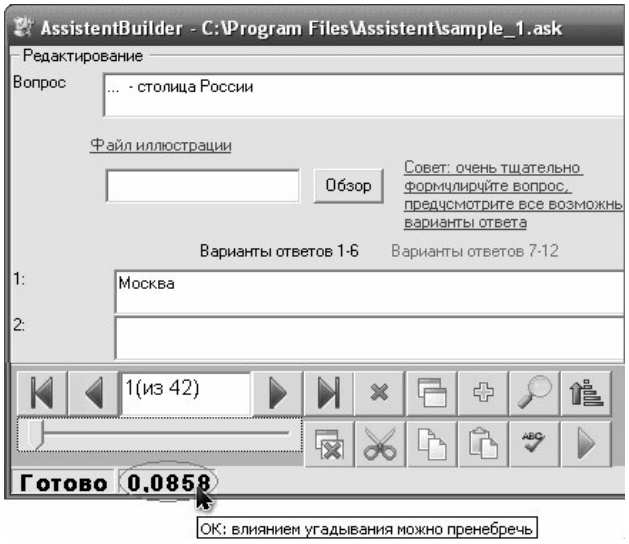


Рис. 6. Анализ влияния случайного угадывания

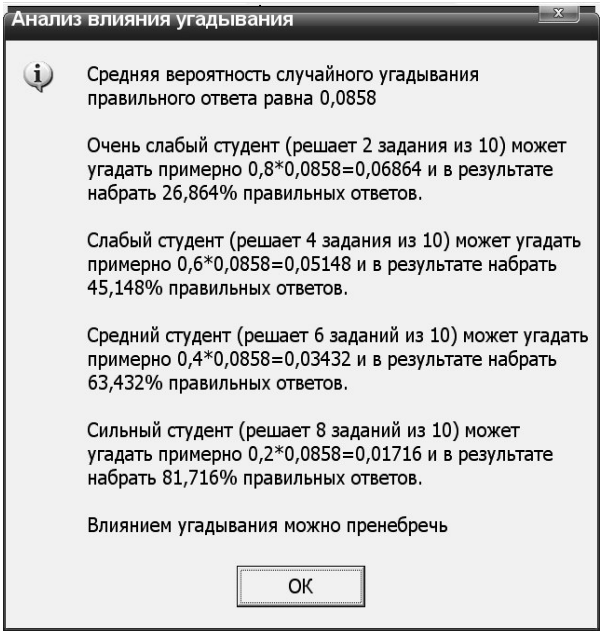


Рис. 7. Полная информация по влиянию случайного угадывания

Если влияние угадывания более существенно, то программа выдаёт рекомендации по его снижению.

Анализ жёсткости алгоритмов оценивания

Тестовый балл, как правило, не используется непосредственно. Часто он переводится в педагогическую оценку, для чего сравнивается с пороговыми значениями некой общепринятой шкалы оценок: зачтено — не зачтено; неудовлетворительно — удовлетворительно — хорошо — отлично и т. д. При этом некоторые считают, например, что за 75% правильных ответов может быть выставлена оценка «хорошо», а за 74,5% — «удовлетворительно». Т.е. за почти одинаковые результаты выполнения теста могут быть выставлены существенно различающиеся оценки, что представляется недостаточно обоснованным. Именно оценка является информаций об успехе или неуспехе, на основе оценки принимается решение о ходе процесса обучения. По оценке студенты судят об уровне своих знаний, а также об объективности педагога. Известно, что оценка приводит к благоприятному воспитательному эффекту только тогда, когда обучаемый внутренне согласен с ней. Ощущение несправедливости полученной оценки ослабляет мотивацию обучения, может привести к возникновению конфликтных ситуаций. Поэтому повышение обоснованности оценки представляется практически значимой задачей.

Важнейшей причиной неточности педагогической оценки является неоднозначность критериев оценивания. Например, оценка «неудовлетворительно» обычно рекомендуется в случае, если обучаемый не знает значительной части программного материала, допускает существенные ошибки, с большими затруднениями выполняет практические задания, задачи. Поскольку каждый из преподавателей имеет своё собственное представление о «значительной части», «существенных ошибках» и «больших затруднениях», то один и тот же ответ разными преподавателями совершенно добросовестно может быть оценён по-разному.

