УДК 37.02

НОВОЕ НАПРАВЛЕНИЕ В ТЕОРИИ ТЕСТИРОВАНИЯ

Ключевые слова: теория тестирования, критический анализ, новая модель тестирования, решение тестовых заданий, однородный во времени стохастический процесс.

Попов А.П.

кандидат физико-математических наук, начальник отдела контроля качества образования Педагогического института Южного федерального университета

1. Анализ параметрических моделей тестирования

Введение

Теория и практика дидактического тестирования насчитывают не один десяток лет и в своем развитии прошли за это время несколько этапов. Один из этапов связан с появлением в начале 60-х гг. прошлого века параметрических моделей, призванных сделать процедуру оценивания трудности тестовых заданий и уровня подготовленности испытуемых более объективной. Наибольшее распространение получили модели Раша и Бирнбаума вместе с сопутствующими методами обработки результатов тестирования. В модели Раша [8] предполагается, что вероятность правильного решения тестового задания зависит от уровня подготовленности испытуемого ϑ и трудности тестового задания δ. В двухпараметрической модели Бирнбаума [9; 10] появляется новый параметр γ – дифференцирующая способность тестового задания. Наконец, в трехпараметрической модели Бирнбаума появляется еще один параметр, позволяющий учесть возможность случайного выбора (угадывания) правильного ответа.

К сожалению, из-за крайне неудачного выбора параметризации обе модели обладают неустранимым недостатком, а именно, как будет показано ниже, оценка уровня подготовленности испытуемого, основанная на общепринятом в математической статистике принципе максимального правдоподобия, в рамках этих моделей не зависит от сложности правильно выполненных заданий.

Критические замечания по поводу параметрических моделей Раша и Бирнбаума целиком относятся и к методу, применяемому в системе ЕГЭ и ЦТ для оценивания уровня подготовленности учащихся, поскольку в основе этого метода лежит все та же модель Раша [2].

1.1. Модель Раша

В модели Раша вероятности правильного (соответственно, неправильного) решения тестового задания равны:

$$p(\vartheta,\delta)=\frac{e^{\vartheta-\delta}}{e^{\vartheta-\delta}+1},\qquad q(\vartheta,\delta)=\frac{1}{e^{\vartheta-\delta}+1}. \tag{1.1}$$

Пусть тест содержит n заданий, трудности которых $\delta_1, \delta_2, \ldots, \delta_n$ предполагаются известными. Введем обозначение χ_i для характеристической функции, которая равна 1 или 0 в зависимости от того, верно или неверно решено i-е задание. Тогда логарифмическая функция правдоподобия:

(1.2)

Необходимое условие максимума функции (1.2), а именно равенство нулю производной по параметру ϑ, приводит к уравнению:

$$\sum_{i=1}^{n} p(\vartheta,\chi_i)=\sum_{i=1}^{n}\chi_i\ . \tag{1.3}$$

В соответствии с принципом максимального правдоподобия уровень подготовленности испытуемого должен определяться именно из этого уравнения. Левая часть уравнения (1.3) является универсальной для данного теста, монотонно возрастающей функцией $f(\vartheta)$, а правая часть равна общему числу правильно выполненных заданий. Таким образом, оценка уровня подготовленности в модели Раша зависит лишь от общего числа правильно выполненных заданий, но не от их трудности.

1.2. Модель Бирнбаума

В двухпараметрической модели Бирнбаума вероятности правильного и неправильного решения тестового задания равны:

(1.4)

Пусть тест содержит n заданий. Будем считать известными не только трудности заданий $\delta_1, \delta_2, \ldots, \delta_n$, но и дифференцирующие способности всех заданий $\gamma_1, \gamma_2, \ldots, \gamma_n$. Логарифмическая функция правдоподобия:

$$\ln(F(\vartheta))=\sum_{i=1}^{n}\chi_i\gamma_i(\vartheta-\delta_i)-\sum_{i=1}^{n}\ln(e^{\gamma_i(\vartheta-\delta_i)}+1). \tag{1.5}$$

Необходимое условие максимума функции (1.5) приводит к уравнению:

$$\sum_{i=1}^{n}\gamma_i p(\vartheta,\delta_i,\gamma_i)=\sum_{i=1}^{n}\chi_i\gamma_i\ , \tag{1.6}$$

из которого должен определяться уровень подготовленности испытуемого.

