Методология

МЕТОДОЛОГИЧЕСКИЙ АНАЛИЗ ТЕОРИЙ ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ[1]

Хои Суен,
Пенсильванский университет США[2]
HoiSuen@psu.edu
Пуи ВА Лей[3],
Пенсильванский университет США
PuiWa@psu.edu

В статье рассматриваются вопросы истории возникновения основных теорий педагогических измерений, проводится сравнительный анализ их достоинств и недостатков. Главное внимание уделено Generalizability Theory[4].

Ключевые слова: *история и теория измерений, классическая статистическая теория, Generalizability Theory, Item Response Theory.*

Введение

Статистические теории педагогических и психологических измерений имеют более чем столетнюю историю развития. В 1904 году Чарльз Спирман[5] разработал основы первой статистической модели оценки истинного балла испытуемых. В последующие десятиле-

1 Перевод с английского кандидата физ-мат. наук Светланы Янченко.

2 *Hoi K. Suen.* Pennsylvania State University, U.S.A.

3 *Pui-Wa Lei.* Pennsylvania State University, U.S.A

4 На русском языке эту теорию можно назвать расширенной статистической теорией измерений, в смысле сравнения с классической статистической теорией педагогических и психологических измерений (*Прим. ред.*).

5 *Spearman C.E.* «General intelligence» objectively determined and measured. American Journal of Psychology, 5, 201–293, 1904.

тия появились новые исследования. Харольду Галиксену[6] удалось обобщить результаты всех этих работ в целостную теоретическую систему взаимосвязанных теорем и уравнений, которая стала известна как классическая статистическая теория измерений.

С начала 50-х годов прошлого века Фредерик Лорд[7] и датский математик Георг Раш[8] (опубликовано на англ языке позже, в 1960 г.) независимо один от другого работали над созданием другой теории измерений, известной сегодня под названием Item Response Theory (IRT)[9]. В последние три десятилетия интенсивно развивались различные варианты IRT. Наиболее заметными из них являются Rasch model for rating scales[10], что можно перевести как модель Раша для градуированных шкал оценивания, the Faceted Rasch model[11], а также многомерные модели этой теории измерений — Multidimensional Item Response Theory[12]. В 1963 году Ли Кронбах и его коллеги[13] создали новую теорию и назвали её Generalizability Theory (GT)[14].

Item Response Theory (IRT)

Вопреки тому факту, что Расширенная Статистическая Теория (РСТ) педагогических измерений появилась позже IRT, «современной» теорией тестов тем не менее называют IRT. Вероятно, это происходит из-за появления новых вариантов IRT, возможности применения этой теории для организации адаптивного компьютерного тестирования, а также по многим другим причинам[15].

В последнее время IRT привлекает наибольшее внимание специалистов по педагогическим и психологическим измерениям. Одна из возможных причин популярности этой теории — это её применение в широко известных тестовых службах. Среди них — Национальный совет развития образования (National Assessment of Educational Progress, NAEP), служба по разработке теста для оценки способности к овладению образовательными программами различного уровня сложности (Scholastic Aptitude Tests, SAT), а также американская аттестационная служба the Graduate Record Examination (GRE).

Кроме этих служб IRT используется в таких масштабных международных сравнительно-оценочных исследованиях, как третье международное исследование уровня подготовленности по математике и естественным наукам (Third International Math and Science Survey, TIMSS), а также в международной программе оценки качества

6 *Gulliksen H.* Theory of mental tests. New York: Wiley. 1950.

7 *Lord F.M.* A theory of test scores. Psychometric Monographs, No. 7, 1952.

8 *Rasch G.* Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research. 1960.

9 Подходящего названия Item Response Theory в русском языке пока нет. По существу это математико-статистическая теория измерения уровня подготовленности испытуемых и параметров тестовых заданий. Кратко её можно назвать математико-статистической теорией измерений.

10 *Wright B.D., Masters, G.N.* Rating scale analysis. Chicago: MESA Press. 1982.

11 *Linacre J.M.* Many-faceted Rasch measurement. Chicago: MESA. 1989.

подготовленности студентов (the Programme of International Student Assessment, PISA). Другие возможные причины повышенного внимания к IRT — это множество новых статистических разработок данной теории, опирающихся на возможности их активного применения в вычислительных и образовательных технологиях.

В фокусе IRT — определение латентной переменной величины, которая влияет на результаты испытуемых в их ответах на задание теста. Определяется такая переменная посредством вероятностной модели измерения, в которой вероятность правильного ответа испытуемого на задание теста рассматривается как функция от уровня подготовленности испытуемого и меры трудности задания. Графически эта вероятность представляется в виде логистической кривой (огивы). При определении латентной переменной применяются и другие методы, включая различные варианты факторного анализа.

