Метод экстремальных примеров

Method of the Extreme Examples

БАГДАСАРЯН АРСЕН ГЕВОРКОВИЧ
Заместитель директора Центра оценки и тестирования при правительстве Республики Армения, д-р физ.-мат. наук, профессор
E-mail: angenabg@gmail.com
Ереван, Армения

ARSEN BAGHDASARYAN
Deputy Director of Assessment and Testing Center under the Government of Armenia, Doctor of Science (Mathematics), Professor
E-mail: angenabg@gmail.com
Yerevan, Armenia

АННОТАЦИЯ. В статье предлагается метод экстремальных примеров (МЭП) как способ обоснования утверждений в тестологии. Дается применение метода на конкретных примерах о важнейших тестологических понятиях.

КЛЮЧЕВЫЕ СЛОВА: тестология, валидность теста, надежность теста.

ABSTRACT. The article describes the method of the extreme examples (MEE) that can be used to justify the testological statements. The application of the suggested method is provided for some of the important testological concepts.

KEY WORDS: testology, test validity, test reliability.

Как известно, в математике примеры особо ничего не доказывают, можно столкнуться с ситуацией, как в известном анекдоте: «проверили на 99 примерах, что числа 1, 2… 99 меньше 100, и решили, что все числа меньше 100». Разве что контрпримеры, с помощью которых можно доказывать отрицательные утверждения (опровергать прямые): «пример числа 1000 доказывает, что не все числа трехзначные».

К сожалению, в тестологии редко удается какое-либо утверждение четко доказать. Но здесь примеры играют несколько иную роль, чем в математике. В тестологии примеры могут являться если не строгими доказательствами тех или иных тестологических утверждений, то вполне убедительными аргументами в пользу истинности утверждения. Здесь пример играет не столько роль опровергающего аргумента (контрпримера), сколько убеждающего и почти доказательного. Нередко, чтобы указать на суть проблемы, приходится приводить в пример гипотетические, экстремальные ситуации, которые в жизни никогда не происходят в чистом виде, но часто складываются в аналогичном, более реалистичном виде с менее очевидным проявлением проблемы. Это не уменьшает убедительности примера, поскольку в данном случае пример есть не свидетельство, а средство обозначения проблемы. Ситуация напоминает лупу или микроскоп – видимые размеры объекта (проблемы) не совпадают с реальными, но дают возможность обнаружить невидимые, но реально существующие предметы и явления. Предлагаемый подход мы называем *методом экстремальных примеров* (МЭП) и рассматриваем как средство обнаружения, проверки, обоснования и доказательства тестологических утверждений. Рассмотрим применение метода на конкретных примерах о важнейших тестологических понятиях (см. [1, 2]).

1. СУЩЕСТВОВАНИЕ ТЕСТОЛОГИИ КАК НАУЧНОЙ ДИСЦИПЛИНЫ

К сожалению, сегодня приходится доказывать и само существование тестологии как научной дисциплины, хотя существует она с XIX века. В нашем обществе вряд ли кто-то, если он не физик, будет рассуждать о тонкостях, например, ядерной физики. Это понятно, но редко кому в голову приходит, что он точно так же не разбирается в тестологии (бизнесе, политике, футболе), что это тоже есть научная дисциплина со своими закономерностями и соответствующей теорией, о которой он не имеет представления, и что вряд ли его дилетантские суждения разумны и научно обоснованны. Такому человеку бессмысленно, почти невозможно что-то доказывать в процессе научной дискуссии (у него нет знаний, он не поймет, о чем вы говорите, но есть свое мнение и, что еще хуже, предубеждения). В этой ситуации гораздо эффективнее работает МЭП. Думающего человека гораздо эффективнее убеждают примеры, чем научные утверждения по незнакомой тематике в незнакомой терминологии.

Приведу пример даже не гипотетический, в данной ситуации я оказался сам лично. Всю жизнь (в школе, в университете) в качестве иностранного языка я изучал немецкий и абсолютно не знал английский (кроме латинских букв). Лет 20 назад я впервые пошел на курсы английского языка, не имея никаких знаний. Весь курс состоял из семи уровней, и организаторы настояли, чтобы я прошел тест для определения моего уровня (я был уверен, что мой уровень нулевой и мне надо пойти на «Курс № 0»). Я решил соответствующий тест, и по его итогам мне предложили пойти на «Курс № 5». Естественно, я отказался и начал изучение языка с самого начала – с нулевого курса. Данный случай наглядно показывает, как можно составить настолько неудачный тест, по итогам которого даже не имеющий никаких знаний участник из семи возможных попадает на пятый уровень вместо нулевого. (Ведь моя оценка была завышена не потому, что я «умный», а потому, что тест никчемный.) Было бы намного хуже, хотя более реалистично, если бы мой истинный уровень был не нулевой, а, например, соответствовал бы «Курсу № 3», тогда бы я не смог обнаружить ошибку измерения и пошел бы на «чужой» «Курс № 5».

