Тесты — и не роскошь, и не идеал

Константин КОРСАК, заведующий отделом Института высшего образования АПН Украины, кандидат физико-математических наук, доцент

В системах образования Украины и России понятие «оценка» трактуется как определение и выражение в условных единицах — баллах, а также в оценивающих суждениях учителя степени усвоения учениками знаний, умений и навыков соответственно требованиям школьных программ, уровня старательности и состояния дисциплины.

В предлагаемой вашему вниманию статье автор рассматривает более широкое понятие — «педагогические измерения», обозначая им любые способы оценивания и обследования учащихся.

От экзаменов и оценок — к комплексным педагогическим измерениям

Если в прошлом во времена 4–7-годичной продолжительности обязательного образования процедура оценки была предельно простой, то в конце XX в. ситуация коренным образом изменилась — жизнь потребовала создания сложных систем педагогических измерений. Это результат общей эволюции обществ, экономик и культур.

К сожалению, ни одна из имеющихся систем педагогических измерений не считается совершенной, они хранят немало унаследованных от прошлого рудиментов, подвергаются критике и часто реформируются. Главная причина изменений — увеличение разрыва между реальными возможностями и достижениями национальных систем педагогических измерений и требованиями времени. Перечислим основные факторы «давления» внешних обстоятельств на системы образования и средства для педагогических измерений.

1. Прогресс в научной и производственной сферах уменьшил среднюю продолжительность «жизни» массовых профессий, — исчезают одни и появляются десятки новых (например, в сфере информационных технологий в США их уже свыше 120, а в Украине пока не больше 10). Сложно часто менять учебные планы и средства измерений, оценивания.

2. Образование для избранных сменяется всеобщим и очень длительным (в среднем — 18–20 лет). Оценки не могут больше оставаться инструментом для отбора и дифференцирования учащихся, они обязаны стать гуманными и демократичными.

3. В начале XX в. школьные программы содержали мало информации, учитель располагал достаточным временем на опрос и проведение контрольных работ. В наши дни ситуация противоположна — программы перегружены, не хватает времени на опросы и экзамены.

4. Для прошлого характерна адаптационная доктрина воспитания и обучения, мало учитывающая права и интересы учащихся и студентов. В этих условиях педагогические измерения очень упрощались, игнорировались их психологические аспекты.

5. Система образования в прошлом готовила «послушные винтики» со стандартным объёмом знаний. В наше время «идеальный выпускник» принципиально другой, объём знаний утратил приоритет. Теперь самыми важными считаются умение быстро и эффективно учиться, способность адаптироваться к новым условиям и превратностям жизни, желание непрерывно совершенствоваться. Именно эти личностные качества старая система экзаменов измеряла плохо или не определяла вовсе.

6. Обострилась проблема профессиональной оценки во время найма на работу, контроля над процессом самообучения работников (элементы Life-long Education). Некоторые страны (Япония, США и другие) решают эту проблему сравнительно успешно, в Украине же качество работника определяют по старинке, без всяких там «компутеров».

7. Глобализация быстро формирует открытый мировой рынок не только для товаров, но и для квалифицированной рабочей силы. Национальные системы образования просто обязаны учитывать это и готовить молодёжь к международной конкуренции. Это налагает дополнительные требования к средствам и способам педагогических измерений.

Критерии качества педагогических измерений

В начале 90-х гг. XX в. сформировался «мировой стандарт» среднего образования (12 лет школы и три года профильного обучения), включающий (пока как факультативное) требование к педагогическим измерениям. Учителя должны предоставлять данные не только об отдельных учениках и студентах, но и относительно *состояния, эффективности и качества системы национального образования*.

Все связанные с обучением и воспитанием параметры (данные и характеристики) делятся на три большие группы:

1. элементы с очевидными количественными признаками и измерениями: число лет обучения, рабочих дней, учебных часов, решённых во время контрольной работы заданий или задач, ошибок в изложении или диктанте и т.п.;

2. элементы с условными количественными характеристиками. Их оценка имеет относительный характер, ведь сугубо качественному признаку или характеристике в этом случае приписывается число или некоторое количественное выражение (представление уровня или объёма знаний оценкой в определённой шкале баллов; создание рейтинговых списков лучших школ, университетов или иных учреждений);

3. качественные элементы, которые не удаётся достаточно точно охарактеризовать количественными параметрами, поэтому им даётся субъективная оценка (старательность, послушание или воспитанность, гражданская позиция, активность и другие).