Часто пишут о достоинствах IRT, но сравнительно редко — о её недостатках. Для начала, результаты применения IRT очень чувствительны к нарушениям исходных предпосылок этой теории. Кроме того, определение параметров заданий с достаточной точностью требует довольно большой выборки испытуемых, минимум от 200 до 1000 человек, в зависимости от применяемой вероятностной модели. В то же время классическая статистическая и расширенная статистическая теории оказываются более устойчивыми к нарушениям исходных положений. К этому преимуществу двух последних теорий можно добавить ещё одно — нет необходимости в больших выборках для оценки интересующих параметров заданий и теста в целом.

С точки зрения методологии, все три упомянутые теории — классическую, расширенную статистическую и IRT трудно сравнивать, потому что они имеют различные предметы исследования.

В центре внимания IRT — значение латентного параметра испытуемого, явно ненаблюдаемого. Соответственно, главной целью становится оценка этого параметра. В фокусе же классической и расширенной статистической теорий измерения — оценка реально наблюдаемого тестового балла испытуемого. Главная цель последних двух теорий — качество наблюдаемого тестового балла, оцениваемое посредством коэффициентов надёжности и среднеквадратических ошибок измерения. Различие между рассматриваемыми теориями проявляется также в том, что основной единицей для анализа в IRT является тестовое задание, в то время как

12 *McDonald R.P.* Nonlinear factor analysis. Psychometric Monographs, No. 15. 1967.

13 *Cronbach L.J., Rajaratnam N., Gleser G.C.* Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137–163. 1963.

14 На русском языке Generalizability Theory можно назвать Расширенной Статистической Теорией (РСТ) педагогических измерений.

15 По мнению редактора, IRT понятна математикам, она не требует от них знаний классической и других теорий. Что может быть главной причиной увлечения IRT с их стороны.

для классической и расширенной статистических теорий — весь тест в целом.

Классическая статистическая теория измерений

Несмотря на столь повышенное внимание разработчиков тестов и исследователей в сфере образования к IRT, классическая статистическая теория измерений (Classical Test Theory) всё ещё остаётся в качестве основной «рабочей лошадки», используемой в разработке большинства коммерческих и педагогических тестов. Особенно при реализации небольших тестовых исследовательских программ.

В основу классической теории положена идея наличия во всяком измерении так называемого истинного компонента тестового балла испытуемых. Соответственно, теория начинается с предположения о том, что реально наблюдаемый тестовый балл X любого испытуемого состоит из истинного компонента измерения τ и ошибочного компонента (погрешности) измерения ε. Данное предположение записывается как формальное исходное утверждение теории:

$$X = \tau + \varepsilon. \qquad (1)$$

В классической статистической теории ошибка измерения есть мера отклонения наблюдаемого тестового балла X от истинного компонента измерения τ.

Далее делается второе предположение — об отсутствии связи между истинными и ошибочными компонентами измерения на множестве испытуемых, следствием которого является равенство

$$\sigma_X^2 = \sigma_\tau^2 + \sigma_\varepsilon^2. \qquad (2)$$

Результат деления левой и правой частей равенства (2) на дисперсию наблюдаемых тестовых баллов (σ_X^2, левую часть равенства) приводит к идее коэффициента надёжности результатов измерения. Слева получаем единицу, а справа — сумму двух отношений. За значение коэффициента надёжности тестовых результатов ρ_{xx}принимается первое отношение — дисперсии истинных компонентов измерения к общей дисперсии тестовых баллов испытуемых (равенство 3), где ρ_{xx} может принимать значения от нуля до единицы:

$$\rho_{xx} = \frac{\sigma_\tau^2}{\sigma_X^2}. \qquad (3)$$

В равенстве (3) эмпирически наблюдаемо только значение знаменателя дроби — дисперсии наблюдаемых тестовых баллов испытуемых. Дисперсию истинных компонентов измерения приходится определять косвенным образом, на основе тех или иных предположе-

ний о статистическом распределении (вариации) каждого компонента измерений.

Обычным подходом для оценки вариации истинных компонентов измерения считается предположение о параллельности заданий вариантов одного и того же теста. Если статистические характеристики параллельных вариантов теста отвечают определённым, достаточно жёстким, условиям (см. подробнее в книге Crocker & Algina, 1986)[16], то тогда коэффициент корреляции Пирсона исходных тестовых баллов испытуемых по двум вариантам одного и того же теста считается математически эквивалентным коэффициенту надёжности результатов теста. Подобный подход положен в основу методов определения надёжности тестовых результатов как повторное тестирование одних и тех же испытуемых одним и тем же вариантом теста (test-retest method), использование эквивалентных вариантов одного и того же теста (the equivalent forms method) и т.н. методы определения внутренней состоятельности теста (the internal consistency methods).

