УДК 001.891:65.011.56

МОДЕЛИ ТЕСТИРОВАНИЯ ЗНАНИЙ И МЕТОДЫ ОЦЕНКИ НАДЕЖНОСТИ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

О.Ю. Чередниченко
Кандидат технических наук, доцент*
Контактный тел.: (057) 707-64-74
E-mail: marxx75@mail.ru
С.И. Ершова
Старший преподаватель*
Контактный тел.: (057) 707-64-74
E-mail: esi@kpi.kharkov.ua
О.В. Янголенко
Аспирантка*
Контактный тел.: 098-438-11-14
E-mail: olga_ya26@mail.ru
Т.Н. Запорожец*
Контактный тел.: 095-702-91-03
E-mail: tatysya110589@mail.ru
*Кафедра автоматизированных систем управления
Национальный технический университет «Харьковский политехнический институт»
ул. Фрунзе, 21, г. Харьков, Украина, 61002

Наведено порівняльний аналіз існуючих статистичних методів оцінювання знань. Розглянуті різні моделі тестів. Проаналізовані основні методи та способи оцінки надійності результатів тестування

Ключові слова: тестування знань, модель тесту, коефіцієнт надійності

Приведен сравнительный анализ существующих статистических методов оценки знаний. Рассмотрены различные модели тестов. Проанализированы основные методы и способы оценки надежности результатов тестирования

Ключевые слова: тестирование знаний, модель теста, коэффициент надежности

The comparative analysis of existing statistical methods of knowledge estimation is given. Different test models are considered. The basic methods and ways of test reliability estimation are analyzed

Keywords: knowledge testing, test model, reliability coefficient

Введение

На сегодняшний день все высшие учебные заведения Украины вовлечены в процесс реформирования системы образования и приведения ее к европейским стандартам. На этом фоне растет роль педагогического контроля, который имеет целью выявление и оценивание результатов учебной деятельности студентов.

В отличие от традиционных субъективных оценок, выставляемых преподавателями (в 5-тибальной, 12-тибальной шкале или от A до F), тестирование позволяет получить объективный результат измерения. Педагогическое измерение предполагает количественное сопоставление оцениваемых знаний, умений и навыков студента с некоторым эталоном этих свойств с помощью контрольных заданий по проверяемому содержанию предметной области. Тестирование дает возможность статистически анализировать результаты образования учащихся.

Актуальность тестирования обусловлена его преимуществами перед другими методами педагогического контроля. Помимо объективности измерения качества учебных достижений, речь идет и о точности измерений, наличии единых требований для всех студентов, совместимости тестирования с другими современными образовательными технологиями, например, дистанционным образованием.

В основе моделирования и параметризации процесса тестирования лежит множество математических методов, основные из которых: теория вероятности, математическая статистика, дисперсионный и регрессионный анализ. Типичные задачи оценки знаний методом тестирования, их формализованное описание и классификация даны в работах Дж. Раша [1], А. Бирнбаума, Дж. Мастерса [2], В.С. Аванесова [3-5], М.Б. Челышковой [6], А.Н. Майорова [7], А.А. Маслака [8], Ю.М. Неймана и В.А. Хлебникова [9].

Целью данного исследования является системный анализ существующих моделей тестирования и методов оценки надежности тестов.

Как показал анализ, теоретическую основу для создания и использования тестов составляют две теории: классическая теория тестов (Classical Test Theory – CTT) [6, 10] и современная теория измерений (Item Response Theory – IRT) [1]. Эти теории научно обосновывают способность теста быть измерительным инструментом качества подготовки учащихся, предоставляя математический аппарат для статистической обработки результатов тестирования.

Модели классической теории тестов (CTT)

Классическая теория тестов основывается на следующих базовых предположениях [10]:

1) $Y_i = \tau_i + \varepsilon_i$;

2) $\mathrm{Cov}(\tau_i, \varepsilon_j) = 0$;

3) $E(\varepsilon_i) = 0$;

4) $\mathrm{Var}(Y_i) = \mathrm{Var}(\tau_i) + \mathrm{Var}(\varepsilon_i)$;

5) $\mathrm{Rel} = 1 - \frac{\mathrm{Var}(\varepsilon_i)}{\mathrm{Var}(Y_i)}$ или $\mathrm{Rel} = \frac{\mathrm{Var}(\tau_i)}{\mathrm{Var}(Y_i)}$.

