Использование результатов тестирования для оценки качества образования: за и против

Тад Драммонд, Американские Советы по международному образованию

После того как в Кыргызстане были введены стандартизированные тесты, возник вопрос о том, как можно использовать результаты этих тестов. Можно ли применять их для оценки системы образования в целом? Или для сравнения качества работы отдельных школ? Для оценки работы отдельных классов?

Эти вопросы вызывают оживленную дискуссию во всех странах, где применяется тестирование: в Великобритании, США и в Европе. В данной статье мне хотелось бы обсудить границы, в которых можно применять результаты тестов для выведения заключений о качестве образования[1], а также представить краткое описание американского подхода, доказавшего свою эффективность при оценке эффективности школы,[2] – системы "добавленной стоимости" в штате Теннеси, США. В этом штате результаты тестирования применялись для оценки эффективности школ на всех уровнях системы образования. Однако мне хотелось бы особо отметить, что результаты тестов сами по себе обычно НЕ дают нам достаточно информации, чтобы судить о качестве образования или эффективности школы. И, несмотря на то что система добавленной стоимости применяется в некоторых других странах, в Кыргызстане необходимо проделать огромную предварительную работу, прежде чем использовать ее, какой бы привлекательной она ни казалась.

Почему возникает желание использовать результаты тестов для того, чтобы делать выводы о качестве образования в отдельной школе, области, даже в отдельном классе? Прежде всего, в некоторых образовательных системах они являются единственным надежным инструментом оценки достижений учащихся. Если тестирование проводится профессиональным агентством, то оно предоставляет объективную оценку. Тестовые организации обычно имеют в своем распоряжении сложные научные инструменты для необходимого анализа всех компонентов теста, самих тестов и их результатов, что позволяет считать их результаты более надежными и валидными. Благодаря этому,

[1] "Качество образования"- это общий термин, который включает в себя много значений. Однако я использую этот термин в рамках его употребления среди педагогов в Кыргызстане.

[2] В данном контексте под термином "эффективность школы" подразумевается умение школы способствовать академическому приросту ученика. Это понятие более специфично, чем понятие "качество образования", и его можно оценить с помощью эмпирических показателей.

тестирование становится привлекательным способом сравнения результатов разных школ и их достижений в образовании. Их можно легко собрать, проанализировать и интерпретировать, используя существующие базы данных. Они также являются важным показателем для учащихся и родителей (особенно если речь идет о тестировании высокого значения). Естественно, что общественность, и особенно родители, имеют право знать, где и чему должны обучаться дети. И, наконец, правительство, со своей стороны, имеет право знать, какие результаты были достигнуты благодаря вкладу государства в систему образования. На основании всех этих соображений появляется естественное желание применять, сравнивать и делать выводы о качестве образования или эффективности школ по результатам тестирования. Однако, позволяют ли нам результаты тестирования делать выводы о качестве образования в тех или иных школах?

Основным аргументом против того, чтобы считать результаты тестов надежным показателем качества образования, является то, что при акцентировании внимания только на ***итоговых*** тестовых результатах упускаются из виду другие факторы, влияющие на успех или неуспех учащегося в тестировании. Эти факторы очень разнообразны: социально-экономические условия школы, неодинаковый стартовый уровень учащихся, уровень финансирования школы, обеспечение нужными материалами, подход учителей, мотивация, физическая инфраструктура и т.д. Многочисленные исследования показали, что данные факторы могут оказать влияние на итоги тестирования.[3] Следовательно, сравнение школ между собой на основании результатов тестирования без учета вышеперечисленных факторов является неуместным.