Другой подход к определению коэффициента надёжности тестовых результатов основан на предположениях менее строгих, чем, например, предположение о параллельности заданий двух вариантов теста[17]. Такого рода ослабленные предположения в западной литературе получили названия существенно τ-эквивалентные допущения (essentially τ-equivalent assumptions). Один из наиболее распространённых методов оценки надёжности тестовых результатов — это расчёт коэффициента альфа Л. Кронбаха, в основу которого положены именно такие предположения. Коэффициент альфа считается по формуле

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_X^2} \right), \qquad (4)$$

где k — число заданий разрабатываемого теста; σ_i^2 — значение дисперсии баллов, полученных множеством испытуемых по i-тому заданию теста; σ_X^2 равен значению дисперсии наблюдаемых тестовых баллов испытуемых, полученных испытуемыми по всем заданиям теста.

Расчёт оценки надёжности тестовых результатов позволяет далее определить значение стандартной погрешности измерения, выражаемой в виде так называемой стандартной ошибки измерения (σ_E)

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{xx'}}. \qquad (5)$$

Стандартная ошибка измерения является усреднённым значением погрешности измерения, если наблюдаемые тестовые баллы испытуемых рассма-

16 *Crocker L., Algina J.* Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston. 1986.

17 Например, можно принять, что истинные компоненты вариантов заданий или теста равны, а ошибочные компоненты измерения могут быть неравными (Прим. переводчика).

тривать как результаты тестирования группы испытуемых. Эту стандартную ошибку можно использовать для определения значений доверительных интервалов, в пределах которых могут находиться истинные значения тестовых баллов испытуемых. Чем больше значение коэффициента надёжности тестовых результатов, тем меньше доверительный интервал, в котором могут варьировать значения истинных компонентов измерений.

Расширенная статистическая теория измерений (Generalizability Theory, GT)

Расширенную статистическую теорию некоторые авторы рассматривают как родственную классической статистической теории. Обе они имеют общий предмет — оценку истинного и ошибочных компонентов измерения с целью определения надёжности тестовых результатов. И действительно, обе теории занимаются качеством получаемых при измерении тестовых баллов, причём делают это без обращения к идее оценки испытуемых или заданий на оси латентных переменных. К тому же эти теории применимы к обработке данных на выборках небольшого объёма. У этих двух теорий есть и другие общие элементы, равно как есть и важные различия.

Концептуально GT можно рассматривать как расширяемую статистическую теорию измерений, которая позволяет вести поиск не одного, а нескольких источников погрешностей измерения[18], систематических[19] и случайных. В отличие от равенства (1), где во внимание принимается значение наблюдаемого тестового балла, полученного по всем заданиям теста (X), в GT принимается другая модель измерения. Здесь любой наблюдаемый балл X любого испытуемого, полученный при ответе на любое отдельное задание теста, представляется такой моделью:

$$X = \mu + \tilde{\mu_p} + \tilde{\mu_i} + \tilde{\mu_{pi}}, \qquad (6)$$

где: X — наблюдаемый балл испытуемого, полученный по заданию i; μ — средняя арифметическая всех τ-значений испытуемых в генеральной совокупности. На английском языке значение μ имеет специфическое название the universe score; $\tilde{\mu_p}$ — отклонения значения испытуемого p от значения μ; $\tilde{\mu_i}$ — отклонение значения балла испытуемого по заданию i от балла в генеральной совокупности; $\tilde{\mu_{pi}}$ — остаточная случайная ошибка.

Эта модель может быть расширена в зависимости от наличия других возможных ис-

18 В авторском тексте стоит трудно переводимое предложение: Conceptually, the Generalizability Theory may be considered by-and-large a multi-faceted extension of the Classical Theory — аналог многофакторного дисперсионного анализа.

19 Что очень важно, потому что все известные теории измерений в педагогике и психологии принципиально имеют дело только со случайными ошибками измерения.

точников погрешностей измерения, при наличии оснований для предположения о таковых источниках. Такого рода возможные варианты источников погрешностей в рамках GT называются фасетами[20].

В G-теории исследователи подходят к измерениям и оценкам, как к многогранному явлению. Например, при оценке сочинения получаемая испытуемым оценка содержит в себе личностный компонент испытуемого, специфический компонент задания (насколько тема соответствует уровню подготовленности испытуемого), и компонент, зависимый от оценивающего эксперта. Каждый из упомянутых компонентов относятся к тем аспектам измерения, которые не входят, как и всякие ошибки, в цель измерения. В приведённом примере компоненты задания и эксперта называются фасетами оценивания, потому что они не представляют собой цель измерения.

В предположении, что значения всех членов правой части равенства (6) на множестве испытуемых взаимно не коррелируют, можно записать

$$\sigma_X^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi}^2. \quad (7)$$

В GT каждый компонент вариации в формуле (7) рассчитывается непосредственно по эмпирическим данным, чего нельзя сказать о классической теории, где значение вариации истинной части тестовых баллов приходится оценивать опосредованно, на основе идеи параллельных вариантов теста или идеи о τ-эквивалентных заданиях теста. Компоненты формулы (7) рассчитываются посредством дисперсионного анализа, основанного на предположении, что данные удовлетворяют требованиям выборки из генеральной совокупности или предположению о возможности извлечения параллельных выборок из одной и той же генеральной совокупности. Соответствующие компоненты дисперсии, оцениваемые посредством дисперсионного анализа, становятся основой для вычисления коэффициентов надёжности измерений.