Предположение 1 утверждает, что эмпирически полученный результат измерения (Y_i) представляет собой сумму истинного результата измерения (τ_i) и ошибки измерения (ε_i). Величины τ_i ε_i и обычно неизвестны. Из предположения 4 о том, что дисперсия полученных тестовых баллов равна сумме дисперсий истинных и ошибочных компонентов, вытекает предположение 5 об оценке надежности теста.

Надежность представляет собой важнейшую характеристику теста. Надежность отражает точность тестовых измерений и устойчивость тестовых результатов к действию случайных факторов [6]. Высокая надежность означает высокую повторяемость результатов тестирования в одинаковых условиях.

Чтобы определить надежность теста с помощью эмпирически оцененных параметров, основные постулаты должны быть дополнены предположениями, определяющими модель оценивания. Наиболее важными предположениями согласно [10] являются:

6) τ-эквивалентность: $\tau_i = \tau_j$;

7) существенная τ -эквивалентность: $\tau_i = \tau_j + \lambda_{ij}$, $\lambda_{i,j} \in \Re$;

8) τ-однородность: $\tau_i = \lambda_{ij0} + \lambda_{ij1}\tau_j$, $\lambda_{ij0}, \lambda_{ij1} > 0$;

9) некоррелированные ошибки: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$;

10) равные дисперсии ошибок: $\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(\varepsilon_j)$.

Предположения 1 и 2 разными путями устанавливают то, что два теста измеряют одну и ту же область знаний. Это предположение является чрезвычайно важным для того, чтобы сделать вывод о степени надежности из различий между двумя измерениями знаний в одной и той же предметной области одного человека.

Предположение 1 подразумевает идеальную идентичность двух переменных истинных результатов. Предположение 2 позволяет двум истинным результатам отличаться на значение положительной константы. Согласно предположению 3 два теста измеряют знания в одной и той же области, устанавливая, что между истинными результатами есть линейная зависимость. Предположение 4 допускает, что ошибки измерения, относящиеся к разным тестовым оценкам, не коррелируются. Предположение 5 о равных дисперсиях ошибок позволяет говорить о том, что два теста дают одинаково хорошие результаты измерений.

Комбинации данных предположений позволяют определить наиболее важные модели тестов [6, 7, 10, 11]:

– параллельные тесты (предположения 1, 4 и 5);

– существенно τ-эквивалентные тесты (предположения 2 и 4);

– τ-однородные тесты (предположения 3 и 4).

Наиболее простым набором предположений определяется модель параллельных тестов. Два теста и Y_j параллельны, если они τ-эквивалентны, их ошибки не коррелируют и они имеют одинаковые дисперсии ошибок. Предположение 1 подразумевает, что существует однозначно определенная латентная переменная уровня знаний, идентичная для каждой истинной оценки.

Поэтому можно упустить индекс i и обозначить эту латентную переменную η. Тогда можно записать, что $Y_i = \eta + \varepsilon_i$.

Для параллельных тестов теоретические параметры могут быть рассчитаны по параметрам, характеризующим распределение хотя бы двух тестовых оценок, то есть теоретические параметры идентифицируются в данной модели при $m \geq 2$, m – количество имеющихся тестовых оценок. Математическое ожидание переменной η равно математическому ожиданию каждого из тестов, несмотря на то, что дисперсия η может быть рассчитана из ковариации двух разных тестов. Дисперсия $\mathrm{Var}(\varepsilon_i)$ ошибки измерения рассчитывается как разница $\mathrm{Var}(Y_i) - \mathrm{Cov}(Y_i, Y_j), \quad i \neq j$.

Таким образом, модель параллельных тестов владеет следующими свойствами:

11) $E(\eta) = E(Y_i)$;

12) $\mathrm{Var}(\eta) = \mathrm{Cov}(Y_i, Y_j), \quad i \neq j$;

13) $\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i) - \mathrm{Cov}(Y_i, Y_j), \quad i \neq j$;

Модель существенно τ-эквивалентных тестов имеет меньше ограничений, чем модель параллельных тестов. Два теста Y_i и Y_j существенно τ -эквивалентны, если их истинные оценки разнятся на положительную константу и их ошибки не коррелируют. Предположение 2 подразумевает, что существует латентная переменная η, которая является преобразованием каждой истинной оценки: $\eta = \tau_i + \lambda_i, \quad \lambda_i \in \Re$. Латентная переменная η однозначно определена для преобразования, поэтому необходимо фиксировать шкалу ее измерения.