Но левая часть уравнения (1.6) является универсальной для данного теста, монотонно возрастающей функцией $f(\vartheta)$, а правая часть равна суммарной дифференцирующей способности всех правильно выполненных заданий. Следовательно, в модели Бирнбаума оценка уровня подготовленности испытуемого зависит от суммарной дифференцирующей способности правильно выполненных заданий, но никак не связана с их трудностью.

Трехпараметрическая модель Бирнбаума [7], в которой появляется еще один параметр, позволяющий учесть возможность случайного выбора правильного ответа в заданиях закрытого типа, также не избавлена от описанного выше недостатка.

1.3. Обсуждение и выводы

Следует напомнить простую, но чрезвычайно важную истину: моделирова-

ние – не цепочка произвольных, чисто умозрительных построений, но попытка наиболее адекватного описания изучаемого объекта, а потому любая модель должна удовлетворять, по меньшей мере, двум основным требованиям:

1) модель должна быть внутренне согласованной, свободной хотя бы от явных противоречий;
2) все лежащие в основе модели предположения без исключения должны допускать непосредственное сравнение с эмпирическими данными.

Как следует из изложенного выше, требованию внутренней согласованности и непротиворечивости модели Раша и Бирнбаума не удовлетворяют. Следует признать, что ни о каком сравнении этих моделей с эмпирическими данными также не может быть и речи. Приняв какую-либо из этих моделей на веру, можно, используя стандартные методы математической статистики, получить точечные (или интервальные) оценки значений параметров выбранной модели, но нет ни одной закономерности, предсказываемой этими моделями, которая допускала бы прямое сопоставление с эмпирическими данными.

Можно было бы видоизменить модели Раша и Бирнбаума, чтобы избавить их от явных противоречий. Но честнее будет отказаться вообще от использования параметрических моделей в теории тестирования и начать поиски принципиально новых подходов к решению задачи объективной оценки трудности тестовых заданий и уровня подготовленности испытуемых.

2. Новая математическая модель тестирования

Введение

Главной причиной недостатков существующих моделей тестирования [2; 7–10] является то, что ни одна из них не рассматривает тестирование как протекающий во времени процесс. При бланочной форме тестирования это принципиально невозможно, и в результатах тестирования отражается только правильность выполнения тестовых заданий. Компьютерное тестирование предоставляет более широкие возможности, но они до сих пор в полной мере не использовались, и при реализации компьютерного тестирования под кальку копировались методы бланочного тестирования.

2.1. Процесс поиска решения как пуассоновский процесс

В работе [4] впервые была описана разработанная нами новая модель тестирования. Вопреки сложившимся в теории и практике тестирования традициям, в рамках этой модели тестирование рассматривается именно как протекающий во времени реальный процесс. Поэтому при проведении процедуры тестирования фиксируется не только правильность выполнения, но и время выполнения каждого задания, поскольку время выполнения зависит как от трудности заданий, так и от уровня подготовленности испытуемых. В основе модели лежит довольно естественное предположение о том, что поиск решения задания является однородным во времени (или пуассоновским) случайным процессом.

Однородные во времени случайные процессы относятся к классу безгранично делимых распределений. Хорошо известным в математической статистике примером безгранично делимого распределения [1; 6] является гамма-распределение с плотностью вероятности:

$$f(\sigma,\lambda,t)=\frac{(\lambda t)^{\sigma-1}}{\Gamma(\sigma)}\cdot e^{-\lambda t}\,\lambda. \qquad (2.1)$$

Интерпретация параметров гамма-распределения зависит от специфики конкретной прикладной задачи. В нашем случае безразмерный параметр σ

можно отождествить с трудностью задания, а имеющий размерность обратного времени параметр λ – с уровнем подготовленности испытуемого.

Гамма-распределение удовлетворяет интегральному уравнению типа свертки, в которое намеренно введена дельта-функция, чтобы придать ему наиболее симметричный вид:

$$f(\sigma'+\sigma'',\lambda,t)=\int_0^t\int_0^t f(\sigma',\lambda,t')\times \times f(\sigma'',\lambda,t'')\delta(t-t'-t'')dt'dt''. \quad (2.2)$$

Смысл уравнения состоит в том, что если t' – время, необходимое для правильного решения задания трудности σ', а t'' – время, необходимое для правильного решения задания трудности σ'', то сумма этих величин $t = t' + t''$ совпадает со временем, которое требуется для решения задания суммарной трудности $\sigma = \sigma' + \sigma''$. Именно это служит обоснованием предлагаемой нами интерпретации параметров гамма-распределения.