Исходные тестовые баллы испытуемых могут быть интерпретированы как нормативно-ориентированные, когда баллы каждого испытуемого сравниваются с баллами других испытуемых в интересующей группе. Такой подход может быть назван сравнительной моделью измерения. Или эти же баллы можно интерпретировать как критериально-ориентированные, где каждый результат соотносится с каким-либо заранее определённым критерием или согласованным стандартом учебных достижений. Критериально-ориентированная интерпретация результатов тестиро-

20

Примечание редактора: когда в 70-х годах прошлого века вводилось понятие «фасет», в текстах на русском языке для обозначения вариантов одного и того же задания, я опирался на идею Л.Л. Гуттмана о фасете как форме записи множества параллельных вариантов одной и той же теории. Теперь мы видим расширение этого понятия и на случай дополнительных систематических и случайных источников погрешностей измерения, как в случае проведения многомерного дисперсионного анализа. (*В.А.*).

вания соответствует абсолютной модели измерения.

Для случая, когда используется N_i число заданий, а подходящей интерпретацией считается нормативно-ориентированная, коэффициент надёжности измерений, вычисляемый по модели (6) и (7) и называемый G-коэффициентом, вычисляется по формуле (8)

$$\hat{\rho}^2_{xt} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pi}}{N_i}}. \qquad (8)$$

Стандартная ошибка коэффициента надёжности тестовых результатов вычисляется по формуле $\sqrt{\frac{\sigma^2_{pi}}{N_i}}$. Для измерения с критериально ориентированной интерпретацией результатов G-коэффициент вычисляется по формуле

$$\hat{\rho}^2_{xt} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_i}{N_i} + \frac{\sigma^2_{pi}}{N_i}} \qquad (9)$$

со стандартной ошибкой такого коэффициента, определяемого из выражения $\sqrt{\frac{\sigma^2_i}{N_i} + \frac{\sigma^2_{pi}}{N_i}}$.

Формулы 6–9 могут быть изменены в зависимости от избираемого подхода к проведению измерений, определяемых источников вариации и от нужной интерпретации результатов. Например, в ситуации, когда в качестве единственного источника вариации систематической ошибки измерения являются различия в оценках экспертов, а не заданий, формулы 6–9 вполне применимы. Тогда подстрочный индекс *i*, относящийся к заданиям, во всех формулах будет заменен на индекс, относящийся к экспертам.

Для чуть более сложного случая, когда потенциальные источники систематических ошибок включают одновременно экспертов и задания, наблюдаемый исходный балл оценки испытуемого, даваемой определённым экспертом, может быть представлен структурно, как:

$$X = \mu + \tilde{\mu_p} + \tilde{\mu_r} + \tilde{\mu_i} + \tilde{\mu_{pr}} + \\ + \tilde{\mu_{pi}} + \tilde{\mu_{ir}} + \tilde{\mu_{pir}}, \qquad (10)$$

где, как и ранее, X означает наблюдаемый балл испытуемого, полученный им за выполнение задания (или за решение задачи); μ — средний балл (математическое ожидание) для генеральной совокупности всех мыслимых подобных заданий или задач, равно как и для всех мыслимых испытуемых; $\tilde{\mu_p}$ — отклонение средней арифметической оценок экспертов для данного испытуемого p от средней арифметической в $\tilde{\mu_r}$ – отклонение среднего арифметического значения баллов, полу-

ченных множеством всех испытуемых по всем заданиям, у одного эксперта, от средней арифметической генеральной совокупности; $\tilde{\mu}_i$ — отклонение от генерального параметра среднего арифметического балла испытуемых, данного им в каждом задании всеми экспертами; $\tilde{\mu}_{pi}$ — представляет результат взаимовлияния испытуемых и заданий; $\tilde{\mu}_{pr}$ — представляет результат взаимовлияния испытуемых и экспертов; $\tilde{\mu}_{ir}$ — результат взаимовлияния между заданиями и экспертами; $\tilde{\mu}_{pir}$ — результат взаимовлияния между испытуемыми, заданиями и экспертами.