Возможно, самое главное заключается в том, что оценка конечного результата работы школы не учитывает разницу в изначальном уровне подготовки учеников, хотя их стартовый уровень сильно влияет на результаты тестов. В большинстве стран мира школы находятся в разных условиях. Есть районы, где большинство родителей имеют высшее образование, а есть и такие, где проживают в основном люди без образования. Дети приходят в школы с разным уровнем подготовки: некоторые из них уже могут читать и писать, другие – нет. Уровень образования родителей очень часто влияет на уровень подготовки детей к школе. В некоторых странах дети порой приходят в школу не только не умея читать, но и не понимая, что им говорит учитель, так как дома они говорят на другом языке, и начинают изучать новый язык только в первом классе. Таким образом, изначальный уровень учеников отличается в разных школах (и даже классах).

[3] Больше информации о влиянии различных факторов на результаты тестирования можно найти в: Travers, R.M.W. (1981). In J. Millmen (ed.), *Handbook of Teacher Evaluation*. Beverly Hills: Sage.

Кроме того, немалое значение имеет система набора учеников в школу: в некоторых школах, чтобы быть принятыми в первый класс, ученики должны сначала пройти более или менее сложный отбор. У гимназий, лицеев и т.д. могут быть более высокие стандарты, и отнюдь не все ученики могут быть допущены к обучению в этих учебных заведениях. Другие школы принимают всех учеников, независимо от их подготовленности или академических способностей.

Турецкий лицей «Себат» в Кыргызстане отбирает учеников после 6 класса с помощью тестирования, основываясь на их академических способностях и применяя конкурсный подход. Даже если все остальные условия равны, будет ли справедливо сравнивать результаты тестов выпускников турецких лицеев с результатами выпускников школ, в которых не существует отбора учеников? Безусловно, нет. И турецкие лицеи – это не единственные школы в Кыргызстане, тщательно отбирающие учеников. Существуют и другие школы, как частные, так и государственные, которые следуют такому же принципу. Есть школы, принимающие почти всех учащихся, но требующие определенных «взносов» для поступления, тем самым устанавливая социально-экономический барьер, благоприятствующий детям из состоятельных семей.

В общем, для того, чтобы сравнивать качество работы школ, необходимо учитывать первоначальную подготовленность учеников, а это довольно сложно осуществить на практике. Несмотря на наше желание сделать выводы и сравнения, мы почти всегда в каком-либо смысле будем сравнивать «яблоки и апельсины», то есть абсолютно разные, несопоставимые вещи. В настоящее время ни Национальный центр тестирования при Министерстве образования, ни Центр оценки в образовании и методов обучения не имеют ни возможности, ни цели развивать необходимые инструменты для учета вышеназванных факторов. Они лишь сообщают итоги тестирования.

Другая причина, почему мы не должны придавать излишний вес результатам тестов и делать по ним заключения об уровне образования, - это то, что мы можем забыть о других важных показателях- физическом и психологическом состоянии учащихся, их уверенности в своих силах и т.п.. Несомненно, некоторые знания и навыки можно будет измерить, а вот другие – нет. Еще один довод против придания чрезмерно большого значения результатам тестов – учителя могут начать «учить для теста», уделяя недостаточно внимания предметам, не входящим в содержание теста. Это особенно касается тестов высокого значения, каким и является Общереспубликанский тест.

К сожалению, неверное толкование и применение результатов тестов некоторыми работниками образования в Кыргызстане привело к тому, что учителя неправильно понимают, как нужно на самом деле использовать эти результаты. В настоящее время

повсеместно учителя воспринимают результаты всех тестов как показатели качества образования в их школах. К примеру, некоторые из директоров школ, с которыми Американские Советы работали в 2004 году, признаются, что они боялись позволять всем своим выпускникам принимать участие в Общереспубликанском тестировании, так как, по их мнению, школа или даже весь район после тестирования могли получить взыскание за плохую работу.[4]

Существует также опасность и того, что если по результатам теста будут судить о работе школы в целом, школа будет стремиться уделять больше внимания «средним» ученикам, что отразится на работе с сильными или, наоборот, с отстающими ребятами (на которых они не возлагают больших надежд). Школы будут пытаться показать, что они хорошо обучают. Появляется определенный риск, что время и силы будут потрачены только на тех, кто способен улучшить результаты ценой минимального количества потраченных на них средств и времени. Более «трудные» дети или те, которые все равно наберут высокие баллы на тестировании, получать меньше внимания, чем они заслуживают.