Развитие теории и практики оценки достижений учащихся и мониторинга качества деятельности учебных заведений — важная проблема. В Украине она актуальна в связи с тем, что реформируется система образования, а достижение её целей тормозится недостатками существующих методов оценивания и их чрезмерным субъективизмом. В России продолжается оживлённая полемика по поводу введения «Единого государственного экзамена» — централизованного тестирования всех выпускников школ страны. Сторонники единых экзаменов обращаются к существительному «прогресс», скептики же резонно указывают на тот факт, что почти все выпускники школ и при старой системе поступили в 2001 и 2002 гг. в вузы — зачем же понапрасну будоражить всю страну?

При выборе новых методов оценки и обоснования их преимуществ над существующими следует опираться на научные достижения теории измерений, в частности, последовательно учитывать *критерии качества педагогических измерений*. Самые важные из них — ***объективность, надёжность, валидность и точность***. Из соображений лаконизма ограничимся их краткой характеристикой и сконцентрируем внимание на анализе способов и методов оценивания.

Объективность — антипод *субъективности* с её своенравием, импульсивностью и непредсказуемостью, хаотическими реакциями на второстепенные факторы, влиянием предубеждений, навязанных стереотипов и пр. По-настоящему объективное педагогическое измерение даёт результаты, которые не зависят от состояния, личных черт характера и количества тех, кто его проводит.

Обеспечить объективность оценивания, как показывает практика, можно лишь максимальной стандартизацией её проведения. В этом смысле понятия «объективность» и «стандартизация» почти тождественны. Действительно, объективность процедуры измерения возможна лишь при одинаковых условиях для всех участников, следовательно, при максимальной стандартизации процесса оценивания. Эта процедура должна дополняться объективностью обработки данных и интерпретации полученных результатов.

Надёжность метода измерения определяется уровнем устойчивости результатов, их

повторяемостью во время дополнительных измерений в стандартных условиях.

Степень надёжности метода определяется с помощью коэффициента надёжности. *Коэффициент надёжности* равен коэффициенту корреляции (R) между результатами, полученными одинаковым методом и при одинаковых условиях. Он количественно характеризует уровень совпадения результатов измерений, проведённых в одинаковых (стандартных) условиях. Степень надёжности зависит от: объективности метода, параметров способа измерения, стабильности измеренной характеристики.

Очень важна стабильность измеряемых характеристик, поскольку в сфере образования они часто нестойки, зависят от внутренних и внешних факторов и изменяются с течением времени. Например, педагогам хорошо известны факты влияния на результаты оценки психического и физического состояния экзаменуемого (тестируемого).

Валидность (от лат. *validus* — обоснованный, соответствующий действительности, пригодный) — сложная комплексная характеристика измерений. Валидность измерения — обязательная предпосылка уверенности педагога в том, что действительно измеряются знания учащихся, а не что-то другое. Ведь был в истории образования Франции опыт применения тестов США по математике, «измерявших» не столько глубину знаний, сколько устойчивость и силу нервной системы учеников.

Следовательно, чем выше валидность метода измерения, тем точнее полученные данные.

Для большинства педагогических измерений валидность метода характеризуют более узкими критериальными признаками — *валидностью содержания; соответствия и прогноза.*

Валидность содержания — это соответствие требований, которые мы выдвигаем к измерению, содержанию и сущности объекта измерения. Валидность содержания часто иллюстрируют примерами педагогических измерений уровня успешности. В этом случае содержание определяется планом и методами обучения, а требования — совокупностью знаний, умений и навыков выпускников. Нарушение необходимого соответствия между требованиями и содержанием обучения приводит к недостоверности результатов измерения.