2.2. Обработка результатов

Для оценки трудности заданий и уровня подготовленности студентов, участвующих в тестовом испытании, мы используем принцип максимального правдоподобия Фишера. В данном случае функция правдоподобия равна:

$$F(\sigma,\lambda)=\prod_{i=1}^{n}\prod_{j=1}^{N}(\chi_{i,j}\cdot f(\sigma_i,\lambda_j,t_{i,j}) +1-\chi_{i,j}), \quad (2.3)$$

и содержит неизвестные параметры, значения которых определяются из условия максимума функции правдоподобия.

На практике, однако, удобнее оценивать значения параметров из условия максимума логарифмической функции правдоподобия:

$$\ln(F(\sigma,\lambda)) = \sum_{i=1}^{n}\sum_{j=1}^{N}\chi_{i,j}\ln(f(\sigma_i,\lambda_j,t_{i,j})). \quad (2.4)$$

Условия максимума функции (2.4) приводят к уравнениям:

$$\psi(\sigma_i)=\frac{\sum_{j=1}^{N}\chi_{i,j}\ln(\lambda_j t_{i,j})}{N_i}, \quad i=1,2,\ldots,n, \quad (2.5)$$

$$\lambda_j=\frac{\sum_{i=1}^{n}\chi_{i,j}\sigma_i}{\sum_{i=1}^{n}\chi_{i,j}t_{i,j}}, \quad j=1,2,\ldots,N, \quad (2.6)$$

служащим основой определения значений трудности тестовых заданий и уровня подготовленности тестируемых.

В уравнение (2.5) входит пси-функция Эйлера:

$$\psi(\sigma)=\frac{\Gamma'(\sigma)}{\Gamma(\sigma)}. \quad (2.7)$$

Помимо этого, в уравнении (2.5) введено обозначение

$$N_i=\sum_{j=1}^{N}\chi_{i,j} \quad (2.8)$$

общего числа студентов, правильно выполнивших i-е задание.

Содержание уравнения (2.6) предельно просто: уровень подготовленности испытуемого оценивается как отношение суммарной трудности правильно выполненных заданий к времени, затраченному на их решение.

Назовем безразмерную величину $\tau_{i,j} = \lambda_j \cdot t_{i,j}$ приведенным временем, затраченным j-м студентом на выполнение i-го задания. Тогда уравнение (2.5) означает, что пси-функция трудности задания совпадает со средним значением логарифма приведенного времени, затраченного испытуемыми на поиск его правильного решения.

Уравнения (2.5–2.6) допускают решение методом итераций, причем точность решения порядка 0,1% достигается при 10–25 итерациях. Укажем,

как следует выбирать стартовые значения параметров. Используя формулы для моментов гамма-распределения [6]:

(2.9)

нетрудно получить явные выражения для математического ожидания времени решения тестовых заданий и его дисперсии:

$$\langle t\rangle = \frac{\sigma}{\lambda},\ D = \left\langle t^2 \right\rangle - \langle t\rangle^2 = \frac{\sigma}{\lambda^2}, \quad (2.10)$$

откуда следует простая оценка трудности тестового задания:

(2.11)

По известным формулам математической статистики [там же] найдем выборочные оценки среднего значения и дисперсии времени правильного решения i-го задания:

$$H(\sigma,\lambda,\nu) = \int_0^\infty f(\sigma,\lambda,t)t^\nu dt = \frac{\Gamma(\sigma+\nu)}{\Gamma(\sigma)\lambda^\nu},$$

$$\langle t\rangle_i \approx \frac{\sum_{j=1}^{N} \chi_{i,j} t_{i,j}}{N_i},$$
$$D_i \approx \frac{\sum_{k=1}^{N_i} \chi_{i,j}\left(t_{i,k} - \langle t\rangle_i\right)^2}{N_i - 1}. \quad (2.12)$$

Набор стартовых значений параметров получается в результате подстановки выборочных оценок (2.12) в формулу (2.11), с последующей подстановкой полученных значений трудности тестовых заданий в формулу (2.6).

