**Теория педагогических измерений**

Качество образования не может быть выше качества профессиональной подготовки учителей.

Андреас Шляйхер, математик, координатор PISA

|  |
| --- |
| Теория педагогических  измерений  – это прикладная научная теория, сформировавшаяся на стыке психолого-педагогической науки, теории систем, теории измерений (теории шкалирования), общей статистики, математики (теории вероятностей, регрессионнно-корреляционного и факторного анализа), общей логики, социологии, лингвистики, философии.  Одновременно  теорию педагогических измерений можно рассматривать и как процесс практической деятельности, нацеленный на получение объективированных оценок уровня текущей и итоговой академической подготовленности обучаемых, и как часть общего процесса педагогического оценивания, имеющего на Западе название «evaluation». «Educational Measurement» – семантически близкий эквивалент названия «Педагогические измерения».  Основной предмет рассматриваемой науки – разработка стандартизированных технологий, качественных инструментов (в том числе педагогических тестов) для измерения и оценивания (мeasurement and evaluation) уровня академической подготовленности обучаемых, а также качества обучения в целом.  Именно в этом и цель, и главный смысл педагогических измерений. |

**1. Основные понятия теории педагогических измерений.**

Согласно наи­более распространенному определению «измерение» – это процеду­ра приписывания чисел некоторым характеристикам объектов в соответствии с определенными правилами *(иначе, количественное сопоставление свойств объекта с эталоном, принятым в качестве единицы измерения, В.Т.[[1]](#footnote-1))*.

По мере развития психолого-педагогической науки воз­никла потребность во введении **качественных оценок** для величин, отличающихся по степени проявления того или иного свойства**.** Например, классифицирующие обучающихся понятия в нашем образовании («гений», «умница», «знающий», «светлая голова» «подготовленный», «ботаник», «абсолютный ноль» и др.) являются качественными оценками, дифференцирующими обучающихся по уровню знаний, и определяются субъективно учителем или группой учителей.

Нередко качественные оценки выражают с помощью чисел, которые выбирают на основе экспертных суждений, рекомендаций, соглашений. Приписываемые числа могут трактоваться по-разному. Так, в традиционном для образовательных организаций России педагогическом оценивании в форме «отметок», представляющем **социально - педагогический ранжир** (фактически качественную дискретную шкалу), несмотря на общепринятые нормы, у большинства учителей есть свои представления о том, за что нужно ставить «5», «4», «3» и т.д. При этом, в силу латентности измеряемых переменных оцениванию подвергаются не сами характеристики обученности детей, а их эмпирические референты**.**

**2. Компоненты процесса педагогических измерений.**

|  |
| --- |
| Процесс педа­гогических измерений включает: |
| - выбор предмета измерения (латентных характеристик объек­тов); |
| - выбор эмпирических референтов (наблюдаемых характерис­тик объектов); |
| - разработка измерительных процедур (технологий); |
| - конструирование и использование измерительных инструмен­тов; |
| - выбор шкалы (если измеряемая переменная одна) или шкал (если измеряют более одной переменной при многомерных изме­рениях); |
| - построение отображения результатов измерения на шкалу по определенным про­цедурам и правилам (такова региональная критериальная модель РКМ перевода первичных баллов в тестовые и традиционные); |
| - обработку, анализ и интерпретацию результатов измерения. |

В силу неизбежности ошибок измерения (**первый постулат классической тестологии**) оцениваемые (наблюдаемые) характе­ристики объектов могут принимать значе­ния лишь близкие к истинным. Любые отклонения от стандартизированных условий из­мерения, обработки, анализа и интерпретации полученных резуль­татов еще больше увеличивают ошибки измерения, которые представляют наи­большую опасность в эмпирических науках.