Валидность соответствия — совпадение результатов измерения одного признака различными методами. Для обеспечения валидности соответствия на начальной стадии углублённого изучения возможностей нового метода измерения сравнивают полученные с его помощью результаты с данными эталонных способов. Если инновационность нового метода высока, а эталон отсутствует, то валидность соответствия оценивается сравнением с уже существующими методами, которые удовлетворяют критериям качества.

Валидность прогноза — соответствие полученных результатов прогнозируемым. Каждый современный метод педагогического измерения даёт возможность получить результаты лишь с ограниченной точностью.

Точность метода определяет минимальную или систематическую ошибку, с которой можно провести измерения. Теория ошибок исходит из того, что при условии устранения систематических ошибок постороннего характера колебания результатов измерения подчинены чётким статистическим закономерностям. Это и даёт возможность количественно определить меру точности и пользоваться ею в дальнейшем.

Качество современных способов педагогических измерений

Существует много способов педагогического измерения (оценки). Среди наиболее распространённых — **наблюдения; устная форма проверки знаний** (устное опрашивание**); письменная форма проверки знаний** (письменные работы**); собеседование в виде интервью; тестирование.** Ещё одним способом измерения учебной успеваемости считают анкетирование, но оно не является самостоятельным, оставаясь вариантом тестирования. К тому же правильнее будет отнести его не к сфере контроля, а к диагностике.

Наблюдение — наидревнейший способ оценки и контроля. Его преимущество в том,

что оно даёт возможность увидеть объект в общих чертах и в реальном виде. Результативность и точность выводов, сделанных на основе наблюдения, зависят от личных качеств наблюдателя, а также от многих посторонних факторов. Этот способ оценки имеет большое значение для педагогов. Вместе с тем наблюдение не удовлетворяет требованиям ни одного из приведённых выше критериев качества педагогических измерений уровня знаний учеников и студентов.

Устная форма проверки знаний — тоже очень старая и распространённая форма педагогических измерений, особенности которой обобщены в таблице 1.

Таблица 1. Положительные и отрицательные черты устной формы проверки знаний

Положительные черты
Большая продолжительность существования и массовая распространённость
Простота и доступность, неприхотливость к условиям проведения
Получение достаточно надёжных данных о речевом развитии ученика или студента
Получение представления об общем развитии ученика или студента
Отрицательные признаки
Высокая субъективность процесса измерений и интерпретации результатов. Невозможность обеспечения стандартизации условий измерения
Низкий уровень надёжности, коэффициент которой редко превышает 0,5
Невозможность обеспечить валидность измерения, в первую очередь — валидность содержания
Малая или недостаточная точность, которую не удаётся существенным образом повысить заменой традиционной четырёхбалльной шкалы на 10–20 — или 100-балльную
Чрезмерные затраты времени на измерения в больших группах учащихся (в особенности — при использовании многобалльных шкал)
Недопустимо высокий уровень психологических стрессов во время серьёзных устных измерений
Невозможность апелляций, воспроизводства конфликтных моментов и ситуаций и их анализа.

Специалисты доказывают несоответствие устного опрашивания всем современным критериям качества — объективности, надёжности и валидности, указывают на «грубость», малую распознавательную способность, которую не удаётся повысить ни применением шкал с большим количеством ступенек, ни расширением состава экзаменационных комиссий. Даже лучшие устные формы контроля знаний (сдача экзамена комиссии или группе преподавателей, повторное независимое опрашивание и пр.) не гарантируют объективности процесса измерения и интерпретации его результатов. Субъективизм здесь чересчур высок, ответ неодинаково оценивается различными экзаменаторами, что приводит к чрезмерным колебаниям оценок. Последнее — следствие недостаточной надёжности и точности традиционного устного метода измерения, вытекающее из нарушения критерия объективности. Коэффициент надёжности изменяется от $R = 0{,}4$ или меньших значений до (приблизительно) $R = 0{,}6$. Его вычисляют как коэффициент корреляции между оценками двух различных экзаменаторов или комиссий, работающих с определённым классом или студенческой группой по одному и тому же предмету, с одинаковыми учебными целями и критериями.

К перечисленным недостаткам устной формы измерения знаний прибавляется невозможность выполнения критерия валидности. При допустимой продолжительности устного опрашивания не обеспечивается валидность содержания (ибо несколько вопросов в билете и 5–10 дополнительных не охватывают всего содержания дисциплины).

