Тема: «Оценка критериальной валидности педагогических тестов по биологии,
используемых в школе».

Несмотря на все расширяющуюся практику массового тестирования, эта технология еще не стала обоснованной составляющей образовательного процесса, поскольку отношение педагогов к методам и процедурам тестирования результатов учебных достижений, в силу различных причин, далеко не однозначно. Одной из этих причин является крайне слабая проработанность вопроса о соответствии результатов оценки учебных достижений с использованием различных методов, в том числе и тестовых. Обычно в отечественной и зарубежной тестологии степень соответствия между результатами тестирования и применения традиционных методов контроля знаний определяется как коэффициент критериальной валидности (пригодности теста для решения поставленной задачи). Общетеоретических сведений об этой и других статистических характеристиках качества тестов в литературе приведено немало, но что касается конкретных примеров применения этой технологии оценки, то они практически отсутствуют.

В связи с этим, основной целью работы было разработать систему контрольных заданий в традиционной (письменные контрольные работы на разных этапах обучения) и в тестовой форме по биологии для школьников и, после параллельного применения, оценить степень согласованности получаемых этими методами педагогических результатов.

Для достижения этой цели нам необходимо решить следующие задачи:

– проанализировать имеющиеся в литературе сведения по теории и практике применения педагогических тестов, необходимые для последующего анализа получаемых результатов;

– разработать репрезентативную для данного рода работ систему заданий в традиционной и тестовой форме по биологии для учащихся 5-11 классов;

– провести апробацию контрольных материалов в течение нескольких этапов обучения в школе

– оценить статистические характеристики качества примененных тестов, включая определение коэффициентов критериальной валидности.

Перед выполнением данной работы была выдвинута гипотеза: разработанные тесты имеют достаточное качество и позволяют оценивать в полной мере качество знаний учащихся, следовательно, могут в равной мере с традиционными способами проверки применяться в школе.

Для выполнения поставленной цели необходимо было дать определение таким понятиям, как тест и валидность.

Тест – это объективное и стандартизированное измерение, легко поддающееся количественной оценке, статистический обработке и сравнительному анализу.

Ключевыми понятиями тестологии, как одной из методических теорий, считаются: измерение, тест, содержание и форма заданий, надежность и валидность итогов измерения. Кроме того, в тестологии применяются такие понятия статистической науки, как выборочная и генеральная совокупность, средние показатели, вариация, корреляция, регрессия и др. Важное место в тестологии занимают вопросы методики разработки педагогического теста.

Валидность – это характеристика теста, отражающая его способность получать результаты, соответствующие поставленной цели и обосновывающая адекватность принимаемых решений. Оценивается от 0 до 1, чем ближе к 0 – невозможно с помощью теста оценить знания и умения по предоставленной теме или разделу, 1 – охватывает полностью материал темы и раздела. В случае если валидность меньше 0,6, то необходим дополнительный опрос.

После создания теста начинается процесс его валидизации. Валидизация – процесс накопления подтверждений для доказательства валидности теста.

Проблема валидности использования тестов для поверки качества усвоения материалов может решена при практическом применении тестов наравне и одновременно с классическими методами проверки знаний учащихся, таких как контрольные работы, устный опрос и др. по одним и тем же темам.

Для этого была разработана система тестовых заданий и система письменных контрольных работ по биологии для учащихся 5-11 классов за 1 и 3 четверти обучения, 1 полугодие и итоговая за учебный год.

В качестве первичного материла для исследований при использовании тестов служили составленные нами матрицы результатов тестирования школьников

В приведенной матрице в первом столбце содержится номер участника тестирования (после ранжирования становится положением ученика в рейтинге), во втором столбце – фамилия и имя ученика, в последующих столбцах – номера заданий. В ячейках цифра «0» обозначает то, что ученик данное задание выполнил неправильно, а «1» – полностью правильное выполнение. Матрицы обрабатывали с использованием классических приемов квалиметрического анализа результатов тестирования, а также, как показано в учебно-методическом пособии В.А.Полынова.

При этом находили следующие показатели: индивидуальный тестовый балл, средний тестовый балл в классе, трудность каждого задания, дифференцирующую способность задания (по точечно-бисериальному коэффициенту Rpb), стандартное отклонение и надежность результатов тестирования (по Кьюдеру-Ричардсону – RKR-20).

