А. Н. Алехин, Г. В. Иванов

ПРИНЦИПЫ МЕТРИЧЕСКОЙ ОЦЕНКИ КОМПЕТЕНТНОСТИ СПЕЦИАЛИСТОВ ПРИ РЕШЕНИИ ТЕСТОВЫХ ЗАДАЧ

Новой и практически не разработанной проблемой для обеспечения мониторинга эффективности образовательного процесса стала необходимость текущего и итогового оцениваний таких интегративных характеристик субъекта профессиональной деятельности, как его профессиональные компетенции. Компетенции специалиста следует рассматривать в качестве таких его характеристик, как умения действовать в профессиональных ситуациях, выполняя свои профессиональные обязанности. Наиболее адекватным средством для оценки такого умения является тренажёр. Надо сказать, что тренажёрная подготовка и выполнение задач, максимально приближенных к реальным условиям профессиональной деятельности, давно и прочно доказала свою эффективность при подготовке разных специалистов, общей характеристикой деятельности которых являются высокий уровень ответственности, наличие специальных навыков, сложность и экстремальный характер профессиональных ситуаций. Системы оценки и мониторинга знаний в профессиональном образовании до сей поры ограничивались перечнями контрольных вопросов, успешность ответов на которые позволяла оценить уровень остаточных знаний по различным предметным областям осваиваемой дисциплины. Однако такой подход становится недостаточным там, где объектом оценивания становятся профессиональные компетенции будущего специалиста. К тому же остаточные знания, даже при достаточном их объёме не всегда являются единственным условием профессиональной успешности. В связи с изложенным кафедрой клинической психологии и психологической помощи было предложено направление перспективных научно-исследовательских разработок по созданию автоматизированной системы оценки профессиональных компетенций специалиста на основе анализа решения им как перечней контрольных вопросов, так и нового вида заданий, которые определены нами как ситуационные задачи. Эти исследования были поддержаны и реализуются теперь в рамках государственного заказа Министерства образования РФ и науки (№ 01201357146). На первом этапе научно-исследовательской работы (НИР) были разработаны наборы контрольных вопросов по всем дисциплинам специальности «клиническая психология» и алгоритмы оценки уровня знаний специальности на основании вкладов осведомлённости по ведущим предметным областям. В результате были получены предварительные правила оценивания профессиональных компетенций будущего специалиста. На данном этапе НИР в дополнение к тестовым заданиям были разработаны наборы ситуационных задач, моделирующие на уровне дискуссии с ЭВМ типовые профессиональные ситуации, возникающие в практике клинического психолога при работе в образовательных и медицинских учреждениях. Это задачи диагностики психического развития и актуального состояния, задачи планирования мероприятий психологической профилактики и коррекции нарушений психического состояния и развития, организации и реализации мероприятий психологической профилактики нарушений психического здоровья. Ситуации представляются в вербальной форме, пошагово: от описания ситуации через ее развитие в диалоге к окончательным решениям и выводам. Алгоритм предъявления ситуационных задач предполагает активную обратную связь системы с поведением испытуемого. Правильные решения на каждом шаге развёртывания ситуации позволяют испытуемому продвигаться к решению задачи, ошибочные ответы либо возвращают испытуемого на прежний уровень развития ситуации либо ведут к тупиковым ветвям диалога, за что испытуемый теряет набранные

баллы. Неспособность испытуемого при ряде предъявлений продвинуться в решении задачи прекращает дальнейшее предъявление текущей задачи и формирует итоговую оценку. В данной публикации основное внимание уделяется описанию алгоритмов оценки успешности выполнения тестовых заданий; описанию моделей профессиональных ситуаций будет посвящена другая статья.

Проблема оценивания знаний, умений и навыков в условиях высшего профессионального образования по-прежнему является актуальной, и предлагаемые решения часто содержат в себе скрытые допущения, которые снижают надёжность и валидность оценки.

Кратко рассмотрим историю тестологии, как историю экспериментальной проверки этих допущений.

Каждая наука в своем развитии проходит 3 стадии [6]:

1) стадию накопления эмпирических данных;

2) стадию поиска простых эмпирических закономерностей в данных;

3) стадию формулировки выраженных языком математики гипотез о природе исследуемой реальности.