Теперь представим, что тестирование могло бы быть конкурсного типа с высокими ставками, затрагивающим вопрос жизни и смерти, и участник с нулевым (или даже третьим) уровнем знаний попадает (за счет более достойных участников) на незаслуженный пятый уровень, а ошибка измерения не обнаруживается. Такие случаи в нашей жизни происходят сплошь и рядом, в реальных, ответственных ситуациях. Вряд ли в нашем примере это было целью организаторов, видимо, они даже не имели представления о том, что их тест не работает и дает искаженные результаты, но, конечно же, это не снимает с них ответственности.

Приведенный пример убеждает, что существует понятие качества инструмента оценивания – теста, тогда можно поверить и в существование соответствующей теории его обеспечения – тестологии.

2. СУЩЕСТВОВАНИЕ ПРИЕМЛЕМЫХ ХАРАКТЕРИСТИК ТЕСТОВОГО ЗАДАНИЯ

Поскольку, как мы убедились, существует дисциплина тестология, то она должна иметь и цель своего изучения. Будем считать, что основной целью тестологии является создание качественного, хорошего инструмента оценивания – теста. Естественно считать, что *«хороший» тест должен состоять из «хороших» заданий.* Тогда возникает вопрос: *какие задания считать «хорошими» и как их распознавать?*

Представим, что учитель математики задает классу следующее задание (см. [3]).

Пример 1. *Найти значение следующего выражения: 2 + 2.*
Весь класс решает задание – оно слишком легкое. Учитель решает дать более трудное задание и дает следующее.

Пример 2. *Решить проблему Ферма.*
На этот раз никто не решает задание – оно слишком трудное.

Вывод 1: *существует оптимальная средняя **степень трудности** задания.*

Приведенные примеры имеют общую особенность – они бесполезны для обнаружения подготовленных учеников: задания не имеют дискриминативной способности.

Вывод 2: *существуют задания, которые решает более подготовленный ученик и не решает менее подготовленный, – задания обладают приемлемой **дискриминативностью.***

Таким образом, приведенное применение МЭП убеждает, что: *«хорошее» тестовое задание должно:*
- иметь приемлемую степень трудности;
- иметь приемлемую степень дискриминативности.

3. СОВОКУПНОСТЬ «ХОРОШИХ» ЗАДАНИЙ НЕ ВСЕГДА ЕСТЬ «ХОРОШИЙ» ТЕСТ

Для обоснования этого утверждения представим, что для оценки степени подготовленности учеников по *химии* задается «хороший» тест по *физике* (все задания имеют оптимальную степень трудности и обладают идеальной дискриминативностью). Тем не менее вряд ли такой тест можно считать «хорошим»: удачно прошедший этот тест ученик силен в физике, но не факт, что и в химии тоже, тест это выявить не может. Таким образом, он не соответствует целям оценивания.

4. ВАЛИДНОСТЬ ТЕСТА

Остался открытым вопрос: *какой тест считать «хорошим»?* Понятно, что он должен состоять из «хороших» заданий, но, как показывает предыдущий пример, этого недостаточно. На самом деле предыдущее применение МЭП доказывает чуть больше, чем сформулированное утверждение: *тест должен оценивать то, для оценивания чего он предназначен.* Это свойство теста называется ***валидностью.***

Рассмотрим следующий экстремальный пример, доказывающий важность этого понятия.

Пример 3. *Предположим, что в 5-м классе для разработки теста по математике выделены следующие темы:*
- натуральные числа;
- арифметические операции с натуральными числами;
- восклицательные предложения;
- обыкновенные дроби;
- смешанные числа.

Замечаем, что тема 3 *(восклицательные предложения)* не относится к математике, это раздел русского языка. Следовательно, как бы качественно ни был разработан тест (тестовые задания), он не будет валидным, не будет полноценно оценивать рассматриваемый предметный материал (математику).

Приведенный пример чересчур экстремальный, но следующий, более реалистичный, не отличается по сути.

Пример 4. *Снова предположим, что в 5-м классе для разработки теста по математике выделены темы:*
- натуральные числа;
- арифметические операции с натуральными числами;
- квадратные уравнения;
- обыкновенные дроби;
- смешанные числа.