Дисперсия наблюдаемых экспертных оценок испытуемых по множеству экспертов равна

$$\sigma_X^2 = \sigma_p^2 + \sigma_r^2 + \sigma_i^2 + \sigma_{pr}^2 + \\ + \sigma_{pi}^2 + \sigma_{ir}^2 + \sigma_{pir}^2. \quad (11)$$

На основе равенства (11), в зависимости от цели анализа, можно определить несколько G-коэффициентов. Факторы, подлежащие при этом рассмотрению: число экспертов, число заданий или задач, рассматриваются ли то и другое как источник случайных или систематических погрешностей, а также — какая модель измерения имеется в виду — сравнительная или абсолютная. Наиболее частая ситуация измерения — когда эксперты и задания рассматриваются как случайные переменные величины, а модель измерения — сравнительная. В ситуации применения формул (10) и (11), при N_r числе экспертов и N_i числе заданий G-коэффициент равен

$$\hat{\rho}_{xt}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pr}^2}{N_r} + \frac{\sigma_{pi}^2}{N_i} + \frac{\sigma_{pir}^2}{N_i N_r}}. \quad (12)$$

При применении абсолютной модели измерения G-coefficient вычисляется по формуле

$$\hat{\rho}_{xt}^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_r^2}{N_r} + \frac{\sigma_i^2}{N_i} + \frac{\sigma_{pr}^2}{N_r} + \\ + \frac{\sigma_{pi}^2}{N_i} + \frac{\sigma_{ir}^2}{N_i N_r} + \frac{\sigma_{pir}^2}{N_i N_r}}. \quad (13)$$

По мере роста числа потенциальных источников ошибок измерения модель становится громоздкой, число возможных подходов к оценке данных становится большим, а формул расчёта коэффициентов надёжности результатов и стандартных ошибок измерении становится больше.

Проблема множества источников погрешностей измерений

Несмотря на различия в моделях измерения, формулах расчёта коэффициентов надёжнос-

ти результатов и стандартных ошибок измерения в классической и РСТ теориях, некоторые авторы продолжают считать, что обе эти теории соответствуют одной и той же модели оценивания истинного компонента измерения (true score model). Другие авторы, такие как Рональд Коэн и Марк Свердлик (1999, с. 170), полагают, что расширенной теории не удалось стать эффективной заменой классической статистической теории[21]. Они утверждают, что максимум, чего добилась РСТ-теория — это лучшее понимание методов измерения. Другие авторы, такие как Мередит Галл, Вальтер Борг и Джойс Галл (p. 203), утверждают[22], что РСТ-теория является такой же, как и классическая статистическая теория, но только включает одну дополнительную возможность, позволяющую в ходе процесса измерения оценивать различные источники погрешности измерения.

В ситуации использования обычных стандартизованных бланковых тестов с нормативно-ориентированной интерпретацией результатов, с использованием формы заданий с выбором одного правильного ответа из нескольких предложенных, с частью таких утверждений, при некоторых условиях, можно согласиться. На эмпирическом уровне G-коэффициент может оказаться эквивалентным известному коэффициенту альфа Ли Кронбаха[23], т.н. коэффициенту внутриклассовой корреляции Сирила Хойта[24], и в случае использования дихотомических оценок — единица или ноль — коэффициенту Фредерика Кадера и Марион Ричардсон (KR-20)[25]. Однако в случаях, когда модель измерения усложняется, РСТ-теория остаётся единственной теорией, позволяющей провести качественную проверку точности измерения наблюдаемых тестовых баллов испытуемых. Из чего можно сделать вывод, что это разные теории.

Пример

В сложной измерительной ситуации, такой как оценка результатов абитуриентов и аттестация выпускников, случайные и систематические оценки могут возникать из различных источников. Наблюдаемый балл представляет собой сумму истинного компонента и результата взаимодействия различных источников ошибок измерения. В подобных случаях в классической статистической теории измерений используются различные методы определения надёжности тестовых результатов. Это метод повторного тестирования одним и тем же тестом одной и той же группы испытуемых, метод коррелирования оценок различных экспертов, и

21 *Cohen R.J., Swerdlik M.E.* Psychological testing and assessment: An introduction to tests and measurement. Mountain View, CA: Mayfield. 1999.

22 *Gall M.D., Borg W.R., Gall J.P.* Educational research: An introduction (8th ed.). New York: Longman. 2007.

23 *Cronbach L.J.* Coefficient alpha and the internal structure of tests. Psychometrika, 16 (3), 1951, pp. 297–334.

24 *Hoyt C.J.* Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, pp. 153–160.

25 *Kuder G.F., Richardson M.W.* The Theory of Estimation of Test Reliability // Psychometrika, 2: 1937, pp. 151–160,. В большинстве, если не во всех, изданиях на русском языке фамилия второго автора написана как мужская, т.е. что это формула Ричардсона. Между тем этот автор — женщина.

расчёт т.н. коэффициентов внутренней состоятельности тестовых результатов (internal consistency)[26]. Все эти коэффициенты позволяют посчитать коэффициент надёжности тестовых результатов в зависимости от различных источников погрешности измерений.