Результат тестового контроля определяется по заранее установленным правилам и не зависим от личности преподавателя. Устраняя субъективизм процедуры оценивания, тесто-

вый контроль знаний всё же не гарантирует точность оценки. Необходимо понимать, что тестирование позволяет достичь высокой степени объективности оценки, не гарантируя этого автоматически.

Анализ случайной погрешности результатов тестирования

Погрешность — неизбежная часть любого измерения, и педагогические измерения не являются исключением. В статистике различают три основных вида ошибок: систематические, грубые и случайные.

Систематические ошибки однонаправлено либо преувеличивают, либо преуменьшают результаты измерений. При тестировании причинами систематической погрешности могут стать ошибки в разработке и применении теста. Например, если использовать тест по высшей математике, разработанный для технической специальности, при тестировании студентов гуманитарного вуза, то получим систематическое занижение оценки. Случайное угадывание правильных ответов и недостаточный контроль за испытуемыми (соответственно, использование запрещённых справочных материалов, помощь других лиц и даже подмена тестируемого) увеличивают, по сравнению с истинным, значение тестового балла. Универсальных методов устранения систематических ошибок не существует, общая рекомендация — минимизировать влияние вызывающих систематические ошибки факторов.

Грубые ошибки возникают вследствие просчёта при вычислении тестового балла или неправильной регистрации результата (например, запись оценки в строку экзаменационной ведомости, не соответствующую фамилии тестируемого).

Случайными можно считать ошибки ввода данных; ошибки, вызванные неверным истолкованием условия задания и т.п. Единственно возможный способ объективного учёта случайных погрешностей состоит в определении их статистических закономерностей. Случайные ошибки происходят от различных случайных причин, действующих при каждом из отдельных измерений непредвиденным образом то в сторону уменьшения, то в сторону увеличения результатов.

Каковы источники случайных ошибок в случае тестового контроля? Основная причина — ограниченность числа заданий. Понятно, что чем больше заданий выполняет студент, тем полнее может быть представление о его знаниях. Проведение тестирования основано на формировании ограниченного набора тестовых заданий, что даёт возможность лучше организовать тестирование, обеспечивает быстроту проведения контроля знаний, приводит к экономии затрат труда на получение и обработку информации. Однако ограниченный набор заданий не всегда достаточен для полной проверки структуры и глубины знаний. Возникающие ошибки репрезентативности в сочетании с фрагментарностью знаний части обучаемых могут привести к зависимости тестового балла от того, какие именно задания предложены конкретному студенту («счастливый» и «несчастливый» билет).

Анализ влияния системы оценки правильности ответа

Определённое влияние оказывает широкое распространение двоичной системы оценки правильности ответа на каждое задание (правильно или неправильно, 1 или 0). Ввиду малого объёма отдельного задания сложно

различать степень правильности ответов. В результате неполные или неточные ответы квалифицируются как незнание ответа, что не всегда оправдано. Вместе с тем правильный ответ, оцениваемый максимальным баллом, не всегда соответствует известным критериям оценки «отлично» — точное и прочное знание материала в заданном объёме; исчерпывающее и логически стройное его изложение; умение обосновывать принятые решения, обобщать материал[3].

Педагогическое тестирование можно сравнить с определением площади некоторой фигуры по методу Монте-Карло. Для применения этого метода фигуру вписывают в другую, известной площади (например, в квадрат), и случайным образом «бросают» точки, подсчитывая число попаданий в фигуру. При достаточно большом числе испытаний отношение числа точек, попавших внутрь фигуры, к общему числу точек стремится к отношению их площадей. Тогда квадрат — это область, в которой проверяются знания; фигура неизвестной формы и площади — структура и глубина знаний тестируемого, а точки — тестовые задания. При достаточно большом числе заданий доля правильных ответов p приближается к истинной величине относительного объёма знаний тестируемого.

В таком случае нужно рассматривать доверительный интервал доли правильных ответов — интервал, который с заданной вероятностью α накроет неизвестное значение. Например, доверительный интервал доли правильных ответов $p = 0{,}75 \pm 0{,}05$ при вероятности 0,9 означает, что с вероятностью 90% истинное значение p находится в интервале 0,7...0,8.

При обработке данных будем исходить из того, что погрешности имеют нормальное распределение. Если считать, что погрешность измерения определяется в результате совокупного действия многих малых факторов, действующих аддитивно и независимо друг от друга, то в силу Центральной Предельной Теоремы теории вероятностей погрешность измерения хорошо приближается (по распределению) нормальной случайной величиной.