В этом примере тему 3 *«Квадратные уравнения»* не проходят в 5-м классе (ее изучают позже) – тест снова не будет валидным.

Конечно, приведенные примеры экстремальные – никто для оценивания знаний по химии не задаст тест по физике, в тест по математике не попадут темы русского языка, материал старших классов не будет включен в тест для 5-го класса. Но мало чем по сути от приведенных отличается следующая уже совсем реалистичная ситуация, когда учитель ставит отметки по своему предмету, исходя из поведения ученика в классе: дисциплинированному ставит высокие отметки, а недисциплинированному снижает. Во всех приведенных ситуациях полученная оценка невалидна: она не соответствует предмету оценивания.

5. НАДЕЖНОСТЬ ТЕСТА

Инструмент оценивания – тест, как и всякий инструмент, должен быть надежным. Представим, что изобретены весы, которые выставляют взвешиваемому человеку оценку по какому-либо предмету (например, по математике) и воспроизводят на экране. Было бы странно, вставая на обычные весы, увидеть свой вес, например, 75 кг, а через час – 90 кг. Мы бы не поверили результату измерения – решили бы, что инструмент (весы) не надежен. Конечно, незначительные изменения веса могут быть зафиксированы (например, утром и вечером), но разница показаний (погрешность измерения) должна находиться в разумных пределах. Аналогичную надежность должен продемонстрировать и тест как инструмент оценивания: оценки ученика по тесту при гипотетических многократных тестированиях в идентичных условиях должны разниться друг с другом на приемлемую, достаточно малую величину (стандартная ошибка измерения).

6. ТРУДНОСТЬ ЗАДАНИЯ

Нельзя, оценивая трудность задания, полагаться на личное мнение (на глаз). Глазомер может подвести: казалось бы, с легким заданием на деле ученики могут плохо справляться, а с трудным более успешно. Рассмотрим два примера.

Пример 5. *Вычислить следующий интеграл:* $\int x dx$.

Пример 6. *Вычислить меньшую высоту треугольника со сторонами: 4 см, 8 см, 11 см.*

Если провести социологический опрос на тему «Какое из заданий труднее?», то, скорее всего, большинство опрашиваемых на глаз выберет первое, о вычислении интеграла. Рассмотрим ситуацию подробнее.

Если опрос производится в средней школе, то *первое задание, конечно, труднее,* поскольку ученики не знакомы с понятием неопределенного интеграла и никто из них его не решит. Второе же задание, хотя и требует определенных знаний по геометрии, аккуратных расчетов и выводов, многие решат.

Если же опрос производится среди студентов математического вуза, которые знакомы с понятием неопределенного интеграла, то первое задание решат все. Оно напоминает задание типа «Назови столицу Франции», знающий человек, прочитав требование, сразу дает правильный ответ. Второе же задание правильно решат многие, но некоторые могут ошибиться в расчетах, а другие могут не вспомнить нужные формулы и теоремы из школьной геометрии. Следовательно, *для студентов второе задание труднее.*

Приведенные экстремальные примеры убеждают, что *истинная степень трудности задания обнаруживается исходя из того, какая часть испытуемых правильно его решает.*

Так можно пройтись практически по всей тестологии, на экстремальных примерах обосновывая суть соответствующих понятий и утверждений. Конечно, многие тестологические (в основном психометрические) утверждения строго доказываются на основании методов математической статистики и теории вероятностей (в частности, основные утверждения классической и современной теорий теста, теория надежности, теория выборки, теория проверки гипотез и др.). Но во многих случаях строгого математического доказательства может не существовать или оно может выглядеть чересчур сложным и непонятным для неспециалистов. Часто в таких случаях авторы избегают подробного изложения и сглаживают острые углы. Читателю остается только поверить в истинность такого утверждения. Думаем, в таких случаях применение метода экстремальных примеров может выправить ситуацию и с неменьшей убедительностью свидетельствовать об истинности рассматриваемой закономерности.

СПИСОК ЛИТЕРАТУРЫ

1. *Крокер, Л.* Введение в классическую и современную теорию тестов / Л. Крокер, Дж. Алгина. – М. : Логос, 2012. – 667 с.

2. *Челышкова, М. Б.* Теория и практика конструирования педагогических тестов / М. Б. Челышкова. – М. : Логос, 2002. – 431 с.

3. *Арутюнян, К. В.* Международное исследование по математике и предметам естествознания TIMSS 2003 в Армении / К. В. Арутюнян, А. Г. Багдасарян. – Ереван : Астгик, 2004. – 52 с.