Из-за этих недостатков устное опрашивание в мировой практике для важных экзаменов применяется ограниченно. Но оно незаменимо при аттестации речевого развития, навыков общения, умения выступать и дискутировать. Можно утверждать, что метод будет использо-

ваться и в дальнейшем, но потеряет привилегии квазимонополиста.

Письменная форма проверки знаний тоже имеет продолжительную историю. Убеждение в том, что она универсальна и может применяться во всех случаях, не раз служило основанием для административных запретов устных экзаменов и замены их письменными работами. В самом деле, письменная форма существенным образом превосходит устную по критерию объективности: сравнительно легко унифицируются условия измерений, возрастает объективность оценки, обеспечивается накопление и сохранение данных, облегчается их обработка и т.п. Но широко распространённые варианты письменного измерения чрезмерно субъективны на стадии интерпретации результатов. Это приводит к несоответствию метода письменной проверки знаний критериям надёжности и валидности. В этом устный и письменный экзамены тождественны. Письменная форма педагогических измерений с точки зрения образовательной администрации имеет лишь одно преимущество над устной — уменьшает количество злоупотреблений со стороны экзаменаторов и позволяет легче выходить из конфликтных ситуаций во время обжалования оценки. Мировая практика подтверждает, что письменные работы останутся надёжным способом педагогических измерений в границах сравнительно узкого поля применения.

Обобщение положительных и отрицательных черт письменного педагогического измерения представлено в таблице 2.

Таблица 2. Положительные и отрицательные черты письменной формы проверки знаний

Положительные черты

Достаточно высокий уровень стандартизации условий проведения (унификация измерения)

Незначительный уровень влияния на учащихся субъективных факторов при выполнении письменных работ

Большой опыт проведения письменных измерений и продолжительность сохранения их результатов

Доступность, возможность охватить проверкой большое количество учащихся

Получение данных о когнитивном развитии ученика или студента

Возможность проверки уровня проблемного мышления, общей и специфической грамотности и пр.

Приемлемая продолжительность проведения первой стадии письменного измерения

Отрицательные признаки

Большая затрата времени экзаменаторов на проверку письменных работ

Высокая субъективность в интерпретации результатов

Недостаточный уровень надёжности

Посредственный уровень общей валидности измерения, недостаточный — валидности содержания

Малая точность, которую можно частично повысить лишь увеличением продолжительности измерения, что усилит другие его недостатки

Собеседование как способ оценки знаний применялось издавна, но лишь в последнее время проводится в виде ***интервью.*** В развитых странах интервью проходит в письменной или устной формах и применяется после проведения тестирования как его дополнение. Цель — получить максимальное количество информации о личности респондента (внешний вид, поведение, речь и пр.). Обобщение особенностей собеседования или интервью приведено в табл. 3.

Таблица 3. Положительные и отрицательные черты проверки знаний во время собеседования или интервью

Положительные черты

Возможность получить общее представление об ученике, студенте или претенденте на вакансию

Высокая свобода действий экзаменатора

Получение данных о когнитивном развитии того, с кем проводят собеседование или интервью

Доступность и демократичность метода

Отрицательные признаки

Большие затраты времени экзаменаторов для получения необходимого количества информации

Неудовлетворительная общая объективность из-за слабой унификации условий проведения. Чрезвычайно высокая субъективность в интерпретации результатов

Малая точность, которую можно частично повысить только за счёт увеличения продолжительности измерения, что усилит другие его недостатки

Недостаточный уровень валидности

Недостаточный уровень надёжности

Тестирование. Термин «тестирование» восходит к английскому *test* — экзамен и используется, как утверждает известный французский энциклопедический словарь Larousse, для измерения или оценки природных и приобретённых способностей с целью предвидения поведения или достижений человека в определённых обстоятельствах. Советский Энциклопедический Словарь акцентирует наше внимание на его применении лишь в сфере психологии и педагогики, ибо тест — это «стандартизированные задания, по результатам которых судят о психофизиологических и личностных характеристиках, а также знаниях, умениях и навыках испытуемого».

