АНАЛИТИЧЕСКИЕ ОБЗОРЫ

Е.В. Сивак
(*Москва*)

ИСТОРИЯ ОЦЕНОЧНЫХ ИССЛЕДОВАНИЙ В ОБРАЗОВАНИИ В США: АНАЛИТИЧЕСКИЙ ОБЗОР[1]

В обзоре прослеживаются основные этапы развития оценочных исследований в образовании в США: описывается процесс профессионализации этой области, отмечаются основные изменения в методологии (в том числе развитие методологии эксперимента и альтернативных методов), задачах оценивания, приводятся результаты наиболее значимых исследований.

Ключевые слова: оценочные исследования в образовании, эксперименты, квазиэксперименты.

Оценочными исследованиями называют особый тип прикладных социальных исследований, в которых методология социальных наук используется для изучения эффективности и других важных аспектов социальных программ (управления и результатов, дизайна программ, концептуализации социальной проблемы). Особенный интерес представляет история оценочных исследований в сфере образования в США: в этой стране начало оценочным исследованиям было положено в середине XIX в. Происходившие на протяжении всего этого времени изменения в представлении о том, что такое оценочное исследование, и в мето-

[1] **Елизавета Викторовна Сивак** – аспирант кафедры анализа социальных институтов НИУ ВШЭ, младший научный сотрудник международной научно-учебной лаборатории институционального анализа экономических реформ НИУ ВШЭ. E-mail: Elizaveta.sivak@gmail.com.

дологии оценивания дают возможность пронаблюдать и отметить многие части спектра разнообразных оценочных исследований, проводимых сегодня в разных странах.

Для того чтобы упорядочить изложение, можно выделить три наиболее крупных периода в истории оценочных исследований в образовании (разумеется, возможна и иная периодизация): 1) появление интереса к оцениванию в образовании, первые исследования, 2) развитие эмпирических исследований, 3) период профессионализации и дальнейшего развития. Каждый период характеризуется своим особым подходом к оцениванию, определением основных задач и ключевыми исследованиями[1].

Появление интереса к исследованиям образования (1850–1900-е годы)

Первая попытка оценивания

В XIX в., в период индустриальных революций и изменения структуры социальной сферы, в США и Великобритании начинаются первые попытки образовательных реформ и введения социальных программ. В середине XIX в. в США появляются первые журналы, посвященные вопросам образования[2]. В этот период этими вопросами занимались различные государственные комиссии – Президентская комиссия по школьному финансированию, правительственные комиссии, местные советы по образованию. Комиссии оценивали

[1] В обзоре отмечаются исследования, в которых изучаются образовательные учреждения разных уровней (дошкольного, школьного, университетского), отдельно каждый уровень образования не рассматривается, так как в этом нет необходимости для достижения целей данной работы – описания основных этапов развития оценочных исследований в образовании и их методологии.

[2] С 1855 по 1881 г. Г. Бернард (H. Barnard) издавал *American Journal of Education*, а с 1893 г. начал издаваться журнал *School Review*, который в 1979 г. поменял название на *American Journal of Education*.

различные образовательные нововведения, собирая данные от школ. Но эти оценочные исследования были скорее фикцией, так как использовались не для корректировки программ, а скорее для того, чтобы оправдать существовавшую государственную образовательную политику и сделать школу более подотчетной [1].

Первая попытка произвести оценивание в образовании в США была предпринята в 1845 г. в Бостоне. Совет по образованию Бостона заменил устные экзамены в школах на письменные. Устные экзамены, по мнению членов совета, стали неудобными в условиях роста числа учеников и, кроме этого, были несправедливыми, так как не позволяли стандартизировать процедуру экзамена. В такой замене была также и политическая подоплека: нужно было найти легко измеримый, формализованный показатель работы школы для того, чтобы можно было сравнивать школы и Совет по образованию мог сам назначать директоров и отстранять от работы несогласных с советом директоров[1] [2]. Введение письменных экзаменов позволило ранжировать учеников по оценкам, считать средние оценки, т.е. это был очередной шаг к тестированию[2]. Эта попытка измерения работы школ положила начало традиции использовать оценки школьников как важнейший источник данных при оценивании эффективности школы или образовательной программы.