Можно фиксировать один из коэффициентов λ_i (например, $\lambda_i = 0$) или математическое ожидание η (например, $E(\eta) = 0$).

Таким образом, свойства существенно τ-эквивалентных тестов при фиксировании шкалы η $E(\eta) = 0$ следующие:

14) $\mathrm{Var}(\eta) = \mathrm{Cov}(Y_i, Y_j), \quad i \neq j$;

15) $\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i) - \mathrm{Cov}(Y_i, Y_j), \quad i \neq j$;

Два теста Y_i и Y_j τ-однородные, если их истинные оценки являются положительными линейными функциями друг друга и их ошибки не коррелируют. Предположение 3 подразумевает, что существует латентная переменная η такая, что каждая истинная оценка является положительной линейной функцией дру-

гой истинной оценки, то есть $\tau_i = \lambda_{i0} + \lambda_{i1}\eta$, $\lambda_{i0}, \lambda_{i1} \in \Re$, $\lambda_{i1} > 0$ или $Y_i = \lambda_{i0} + \lambda_{i1}\eta + \varepsilon_i$.

Латентная переменная η однозначно определена для положительных линейных функций. Поэтому для этой модели также необходимо фиксировать шкалу η. Этого можно достичь, фиксируя пару коэффициентов (например, $\lambda_{i0} = 0$ и $\lambda_{i1} = 1$) или математическое ожидание и дисперсию (например, $E(\eta) = 0$ и $Var(\eta) = 1$).

Все параметры модели τ-однородных тестов определяются, если есть хотя бы три разных теста, для которых предположения 3 и 4 выполняются. Следующиесвойствасформулированыпредполагая,что $E(\eta) = 0$ и $Var(\eta) = 1$. Другие способы фиксации шкалы η приведут к другим формулам.

Свойства τ-однородных тестов при фиксировании шкалы η $E(\eta) = 0$ и $Var(\eta) = 1$ выглядят следующим образом:

16) $\lambda_{i1} = \sqrt{\dfrac{Cov(Y_i, Y_j)\,Cov(Y_i, Y_k)}{Cov(Y_j, Y_k)}}$, $i \neq j, i \neq k, j \neq k$;

17) $Var(\varepsilon_i) = Var(Y_i) - \lambda_{i1}^2$;

Таким образом, CTT позволяет выразить уровень знаний с помощью вероятностных распределений соответствующих латентных переменных и является распространенным инструментом создания тестов.

Модели Item Response Theory (IRT)

CTT имеет серьезный недостаток: измерение знаний испытуемых зависит от характеристик тестовых заданий. В этой ситуации сложно сравнить испытуемых, которые прошли тесты, отличающиеся хотя бы на одно задание, или сравнить задания, которые даются разным группам испытуемых. IRT была разработана для того, чтобы справиться с этим недостатком.

К тому же, IRT может быть использована для прогнозирования свойств всего теста с помощью свойств его заданий, а также для манипуляций с частями теста с целью достичь заданных свойств измерения [12].

IRT предлагает широкий выбор моделей для тестов с дихотомическими и многовариантными заданиями. В IRT устанавливается связь между двумя множествами значений латентных переменных [11]. Первое множество составляют значения латентной переменной, определяющей уровень подготовленности испытуемых θ_i, где i - номер испытуемого ($i = \overline{1,N}$, N – количество испытуемых). Второе множество составляют значения латентной переменной, характеризующей трудность j-го задания β_j ($j = \overline{1,M}$, M – количество заданий в тесте).