Аналитически доверительный интервал доли правильных ответов записывается в виде

$$p \pm \Delta p = p \pm \varepsilon \frac{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}}{\sqrt{m(m-1)}}, \qquad (12)$$

где Δp — погрешность определения доли правильных ответов, вызванная действием случайных факторов; σ — среднее квадратичное отклонение результатов выполнения i-го задания x_i от среднего значения $\bar{x}$; m — число заданий; $s_{\bar{p}}$ — среднее квадратичное отклонение доли правильных ответов от истинного значения; ε — табличный коэффициент для заданного значения вероятности α ($\alpha = 0{,}68$ соответствует $\varepsilon = 1{,}0$; $\alpha = 0{,}90$ соответствует $\varepsilon = 1{,}65$; $\alpha = 0{,}997$ соответствует $\varepsilon = 3{,}0$).

Простой анализ выражения (12) показывает, что случайная погрешность зависит от однородности результатов выполнения отдельных заданий и количества заданий m. Нетрудно заметить, что если все ответы правильны ($x_i = \bar{x}$, $\sigma = 0$), то случайная погрешность равна нулю. Аналогично $\Delta p = 0$ в случае, когда ответы полностью неверны

[3] *Буланова-Топоркова М.В.* и др. Педагогика и психология высшей школы. Ростов-на-Дону: Феникс, 2002. 544 с.

($\boldsymbol{x} = \bar{x} = \boldsymbol{\sigma} = = 0$). Это означает, что случайная погрешность отсутствует только в этих двух крайних случаях.

Очевидно, что максимальное значение $\Delta\boldsymbol{p}$ принимает в случае использования двоичной системы оценивания правильности ответа (правильно или неправильно) при условии равенства количества правильных и неправильных ответов ($\boldsymbol{x_i} = 0;1;0;1;0;1...$):

$$\Delta p_{\max} = \mathrm{e}\frac{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}}{\sqrt{m(m-1)}} = \mathrm{e}\frac{\sqrt{0,5^2 m}}{\sqrt{m(m-1)}} = \frac{\mathrm{e}}{2\sqrt{m-1}}. \quad (13)$$

Например, случайная погрешность доли правильных ответов для теста с $\boldsymbol{m} = 20$ заданиями при доверительной вероятности 0,68 не превышает $\Delta\boldsymbol{p}_{\max} = 0,11$ (или 11%), при $\boldsymbol{m} = 50$ $\Delta\boldsymbol{p}_{\max} = 0,07$, при $\boldsymbol{m} = 200$–$0,035$. Из этого уравнения легко получить зависимость для расчёта количества заданий, гарантирующего, что случайная погрешность не превысит заданного значения:

$$m = \frac{\mathrm{e}^2}{4\Delta p_{\max}^2} + 1. \quad (14)$$

Так, для обеспечения случайной погрешности не более 0,05 (5%) при указанной доверительной вероятности требуется 101 задание. Графически зависимость $\boldsymbol{m} = \boldsymbol{f}$ ($\Delta\boldsymbol{p}$, ε) представлена на рис. 8.

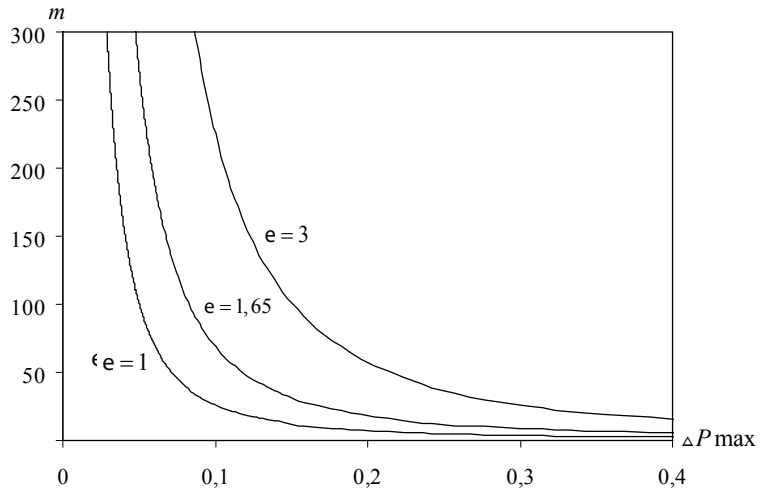


Рис. 8. Зависимость количества заданий теста от максимально допустимой величины случайной погрешности