Дж. Райс и первый эксперимент в оценочных исследованиях

Между 1887 и 1898 г. Дж. Райс (J. Rice) в США провел исследование, которое считается первым в оценивании образования [1]. Его

[1] Х. Манн (H. Mann), один из членов совета, в тот период хотел бороться с директорами, которые отказывались от предложенной Манном отмены телесных наказаний.

[2] Первый шаг – введение количественной оценки за экзамен в 1792 г. У. Феришем (W. Farish) (до этого использовались только качественные суждения) [2]. Это был важный шаг, потому что впервые на экзаменах начали оценивать не риторику и стиль, а технические компетенции.

целью было изучение методов преподавания и улучшение управления школами. Райс провел сравнительное исследование эффективности «зубрежки» в обучении правописанию в нескольких образовательных округах, используя результаты теста (одинакового в разных округах) как показатель результативности образовательной программы по улучшению правописания. Он обнаружил, что нет значимых различий в результатах теста между двумя системами обучения: той, где 200 мин. в неделю тратится на заучивание написания слов, и той, где этому уделяется не больше 10 мин. в неделю. Эти результаты привели в итоге к пересмотру пользы от заучивания в обучении правописанию и изменению учебных планов.

Райс также описал показатели, по которым можно судить о качестве работы школы: как выглядит классная комната, каково отношение учителя к ученикам, как проходит устный опрос, как устроена *busy-work* (задания, которые не дают никаких знаний и даются ученикам только в воспитательных целях – для того чтобы они были чем-то заняты), как учителя отвечают на некоторые общие педагогические вопросы, посещают педагогические собрания и что делают для собственного интеллектуального развития [3]. По факту Райс оценивал работу школы только по двум параметрам: насколько здесь развито «механистическое» обучение («зубрежка» четких и определенных фактов) и насколько представлены принципы научного, нового, с точки зрения Райса, образования – педагогики, основанной на психологических принципах обучения. Эти принципы заключались в том, что обучение не должно состоять в усвоении фактов в уже готовом виде или в групповой бесполезной работе в классе (*busy-work*); нужна такая педагогика, которая бы стимулировала интерес учеников к получению знаний и их понимание разных вопросов, а не заучивание фактов. Райс (которого называют предшественником «прогрессивного» образования) считал, что для учителей необходимо проводить специальные тренинги, чтобы обучать их преподаванию в соответствии с принципами «нового образования».

Несмотря на то что исследования Райса относят скорее к журналистским, чем к научным, из-за «разоблачительного» стиля его работ [4], его роль трудно недооценить – им была сделана первая попытка эксперимента в оценочном исследовании в образовании.

Развитие эмпирических исследований (1900 – конец 1950-х годов)

Стандартизованные тесты

Появление в начале XX в. научного менеджмента повлияло и на управление образованием: усилились требования к систематизации, стандартизации и эффективности школ. Акцент на эффективности виден, например, в пятнадцатом ежегодном сборнике Национального общества изучения образования (*National Society for the Study of Education* (NSSE)) под названием «Стандарты и тесты измерений эффективности школ и школьных систем» [5], где предлагались тесты на чтение и письмо, стандарты оценки учителей и администраторов школы. Началось распространение идеи подотчетности школ [2]. Во многих школах в этот период были проведены исследования эффективности учителей и школы с использованием разных критериев эффективности – бюджет школы, затраты на ученика, доли отчисленных, количество учеников на одного учителя и др.

Эффективность и качество обучения оценивались, в отличие от исследований Райса, по результатам учеников, а не только по «вкладу» учителей и школы в эти результаты (что измерялось такими параметрами, как бюджет школы и т.д.). Для этого применялись разработанные «объективные» тесты по арифметике, правописанию и др., чтобы определить качество преподавания. Например, в 1901 г. был впервые применен тест-предшественник SAT (*scholastic aptitude test*; тест на аналитические способности),

который используется в США до сих пор как экзамен при поступлении в колледжи.

Наибольшее распространение эти тесты получили после Первой мировой войны. До 1930 г. исследования эффективности и тестирования инициировались школьными округами, комитетами учителей и специальными бюро и департаментами школьного округа. Доля учеников, сдавших тест, служила критерием, по которому учителя могли судить, «дотягивают» ли их классы до среднего уровня по городу [6].