Классические определения тестирования в психологии подчеркивают: а) эмпиричность оценки; б) определение личных признаков и качеств через использование количественных показателей. **В педагогике классическим и вместе с тем достаточно полным считают «критериальное» определение К. Ингекампа:** ***«Тестирование — это метод педагогической диагностики, с помощью которого выбор поведения, презентующего предпосылки или результаты учебного процесса, должен максимально отвечать принципам сопоставления, объективности, надёжности и валидности измерений. Он должен пройти обработку и интерпретацию и быть приемлемым для применения в педагогической практике».***

Тестирование нельзя рассматривать как идеальный метод, исключая на этом основании все иные. Хотя нет сомнений — при надлежащей предварительной подготовке именно тесты лучше других средств удовлетворяют основные методические критерии качества, обеспечивают приемлемую объективность всех трёх главных стадий процесса оценки — измерения, обработки данных и их интерпретации. Коэффициент надёжности — от 0,7 до 0,9, что следует признать достаточно высоким показателем.

Хорошо подготовленное тестирование даёт возможность удовлетворить и критерий валидности. Знания оцениваются по объёму и полноте, системности, обобщённости и мобильности. Последние характеристики определяются с помощью теста соответствующей сложности, а объём знаний — путём получения ответов на определённое количество вопросов.

Важным этапом при использовании любых способов педагогического измерения является ***оценивание*** — процедура, заключающаяся в конвертации полученного во время измерений первичного результата в нормированную шкалу баллов — «оценку». Методика оценивания состоит в определении алгоритма выполнения этой процедуры.

В основе современной процедуры оценки результатов тестирования лежит теория шкалирования**,** обеспечивающая получение достоверных данных об уровне знаний, умений и на-

выков учащихся. Для использования преимуществ этого метода достаточно перейти от старой четырёхбалльной к многобалльной шкале оценки, лучшим вариантом которой остаётся 11-балльная (0–10).

Общие особенности тестирования как формы педагогических измерений приводятся в таблице 4.

Таблица 4. Положительные и отрицательные черты тестовой формы проверки знаний

Положительные черты

Высокая объективность процесса измерений и интерпретации результатов

Возможность обеспечения стандартизации условий измерения

Приемлемый уровень надёжности, коэффициент которой может достигать 0,9

Возможность обеспечить валидность измерения, в первую очередь — валидность содержания

Достаточная точность, которую можно повысить заменой традиционной четырёхбалльной шкалы на более протяжённую

Незначительные затраты времени на измерения в больших группах учеников и студентов

Незначительный уровень влияния субъективных факторов во время измерений

Лёгкость обеспечения продолжительного сохранения измерений результатов и автоматизации их обработки

Облегчение процесса интеграции государственной системы образования в европейскую, благоприятствование мобильности учеников, студентов, просвещенцев, специалистов всех профилей

Отрицательные признаки

Необходимость обоснованного изменения психологии воспитания и обучения, связанная с переходом к высшему уровню состязательности и индивидуализма

Замена учебников, рассчитанных на устное опрашивание, новыми, ориентированными на тестовую форму проверки знаний

Значительные затраты времени на первичную подготовку качественных материалов для проведения измерений

Необходимость преодоления сопротивления и комплекса предубеждений приверженцев старых методов педагогических измерений

Малое количество специалистов по тестированию в системе образования, что замедлит процесс перехода на современное тестирование

Шкалы оценки различных стран мира

Как ни парадоксально, *но шкалы оценки не играют значительной роли в педагогических измерениях*. Об этом свидетельствует непреложный факт — чрезвычайное разнообразие шкал оценок и замедленность процесса формирования мирового стандарта. В качестве доказательства, привожу табл. 5, содержащую данные о шкалах нескольких десятков стран, используемых ими при проведении важных экзаменов.