Испытуемым выставляются оценки за суммарную трудность правильно решенных заданий:

$$\alpha_j \approx \sum_{i=1}^{n} \chi_{i,j}\sigma_i, \quad (2.13)$$

а затем приводятся к 100-балльной шкале нормировкой на наилучший результат в группе, что удобно при переводе оценок в привычную 5-балльную шкалу.

Описанная процедура в ходе тщательной и всесторонней проверки показала свою работоспособность и надежность. В Педагогическом институте ЮФУ разработан комплекс программ, предназначенных для сетевого компьютерного тестирования, содержащий клиентскую программу, редактор баз тестовых заданий, а в последней версии также модуль, реализующий упрощенную схему обработки результатов тестирования. Индивидуальные тесты формируются путем случайной выборки из базы тестовых заданий, при этом обеспечивается равномерность выборки тестовых заданий из каждого блока. В процессе тестирования в базе результатов фиксируется не только правильность выполнения всех заданий, но и время работы над каждым тестовым заданием.

2.3. Сравнение модели с эмпирическими данными

Здесь приведены результаты тестирования по элементарной математике студентов первых четырех курсов отделений физики, математики и информатики ПедИ ЮФУ, проведенного в осеннем семестре 2006 г. При тестировании была использована база тестовых заданий, созданная автором статьи.

В отличие от параметрических моделей, предлагаемая нами модель тестирования допускает прямую опытную проверку. В частности, возможно сравнение теоретического распределения времени решения тестовых заданий с эмпирическими данными.

На рис. 1–2 приведены результаты такого сравнения для двух наугад выбранных заданий (время указано в секундах). В тестировании участвовали 176 студентов, но из-за различных нарушений режима тестирования к обработке были приняты 145 результатов.