Но этот подход проблематичен как с теоретической точки зрения, так и практической. Во-первых, возникает необходимость обращаться к разного рода дополнительным идеям, такое, например, какой может быть надёжность оценок различных экспертов, дающих оценки одним и тем же объектам. Или идея стабильности результатов испытуемых, а также идея внутренней состоятельности (когерентности) результатов испытуемых. Всё это вопросы определения надёжности измерений, не связанные непосредственно с классической статистической теорией измерений. Эта теория в состоянии лишь обобщённо представить все различные ошибки измерения в одну общую ошибку.

Во-вторых, из сказанного выше вытекает, что существуют различные типы коэффициентов надёжности результатов, что не соответствует положениям самой классической теории измерений. Такие характеристики, как стабильность результатов испытуемых, схожесть оценок экспертов и внутренней состоятельности нужны как предварительные условия оценки надёжности результатов при ведущей идее возможной параллельности двух и более вариантов одного и того же теста[27]. И наконец, если при опоре на классическую статистическую теорию получаются различные значения коэффициентов надёжности измерений, то, соответственно, получатся и различные значения стандартных ошибок измерения. В таких случаях возникает вопрос — какую из возможных значений стандартных ошибок измерения надо взять для построения доверительного интервала для значения исходных тестовых баллов испытуемых[28]?

Поскольку эмпирически наблюдаемый тестовый балл содержит в себе погрешности измерения из различных источников, а также результатов взаимодействия различных источников погрешностей, то желательно, чтобы при построении доверительных интервалов для тестовых баллов учитывались бы стандартные ошибки всех источников погрешностей измерения. Как замечательно показал в своей работе M.D. Reckase (1995), надёжность оценок экспертов не равна надёжности исходных результатов измерения[29]. Кроме того, классическая теория не даёт ни понятийного аппарата, ни статистического механизма для опреде-

26

Это коэффициенты типа KR-20, и коэффициента альфа Ли Кронбаха.

27

На английском и русском языках часто говорят о параллельных формах тестов, но понимать такие слова лучше как параллельные варианты одного и того же теста.

28

В отличие от тестовых баллов, определяемых в IRT, где ошибка измерения заметно зависит от меняющихся значений информационной функции, в классической статистической теории измерений значение стандартной ошибки измерения принимается одинаковым для всех испытуемых. Эта ошибка вычисляется по формуле (5).

29

Reckase M.D.
The reliability of ratings versus the reliability of scores. Educational Measurement: Issues and Practice, pp. 14, 31. 1995.

ления величины надёжности из комбинированных источников ошибок измерения. Нельзя сказать, что классическая теория запрещает существование различных источников ошибок измерения; но они не согласуются с ней ни понятийно, ни статистически. (I.I. Bejar, 1983; Hoi K. Suen. 1991).

Для того, чтобы показать возможности G-теории по сравнению с классической теорией, не способной учесть множество источников погрешностей измерения, можно взять небольшой пример моделированной измерительной ситуации, в которой 2 эксперта дважды оценивают выполнение пяти заданий (или задач) у 100 испытуемых. Выполнение каждой задачи оценивалось по пятибалльной шкале, от 0 до четырёх. Исходные баллы испытуемых генерировались в соответствии с моделью градуированного оценивания Джефри Мастерса (Master's partial credit)[30], с границами между оценками, взятыми в соответствии с моделью нормального распределения.

Исходные баллы за выполнение задач у первого эксперта, в первом опыте, были получены на основе вероятностей получения заданного балла всеми испытуемыми. Уровень подготовленности испытуемых был принят в соответствии с нормальным распределением. Исходные баллы для второго эксперта, во втором опыте, генерировались с учётом предположения, что в оценках первого эксперта содержались ошибочные компоненты. В табл. 1 представлены следующие коэффициенты надёжности результатов:

Как видно из табл. 1, для каждого эксперта и каждой ситуации имеется коэффициент альфа. Оценки варьируют от 0,394 до 0,828 при среднем значении корреляции 0,576 у двух экспертов по двум ситуациям. Для каждого эксперта имеется также два коэффициента надёжности оценок, коэффициенты стабильности или повторного оценивания, с последующей корреляцией. Среднее значение коэффициента повторного оценивания (test-retest) равняется 0,660, что несколько выше, чем среднее значение коэффициента альфа, равное 0,576. Посчитано два коэффициента надёжности оценок экспертов, по одному на каждую ситуацию. Средний коффициент надёжности оценок экспертов 0,719 является наибольшим среди других средних оценок. Другими словами, имеется по меньшей мере один коэффициент, учитывающий каждый источник ошибок, в соответствии с классической статистической теорией измерения. К сожалению, размах оценок надёжности довольно заметный, в зависимости от источника погрешности.

30

Masters G.N. A Rasch model for partial credit scoring. Psychometrika, 47, 1982. pp. 149–174.