Алгоритм уточнения оценки при компьютерном тестировании

Вернёмся к задаче перевода результата тестирования в качественные показатели типа «хорошо», «удовлетворительно» и т.п. При таком переводе статистически неразличимые результаты могут привести к разным оценкам. Так, например, доли правильных ответов $\boldsymbol{p} = 0,59$ и $\boldsymbol{p} = 0,61$ при ошибке $\Delta\boldsymbol{p} = 0,05$ соответ-

ствуют практически одинаковым интервалам 0,54...0,64 и 0,56...0,66. Однако при пороговом значении для удовлетворительной оценки R_3 = = 0,6 оценки будут кардинально отличаться — первый обучаемый получит «неудовлетворительно», а второй — «удовлетворительно». Реально же данная ситуация означает, что оценка лежит в пределах от «неудовлетворительно» до «удовлетворительно». Что делает в таких случаях опытный преподаватель? Для уточнения оценки задаёт дополнительные задания. Если при бланковом тестировании подобное организовать сложно, то при компьютерном тестировании вполне возможно реализовать выдачу дополнительных заданий для уточнения оценки.

На рис. 9 схематично показано сопоставление результата тестирования в виде доверительного интервала p со шкалой оценивания ($0...R_3$ — «неудовлетворительно», $R_3...R_4$ — «удовлетворительно», $R_4..R_5$ — «хорошо», свыше R_5 — «отлично»).

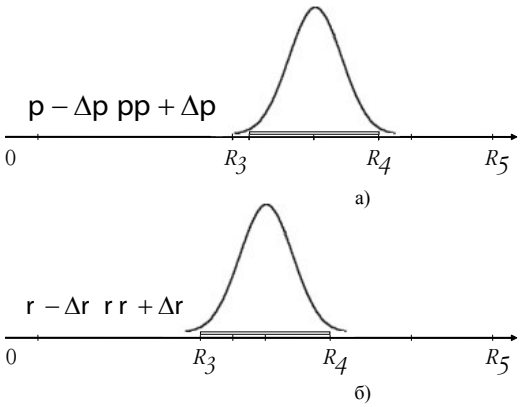


Рис. 9. Сравнение доверительного интервала доли правильных ответов с пороговыми значениями шкалы оценивания

В случае, когда доверительный интервал полностью помещается между двумя соседними значениями шкалы оценивания, можно утверждать, что с вероятностью не меньшей α результат соответствует оценке R_i. Так, на рис. 9а доверительный интервал доли правильных ответов располагается между значениями R_3 и R_4. Следовательно, результат выполнения теста оценивается на «удовлетворительно».

Возможен также вариант, когда значение шкалы оценивания окажется внутри доверительного интервала (рис. 9б). Возникает неоднозначность: с вероятностью P_1 = = $P(p-\Delta p < p < R_3)$ результат соответствует оценке «удовлетворительно», а с вероятностью $P_2 = P(R_3 \le p < p+\Delta p)$ результат соответствует оценке «хорошо».

Очевидно, что при близких значениях p и R_i вероятности примерно равны $P_1 \approx P_2$.

Следовательно, в таком случае равновероятны две разные оценки, что существенно затрудняет оценивание ответа.

Вероятность попадания результата тестирования $\boldsymbol{p}$ в промежуток $[\boldsymbol{p_1}, \boldsymbol{p_2}]$ в предположении нормального распределения[4]:

$$P = \int_{r_1}^{r_2} \frac{1}{\sqrt{2\pi}s} e^{-(p-\bar{p})^2/(2s^2)} dp = F\left(\frac{p_2 - p}{s}\right) - F\left(\frac{p_1 - p}{s}\right), \quad (15)$$

где $\boldsymbol{F}$ — функция нормированного и централизованного нормального распределения (функция Лапласа).

Принимая доверительную вероятность равной 0,68, получим:

$$\Delta p = e \cdot s_{\bar{p}} = s_{\bar{p}}, \quad (16)$$

$$P_1 = P(p - \Delta p < p < R_i) = F\left(\frac{R_i - p}{s_{\bar{p}}}\right) - F\left(\frac{(p - \Delta p) - p}{s_{\bar{p}}}\right) = F\left(\frac{R_i - p}{s_{\bar{p}}}\right) + F(1), \quad (17)$$

$$P_2 = P(R_i < p < p + \Delta p) = F\left(\frac{p - (p + \Delta p)}{s_{\bar{p}}}\right) - F\left(\frac{R_i - p}{s_{\bar{p}}}\right) = F(1) - F\left(\frac{R_i - p}{s_{\bar{p}}}\right). \quad (18)$$

Рассмотрим пример. В таблице 1 представлены результаты тестирования трёх студентов (расчёты проведены для $\alpha = 0{,}68$; результаты выполнения заданий $\boldsymbol{x_i}$ для простоты представлены целыми числами).