Формулы для предсказания эмпирических данных, выводимые из такого рода гипотез, обычно сложнее a posteriori сформулированных закономерностей стадии 2, но они сводят исследуемый предмет к взаимодействию его более простых компонент, описываемому фундаментальными законами и в этом смысле объясняют реальность.

Рассмотрим, как проходила эти стадии тестология.

Тестология — наука о свойствах тестов [10]. В самом общем виде тестология изучает взаимодействие человека с наборами вопросов в аспекте эффективности ответов на них.

Рассмотрим типичный тест для проверки знаний. Тест состоит из вопросов. К каждому вопросу прилагаются варианты ответа, один из которых правильный, а остальные отвлекающие — так называемые дистракторы. Тест предъявляется выборке экзаменуемых. Результат тестирования выборки представлен таблицей. Строки этой таблицы — испытуемые. Столбцы таблицы — экзаменационные задания. Если i-й испытуемый ответил правильно на j-й вопрос, клетка — пересечение i-й строки и j-го столбца заполняется единицей. Если ответ был неправильный, клетка заполняется нулем.

По данной таблице (матрице ответов) для каждого испытуемого можно подсчитать его первичный балл — число правильных ответов. Кажется необходимым, чтобы испытуемый с большим уровнем знаний имел больший первичный балл. Также можно рассчитать первичный балл задания — число испытуемых в выборке, верно ответивших на данное задание. Задание с меньшим первичным баллом является более трудным (так как на него смогли правильно ответить меньше людей) [8].

Таким образом, вводится характеристика испытуемого — уровень знаний и характеристика задания — трудность.

Однако разработанные на стадии 1 развития тестологии первичные баллы как средство измерения этих характеристик не безупречны — мера трудности задания зависит от уровня обученности испытумых, принявших участие в тестировании. Мера обученности испытуемого, напротив, зависит от того, какие задания использовались — трудные или легкие.

На стадии 2 было замечено, что если построить гистограмму, на которой по горизонтальной оси отложен первичный балл испытуемого, а по вертикальной — то, какую долю от общей выборки составляют испытуемые, набравшие данный балл, получится колоколообразный график, хорошо описываемый функцией нормального распределения.

Практическим применением данного факта стало выражение способности испытуемого через отклонение его первичного балла от среднего по выборке. Это отклонение

делилось на рассчитанное по выборке стандартное отклонение первичных баллов (меру разброса первичного балла от испытуемого к испытуемому). Хотя такой подход делал результат независимым от конкретного набора экзаменационных вопросов, зависимость от выборки оставалась: один и тот же студент при сравнении с неуспевающей выборкой выглядел успевающим, но при сравнении с успевающей — наоборот.

Тогда же было предложено характеризовать каждое задание так называемой характеристической кривой. Исходная большая выборка разбивалась на подвыборки. Каждую подвыборку составляли испытуемые, получившие один и тот же первичный балл. Таким образом, имелись подвыборки «единичников» (ответивших только на один вопрос), «двоечников», «троечников» и т. д. Для каждой подвыборки вычислялся процент лиц, получивших данный первичный балл и правильно ответивших на определенный вопрос (например, на вопрос № 1). Далее для каждого вопроса строилась характеристическая кривая, абсцисса которой — первичный балл испытуемых данной подвыборки, а ордината — процент испытуемых данной подвыборки, правильно ответивших на данный вопрос [1].

Этот процент рассматривается как оценка вероятности того, что испытуемый с данным уровнем успеваемости (троечник, четверочник и т. д.) ответит на данный вопрос. То есть характеристическая кривая вопроса связывает уровень успеваемости испытуемого с вероятностью ответить на данный вопрос. Было обнаружено, что характеристические кривые заданий — монотонно возрастающие.

На стадии 3 развития тестологии датский математик Г. Раш попытался найти аналитическое выражение для характеристической кривой задания (так называемая Item Response Theory — IRT). В рамках модели Раша для конкретного человека существует вероятность ответить на конкретный вопрос.