В данной работе возникла необходимость выставления по результатам тестирования качественной оценки по четырехбалльной системе («двойка», «тройка», «четверка», «пятерка»), что в данном случае означало перевод тестового балла в оценку. Для этого существует множество в разной степени обоснованных подходов, в том числе и сложные математизированные процедуры.

В данном слаче мы воспользовались достаточно простым из математически обоснованных подходов, который разработан и рекомендуется в использовании АСТ-Центром. Формула представлена на слайде.

В ходе технической части работы, нами была произведена полная статистическая обработка 20 тестовых матриц размером, в среднем, (16-24)Х(6-10) каждая (т.е. 16-24 задания выполнили 6-10 человек в матрице). Если в школьной группе было менее 6 человек (так как работа проводилась в маленькой сельской школе) – то матрицу по ней не составляли (в это число вошли 8 и 10 классы школы с 4 и 2 учениками в классе)

На следующем этапе работы мы провели оценку качества каждого теста.

Величины показателей качества тестовых заданий представлены на слайде.

При этом мы нашли и обозначили номера всех заданий, требующих переработки.

Для сравнительной характеристики успешности выполнения всех видов контрольных заданий (тест - контрольная работа) находили среднюю арифметическую оценку с округлением до десятых для всех классов.

В качестве основного показателя для проверки согласованности результатов, получаемых различными методами контроля знаний, и возможности их взаимозамены рассчитывали коэффициенты критериальной валидности, как величины корреляции между результатами использования традиционных форм контроля и тестовых.

Для анализа качества тестовых материалов мы применили интегрированный подход, дающий более надежный результат.

По статистическим критериям общие результаты анализа качества использованных тестовых материалов по биологии после апробации были следующими:

1. Средний балл, набираемый учениками составил (по всем тестам) 71% от максимального. Соответственно, в целом ученики справлялись с предоставленными заданиями весьма успешно. Данное прохождение тестов считать успешным только для учеников, но не для качества самих тестов, поскольку согласно классическому правилу для нормативно-ориентированных тестов показатель выполнения их должен составлять около 50±10%.

2. Средняя трудность (доля неправильных ответов на задания) всех тестов оказалась равной приблизительно 0,30. Учитывая, что в ориентированных на рейтинг тестах этот показатель должен быть близок к 0,50 (средний по трудности тест), составленные для данных учеников тесты несколько не вписывались в этот критерий, т.е. являлись слишком легкими.

3. Средний точечно-бисериальный коэффициент (Rpb), подсчитанный сначала для каждого теста в отдельности, а затем и по всем тестам, свидетельствующий о способности заданий теста дифференцировать учеников по уровню подготовки, оказался равен 0,32. Это значение в достаточной мере превышает установленный критический уровень для нормативно-ориентированных тестов (0,20), что свидетельствует о вполне удовлетворительном качестве большинства разработанных заданий в тестах. По величине Rpb нами были обнаружены в каждом тесте конкретные задания, требующие переработки, т.е. «бракованные» (см. табл.2). В основном, это были слишком легкие задания, а также неудачно построенные, с которыми как «слабые» (набравшие минимум баллов по всему тесту), так и «сильные» ученики справлялись одинаково. На основании этого мы предполагаем дальнейшую переработку этих заданий или полную замену, причем, в сторону увеличения их сложности .

4. Как уже было сказано, все задания с Rpb ниже 0,2 были идентифицированы в каждом тесте, на основании чего был подсчитан средний процент таких заданий во всех тестах. На их долю приходилось в среднем по всем тестам, около 31%, что также свидетельствует о качестве разработанных тестов по этому критерию близком к достаточному, хотя указывает на необходимость дальнейшей оптимизации теста. Идеальный тест не должен был бы содержать таких заданий, но реально, валидным признается тест, содержащий не более трети (33%) «бракованных» по дифференцирующей способности заданий [25]. Наиболее неудачными в этом отношении оказались все примененные тесты для 11 класса, требующие более значительной переработки (в сторону увеличения сложности).

5. Надежность результатов тестирования по внутренней согласованности была определена по классическому критерию RKR-20. Среднее значение надежности по всем тестам, как и ожидалось, оказалось слишком близки к критическому (0,71 при норме – 0,70). Связано это было, прежде всего, с тем, что тесты действительно оказались довольно легкими для испытуемых. Уменьшение длины теста приводит к снижению его надежности, но увеличивать количество заданий с 16-24 до, например, 30-40 для данных целей тестирования школьников нам представляется нецелесообразным. Тем не менее, большинство предложенных тестов вписались в данный критерий, что свидетельствует о их удовлетворительном качестве.