Поэтому для исследователя совершенно обязательно доказательство точно­сти **(надежности**) результатов измерения *(педагогическое тестирование в рамках технологии РКМ в этом смысле существенно отличается от традиционных оценочных процедур и таких оценочных средств, как контрольные, самостоятельные, зачеты, ВПР и т.п., В.Т.)*.

|  |
| --- |
| Надежностью *(reliability) результатов тестирования*называется характеристика точности результатов измерений и их устойчивости к действию случайных факторов (термин «надежность» часто, хотя и не совсем верно, используют по отношению к тесту). |

Есть еще одна характеристика качества результатов тестирования –**валидность.** Валидность результатов педагогических измерений – это характеристика адекватности результатов измерения постав­ленной цели создания теста. Один из видов валидности **– содержательная** (контент-валидность**)**, которая в педагогических измерениях является приоритетной. Определение степени валидности других видов повышают доверие к проводимым измерительным процедурам.

В основе работы экспертов, оценивающих степень валидности, обычно лежит анализ полноты, значимости, правильности пропорций со­держания теста и его соответствия запланированным для провер­ки видам учебной деятельности *(с этой целью для педагогических работников, принимающих участие в РКМ в качестве экспертов по оценке контент – валидности тестов, разработаны специальные рекомендации, В.Т.)***.**

**Надежность и валидность явля­ются взаимосвязанными свойствами, характеризующими различ­ные аспекты качества педагогических измерений**.

Оценка истинных баллов *(true scores)*испытуемых – главная цель всех, кто создает или применяет педагогические тесты. Но, так как **любые результаты** тестирования всегда содержат в себе ошибочные компоненты, то приходится заменять истинные баллы (параметры испытуемых) наиболее достоверными оценками.

Существует **концептуальная формула для коэффициента надежности** (*см. лекции и методические пособия, В.Т.* <https://coko60.ru/biblioteka>). Формулы определения надежности «К-R» (Кьюдера-Ричардсона) – это производные формулы, которые используются на практике для тестов РКМ, включающих только дихотомические задания.

**3. Факторы, влияющие на надежность теста как инструмента измерения**.

* **Чем выше содержательная однородность (гомогенность) теста, тем он надежнее.** Этот вывод представляет особую важность для относительно коротких тестов от **15 до 35** заданий *(гомогенность теста и оптимальность длины остаются дискутируемой проблемой в теории педагогических измерений, В.Т.)*.
* **Надежность измерений повышается с увеличением длины теста***.* Этот формальный вывод не всегда согласуется с реальными возможностями учащихся. По мере роста длины теста повышается утомляемость и снижается мотивация к выполнению заданий, что в совокупности ведет к росту ошибки измерения. Поэтому при выборе оптимальной длины теста разработчики анализируют груп­пу факторов, среди которых высокая дисперсия тестовых баллов, нормальный характер их распределения, форма используемых заданий, возраст учащихся и время выполнения теста, выбранное в соответствии с целями тестирования и физиологическими воз­можностями учащихся.

**3. Правильная композиция заданий теста.**

«…она *<* композиция *>*, как и в создании произведений искусства, зависит не только от оригинальности идеи и сюжета, но и от мастерского владения **формой**. Вот почему хорошо сделанное *(сконструированное, В.Т.)* задание – это всегда лучшее на текущий момент содержание и наилучшая форма; то и другое – результат, появляющийся вследствие осознания важности их взаимосвязи!» – утверждает доктор педагогических наук В.С.Аванесов.Качественные характеристики результатов тестирования во многом зависят от органичности такой взаимосвязи.

**4. Измерительный инструмент**.

Измерительный инструмент вклю­чает два компонента:

Первый компонент – само измеряю­щее устройство, роль которого в педагогических измерениях чаще всего выполняет  **педагогический** **тест** ( в РКМ - это гомогенный стандартизированный педагогический тест с критериально-ориентированной интерпретацией результата).

Второй компонент  – **шкала,** которая служит для фиксации ре­зультатов измерения и на которой откладываются оценки измеряемой переменной.

При измерениях с высокой надежностью и валидностью шкала адекватно отображает оцениваемые характеристики и представляет их без существенных искажений.