Вводится две величины: уровень знаний данного человека θ и трудность данного задания δ. Вероятность того, что данный человек правильно ответит на данное задание, дается функцией:

$$P(\text{правильного\_ответа}) = \frac{e^{\theta}}{e^{\theta} + e^{\delta}}.$$

Или эквивалентно:

$$P(\text{правильного\_ответа}) = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}}.$$

Зависимость вероятности решения конкретного задания от того, какой у человека уровень знаний, задает характеристическую кривую пункта (теперь уже теоретически). Ранее нами проводились теоретические разработки по выводу модели Раша из математической формулировки теории условных рефлексов.

Сложность пункта δ соответствует такому уровню знаний у человека, при котором этот человек отвечает на пункт правильно с вероятностью 50%.

Классическая модель Раша кроме основной формулы содержит еще несколько допущений.

Одномерность латентного континуума. Если континуум, на котором откладываются величины уровня знаний и сложности, реально существует, то он одномерен. Модель Раша в классической формулировке не охватывает ситуаций, в которых один вопрос апеллирует сразу к двум видам знаний (политематичен), например, к знаниям по физике и по биологии.

Независимость ответов на вопросы. Вероятность того, что данный человек ответит на данный конкретный вопрос не зависит от того, как он ответил на другие вопросы. Модель Раша успешно описывает ответы на такие вопросы, где правильный ответ успешно находится по ассоциативной связи между вопросом и ответом. Так что, если скажем для правильного ответа на вопрос № 5 необходимо знать правильный ответ на вопрос № 3, модель Раша не сможет правильно описать вероятности ответов, порождаемые таким, состоящим из последовательной активации двух ассоциативных связей процессом поиска.

Алгоритмы. Мы создали компьютерную программу «АнализIRT.exe», которая по матрице ответов испытуемых вычисляет трудность каждого задания и уровень успеваемости каждого испытуемого.

Вначале выборка испытуемых разбивается на подвыборки «нулевиков», «единичников», «двоечников» и т. д.

Для каждого вопроса вычисляется вероятность правильного ответа p, отдельно для каждой подвыборки.

В IRT литературе [8] доказано, что люди, набравшие одинаковый первичный балл имеют и одинаковый уровень способности — θ. Следовательно, вероятность p — найденная по подвыборке, применима к каждому человеку данной подвыборки. Следовательно, эту вероятность можно выразить формулой Раша (формулой, которая описывает именно единичного человека).

$$P(\text{правильного\_ответа}) = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}}.$$

$$\frac{P}{1-P} = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} \Big/ \left(1-\frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}}\right)$$

$$1-\frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} = \frac{e^{(\theta-\delta)}+1-e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} = \frac{1}{1+e^{(\theta-\delta)}}$$

$$\frac{P}{1-P} = \frac{e^{(\theta-\delta)}}{1+e^{(\theta-\delta)}} * \frac{1+e^{(\theta-\delta)}}{1}$$

$$\frac{P}{1-P} = e^{(\theta-\delta)}$$

Тогда:

$$\ln\left(\frac{P}{1-P}\right) = \theta-\delta$$

Может случиться так, что какие-то из величин p будут равны 1. Это приведет к делению на 0 в вышеприведенной формуле, p по модели Раша теоретически никогда не может быть равно 1. Однако эмпирическая частота правильных ответов на вопрос является только статистической оценкой вероятности p. Поэтому частота может отклоняться от значения вероятности. Соответственно относительная частота p, равная 1, рассматривается как приблизительная оценка некоей теоретической вероятности p, близкой к 1, но не равной 1. Исходя из этого вывода, значения относительных частот, равные 1, заменяются на значение 0,995. Это значение дает разность θ–δ равную 5,29 логита.

Может случиться так, что какие-то из величин *p* будут равны 0. Это приводит к выражению ln0, которое не существует. Такие значения *p* заменяются величиной 0,005, что дает разность –5,29 логита.

По формуле программа из вероятностей вычисляет величины разностей θ–δ для каждого сочетания уровень знаний-вопрос. Если имеется *K* вопросов и *N* возможных уровней знаний, в рассмотрение берутся *K***N* уравнений вида θ_i–δ_j.