Может возникнуть и другая долгосрочная проблема, если придавать чрезмерное значение результатам теста: между школами, показавшими хорошие и не очень хорошие результаты, возникнет своего рода систематическое неравенство. В школах с более низкими показателями это может плохо отразиться на моральном состоянии учеников и их стремлении учиться – происходит «геттоизация». Родители захотят отдавать своих детей только в те школы, которые имеют более высокие показатели по результатам тестирования. «Сильные» школы становятся сильнее, «слабые» - слабее.

Давайте рассмотрим некоторые отчеты результатов Общереспубликанского тестирования 2003 года в Кыргызской Республике, например в Свердловском районе. Во-первых, данные по отдельным школам несопоставимы хотя бы из-за неравного количества участников теста. Некоторые школы дают данные по 5-8 участникам, некоторые – по 30-40. В этом случае нельзя забывать, что на средний балл влияет количество участников. Чем ниже количество участвующих, тем сильнее риск отклонения (очень высокие или очень низкие баллы), которые влияют на общий результат. В любом случае, даже если количество учеников одинаково, необходимо принять во внимание и другие факторы. Три из шести самых лучших школ – это турецкие лицеи. Как уже упоминалось ранее, в этих

[4] Создание Общереспубликанского тестирования не имело своей целью определение эффективности школ или качества образования. Целью этого тестирования являлось дифференцирование между когнитивными способностями абитуриентов в целях отбора в высшие учебные заведения. В общем, для того, чтобы тест был валидным, он должен иметь одну цель. Смешивание более чем одной цели тестирования (например, отбор в высшие учебные заведения с измерением школьных достижений) не рекомендуется, поскольку может привести к проблематичному измерению.

лицеях принята выборочная система приема учащихся. Эти школы хорошо оснащены, имеют устойчивые преподавательские ресурсы и материальную базу. Безусловно, следует ожидать, что результаты этих учащихся будут выше, так как материальная база их школ лучше. Отсюда следует, что мы не имеем права делать подобные сравнения с другими учебными заведениями.

Таким образом, необходимо очень тщательно подходить к определению «эффективности школы» или качества образования в целом. Если не учитывать «невидимые факторы» при проведении анализа результатов тестов, это приведет к неточным и ненаучным заключениям. Для политиков очень важно понимать, как применять результаты тестов, также как и не применять их.

Распределение результатов (среднее значение тестовых баллов секций основного теста) Общереспубликанского тестирования 2003 г. по школам Кыргызской Республики

Область: г. Бишкек

областной центр/районный центр/район: Свердловский

		среднее значение тестовых баллов				
наименование школы	число участников тестирования	среднее значение суммы тестовых балов	математика	аналогии и завершение предложений	чтение и понимание	практическая грамматика родного языка
сш КРСУ	30	177.30	46.83	48.40	42.99	38.55
школа-гимназия №12	40	170.95	46.24	48.32	40.65	35.26
Кыргызско-турецкий женский лицей «Айчурек»	24	167.46	49.23	45.05	38.38	34.47
Кыргызско-турецкий мужской лицей им. Ч. Айтматова	33	161.09	49.77	41.29	37.53	32.30
частная школа-комплекс "Звездочка"	8	158.13	42.01	42.30	38.24	35.32
Республиканский Анадолийский кыргызско-турецкий лицей	28	157.54	43.68	36.64	39.69	37.05
гуманитарная гимназия №23 им. Гете	33	146.15	37.55	41.43	36.34	30.45
Бишкекский коммерческо-экономический техникум МАУФБиП	5	143.00	36.92	40.56	33.61	31.56
эколого-экономический лицей №65	41	141.00	39.45	40.64	33.29	27.33
Национальная компьютерная гимназия №5	156	140.83	36.08	32.62	36.24	35.49
учебно-воспитательный комплекс №67	118	139.21	35.21	38.40	34.79	30.52
сш №66	75	136.68	34.22	35.89	34.23	31.99
сш №52	36	136.53	35.45	37.55	32.66	30.48
сш №1	58	134.81	35.72	38.86	32.23	27.69
сш №51	39	134.26	34.58	39.03	32.50	27.99
сш №46	15	132.13	35.78	35.13	31.48	29.40
сш №4	8	130.50	38.38	36.53	29.27	25.78
сш №33	29	130.21	36.05	36.34	31.18	26.38
школ-гимназия №38	82	128.98	33.58	32.24	31.60	31.20