Таблица 5. Шкалы измерений, которые применяют в различных странах на выпускных и иных важных экзаменах

0–100 баллов

Афганистан, Бангладеш, Бахрейн, Белиз, Боливия, Бурунди, Гватемала, Гаити, Гондурас, Заир, Замбия, Израиль, Индия, Иордания, Ирак, Иран, Кения, Китай, Коста-Рика, Либерия, Мексика, Мьянма, Никарагуа, Нигерия, Объединённые Эмираты, Пакистан, Панама, Перу, Руанда, Саудовская Аравия, Сирия, Судан, США, Тайвань, Египет, Эфиопия, Фиджи, Филиппины, Шри-Ланка, ПАР, Южная Корея, Япония

30 баллов

Италия

20 баллов

Венесуэла, Бельгия, Ливан, Иран, Португалия, Франция и 14 франкоязычных стран Африки

15 баллов

Алжир, Марокко, Тунис

13 баллов

Дания

0–12 баллов

Уругвай

1–12 баллов

Украина

0–10 баллов

Албания, Аргентина, Вьетнам, Греция, Исландия, Мексика, Нидерланды, Румыния, Сальвадор, США, Турция, Швейцария, Югославия и пр.

6 баллов

Австралия, Болгария, Германия, Польша

5 баллов

Австрия, Венгрия, Великобритания и 13 стран Содружества, Иран, Катар, Куба, Латвия, Монголия, Таиланд, Хорватия

4 балла (2, 3, 4, 5)

Россия и большинство стран СНГ, Испания, Китай, Сингапур, Чехия

3 балла (3, 4, 5)

Непал

Примечание: в некоторых странах используются несколько шкал (как, например, в США)

Из доступных данных известно, что ни одна страна мира не использует шкалу 1–12, имеющую чётное количество баллов. С такой шкалой Украина становится оригинальной, но радоваться этому обстоятельству не стоит: надежды на присоединение к ней стран Европы и мира и преобразования её шкалы в мировую нет.

Объём статьи не позволяет детально осветить основы теории шкалирования, укажем лишь несколько важных обстоятельств:

1) При условии качественного измерения гистограмма оценок имеет форму симметричной одновершинной кривой распределения Гаусса. Следовательно, количество баллов в шкале должно быть нечётным (1–5 или 0–10) и иметь в центре один из баллов (соответственно 3 и 5). Украинская шкала в центре имеет баллы 6 и 7, поэтому по законам математической статистики не подходит для измерения характеристик учеников.

2) Шкалы с большим количеством баллов (101 балл и больше) предусматривают использование тестов с огромным количеством вопросов и задач. Они получили широкое распространение в США.

Идеальны ли многобалльные шкалы и система тестов? Конечно, нет. Не только европейские учёные-педагоги, имеющие собственные образовательные традиции и гордость, но и сами американцы без видимых усилий находят недостатки в абсолютизации тестирования и игнорировании иных способов педагогических измерений. Не удивительно, что страны Европы с традициями использования одинаковых учебных планов во всех школах пока только в исключительных случаях обращаются к американским тестам и шкалам.

Изменение шкал — ***достаточно дорогой процесс***, требующий больших изменений. Например, введение в украинских школах 12-балльной шкалы в несколько раз увеличивает затраты времени на опрашивание и распределение учеников по уровням знаний (опрос учителей привёл автора к выводу, что затраты времени увеличились в 2–4 раза). Напряжённость труда учителя заметно выросла, что было бы справедливо компенсировать повышением зарплаты. К сожалению, авторы 12-балльного нововведения как раз об этом и не подумали.

*** * ***

Попыткам реформирования системы педагогических измерений должен предшествовать период углублённого ознакомления с практикой проведения государственных экзаменов в развитых странах. Пренебрежение этим проверенным в европейских реформах правилом и азартное желание ускорить процессы изменения в сфере образования угрожают обернуться ошибками.

Педагоги США изобрели и применили тесты «американского» образца потому, что из-за отсутствия жёсткого регулирования средние школы в этой стране имели (и имеют) чересчур разнообразные программы и комплекты предметов. Университеты были вынуждены способствовать появлению структур, проводящих независимую оценку всех потенциальных абитуриентов. Наши программы обучения столь стандартизированы, что «американские» тесты в школе выглядят чрезмерной роскошью и почти неуместны. Экономически эффективнее и методически правильнее усовершенствовать собственную систему экзаменов, учитывая опыт европейских стран.