**5. Объективность педагогических измерений.**

Может ли быть результат педагогических измерений абсолютно объективным? Появление пер­вых стандартизованных тестов в образовании вызвало массовую позитивную реакцию, поскольку первоначально они рассматривались как средство получения «объективных» оценок подготовки­ обучаемых. По мере развития теории педагогических измерений и накопления опыта применения тестов пришло по­нимание того, что абсолютная объективность – это **недостижи­мая** характеристика любых результатов педагоги­ческих измерений в силу существования ошибочных компонен­тов, неизбежно смещающих оценки. Поэтому при использова­нии педагогических тестов можно говорить лишь о высокой или низкой степени объек­тивности, проявление которой связано с величиной на­дежности тестовых результатов и степенью стандартизации всех этапов тестирования.

Наиболее полно трактовка термина «объективность измерений» представлена в исследованиях **Е. Вебстера (E.Webster),** предложившего **восемь** толкований этого понятия (*три из них не­посредственно относятся к педагогическим измерениям, В.Т.)*:

|  |
| --- |
| **1. Процедурная объективность**. Под процедурной объективностью понимается независимость результатов тестирования от субъективных суж­дений и влияний на процесс тестера, администрирующего тест. Эта независимость обеспе­чивается благодаря равенству условий (стандартов) тестирования (стандарт конструирования, администрирования, обработки и интерпретации результатов). |

|  |
| --- |
| **2. Традиционная объективность**. Второе углубленное понимание объективности измерений рассматривается в классической теории тестов и основывается на понятиях «наблюдаемый балл» и «истинный балл», отличающихся друг от друга (первый постулат классической тестологии**)** на величи­ну ошибки измерения**.** |

В тех случаях, когда ошибка измерения не превышает выбранных пределов точности измерений, говорят о высокой объективности результатов тестирования, а оценки испытуемых принимают (*условно, В.Т.*) за их истинные баллы.

Вероятностный характер наблюдаемых результатов выполне­ния теста обусловлен влиянием случайных и неслучайных (систематических) ошибок измерения. К случайным факторам можно отнести настроение испытуемого, поведение экзаменатора, обстанов­ку при тестировании в классе и многое другое, то есть то, что учесть и пред­видеть при тестировании попросту невозможно.  Таким образом, такое понимание объективности измерений требует обязательного оценивания вели­чины ошибки измерения.

**3. Инвариантная (специфическая) объективность**.

Третья трак­товка объективности основана на так называемой **современной** теории конструи­рования тестов –  **Item Response Theory (IRT),** которая имеети другое, более общее и точное название, но менее употребляемое название **- Latent Trait Theory (LTT),** которое переводится как математическая теория измерения латентных качеств личности.

Эта теория нередко ошибочно провозглашается как «единственная» и «современная» теоретическая основа педагогических измерений. Между тем, два сравнительно новых варианта классической теории измерения – **Strong True Score Theory и Random Sampling Theory,** а также **Theory of Generalizability** не менее современны, чем **IRT.**

|  |
| --- |
| Преимущества **IRT (теории моделирования и параметризации педагогических тестов),** позволяющие оценить подготовленность обучаемых независимо от трудности заданий теста, приводят к достижению так называемой **инвариантной объективности** измерений, которая предпочтитель­нее объективности, обеспечиваемой классической теорией тестов. |

**Современная** теория измерений появилась в середине XX в. Она строится на более строгой аксиоматической основе. **Измерение** трактуется как **кон­струирование числовой функции**, **осуществляющей изоморфное отображение** некоторой **эмпирической структуры в числовую структуру**, соответствующим образом подобранную **(***изоморфизм – понятие математики, которое опреде­ляет ряд условий взаимно однозначного отображения двух мно­жеств с сохранением их свойств в процессе такого отображения,* В.Т.).