Система «Добавленной стоимости» или «Измерения эффективности школ и учителей»

Мне бы хотелось осветить некоторые недавние попытки использовать тестовые баллы для определения того, насколько отдельные школы или даже классы были эффективны в повышении уровня академического роста студентов. Вместо общих терминов (например, «качество образования»), будем говорить о конкретном воздействии

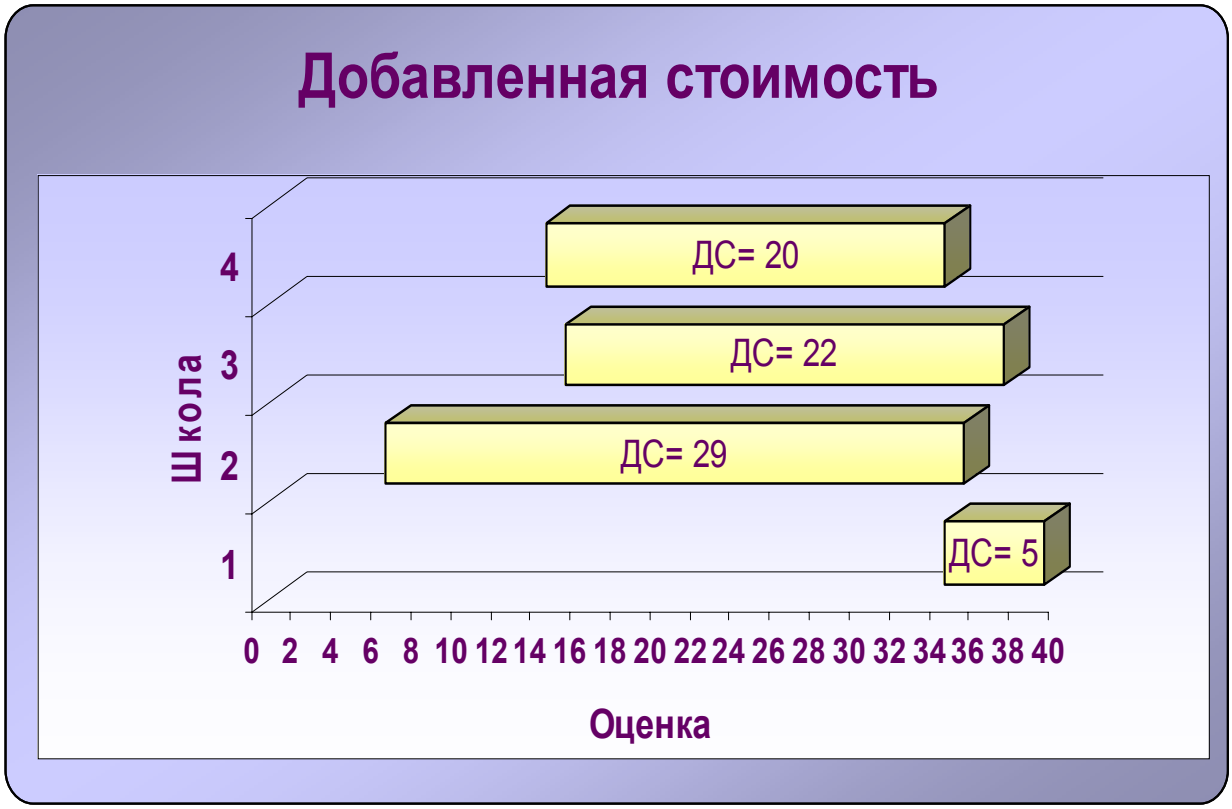
школы и учителей на академический прирост учащихся.[5] В течение 1970-х и 1980-х годов, применение тестовых баллов активно обсуждалось всеми тестовыми и образовательными организациями в Соединенных Штатах Америки. Повсеместно было принято, что баллы тестов не могут быть использованы для сравнения эффективности работы школ и для выведения категорических заключений об эффективности учителей. Однако ученые смогли определить и выделить некоторые факторы, которые позволили им создать новые методы применения результатов тестов для сравнения успешности школ и учителей в воздействии на академический прирост учащихся. При условии, что все решающие факторы поддаются контролю, можно разработать системы, которые будут служить инструментом для оценки не только школы, но даже определенных классов.