Г. Раш предположил, что уровень подготовленности испытуемого θ_i и уровень трудности задания β_j размещены на одной шкале и измеряются в одних и тех же единицах – логитах [1, 13]. Аргументом функции успеха испытуемого является разность $(\theta_i - \beta_j)$. Поскольку модель Раша описывает вероятность успеха испытуемого как функцию одного параметра $(\theta_i - \beta_j)$, то иногда ее называют однопараметрической моделью IRT [11]. Модель Раша определяет вероятность правильного ответа следующим образом [12]:

$$P(X_{ij} = 1 | \theta_i, \beta_j) = f(1, \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)},$$

где X_{ij} – ответ испытуемого i на задание j (равен 1, если – верно, 0 – если неверно);

θ_i – латентная переменная подготовленности испытуемого i;

β_j – латентная переменная трудности задания j.

Если тест содержит задания с различной дифференцирующей способностью, то однопараметрическая модель Раша не может описать такие эмпирические данные. То есть задания отличаются не только трудностью, но и тем, насколько хорошо они оценивают латентную переменную подготовленности испытуемых. Для преодоления этой трудности А.Бирнбаум ввел еще один параметр дифференциации – α (item discrimination parameter) [11]. Этот параметр имеет эффект усиления важности разницы между подготовленностью испытуемого и сложностью задания. Согласно двухпараметрической модели Бирнбаума вероятность правильного ответа равна [13]:

$$P(X_{ij} = 1 | \theta_i, \alpha_j, \beta_j) = \frac{1}{1 + \exp \alpha_j(\beta_j - \theta_i)}.$$

Изначально IRT была разработана для заданий, где не предполагались угадывания. Но если речь идет о возможности выбора между несколькими вариантами ответов в тестовых заданиях, можно ответить на какие-нибудь задания наугад правильно. Не имея никаких знаний в материале, человек может случайно угадать правильный ответ по крайней мере на 1% заданий.

Задания могут различаться тем, насколько они позволяют угадать ответ при низком уровне знаний материала. На основании этого в трехпараметрическую модель Бирнбаума вводится параметр угадывания γ [13]:

$$P(X_{ij} = 1 | \theta_i, \alpha_j, \beta_j, \gamma) = \gamma + \frac{1 - \gamma}{1 + \exp \alpha_j(\beta_j - \theta_i)}.$$

Модели Раша и Бирнбаума относятся к тестам с дихотомическими заданиями. IRT гораздо легче адаптируется к переходу от дихотомических заданий к многовариантным, чем CTT. К основным моделям IRT для тестов с многовариантными заданиями относят:

– Graded Model (GM);
– Nominal Model (NM);
– Partial Credit Model (PCM);
– Rating Scale Model (RSM).

В модели со ступенчатыми или упорядоченными ответами Graded Model [13] ответ может быть ранжирован на шкале баллов, например, от слабого (0) до отличного (9). В GM логистическая функция определяет вероятность того, что ответ будет получен в категории K или выше. Для упорядоченных

категорий ответов $X_{ij}=k$, k=1,2,...m, где m соотносится с наивысшим значением θ, вероятность правильного ответа равна:

$$P(X_i = k|\theta) = \frac{1}{1+\exp(-\alpha_i(\theta-\beta_{jk}))} - \frac{1}{1+\exp(-\alpha_i(\theta-\beta_{jk+1}))}.$$

Кривая P $(X_i=k|\theta)$ будет немонотонной, за исключением первой и последней категории ответов задания. Для первой категории k=1 характеристическая кривая будет монотонно убывающей логистической функцией с нижним пороговым значением $P(X_i = 1|\theta) = 1 - \frac{1}{1+\exp(-\alpha_i(\theta-\beta_{j2}))}$. Для последней категории k=m характеристическая кривая будет монотонно возрастающей логистической функцией с верхним пороговым значением $P(X_i = m|\theta) = \frac{1}{1+\exp(-\alpha_i(\theta-\beta_{jm}))}$.

Nominal Model [13] является альтернативой GM для многовариантных заданий, не требуя заранее никакой зависимости порядка взаимоисключающих категорий и переменной подготовленности θ. NM характеристической кривой для категорий $u = 1,2,...,m_j$ задания j выглядит:

$$P(u_j = x|\theta) = \frac{\exp(\alpha_{jx}\theta + c_{jx})}{\sum_{k=1}^{m}\exp(\alpha_{jk}\theta + c_{jk})},$$

где α_k – параметр дифференциации;

c_k – пересечения.