В программе для нахождения решения системы уравнений используется многомерный вариант метода наименьших квадратов [4]. Этот метод рекомендуется в специализированной литературе, посвященной IRT [8]. Приняты меры к ослаблению влияния ошибок машинного округления на результат [2].

Дальнейшее уточнение полученных оценок. Решение, получаемое методом наименьших квадратов, вполне достаточно, если мы располагаем результатами, полученными на выборке, скажем в 1000 человек. Тогда каждая из вероятностей *p* скорее всего определена по многочисленной подвыборке (т. е. надежно) и не равна 0 или 1. Однако часто выборка испытуемых бывает не столь многочисленна. Для того чтобы получить максимально возможную в таких случаях точность, в нашей программе предусмотрен алгоритм уточнения решения, полученного методом наименьших квадратов. Для уточнения решения используется метод максимального правдоподобия, предложенный Р. Фишером [5].

Рассмотрим матрицу ответов испытуемых, например, такую:

0100110

1110010

0000001

Модель Раша предполагает, что правильный ответ возникает с определенной вероятностью.

Представим себе, что мы взяли некоторые значения параметров θ для каждого испытуемого в выборке и значения δ для каждого вопроса. Тогда мы можем, пользуясь формулой Раша, вычислить вероятность того, что в определенном месте матрицы ответов стоит именно то число (0 или 1), которое там стоит по результатам эмпирического тестирования. Так вычисляется вероятность того, что в клетке первой строки и первого столбца 0, вероятность того, что в клетке — пересечении первой строки и второго столбца 1 и т. д. Другое допущение модели Раша утверждает независимость ответов одного человека на разные вопросы. Следовательно, перемножив вероятности, относящиеся к ответам одного человека на разные вопросы, можно найти вероятность появления именно такой строки матрицы, которая появилась на самом деле при изначально принятых нами значениях параметров.

Понятно, что в условиях отсутствия общения между испытуемыми вероятность порождения человеком данной строки нулей и единиц никак не зависит от того, как отвечали остальные испытуемые. Следовательно, разные строки так же появляются независимо. Перемножив вероятности появления отдельных строк, получим вероятность того, что при данных значениях параметров мы получили бы именно такую матрицу ответов, как получили.

Р. Фишер доказал, что значения параметров (в нашем случае разнообразные θ и δ), при которых *P* (появления данной эмпирической матрицы) достигает максимума, являются наилучшими оценками параметров из тех, что можно получить, основываясь на данной матрице. Величину *P* (появления данной эмпирической матрицы) называют правдоподобием, и обозначают для краткости *L*.

ln L — это монотонно возрастающая функция L. Следовательно, наибольшему значению L соответствует наибольшее значение ln L. Программа ищет такие значения параметров, при которых функция минус ln L достигает минимума.

Наилучшим способом найти значения параметров, при которых –ln L достигает минимума, был бы перебор всех возможных комбинаций параметров (с шагом изменения каждого параметра, скажем 0,01) и вычисление –ln L для каждой комбинации. К сожалению, в данном случае, когда параметров несколько десятков, такая задача непосильна даже для современных ЭВМ.

Для поиска минимума функции –ln L в нашей программе применен метод градиентного спуска [9]. Градиент это вектор, компоненты которого — частные производные функции –ln L по отдельным параметрам.

Поскольку формула Раша известна, частные производные могут быть заданны аналитически, как функции параметров θ, δ и матрицы ответов. L можно выразить через первичные баллы людей и заданий, не прибегая к анализу самой матрицы ответов [8]. Анализ матрицы ответов был приведен нами для иллюстрации. Поэтому частные производные могут быть заданы аналитически, как функции параметров θ, δ и первичных баллов. Доказано, что функция –ln L быстрее всего убывает в направлении, противоположном направлению вектора градиента.