Работа профессора Сандерса основывалась на предположении, что хорошие школы и хорошие классы - это те, которые способствуют обучению и развитию детей. Следовательно, он предполагает, что, если *изолировать многочисленные факторы,* влияющие на конечные оценки, можно сделать вывод, какие школы и классы положительно влияют на учеников. Хорошие школы и классы должны уделять больше внимания *академическому приросту*, независимо от стартового уровня. Основной принцип здесь таков – общий уровень академического роста более важен, чем конечный результат. Почему? Дело в том, что продолжать требовать от преподавателей достигнуть одинаковых результатов просто несправедливо и невозможно. Учителя не могут контролировать, все ли ученики поступают в школы хорошо подготовленными, одинаковы ли социально-экономические условия, в которых живут и работают родители их учеников; уровень образования родителей; и во многих случаях они не могут контролировать то, что получаемые государственные или частные средства не отвечают потребностям учреждений и их обеспечению. В то же время нельзя отрицать, что даже в трудных экономических условиях и среди бедных школ есть много хороших школ, чья работа значительно влияет на учеников даже в самых тяжелых условиях. Хотя по многим причинам, перечисленным выше, их итоговые результаты не так высоки, как у соседних школ. С другой стороны, есть ряд школ, хорошо обеспеченных, в которых работа учителей может быть не настолько эффективна.

Однако требовать от учителей эффективного воздействия на академический прирост учащихся можно, причем независимо от стартового уровня. В западной образовательной литературе суммарный уровень академического прироста называют

[5] Под термином “рост” подразумевается среднее значение ожидаемого академического роста за определенный период обучения. Под термином “прирост” понимается отклонение от среднего значения ожидаемого академического роста, которое может быть как положительным, так и отрицательным.

«добавленной стоимостью». Измерение академического прироста дает нам подход, который позволяет учитывать различные факторы, влияющие на работу школы и конечные результаты при оценке эффективности школы или класса на основании результатов тестов. В качестве примера можно привести график, показанный ниже. Какая школа демонстрирует высокие результаты? В какой школе самый большой академический рост? Где более эффективное преподавание?[6]



Одной из наиболее широко известных программ для определения добавленной стоимости в Соединенных Штатах была программа, разработанная профессором Уильямом Сандерсом из Университета штата Теннеси. Целью его работы являлась разработка инструмента измерения того, насколько хорошо школы или отдельные классы способствуют академическому росту учеников. Основной принцип его работы заключался в следующем: оценка работы учителей и школ определяется по тем факторам, которые школы и учителя могут контролировать.

[6] Предполагается, что тестирование проводится в конце и начале определенного периода. Желтые блоки с левой стороны представляют добавленную стоимость, став начальным уровнем, а правая сторона демонстрирует конечные уровни. В этом случае, школа №1 показывает самый лучший результат, но в то же время самый низкий прирост. Школа №2 демонстрирует самый высокий академический прирост учеников, но находится на третьем месте согласно средним конечным результатам после школ № 1 и № 3.