Дополнительные ограничения вводятся для идентификации модели. Сумма каждого набора параметров должна быть равной нулю: $\sum_{k=0}^{m-1}\alpha_k = \sum_{k=0}^{m-1} c_k = 0$.

NM используется, когда между вариантами ответа может быть определен порядок. Другими словами, модель позволяет определить, какой порядок вариантов ответа ассоциируется с высоким уровнем латентной переменной подготовленности. Эта модель также используется для определения местонахождения нейтрального ответа на шкале Лайкерта среди упорядоченных ответов.

Partial Credit Model [2, 14, 15] предполагает, что из упорядоченных категорий задания 0<1<2,...,<m условная вероятность выбора в задании категории x, а не x-1, должна монотонно возрастать на области определения латентной переменной подготовленности.

В PCM вводится параметр задания δ_{ix}, управляющий вероятностью выбрать категорию x, а не $x-1$. Параметр δ_{ix} может рассматриваться как шаговая сложность задания, ассоциируемая с лежащей в основе характеристикой, где категории x-1 и x пересекаются.

Вероятность того, что человек i выберет категорию x с одним из возможных результатов 0, 1, 2,..., m задания j может быть представлена в следующем виде:

$$P(u_i = x|\theta_i) = \frac{\sum_{k=0}^{x}(\theta_i - \delta_{jk})}{\sum_{h=0}^{m_j}\exp\sum_{k=0}^{h}(\theta_i - \delta_{jk})}, \quad x = 0,1,...,m_j.$$

Rating Scale Model [13] получена из модели PCM с ограничением на равные коэффициенты дифференциации для всех заданий. Отличие этой модели в том, что расстояние между шагами сложности от категории до категории внутри каждого задания одинаковы для всех заданий. Модель RSM включает дополнительный параметр λ_j, который располагает задание j на шкале измерения. Функция ответа для безусловной вероятности того, что человек i выберет категорию x с одним из возможных результатов 0, 1, 2,..., m задания j равна:

$$P(u_i = x|\theta_i) = \frac{\exp\sum_{k=0}^{x}(\theta_i - (\lambda_j + \delta_k))}{\sum_{h=0}^{m_j}\exp\sum_{k=0}^{h}(\theta_i - (\lambda_j + \delta_k))}, \quad \sum_{k=0}^{0}(\theta_i - (\lambda_j + \delta_k)) = 0.$$

Данная модель требует, чтобы формат заданий был одинаковым на всей шкале (например, все задания имели бы четыре категории ответов).

Таким образом, теория IRT является альтернативой CTT и позволяет измерять знания испытуемых с помощью линейной шкалы.

Методы измерения надежности в рамках классической теории тестов

Качество теста чаще всего характеризуется надежностью и валидностью. Тест считается надежным, если при повторном выполнении он дает близкие результаты при условии, что подготовка учащегося не изменилась. Надежность характеризует воспроизводимость результатов тестирования, а валидность – это характеристика адекватности теста поставленной цели его создания.

В рамках классической теории тестов были сформированы основные способы оценивания надежности результатов тестирования. Наиболее простым способом является оценка корреляции двух разных тестовых оценок, что соответствует модели параллельных тестов[10]: $Rel = Corr(Y_i, Y_j)$.

Надежность суммарной оценки $S = Y_1 + ... + Y_m$ параллельных тестов считается по формуле Спирмена-Брауна:

$$R(S) = \frac{m \cdot Rel(Y_i)}{1 + (m-1)\cdot Rel(Y_i)}, \quad (1)$$

где m – количество параллельных тестов.

Существует несколько основных методов получения данных, необходимых для расчета надежности тестов, соответствующих модели параллельных тестов [6]. Ретестовый метод основан на подсчете корреляции индивидуальных баллов испытуемых, полученных в результате двукратного выполнения одного теста с интервалом времени в 2-3 недели.

Этот метод оценки надежности прост в вычислениях, но его недостатком является трудность определения временного интервала между проведениями двух тестирований. Близкое по времени повторное тестирование может дать высокую надежность, однако она не будет объективно характеризовать качество теста.

Метод параллельных форм более предпочтителен по сравнению с ретестовым, поскольку он снижает степень влияния свойств человеческой памяти запоминать задания и ответы предыдущего теста. Недостатком данного метода является сложность составления тестов и необходимость доказательства их параллельности.