Рост распространенности стандартизированного тестирования продолжался до 1960-х годов [1]. Тесты, которые изначально разрабатывались и использовались в отдельных университетах, начали получать всё большую распространенность. Это, например, SAT, изначально использовавшийся в колледжах Северо-Восточной части США; тесты на базовые навыки *Iowa Test of Basic Skills* (ITBS) и *Iowa Test of Educational Development* (ITED), разработанные в Центре изучения измерений (*Measurement Research Center*) в Университете Айовы, которым руководил Э. Линдквист, изначально использовались только для оценки студентов Айовы, а затем распространились и на другие учебные заведения. С расширением использования тестов были созданы стандарты тестирования и общие технические рекомендации[1].

В этот период появилось представление о школе как о фабрике, располагающей определенными ресурсами (учителями того или иного уровня квалификации, материальными ресурсами и т.д.), и перерабатывающей их по определенным схемам (учебным планам и методикам) в основной продукт – обученных учеников. Это представление до сих пор превалирует в представлении о должной

[1] В 1954 г. Комитет Американской психологической ассоциации разработал Технические рекомендации для психологических тестов, а в 1955 г. Комитет Американской ассоциации образовательных исследований и Национального совета по измерениям в образовании подготовил Технические рекомендации тестирования результатов учеников.

работе школы (см. например, [7], где описываются возможные варианты переустройства школы по аналогии с японскими автомобильными заводами). Наиболее распространенный метод оценивания работы школ и учителей того времени – стандартизованные тесты – напрямую соответствовали такому представлению о школе.

Альтернативный подход к оцениванию школ: эксперименты

Однако в этот период быстрое развитие тестирования было не единственным усовершенствованием в методологии оценочных исследований. Развивался также и эксперимент, в том числе как метод оценочных исследований. Появились работы Э. Линдквиста об эксперименте как методе изучения образования [8]. Кроме этого, под руководством Р. Тайлера (R. Tyler) было проведено исследование, которое стало первым крупным экспериментом в образовании и единственной серьезной попыткой оценивания эффективности разных моделей образования вплоть до середины XX в.

Р. Тайлер ввел альтернативную концептуализацию оценивания (и закрепил в 1930-х годах сам термин «оценочные исследования в образовании» (*educational evaluation* или *evaluational research in education*)) – не как сравнение средних результатов тестирования с некими «объективными» пороговыми значениями, а как сравнение запланированных и фактических результатов. Это определение ближе к подходу Райса, изучавшего различия в результатах, которые дают разные образовательные методы.

По такой модели было построено и оценочное исследование Тайлера – *Восьмилетнее исследование* (*Eight Year Study*), известное также как *Исследование тридцати школ* (*Thirty-School Study*)[1]. Целью исследования было установление таких связей

[1] По факту исследование шло не восемь лет – с 1933 по 1941 г., а с 1930 по 1942 г., пока не прекратилось финансирование. Название «Восьмилетнее исследование» появилось потому, что изучался опыт обучении в старшей школе и переход в колледж и начало обучения в колледже, что занимало 8 лет [19].

между школой и колледжем, которые бы не мешали, а, наоборот, способствовали экспериментам и реконструкции средней школы, а также определение того, как средняя школа в США может лучше соответствовать потребностям учеников [9].

Вопрос об опыте обучения в школе и о том, какая школа лучше соответствует потребностям учеников, возник из обсуждений в начале 1930-х годов вопроса, насколько эффективно традиционное обучения в старших классах обычных общеобразовательных школ по сравнению с обучением в «прогрессивных» средних школах (*progressive secondary schools*) [1]. Основоположником движения за прогрессивное образование был Дж. Дьюи. Основной принцип прогрессивного образования – обучение должно основываться на опыте, решении практических задач[1]. Учитель-«прогрессивист» не только занимается с учениками зубрежкой и чтением, но также пытается обучать на их опыте. Первая школа-лаборатория, основанная на принципах прогрессивного образования, была создана в 1896 г. при Чикагском университете и такие школы-лаборатории сохранились здесь до наших дней, руководствуясь в своей работе принципом *learning by doing*[2].