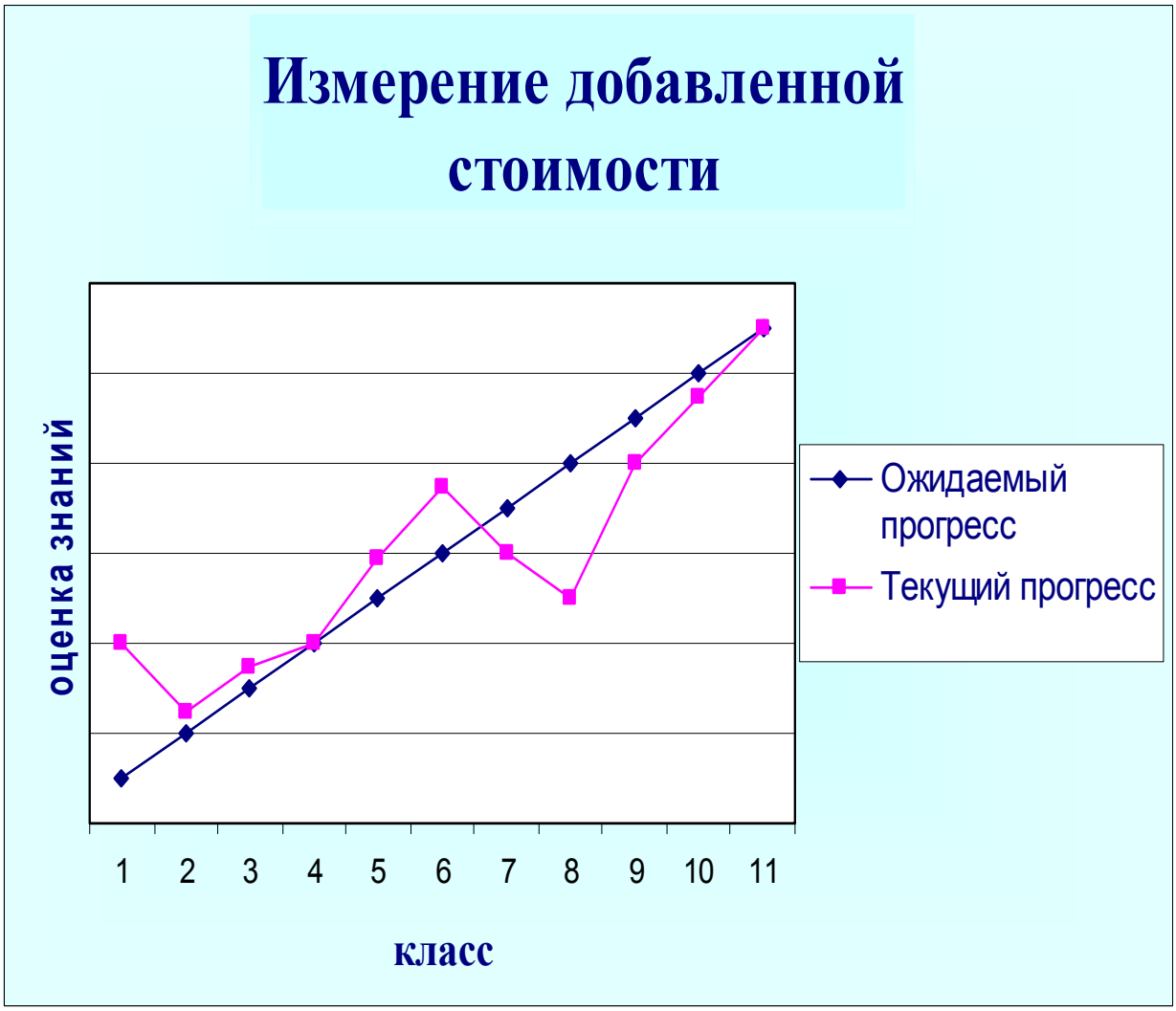
Исследование Сандерса было включено в его государственный план в 1992 году для того, чтобы позволить школам и учителям быть более ответственными за ту работу, которую они проводят. С 1993 года тестирование проводится по математике, навыкам чтения, словесно-логическим навыкам и социальным наукам для учеников с третьего до восьмого классов. Программа Сандерса основана на многовариантном статистическом моделировании, включающем информацию о результатах тестов всех учеников, проходивших тестирование за несколько лет. Результаты тестов всех учеников за все годы тестирования аккумулируются, позволяя отслеживать долгосрочные тенденции в обучении и отмечать отклонения. В настоящее время 5 миллионов введенных в базу данных связывают каждого ученика с его личными показателями, учителями и школами.[7]