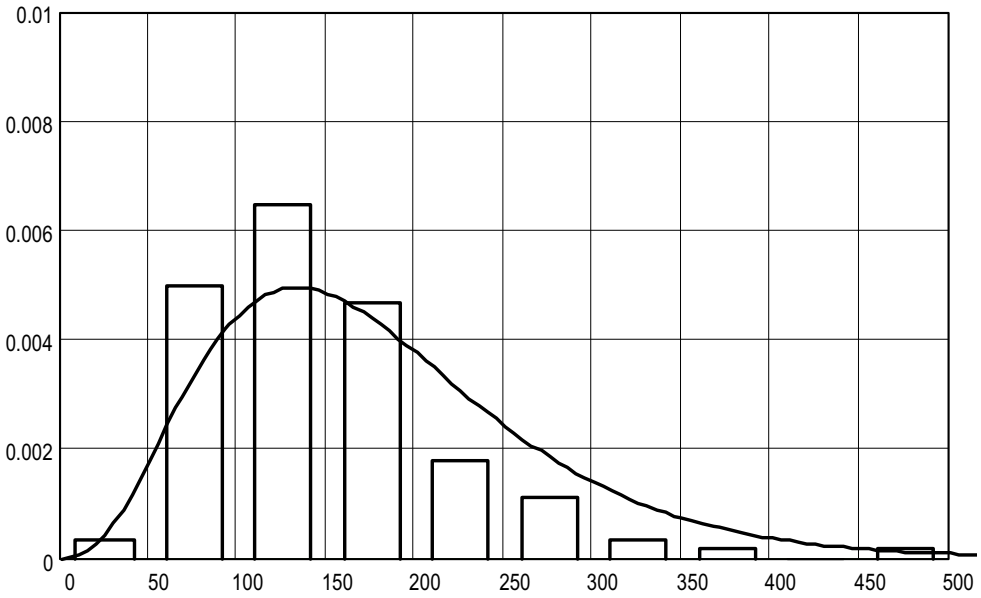

Рис. 1. Распределение времени правильного решения 1-го задания

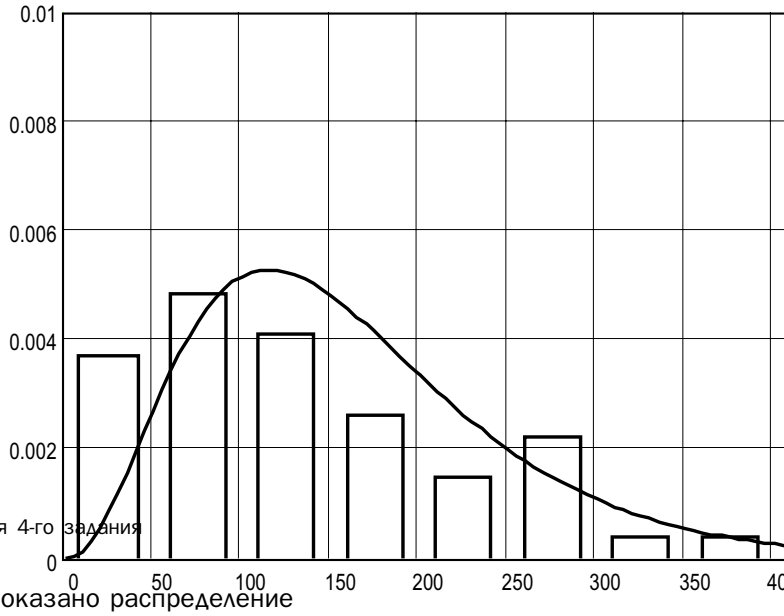

Рис. 2. Распределение времени правильного решения 4-го задания

Объем статистической выборки хотя и не велик, но достаточен, чтобы согласие теоретического распределения (2.1) с эмпирическим распределением времени выполнения каждого тестового задания было четко выраженным (расхождение достигает около 7–8%, что видно и на рис. 1–2).

На рис. 3 показано распределение оценок за общую трудность правильно выполненных заданий. Для сравнения показан график плотности нормального распределения с теми же параметрами.

Статистический анализ показывает, что с надежностью 0,95 гипотезу о

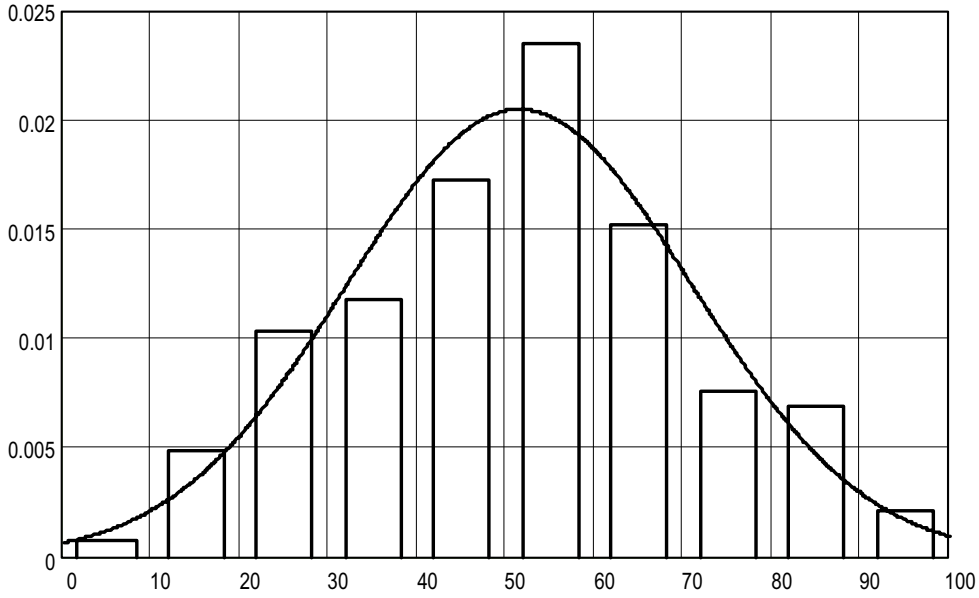

Рис. 3. Распределение оценок за суммарную трудность правильно выполненных заданий

нормальном распределении оценок можно принять.

Приобретенный опыт убеждает в том, что оценка за суммарную сложность правильно решенных заданий объективно оценивает уровень подготовленности и может использоваться при определении рейтинга испытуемых. Так, в рейтинг-листах, составленных по результатам тестирования по «Элементарной математике», деканы факультетов физики и математики и информатики сочли сомнительными не более 4 оценок из общего числа 145 (менее 3%). Следует также отметить, что во всех академических группах (по мнению преподавателей, ведущих занятия в группах) рейтинг-листы возглавили объективно лучшие студенты.

2.4. Краткие выводы

Совпадение предсказываемых моделью теоретических зависимостей с эмпирическими распределениями позволяет считать, что модель адекватно описывает процесс тестирования и позволяет достаточно объективно оценивать трудность тестовых заданий и уровень подготовленности тестируемых. Работа по обобщению, развитию и совершенствованию модели продолжается [3; 5].