Таблица 1

Результаты тестирования

Фамилия	Результаты выполнения заданий x_i	p	σ	$\Delta p = s_{\bar{p}}$	$p \pm \Delta p$
Иванов	1;1;1;1;1;1;1;1;0;1;1;1;1; 1;1; 1;1;1;1	0,95	0,224	0,050	0,9...1,0
Петров	1;1;0;0;1;1;1;1;1;0;1;1;1;1; 1;1; 1;1;1;0	0,80	0,410	0,092	0,71...0,89
Сидоров	1;1;1;0;0;1;0;0;0;0;1;1;1;1;1;0; 1;1;0;1	0,60	0,503	0,112	0,49...0,71

Предположим, заданы пороговые значения для четырёхбалльной шкалы $\boldsymbol{R_3} = 0{,}6$; $\boldsymbol{R_4} = 0{,}75$ и $\boldsymbol{R_5} = 0{,}9$ (иными словами, до 60% правильных ответов — «неудовлетворительно», 60–75% — «удовлетворительно», 75–90% — «хорошо» и свыше 90% — «отлично»). Тогда доверительный интервал результата тестирования испытуемого Иванова (см. табл. 1) $\boldsymbol{p \pm \Delta p} = 0{,}9...1{,}0$ с вероятностью 0,68 превышает $\boldsymbol{R_5} = 0{,}9$ и, следовательно, соответствует оценке «отлично». В табл 1. приводится пример оценок, полученных тремя студентами.

В доверительный интервал результата студента Петрова 0,71...0,89 попадает пороговое значение $\boldsymbol{R_4} = 0{,}75$. Вероятности $\boldsymbol{P_1}$ и $\boldsymbol{P_2}$ в этом случае:

$$P_1 = P(0{,}708 < p < 0{,}75) = F\left(\frac{0{,}75 - 0{,}8}{0{,}092}\right) + F(1) = 0{,}13$$

$$P_2 = P(0{,}75 < p < 0{,}892) = F(1) - F\left(\frac{0{,}75 - 0{,}8}{0{,}092}\right) = 0{,}55$$

Это означает, что с вероятностью 0,55 результат тестирования соответствует оценке «хорошо», а с вероятностью 0,13 — «удовлетворительно».

Ещё менее точно определена оценка Сидорова. На середину доверительного интервала его результата приходится пороговое значение $\boldsymbol{R_3 = 0{,}6}$. Поэтому практически с равной вероятностью результат тестирования может быть интерпретирован как «удовлетворительно», так и «неудовлетворительно», что не поз-

[4] *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1022 с.

воляет определить оценку с приемлемой точностью.

В этих условиях разумным представляется в качестве критерия обоснованности оценки принять вероятность её соответствия данному результату выполнения теста. Если значение шкалы оценивания окажется внутри доверительного интервала, рекомендуется найти разность вероятностей двух оценок. При значении разности меньшей заданной следует выдавать дополнительные задания до тех пор, пока вероятность одной из оценок не окажется существенно большей.

Тогда в основу процедуры оценивания при компьютерном тестировании может быть положен предлагаемый ниже алгоритм.

1. Задаём количество заданий m, максимальное количество заданий с учётом дополнительных m_{max} и минимальную величину разности вероятностей оценок, при которой оценка считается найденной $\Delta P_{min} = |P_2 - P_1| \approx 0{,}3\ldots0{,}4$. Значение доверительной вероятности можно принять постоянным и в настройках теста не задавать.

2. Предлагаем тест из m заданий.

3. После получения ответов на задания теста определяем доверительный интервал доли правильных ответов, который сравниваем с пороговыми значениями R_i шкалы оценок:

1) если доверительный интервал полностью помещается между двумя соседними значениями шкалы оценивания, то оценку можно считать найденной, тестирование окончено;

2) если значение шкалы оценивания окажется внутри доверительного интервала и $|P_2 - P_1| \geq \Delta P_{min}$, то в качестве итоговой принимаем оценку, вероятность которой больше; тестирование окончено;

3) если значение шкалы оценивания окажется внутри доверительного интервала и $|P_2 - P_1| < \Delta P_{min}$, то оценка определена с недостаточной точностью, переходим к пункту 4.