В системе добавленной стоимости инструменты тестирования, используемые для определения стартового и конечного уровней учащихся, должны быть эквивалентными и стандартизированными. Люди, профессионально занимающиеся тестированием с использованием новейших методов статистического анализа тестовых заданий, должны обращать особое внимание на развитие технологий и систем обработки данных по добавленной стоимости. Анализ результатов теста отражает прогресс или регресс каждого ученика в течение трех лет. В этом отношении необходимо отметить два момента. Прежде всего, абсолютные результаты не сообщаются, так как оценивается академический прирост. И для составления отчета по отдельным школам информацию собирают в течение трех лет перед тем, как сделать анализ. Выборка должна быть большой, поскольку маленькая выборка может привести к ошибочной интерпретации данных в результате наличия в этих данных слишком высоких или слишком низких значений. И наконец, результаты работы отдельных классов не доводятся до общественности и используются только внутри школы.

Работа Сандерса сложная и требует применения высокого уровня статистических данных. Для оценки эффективности школ и классов по его методу необходимо, чтобы было соблюдено несколько условий. Во-первых, должен оцениваться «начальный уровень» учеников. Это означает, что стандартизованные тесты должны быть совершенно идентично разработаны и проведены среди всех учеников. Необходимо создать статистические модели, прогнозирующие «нормальный рост», чтобы позволить нам измерить, каким должен быть нормальный академический прогресс для данной группы учеников с определенным начальным уровнем. Данные модели используются в школах для определения прогнозируемых результатов. Затем в конце отчетного периода (года)

[7] Важно отметить, что в развитие науки тестирования необходим большой вклад: измерение добавленной стоимости базируется прежде всего на имеющемся опыте профессионального тестирования.

устанавливается разница между ожидаемыми и полученными результатами и проводится анализ с помощью тестирования. Необходимо отметить, что сравнение конечных уровней не проводится между школами, но в основном измеряется уровень академического прироста учеников за время обучения в школе. На схеме ниже приведена примерная модель, демонстрирующая, что реальный академический прирост может быть больше или меньше ожидаемого год от года.



Обзор результатов ранней работы Сандерса:

Прежде всего, одним из наиболее интересных результатов работы профессора Сандерса является наблюдение о том, что существуют измеримые различия между школами и учителями в отношении эффективности воздействия на академический прирост у учащихся. Данные эффекты имеют тенденцию оставаться постоянными год от года. Также результаты этого анализа тесно соотносились с традиционным оцениванием

работы учителей (открытые уроки, анализ портфолио, собеседования). И учителя, которых оценивали положительно посредством более традиционных методов наблюдения, также показали хорошие результаты в отношении академического прироста учеников (по результатам анализа данных тестирования). Другое существенное достижение заключалась в том, что результаты школ не связывались с их месторасположением. Прогресс учеников также не был связан со знаниями, с которыми они поступали в классы. Согласно Сандерсу, самым значительным результатом его ранней работы был вывод о том, что эффективность преподавания может оказаться измеримой с помощью результатов теста. Учителя имеют больше влияния на развитие учеников, чем ожидалось. Социально-экономические и этнические факторы имеют гораздо меньшее влияние, чем ожидалось.