Работа нашей программы по уточнению оценок параметров, полученных методом наименьших квадратов, происходит следующим образом:

1) исходя из оценок метода наименьших квадратов и первичных баллов людей и заданий, с помощью аналитических выражений для частных производных рассчитывается вектор антиградиента функции –ln L;

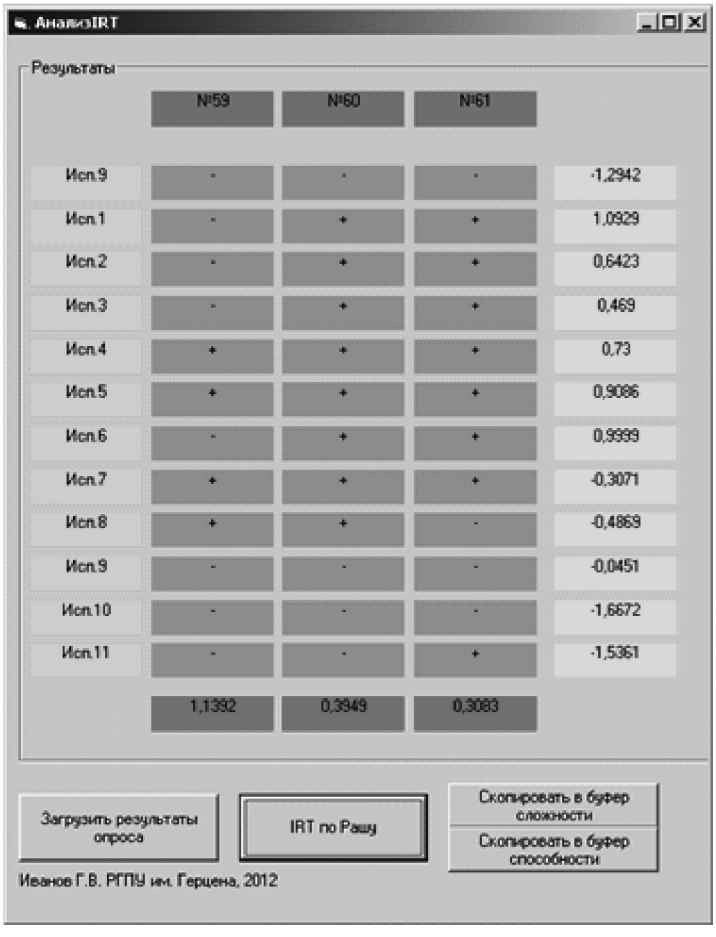
2) рассчитывается евклидова длина вектора антиградиента. Если она меньше 0,01, значит, все частные производные близки к 0 (практически нулевые). Следовательно, мы нашли минимум $\Rightarrow$ цель достигнута;

3) умножив вектор антиградиента на константу 0,01, получаем вектор сдвига. Этот вектор сдвига прибавляется к вектору оценок параметров. Новые оценки параметров вводятся в процедуру расчета антиградиента, и все повторяется сначала. Процесс продолжается до тех пор, пока не будет выполнено условие шага 2.

Графически работу алгоритма можно представить как сползание точки по склону ямы, дно которой — минимальное значение –ln L. Таких ям в рельефе функции правдоподобия может быть несколько, нас интересует самая глубокая. Если бы алгоритм градиентного спуска стартовал из случайно выбранной точки (например, 0, 0, 0…), велик бы был шанс попасть в не самую глубокую яму и принять локальный минимум за глобальный, не найдя наилучшие оценки параметров. Однако мы стартуем из координат, полученных методом наименьших квадратов и лежащих, если не у самого глобального минимума, то, во всяком случае, близко к нему. Следовательно, можно полагать, что точка в самом начале градиентного поиска уже лежит на склоне наиболее глубокой ямы, следовательно, сваливается на дно именно этой, самой глубокой ямы, где алгоритм и прекращается. Координаты дна ямы представляют собой уточненные оценки параметров модели Раша.

Использование. Данные, которые анализирует программа, должны быть представлены в виде файлов *.rh (это текстовые файлы, содержащие названия вопросов, имена испытуемых и матрицу ответов). Дополнительно к параметрам модели Раша программа выводит первичные баллы людей и заданий. И их, и результаты анализа по Рашу можно скопировать в буфер обмена, для дальнейшего анализа в Excel.