Метод расщепления теста предполагает однократное проведение теста. Множество тестовых заданий делится на две половины, например, все четные и нечетные задания. Эти части теста могут рассматриваться как приближение к параллельным формам. Корреляция между результатами двух частей будет надежностью каждой из половин теста, но не всего теста. Для оценки надежности всего теста необходимо использовать формулу Спирмена-Брауна (1), в которой m принимает значение количества частей теста, полученных в результате расщепления.

Недостаток метода расщепления заключается в том, что делить тест можно разными способами. Каждое деление будет давать немного отличающуюся корреляцию между частями теста. Таким образом, можно получить разные оценки надежности даже при выполнении одних и тех же заданий теми же людьми.

В отличие от параллельных тестов, надежность в модели существенно τ-эквивалентных тестов не может быть определена как корреляция между двумя тестами. В этом случае надежность равна [10]: $\mathrm{Rel}(Y_i) = \mathrm{Cov}(Y_i, Y_j) / \mathrm{Var}(Y_i), \quad i \neq j$.

Для существенно τ-эквивалентных тестов надежность их суммарной оценки $S = Y_1 + \ldots + Y_m$ может быть рассчитана с помощью коэффициента Кронбаха:

$$\alpha = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{1}^{m} \mathrm{Var}(Y_i)}{\mathrm{Var}(S)} \right), \qquad (2)$$

где m – количество тестов.

Этот коэффициент является нижней границей надежности S, если речь идет только о некоррелируемых ошибках.

Также коэффициент Кронбаха используется, чтобы избежать деления теста или повторных тестирований. В этом случае речь идет о методах оценки надежности по внутренней согласованности теста [6, 7].

Тогда в формуле (2) $\mathrm{Var}(Y_i)$ – это дисперсия каждого задания теста, $\mathrm{Var}(S)$ – дисперсия всего теста, m – количество заданий теста.

К достоинствам метода, основанного на расчете коэффициента Кронбаха, относится возможность оценить надежность теста, состоящего как из дихотомических заданий, так и из многовариантных заданий, использование всей статистической информации, которую несут задания и легкость расчета. Частным случаем коэффициента Кронбаха для дихотомических заданий является коэффициент Кьюдера-Ричардсона KR20 [16].

Для оценки надежности тестов, реализованных с помощью моделей IRT, также рассчитывается коэффициент Separation Reliability (SR) [16] для испытуемых и заданий. Для испытуемых он отражает то, насколько хорошо множество заданий позволяет разделить уровни подготовленности испытуемых. Для заданий SR отражает, насколько множество испытуемых позволяет разделить уровни трудности заданий теста. SR рассчитывается путем вычитания из единицы отношения среднего квадрата ошибки $\mathrm{MSE}(\eta)$ рассматриваемой латентной переменной η к дисперсии η $\mathrm{Var}(\eta)$, т.е. $\mathrm{SR} = 1 - \frac{\mathrm{MSE}(\eta)}{\mathrm{Var}(\eta)}$.

Таким образом, в основе различных методов расчета коэффициентов надежности лежат достижения CTT.

Выбор метода оценки надежности, вообще говоря, определяется моделью теста.

Выводы

Ограничение CTT состоит в том, что модели этой теории не совсем адекватны для моделирования ответов на отдельные задания теста. Эта задача лучше решается с помощью моделей IRT, которая определяет, как вероятность правильных ответов на одну категорию вопросов зависит от измеряемой латентной переменной.

Второе ограничение заключается в том, что CTT сфокусирована исключительно на ошибках измерения. Преимущества IRT по сравнению с CTT следующие [11]:

18) IRT превращает измерения, выполненные в дихотомических и порядковых шкалах, в линейные измерения, в результате качественные данные анализируются с помощью количественных методов;

19) мера измерения параметров модели Раша является линейной, что позволяет использовать широкий спектр статистических процедур для анализа результатов измерений;

20) оценка трудности тестовых заданий не зависит от выборки испытуемых, на которых она была получена;

21) оценка уровня подготовленности испытуемых не зависит от используемого набора тестовых заданий;

22) неполнота данных не является критичной.