В дальнейших исследованиях, начиная с 1996 года Сандерс открыл, что влияние учителя на ребенка отслеживается на протяжении нескольких лет. Также, совокупный академический прирост учеников 3-8 классов по всему штату не был связан с их этнической принадлежностью. Результаты не были связаны также и с социально-экономическим статусом ученика. Они не зависели и от средних показателей школы. В итоге, положительный прирост был заметен в разных школах независимо от стартового уровня, этнической принадлежности или географического расположения.

Хотелось бы отметить некоторые другие интересные результаты работы профессора Сандерса, например: когда ученики переходят в другое здание школы, у них снижается уровень прогресса (так как в 6 и 7 классах в США ученики обычно меняют здания школ). Некоторые школы стабильно демонстрировали значительный прогресс, но, наблюдая за улучшениями, нужно отметить, что наиболее низкими они были у тех детей, которые справлялись с учебой лучше всех. Только самые эффективные (лучшие 20%) учителя справлялись со всеми уровнями достижений учеников. И наконец, размер класса, вопреки ожиданиям, не являлся значительным фактором.

Согласно У. Сандерсу, проведение подобного рода исследований может натолкнуться на ряд препятствий при применении методов добавленной стоимости. Вот некоторые из них:

- миграция учеников;
- для анализа требуется сложное техническое оборудование;
- необходимо разрабатывать и проводить большое количество тестов;
- на сегодняшний день эти методы не могут быть использованы отдельно от других (открытые уроки, собеседования и т.п.).

Заключение

Несмотря на то, что раннюю работу Сандерса критиковали, со временем его подход начал вызывать интерес. Сандерс получил вознаграждение за свою работу в данной сфере. Однако применение результатов тестов для оценки качества образования остается противоречивым. Даже в штате Теннеси, где его работу в целом признали, традиционные подходы, такие как наблюдение за уроками, интервью и оценка личного дела каждого ученика, широко применяются и до сих пор как основной метод оценки эффективности школ и отдельных классов.

В итоге хотелось бы сказать, что оценка качества образования и эффективности школ и классов – это процесс сложный, трудоемкий и требующий индивидуального, научного подхода. Основной проблемой является то, что в реальной жизни мы никогда не сравниваем «две одинаковые вещи». Использование результатов тестов для оценки качества образования не рекомендуется в большинстве случаев, особенно в общереспубликанском тестировании и тестах, проводимых Национальным центром тестирования. Инструменты оценки (метод добавленной стоимости), необходимые для объективной оценки качества образования, новы для нас, и их внедрение – это сложный и длительный процесс. Развитие подходов к измерению добавленной стоимости необходимо непрерывно исследовать и совершенствовать. Таким образом, на сегодняшний день результаты тестов могут быть использованы только для дальнейшего обучения и развития в партнерстве с учителями, а не для вынесения необоснованных суждений до тех пор, пока мы не разработаем свои методы.

Ссылки:

1. Sanders, W.L. & Horn, S.P. (1995a). Educational Assessment Reassessed: The Usefuleness of Standardized and Alternative Measrues of Student Achievement as Indicators for the Assessment of Educational Outcomes. *Educational Policy Analysis Archives*, 3 (6).
2. Sanders, W.L., & Horn, S.P. (1995b). The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment. In A.J. Shinklfield and D. Stufflebeam (Eds.) *Teacher Evaluation: Guide to Effective Practice* (pp. 337-350). Boston, Ma., USA: Kluwer.

3. Sanders, W.L., (2000). A Value-Added Assessment for Student Achievement Data: Opportunities and Hurdles. *Journal of Personnel Evaluation in Education.* 14:4, pp. 329-339.

4. Sanders. W.L., Saxton, A.M. & Horn, S.P. (1997) The Tennessee Value-Added Assessment System (TVAAS): A Quantitative, Outcome- based Approach to Educational Assessment. In Millman, J. (Ed.), *Grading Teachers, Grading Schools.* Thousand Oaks, CA, USA: Corwin Press.