Таблица 1. **Оценки коэффициентов надёжности и значения стандартных ошибок измерения, посчитанные на основе классической статистической теории измерений**

Источник погрешности	Коэффициент альфа	Стандартная ошибка измерения
Эксперт 1, ситуация 1	0.828	20.423
Эксперт 2, ситуация 1	0.532	20.792
Эксперт 1, ситуация 2	0.549	20.915
Эксперт 2, ситуация 2	0.394	30.182
Среднее значение	0.576	
Источник погрешности	Повторное тестирование с последующим коррелированием (test-retest)	Стандартная ошибка измерения
Эксперт 1	0.814	20.218
Эксперт 2	0.506	30.617
Среднее значение	0.660	
Источник погрешности	Корреляция оценок между экспертами	Стандартная ошибка измерения
Ситуация 1	0.763	20.454
Ситуация 2	0.675	20.404
Среднее	0.719	

Встаёт вопрос — какое из полученных значений надёжности оценок надо использовать для оценки точности исходных баллов испытуемых? И если нужно построить доверительный интервал, то какое значение стандартной ошибки измерения следует брать? В пределах классической теории измерений ответа на этот вопрос нет. Если считать все погрешности измерений случайными, что делается в классической статистической теории, то различающиеся значения ко-

эффициентов надёжности следовало бы рассматривать как различные попытки оценить истинное значение коэффициента надёжности оценок и истинного значения стандартной ошибки измерения.

С другой стороны, если допустить, что полученные различия значений коэффициентов надёжности имеют своей причиной различные источники погрешностей измерения, то тогда истинные значения коэффициента надёжности результатов и стандартной ошибки измерения должно быть получены с учётом информации, содержащейся в различных коэффициентах. Но в классической теории нет такого метода, позволяющего учесть влияние различных источников погрешностей измерения значения. В противоположность этому в G-теории есть возможность учёта таких источников. В табл. 2 представлены четыре варианта измерения, каждое из которых учитывает три источника погрешностей, случайных или систематических.

Как можно видеть из табл. 2, если оценка выполнения 5 случайно отобранных заданий, в одной случайно отобранной ситуации выполняется одним случайно отобранным экспертом, то для сравнительной интерпретации получается коэффициент надёжности 0,471. Полученный коэффициент представляет надёжность оценок в случае, когда все три источника погрешностей рассматриваются вместе. Он заметно ниже среднего коэффициента внутренней состоятельности (internal consistency coefficient), равного 0,576, среднего коэффициента стабильности, равного 0,660 и ниже среднего коэффициента корреляции оценок между экспертами 0,719 — все рассчитанные в соответствии с классической статистической теорией измерения, приведённые в таблице 1. Откуда видно, что G-коэффициенты позволяют учесть различные источники погрешностей, в отличие от классической теории. Не случайно, что значение G-коэффициентов ниже. Соответственно выглядят и стандартные ошибки измерения. Значение такой ошибки 3.38 в табл. 2 представляется единственным обоснованным для построения доверительного интервала для исходных оценок.

Как было отмечено выше, игнорирование идеи несовместимости классической статистической теории с возможностью учёта различных источников погрешности измерений приводит к тому, что приходится вычислять различные коэффициенты надёжности, такие как отмеченные ранее коэффициенты надёжности оценок экспертов, внутренней состоятельности или стабильности тестовых результатов.

Таблица 2. **Оценки коэффициентов надёжности и стандартных ошибок измерения, получаемых на основе G-теории**

Измерительные ситуации и сравнительные источники погрешностей измерений	Сравнительные коэффициенты надёжности на основе G-теории (G-coefficients)	Сравнительные стандартные ошибки измерения
1 случайно отбираемый эксперт, в одной ситуации измерения, по пяти случайно отобранным заданиям	0.471	3.380
1 случайно отбираемый эксперт, в одной ситуации измерения, задания не меняются	0.629	2.831
1 случайно отбираемый эксперт, ситуация измерения не меняется, 5 заданий отбираются в случайном порядке.	0.533	3.173
1 случайно отбираемый эксперт, ситуация измерения не меняется, задания не меняются. Здесь проверяется надёжность оценок экспертов (inter-rater reliability)	0.641	2.783
Источники вариации данных	Абсолютные значения G-коэффициентов	Абсолютные значения стандартной ошибки измерения
1 случайно отбираемый эксперт, в одной случайно выбранной ситуации измерения, по пяти случайно отобранным заданиям	0.259	4.000
1 фиксированный эксперт, в случайно выбранной ситуации измерения, 5 фиксированных заданий (stability)	0.348	3.750
1 фиксированный эксперт, 1 фиксированная ситуация, 5 случайных заданий (internal consistency)	0.356	3.727
1 случайно отбираемый эксперт, 1 фиксированная ситуация, 5 фиксированных заданий (inter-rater reliability)	0.411	3.566

Как показано в табл. 1, имеется четыре различных коэффициента внутренней состоятельности оценок, две различающиеся оценки стабильности результатов и два коэффициен-

та надёжности оценок экспертов. В каждом из перечисленных вариантов оценки коэффициентов надёжности заметно отличаются. Какая из них является более правильной? В пределах классической статистической теории ответ на это вопрос неизвестен. Оценки зависят от конкретных выборок задач, экспертов и ситуаций.