Вместе с тем мы далеки от мысли, что тестирование (в какой бы то ни было форме) может служить универсальным средством оценки знаний, а тем более способностей учащихся.

Нередко приходится слышать, что тестирование способно полностью заменить традиционные формы контроля или что тестирование можно использовать не только как средство контроля подготовки студентов по конкретному разделу той или иной дисциплины, но и как инструмент измерения способностей учащихся или даже как средство обучения или самообучения студентов. Ниже перечислен ряд причин, по которым согласиться с этим мнением нельзя:

1. Содержание многих дисциплин практически не допускает представления в форме тестовых заданий (в качестве примера можно привести такие дисциплины, как живопись, дизайн или начертательная геометрия).

2. Чтобы все испытуемые оказывались в равных условиях, генерируемые программой индивидуальные тесты должны иметь одинаковую трудность. Для этого нужно обеспечить одинаковую трудность заданий в пределах одного блока, что представляется весьма проблематичным для большинства дисциплин гуманитарного цикла (история, литература, экономика, философия, педагогика).
3. В принципе, конечно, можно представить себе обучающую программу, в которую встроены средства тестового контроля, несложно даже написать какую-либо демонстрационную версию такой программы, но создать полноценную обучающую программу по всем разделам даже одной дисциплины представляется непосильной задачей.
4. Наконец, нужно ясно понимать, что любая, даже самая совершенная система автоматического оценивания уровня подготовленности студентов по результатам тестирования оперирует с формальными, чисто внешними признаками (правильность ответов, время решения и т.д.). Поэтому здесь уместнее говорить не об объективности полученных оценок, а лишь об объективизации самого процесса оценивания.

Таким образом, на долю тестирования (в какой бы форме оно ни проводилось) остается входной, текущий и выходной контроль знаний, причем для круга дисциплин, допускающих достаточно глубокую формализацию излагаемого материала. Это прежде всего математика, физика, информатика, дисциплины естественно-научного цикла, некоторые разделы лингвистики и психологии. Но и здесь тестирование следует использовать лишь как некое сито, служащее для предварительного отсева студентов, явно не усвоивших некоего минимума знаний по данному предмету. Впрочем, нужно заметить, что, как показывает опыт, этот отсев иногда может оказаться гораздо более жестким, чем ожидал сам преподаватель, писавший тестовые задания и составлявший базу тестовых заданий.

Литература

1. *Гнеденко, Б.В.* Курс теории вероятностей / Б.В. Гнеденко. М.: Наука, 1988.
2. *Нейман, Ю.М.* Как оценивается уровень подготовленности учащихся по результатам ЕГЭ / Ю.М. Нейман, В.А. Хлебников. М.: Изд-во ЦТ МО РФ, 2003.
3. *Попов, А.П.* Критический анализ параметрических моделей Раша и Бирнбаума / А.П. Попов // Материалы 4-й НМК «Инновационные методы и средства оценки качества образования». М.: Изд-во МГУП, 2006. С. 231–235.
4. *Попов, А.П.* Новая математическая модель тестирования / А.П. Попов, А.А. Богомолов, Л.А. Попова // Наука и образование. 2005. № 3. С. 221–228.
5. *Попов, А.П.* Поиск решения как однородный во времени случайный процесс. Новая математическая модель тестирования / А.П. Попов, А.А. Богомолов, Л.А. Попова // Материалы 4-й НМК «Инновационные методы и средства оценки качества образования». М.: Изд-во МГУП, 2006. С. 236–247.
6. Справочник по теории вероятностей и математической статистике / В.С. Королюк [и др.]. М.: Наука, 1985.
7. *Lord, F.M.* Statistical Theories of Mental Test Scores / F.M. Lord, M. Novic. Mass.: Addison-Wesley Publisher Co. Reading, 1968.
8. *Rasch, G.* Probabilistic Model for Some Intelligence and Attainment Tests / G. Rasch. Chicago: Univ. of Chicago Press, 1980.
9. *Wright, B.D.* Best Test Design / B.D. Wright, M.N. Stone. Chicago: MESA Press, 1979.
10. *Wright, B.D.* Rating scale analysis. Rasch measurements / B.D. Wright, G.N. Masters. Chicago: MESA Press, 1982.