Расчет надежности тестов производится на основе предположений, касающихся различных моделей CTT.

В этом состоит главная заслуга данной теории. IRT для оценки надежности использует результаты классической теории тестов в этой области.

Наиболее гибкий подход к созданию и использованию тестов заключается в совместном использовании обеих теорий.

Классическая теория тестов обеспечивает этап создания теста, а также оценки его качества, а Item Response Theory позволяет получить устойчивые оценки латентных параметров испытуемых и трудности заданий.

Литература

1. Rash G. On Objectivity and Specificity of the Probabilistic Basis for Testing // http://www.rasch.org/memo196x.pdf, 10.09.2011.
2. Masters G. N., Wright B. D. The Essential Process in the Family of Measurement Models [Электронный ресурс] / Режим доступа : \www/ URL: / http://personal.psc.isr.umich.edu/yuxie-web/files/soc543-004/Masters_et_al1984.pdf - 01.09.2011 г. - Загл. с экрана.
3. Аванесов В. С. Item Response Theory: основные понятия и положения [Электронный ресурс] / Режим доступа : \www/ URL: http://testolog.narod.ru/Theory59.html/ — 10.10.2011 г. - Загл. с экрана.
4. Аванесов В. С. Проблема объективности педагогических измерений [Электронный ресурс] / Режим доступа : \www/ URL: http://testolog.narod.ru/Theory34.html/ - 25.09.2011 г. - Загл. с экрана.
5. Аванесов В. С. Метрическая система Георга Раша - Rasch Measurement (RM) [Электронный ресурс] / Режим доступа : \www/ URL: http://testolog.narod.ru/Theory68.html/ - 16.10.2011 г. - Загл. с экрана.
6. Челышкова, М. Б. Теория и практика конструирования педагогических тестов: Учебное пособие [Текст] / М. Б. Челышкова. - М. : Логос, 2002. - 432 с.
7. Майоров, А. Н. Теория и практика создания тестов для системы образования (Как выбирать, создавать и использовать тесты для целей образования) [Текст] / А. Н. Майоров. - М. : Интеллект-центр, 2001. - 296 с.
8. Маслак, А. А. Измерение латентных переменных в социально-экономических системах: Монография [Текст] / А. А. Маслак. - Славянск-на-Кубани : Изд. Центр СГПИ, 2006. - 333 с.
9. Нейман, Ю. М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов [Текст] / Ю. М. Нейман, В. А. Хлебников. - М. : Прометей, 2000. — 168 с.
10. Steyer R. Classical (Psychometric) Test Theory [Электронный ресурс] / Режим доступа : \www/ URL: / http://metheval.uni-jena.de/materialien/ publikationen/ctt.pdf/ - 01.08.2011 г. - Загл. с экрана.
11. Ким, В. С. Тестирование учебных достижений. Монография. [Текст] / В. С. Ким. - Уссурийск : Издательство УГПИ, 2007. — 169 с.
12. Mislevy R. J., Wilson M. R., Ercikan K., Chudowsky N. Psychometric Principles in Student Assessment. International Handbook of Educational Evaluation [Электронный ресурс] / Режим доступа : \www/ URL: / http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.8477/ - 09.10.2011 г. - Загл. с экрана.
13. Reeve B. An Introduction to Modern Measurement Theory [Электронный ресурс] / Режим доступа : \www/ URL: / http://moaweb.nl/bibliotheek/materiaal-bijeenkomsten-1/2009/pretesten-van-vragenlijsten-23-juni/ - 19.08.2011 г. - Загл. с экрана.
14. Masters, G. N. Partial Credit Model [Text] / G. N. Masters ; Encyclopedia of Social Measurement. - Elsevier/Academic Press, 2005. - 3000 p.
15. Masters G. N. The Analysis of Partial Credit Scoring [Электронный ресурс] / Режим доступа : \www/ URL: / http://nccu.edu.tw/~mnyu/Study%20of%20Test%20Theory/The%20analysis%20of%20partial%20credit%20scoring.pdf - 11.10.2011 г. - Загл. с экрана.
16. Wright, B. Measurement Essentials. 2nd edition [Text] / B. Wright, M. Stone. - Wilmington, Delaware : Wide Range, Inc., 1999. - 221 p.