1. Если общее количество выполненных заданий меньше m_{max}, то предлагаем дополнительное задание и переходим к пунк-ту 3. В противном случае в качестве итоговой принимаем оценку, вероятность которой больше; тестирование окончено.

Если тестовые задания оцениваются разным количеством баллов, то долю правильных ответов можно заменить отношением индивидуального тестового балла к сумме баллов за все задания. Особенностью изложенного алгоритма является учёт не столько погрешности результата тестирования, сколько вероятности его соответствия той или иной оценке. Кроме того, методы математической статистики применены для обработки результатов ответа одного обучаемого, а не группы. Такой подход даёт возможность повысить обоснованность каждой конкретной оценки.

Алгоритм уточнения оценки реализован в программе Assistent. Программа моделирует алгоритм работы преподавателя и в спорных случаях автоматически выдаёт дополнительные задания для уточнения оценки.

Для включения этого режима работы в редакторе тестов AssistentBuilder следует нажать кнопку ***Параметры теста,*** выбрать пункт ***Уточнение оценки*** и задать параметры системы оценивания:

• желаемое количество заданий для тестирования;

• максимальное количество заданий при тестировании с учётом дополнительных — предельное количество задаваемых вопросов;

• минимальная величина разности вероятностей оценок, при которой оценка считается найденной. Assistent будет задавать дополнительные задания до тех пор, пока разность вероятностей оценок не станет больше указанной величины или число выполненных заданий не превысит максимальное количество.

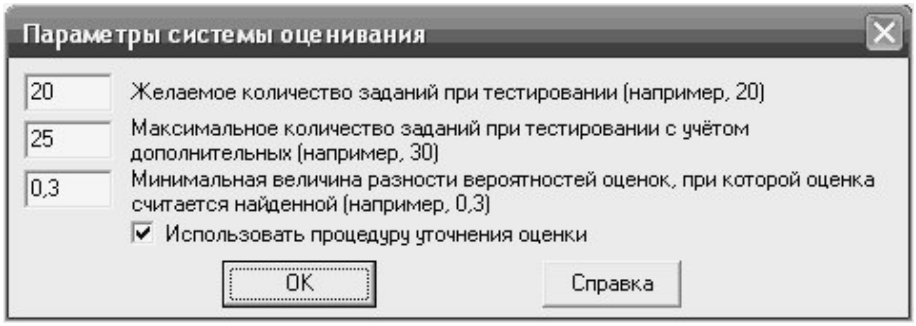


Рис. 10. Параметры уточнения оценки

Определение значений весовых коэффициентов тестовых заданий

Обычно вес тестовых заданий или принимается равным (за каждое задание — одинаковое количество баллов), или назначается разработчиком теста на основе интуиции. Очевидно, что такие значения весовых коэффициентов могут быть далеки от оптимальных.

К сожалению, классическая теория тестов не имеет общепринятого механизма задания значений весовых коэффициентов тестовых заданий[5]. При определении веса тестового задания (т.е. количества баллов, которое можно получить за правильный ответ) преобладает эмпирический подход, основанный на экспертном оценивании[6].

При определении весовых коэффициентов тестовых заданий может учитываться:

• только сложность тестового задания;

• сложность и важность тестового задания.

Очевидно, что объективно сложность тестового задания может быть измерена только статистически — по результатам выполнения этого задания достаточно представительной группой испытуемых, уровень подготовленности которых соответствует уровню подготовленности тех лиц, для контроля знаний которых предназначен тест.

Анализируя статистику ответов, в качестве показателя сложности тестового задания целесообразно принять долю правильных ответов $\boldsymbol{p}$. Чем сложнее тестовое задание (т.е. меньше $\boldsymbol{p}$), тем выше должен быть его удельный вес.

Важность тестового задания для усвоения учебного материала, к сожалению, не поддаётся объективному измерению. Поэтому

---

5 *Челышкова М.Б.* Теория и практика конструирования педагогических тестов: Учеб. пособие. М.: Логос, 2002.

6 *Майоров А.Н.* Теория и практика создания тестов для системы образования. М.: Интеллект-центр, 2002.