Отметим, в завершение, что поскольку наблюдаемые нами ответы зависят от разностей θ–δ к вычисленным программой значениям способностей и трудностей можно прибавить константу, и новые значения будут так же хорошо описывать данные, т. е. шкала, на которой отложены способности и трудности, — интервальная. Поэтому вычисленные программой значения θ и δ нормированы так, что сумма всех θ по выборке равна 0.



Интерфейс программы. Окно данных скроллировано, чтобы показать сложности вопросов (нижняя строка) и способности испытуемых (крайний правый столбец).

Апробация системы. С целью апробации системы было проведено тестирование студентов — будущих клинических психологов на психолого-педагогическом факультете РГПУ им. А. И. Герцена. Матрица ответов студентов по каждой дисциплине анализировалась с помощью программы «АнализIRT.exe». Для каждого студента был оценен уровень знаний, а для каждого вопроса — уровень сложности.

Анализ группировки дисциплин. Педагогика, как и любая наука, изучает связи между явлениями действительности. В случае нашей работы такими явлениями стали знания студентов по разным дисциплинам.

Возникает вопрос, не существует ли общих причин, каждая из которых обусловливает уровень знаний по нескольким дисциплинам. Такие общие причины и представляли бы собой компетенции.

Мы попытались выявить общие причины уровней знаний, подвергнув кластерному анализу матрицу корреляций между уровнями знаний по разным дисциплинам [7].

Результат представлен ниже:

Ясно выделяются 3 «эмпирические компетенции». Первая объединяет нейропсихологию, аномалии развития и патопсихологию. Эта компетенция связанна с работой клинического психолога с тяжелыми заболеваниями. Вторая «эмпирическая компетенция» (клиническая психология, методы клинической психологии) может быть названа отвечающей за общую профессиональную культуру. Третья эмпирическая компетенция может быть названа психотерапевтической, так как в нее входят дисциплины обучающие работе с личностью.

Выводы

1. IRT представляет собой привлекательный подход к педагогическим измерениям, при котором преодолеваются многие недостатки более ранних подходов.

2. Представленное педагогической общественности программное обеспечение может быть с успехом использовано для мониторинга уровня знаний в высшем образовании, заменяя иностранные пакеты (www.ssicentral.com).

3. Кластерный анализ успеваемости указывает на существование «эмпирических компетенций», возможно, не всегда совпадающих с компетенциями Федеральных государственных образовательных стандартов.

Учитывая, что представленные принципы и способы построения компетентностно-ориентированных тестов можно адаптировать к мониторингу знаний по любой дисциплине, четко сформулировав задания, моделирующие реальные профессиональные проблемы в соответствующей предметной области. В соответствии с этим, разработанная методика измерения компетенций обучающихся посредством автоматизированной системы ситуационных задач является новым эффективным инструментом оценки качества образования.

Литература

1. *Аванесов В. С.* Item Response Theory: основные понятия и положения Статья первая // Педагогические измерения 2007. № 2.
2. *Деммель Дж.* Вычислительная линейная алгебра: теория и приложения. М.: Мир, 2001.
3. *Иванов Г. В.* О связи между стохастической теорией тестов и теорией выбора по сходству в аспекте конструирования систем мониторинга знаний // Материалы конкурса научно-исследовательских работ студентов и аспирантов в области информатики и информационных технологий, Белгород: ФГАОУ ВПО «Белгородский государственный национальный исследовательский университет», 2012.
4. *Стренг Г.* Линейная алгебра и ее применения. М.: Мир, 1980.

5. *Тюрин Ю. Н., Макаров А. А.* Анализ данных на компьютере. М.: Инфра-М, 2003.
6. *Лазарев П. П.* Законы физики и законы биологии. Отд. оттиск журнала «Природа». М., 1915.
7. *Наследов А. Д.* Математические методы психологического исследования. Анализ и интерпретация данных: Учебное пособие. СПб.: Речь, 2004.
8. *Нейман Ю. М., Хлебников В. А.* Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей, 2000.
9. *Носач В. В.* Решение задач аппроксимации с помощью персональных компьютеров. М.: Микап, 1994
10. *Фер Р. М., Бакарак В. Р.* Психометрика: Введение. Челябинск: «Издательский центр ЮУрГу», 2010.