Таким образом, по наиболее важным показателям (дифференцирующая способность, доля «бракованных заданий», надежность по внутренней согласованности) качество вновь разработанных нами тестов по биологии для учеников 5-11 классов признается как вполне достаточное для оценки уровня подготовки прошедших тестирование учеников. Тем не менее, эти тесты нуждаются в существенной переработке и в процедуре выбраковки некачественных заданий (номера их указаны в таблице 2), поскольку по критерию трудности они оказались значительно ниже нормы (легкими), что привело к снижению надежности результатов при их применении.

С целью доказательства применимости разработанных тестов для оценки качества подготовки школьников мы использовали наиболее обоснованный, но, ввиду большой технической сложности, чрезвычайно редко применяемый метод – расчет коэффициентов критериальной валидности по результатам двух параллельных мероприятий: теста и традиционной формы контроля (письменная контрольная). Результаты расчетов показаны в таблице, которая представлена на слайде.

Обращает на себя внимание то, что тесты оказались заметно менее строгой формой контроля, поскольку средние качественные оценки по всем тестам в среднем составили 4,0, в то время, как по всем контрольным работам – 3,8, что представляет собой заметную разницу.

Тем не менее, несмотря на меньшую «строгость» примененной для выставления качественных оценок по результатам тестирования (методика перевода тестового балла представлена в главе 2), нами обнаружена для абсолютного большинства контрольных мероприятий сильная положительная корреляция между оценками, полученными учениками за выполнение контрольной работы и результатами тестирования. Рассчитанные нами коэффициенты критериальной валидности (RКрит.валид.) для тестов и контрольных работ колебались в пределах 0,78-0,93, при среднем значении – 0,90 (т.е. сильная положительная связь).

Таким образом, установленные величины RКрит.валид. свидетельствуют о явной правомерности замены примененного традиционного метода оценки знаний (письменная контрольная работа) тестовыми технологиями, поскольку последние позволили в данном случае получить сходный результат.

Тот же вывод следует и при определении соотношения оценок в классе (в % от общего числа), полученных за прохождение теста и написание контрольной работы. Проценты одинаковых оценки были близки и даже совпадали для обоих видов контрольных мероприятий. При этом, в число «пятерочников», «четверочников» и «троечников» по обоим видам контрольных материалов попадали почти во всех случаях одни и те же ученики.

Таким образом, результаты работы, показывают примерно равный оцениваемый уровень знаний учащихся, это свидетельствует о высокой валидности тестовых проверок, и том, что разработанные нами тестовые задания по биологии могут наравне с классическими использоваться для проверки качества подготовки учащихся, либо быть взаимозаменяемыми..

Разработанные тесты имеют достаточное качество и позволяют оценивать в полной мере качество знаний учащихся, следовательно, могут в равной мере с традиционными способами проверки применяться в школе, получила полное подтверждение.

1. Большинство разработанных вариантов тестов по биологии для школьников по результатам апробации можно характеризовать как соответствующие основным статистическим характеристикам качества: достаточно высокая дифференцирующая способность заданий, при среднем Rpb = 0,32, приемлемая доля не тестовых по квалиметрическим характеристикам заданий – около 31% и достаточная надежность по внутренней согласованности, составившая RKR-20 = 0,71.

2. Установлена основная причина некоторого несоответствия критериальным значениям таких статистических характеристик теста, как стандартное отклонение и асимметрия – высокий процент относительно легких заданий с низкой дифференцирующей способностью в тесте, и отсюда, высокий средний балл, набираемый учениками по большинству тестов. Определены номера этих заданий и принято решение по их корректировке в сторону увеличения трудности (или замене их на более сложные).

3. Установленные высокие величины RКрит.валид. и близкое совпадение процента оценок и состава (пофамильно) учеников в каждой «балльной» группе в обоих видах контрольных мероприятий свидетельствуют о полной правомерности замены примененных традиционных методов оценки знаний тестовыми технологиями и наоборот, поскольку они позволили в данном случае получить очень сходные результаты по выставленным ученикам качественным оценкам.