этот показатель может определить составитель теста или эксперт только на основе собственного опыта.

С учётом изложенного, для определения значений весовых коэффициентов тестовых заданий предложена формула:

$$a = \begin{cases} a = \{a_0 (1 - 0,9p) \ \text{при} \ N \geq 30 \\ \dfrac{a_0 (1 - 0,9) N}{30} \ \text{при} \ N < 30 \end{cases} \qquad (19)$$

где a — весовой коэффициент тестового задания, a_0 — начальное значение весового коэффициента тестового задания, N — число испытуемых, выполнивших задание.

Если число испытуемых, выполнивших задание, меньше 30, то возможны существенные ошибки репрезентативности (т.е. опрошенные выполняют данное задание заметно лучше или хуже, чем в целом по всей генеральной совокупности). Для уменьшения влияния ошибок репрезентативности при $N < 30$ введён понижающий коэффициент, равный отношению $N/30$.

Формула (19) достаточно универсальна, так как позволяет реализовать оба известных подхода к определению весовых коэффициентов:

• может учитываться только сложность тестового задания (для этого начальные значения весовых коэффициентов всех заданий следует задать одинаковыми). Например, для всех заданий принять $a_0 = 1$;

• может учитываться сложность и важность тестового задания — сложность будет автоматически определяться по статистике ответов, а важность каждого задания разработчик теста задаёт самостоятельно. Например, для первого задания $a_0 = 1$, для второго задания принять $a_0 = 3$ и так далее.

Таким образом, накопительный статистический анализ результатов тестирования может служить основой для выбора значений весовых коэффициентов. Данный механизм задания значений весовых коэффициентов тестовых заданий реализован в рамках системы автоматизированного обучения и контроля знаний Assistant путём включения подпрограммы, выполняющей перерасчёт весовых коэффициентов по результатам статистики ответов.

Для включения автоматической коррекции весовых коэффициентов тестовых заданий в AssistentBuilder следует нажать кнопку ***Параметры теста*** и отметить пункт ***Коррекция баллов*** (рис. 11).

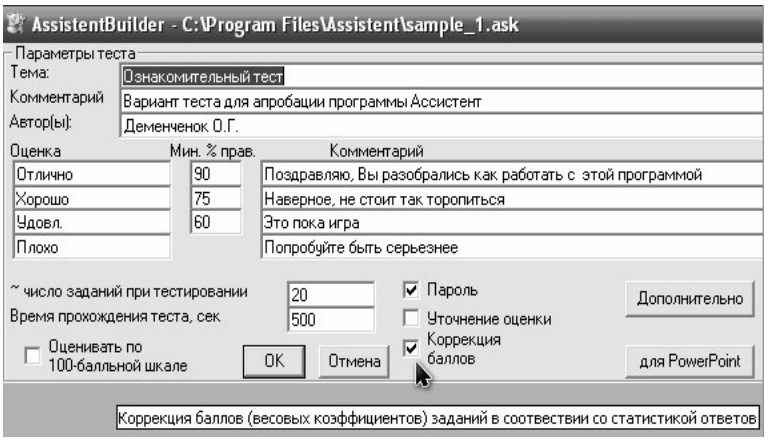


Рис. 11. Вид окна редактора тестов AssistentBuilder (курсором отмечен пункт ***Коррекция баллов***)

Полученные результаты реализованы в рамках программы Assistent (свидетельство о государственной регистрации № 2008610441). Программа отмечена сертификатом корпорации Microsoft, используется в 30 учебных заведениях России, Украины и Беларуси, в том числе:

• в Иркутском государственном университете, Винницком государственном педагогическом университете, Полоцком государственном университете, Государственном техническом университете МАИ, Российском государственном торгово-экономическом университете, Восточно-сибирском институте МВД России, Ярославской государственной медицинской академии, Байкальском поисково-спасательном отряде МЧС России;

• в ГОУ СОШ № 120 г. Санкт-Петербурга, МОУ лицее № 10 г. Батайска, МОУ лицее № 36 г. Иркутска, МОУ СОШ № 1 г. Брянска, МОУ СОШ № 6, 15, 55 г. Иркутска, Общеобразовательной школе № 23 г. Симферополя.

Программа Assistent доступна на сайте www.asksystem.narod.ru, регистрация программы для учебных заведений бесплатна. Автор выражает готовность ответить на вопросы по использованию программы по электронной почте AskSystem@ya.ru.