В противовес этому G-теория даёт возможность оценивать все доступные выборочные данные для оценки вариации среди испытуемых, экспертов, заданий и ситуаций. Коэффициенты надёжности и стандартные ошибки измерения считаются в соответствии с различными источниками погрешностей измерения, что и позволяет делать G-теория. Как показано в табл. 2, применение возможностей этой теории позволяет получить наименее смещённое значение коэффициента стабильности 0.629. Рассматривая задания как единственный источник вариации, получено наибольшее значение коэффициента внутренней состоятельности, равное 0,533; Рассмотрение же вариации между экспертами как естественного источника погрешности измерения, мы получили значение коэффициента 0,641. Таким образом, G-теория позволяет не только теоретически, но и практически получить несколько коэффициентов надёжности результатов, в зависимости от учёта различных источников погрешностей.

Ещё одно интересное различие между классической и G-теорией заключается в том, что первая удобна для нормативно-ориентированной интерпретации результатов, в то время как G-теория удобна как для нормативно-ориентированной, так и для критериально-ориентированной интерпретации тестовых результатов. Все оценки табл. 1 применимы только для случая нормативно-ориентированной интерпретации результатов испытуемых. То же — и с оценками первой половины табл. 2. Абсолютные значения второй половине табл. 2 подходят для критериально-ориентированной интерпретации результатов. В рамках классической статистической теории получение таких оценок невозможно.

Обсуждение результатов

Те, кто смотрят на классическую и G-теорию с позиции их принадлежности к одной и той же модели измерения, называемой «true score model», правы частично в том, что классическая статистическая теория может рассматриваться как специальный случай G-теории, мало реалистичный в строго организованной измерительной ситуации. Оценки, получаемые на ос-

нове классической статистической теории, имеют тенденцию к завышению истинного коэффициента надёжности результатов. Только G-теория в состоянии удовлетворить требованиям различных измерительных ситуаций.

Так же частично правы и те, кто утверждают, что G-теорию надо рассматривать как возможность для прояснения смысла классической теории измерений с добавлением возможностей учесть дополнительные источники погрешностей измерений. И действительно, при применении G-теории исследователь должен рассматривать не один, а несколько источников погрешностей измерения сразу, а также оценить меру погрешности каждого источника ошибочной вариации результатов. Например, если ошибка измерения при оценивании больше зависит от заданий, чем от оценивающих результаты экспертов, то для достижения интересующего уровня надёжности оценок достаточно будет увеличить число заданий, без увеличения числа экспертов.

Однако есть и существенные различия между этими двумя теориями. Во-первых, есть различия в применении статистического аппарата. В тесте, создаваемом по классической теории, предполагается, что дисперсии истинных и ошибочных компонентов измерения, а также корреляции с внешним критерием у параллельных вариантов теста одинаковы. Несколько ослабленное требование к параллельным вариантам теста предполагает равенство дисперсий только истинных компонентов измерения. Такое требование предполагает изменение лексики. Вместо параллельности вариантов теста говорят об эквивалентности его вариантов (Essentially tau-equivalent test). Такого рода требования часто невыполнимы. В G-теории предполагается что параллельные варианты теста являются случайной выборкой из одной генеральной совокупности заданий. Такого рода предположение является более реалистичным и чаще используемым в статистической обработке тестовых результатов.

Другое теоретическое различие заключается в том, что ошибочный компонент измерений в классической теории рассматривается как случайная погрешность неизвестного происхождения. А потому в классической теории невозможно исследовать надёжность результатов в зависимости от таких специфических источников погрешностей, как нестабильность результатов испытуемых, недостаточная внутренняя состоятельность тестовых результатов или в зависимости от меры согласо-

ванности экспертов. Все перечисленные источники могут быть оценены в G-теории.

Практически получается так, что при опоре на классическую теорию измерений все потенциальные источники погрешностей фиксируются, оставляя варьировать только один, ради оценки которого создавалась выборка испытуемых. Но это не реалистично с точки зрения поиска различных возможных смыслов. А потому классическая теория не подходит для более точной оценки уровня подготовленности испытуемых. В противоположность этому, G-теория подходит для множества различных ситуаций оценивания, при этом ограничивать число факторов могут либо затруднения в сборе множества дополнительных данных, либо неадекватность программно-вычислительного обеспечения.

И наконец, при тестировании испытуемых с критериально-ориентированной интерпретацией результатов необходимо учитывать как случайные, так и систематические погрешности. Важно отметить, что аксиоматика измерений по классической теории не учитывает наличие систематических ошибок, а потому эта теория годится для разработки тестов только с нормативно-ориентированной интерпретацией. В то время как G-теория равно подходит как для создания тестов с такой интерпретацией, так и для разработки тестов с критериально-ориентированной интерпретацией.