Для достижения инвариантной (специфической) объективности необходима подгонка данных тестирования к требованиям моделей Раша, Раша-Бирнбаума, теории IRT и длительная серьезная работа над тестом. На практике тестологи часто сталкиваются с тем, что эффект инвариантной объективности либо реализуется со слишком большими затрата­ми, либо не реализуется вообще.

**6. Уровни измерений в образовании**

Типология уровней изме­рения основывается на проявлении свойств, лежа­щих в основе построения шкал. Наиболее общая классификация, предложенная С.Стивенсоном, включает четыре уровня измерений и фиксирует при­сущие им свойства. Согласно такой классификации различают шкалы:

а) дискретные и непрерывные,

б) качественные (шкала наименований, порядковая шкала) и количественные (интервальная шкала и шкала отношений), при этом качественные шкалы иногда называ­ют неметрическими (концептуальными), а количественные – мет­рическими (материальными).

Для каждого уровня измерений су­ществуют группы допустимых преобразований и операций с различными математическими и статистическими величинами, ха­рактеризующими измеряемые признаки (*подробно в лекциях, В.Т.*).

**7. Размерность измерений, одномерные и многомерные конструкты.**

Измерение начинается с постановки цели, в соответствии с ко­торой выбирают одну (одномерный случай) или несколько (мно­гомерный случай) переменных. В последнем случае для обозначе­ния измеряемых характеристик часто используют обобщающий термин – **конструкт.**

|  |
| --- |
| Конструкт– совокупность структурных единиц знаний учебного сегмента, для измерения и оценки которых создается педагогический тест (например, знание исторических дат России, фактологический материал, понимание прочитанного, навыки обучения, способность к анализу текста, логическое мышление, интеллект и т.д.) |

Иногда при про­ведении многомерных измерений создают несколько субтестов, каждый из которых является одномерным. Примером такого подхода является **полидисциплинарный** тест, используемый в отдельных странах и состоящий из набора одномерных субтестов.

Некоторые тестологи не совсем и не всегда понимают, что на самом деле измеряет тест,не проверяют соответствие задуманного конструкта и реальных результатов измерения.

**8. Обработка и анализ данных измерения.**

Последний компонент процесса педагогических измерений, включающий обработку, анализ и интерпретацию данных, служит для выявления **качества** результатов измерения, коррекции тестов и представления полученных данных в форме, удобной для интер­претации и сравнения. Благодаря **сопоставимости** тестовых баллов, достигаемой в процессе обработки, по результатам педагоги­ческих измерений можно выстраивать качественный анализ ре­зультатов и принимать обосно­ванные управленческие решения в образовании.

Бочерашвили В.Т.,

зам. директора ГБОУ ДПО ПО «ЦОКО»,

НМС №5, 28.10.2021г.

Приложение

**Классификация В.С.Аванесова.**

**Частные педагогические теории.**

В качестве одной из таковых могла бы стать логико-педагогическая  теория композиции заданий в тестовой форме. Учитывая непосредственную зависимость качества любого теста от правильной композиции содержания и формы каждого отдельного задания, вряд ли можно признать конструктивными известные попытки недооценки в потребности такого рода теории. Ведь каждое нарушение формы, содержания, того и другого – источники погрешностей измерения. Предмет теории – разработка принципов композиции заданий в тестовой форме, позволяющих творчески соединять содержание каждого задания с подходящей для него формой.

**Общие математические теории педагогического измерения.**

К числу таких теорий надо отнести известную на Западе, но не исследованную в России Random Sampling Theory. Более общей является другая теория – Theory of Generalizability, также не исследованная в России. Это самые общие теории, сосредоточившие в себе мощный статистический аппарат исследования точности результатов измерения, возможности использования дисперсионных комплексов для оценки достоверности влияния различных источников погрешностей при измерении латентых качеств личности.

**Специальные математические теории педагогического измерения.**

К эффективности каждого задания с последующим решением о включении их в тест, в банк заданий можно отнести математико-педагогическую теорию тестовых заданий, в которую целесообразно включить часть положений  классической теории тестов, известной на Западе под названием «Item Analysis», а также постулаты, модели и вычислительные методы IRT. Только взятые вместе, они позволяют проводить тщательный анализ эффективности каждого задания с последующим решением о включении их в тест, в банк заданий или об исключении их из разработки.

**В.С.Аванесов о теории педагогических измерений.**

В современной литературе нет ясности относительно теоретической основы педагогических измерений. Раньше в качестве такой основы ошибочно рассматривались статистические теории. Ситуация несколько выправилась после работ Лорда. Однако затем она вновь ухудшилась под влиянием сторонников теории измерений, называемой Item Response Theory (IRT). Последняя имеет и другое, более общее и точное название, но менее употребляемое название – Latent Trait Theory (LTT), что переводится как математическая теория измерения латентных качеств личности.

В России IRT нередко ошибочно провозглашается как «единственная» и «современная» теоретическая основа педагогических измерений. Между тем, два сравнительно новых варианта классической теории измерения – Strong True Score Theory и Random Sampling Theory, а также Theory of Generalizability не менее современны, чем IRT.

В фокусе изучения IRT находятся, строго говоря, не тесты, как системы заданий, а отдельные задания. Вот почему, скорее, это не современная теория тестов, а математическая теория оценки качества тестовых заданий и уровня подготовленности испытуемых. Если взять за основу классификации, например, философские понятия общего, особенного и единичного, то можно предложить такую классификацию теорий этого направления науки:

**Общая теория педагогических измерений.**

Контуры общей теории педагогических измерений удачно просматривались в работах R.Ebel и других педагогов-специалистов по педагогическим измерениям. Последовавшее затем увлечение большинства американских авторов математической стороной проблемы измерения отвлекло их от построения целостной общей педагогической теории измерения.

Главным предметом такой теории должны стать разработка и применение заданий в тестовой форме, тестовых заданий и тестов для активизации учебного процесса, научной организации тестовой формы самоконтроля в процессе самообразования. Настоящее тестирование начинается, развивается и прекращается вместе с учебным процессом. Отстраненное от такого процесса тестирование неизбежно становится объектом бюрократических искажений, несовместимых с тестовой культурой.

**Частные педагогические теории*.***

С общей теорией соотносятся частные теории. В качестве одной из таковых могла бы стать логико-педагогическая *теория* *композиции заданий в тестовой форме.*Учитывая непосредственную зависимость качества любого теста от правильной композиции содержания и формы каждого отдельного задания, вряд ли можно признать конструктивными известные попытки недооценки в потребности такого рода теории. Ведь каждое нарушение формы, содержания, того и другого – источники погрешностей измерения. Прообразы такой теории можно обнаружить в работах многих авторов второго поколения западных специалистов и в публикациях на русском языке. Предмет теории – разработка принципов композиции заданий в тестовой форме, позволяющих творчески соединять содержание каждого задания с подходящей для него формой.

**Общие математические теории педагогического измерения*.***

К числу таких теорий надо отнести известную на Западе, но не исследованную в России Random Sampling Theory. Более общей является другая теория – Theory of Generalizability, также не исследованная в России. Это самые общие теории, сосредоточившие в себе мощный статистический аппарат исследования точности результатов измерения, возможности использования дисперсионных комплексов для оценки достоверности влияния различных источников погрешностей при измерении латентых качеств личности.

**Специальные математические теории педагогического измерения.**

К ним можно отнести частную математико-педагогическую *теорию тестовых заданий*, в которую целесообразно включить часть положений  классической теории тестов, известной на Западе под названием «Item Analysis», а также постулаты, модели и вычислительные методы IRT. Только взятые вместе, они позволяют проводить тщательный анализ эффективности каждого задания с последующим решением о включении их в тест, в банк заданий или об исключении их из разработки.

1. Комментарии, выделенные курсивом принадлежат автору подготовленных материалов В.Т. Бочерашвили [↑](#footnote-ref-1)