В результате обсуждения эффективности традиционных и «прогрессивных» школ, ведущие колледжи начали отказывать выпускникам «прогрессивных» средних школ в приеме, потому что эти выпускники не изучали определенных курсов. «Прогрессивные» школы были нацелены на изменение школы, но не на снижение шансов учеников при поступлении в колледж, поэтому было инициировано *Восьмилетнее исследование*. В 1932 г. был предложен эксперимент, в ходе которого более 300 колледжей согласились

[1] Но не на любом опыте – опыт и обучение нельзя приравнять друг к другу, так как опыт может и препятствовать обучению, когда мешает приобретению нового или искажает имеющийся опыт [11].

[2] Сайт школ-лабораторий Университета Чикаго [on-line]. URL: http://www.ucls.uchicago.edu/.

отказаться от своих традиционных требований к поступающим для выпускников из 30 прогрессивных школ. Результаты обучения в *high school* и колледже сравнивались для выпускников этих прогрессивных школ и выпускников традиционных *secondary schools* [9]. Каждому выпускнику одной из 30 прогрессивных школ ставился в соответствие выпускник обычной школы, поступивший в тот же колледж. Соответствие отслеживалось по следующим критериям [10]: пол, возраст, семья (профессия родителей), результаты SAT (так как результаты этого теста не зависят от модели обучения в школе), интерес к учебным курсам, предполагаемая профессия.

Результаты исследования показали, что высокие достижения студентов в колледже не были связаны с преподаванием определенного набора предметов в старших классах школы [10]; поэтому требования, которые выдвигали колледжи к выпускникам прогрессивных школ, были признаны неадекватными. *Восьмилетнее исследование* продемонстрировало, что колледжи могут получать необходимую им информацию при приеме студентов, изучая результаты стандартизованных тестов, а не обязывая школы придерживаться определенных учебных планов [10]. Исследование также показало, что эксперимент со школьными планами не приводит к снижению конкурентоспособности выпускников прогрессивных школ при поступлении в колледж [10].

Период профессионализации (с начала 1960-х годов)

Основные черты профессионализации оценочных исследований

Историю оценочных исследований как отдельной профессиональной области обычно отсчитывают с начала 1960-х годов. Выбор этой точки отсчета объясняют тем, что в это время в США возросло число таких исследований после реализации масштабных государственных социальных программ (см., например: [12]).

Вероятно, это не совсем полное обоснование – 60-е годы можно считать периодом выделения оценочных исследований как самостоятельной области главным образом потому, что в этот период появились метаисследования с обсуждением методологии оценочных исследований, а также некоторые другие изменения.

1. Появляются профессиональные организации по оцениванию[1].

2. В крупных университетах возникают новые образовательные программы по оцениванию программ. Университеты начали предлагать курсы по методологии оценивания. Несколько университетов (Университет Иллинойса, Стэнфордский университет, Университет Калифорнии в Лос-Анджелесе, Университет Миннесоты и др.) разработали и продолжали совершенствовать специальные магистерские программы по оцениванию; Министерство образования спонсировало национальную программу по обучению оцениванию.

3. Начинают публиковаться специальные журналы, посвященные оценочным исследованиям в целом и в образовании в частности – *Evaluation Review, American Journal of Evaluation, Educational Evaluation and Policy Analysis, Studies in Educational Evaluation*. До этого профессиональных журналов и другой литературы по оцениванию не было, за исключением неопубликованных статей, которые циркулировали по неформальным сетям среди практиков.

---

[1] Например, Американская ассоциация оценивания (*American evaluation association*), отдел исследований и оценивания в школах (*Division of Research, Evaluation, and Assessment in Schools*) Американской ассоциации по исследованиям в образовании (*American Educational Research Association*, AERA), Общество по оценочным исследованиям (*Evaluation Research Society*) и др. В конце 1920–1930-х годах также существовали институты, которые занимались исследованиями в образовании (например, Педагогический колледж в Колумбийском университете, которым руководил Дж. Стрейер (G. Strayer)), но оценивание было в основном делом местных школьных округов и образовательных комитетов, а не крупных профессиональных организаций [1].

В конце 1970-х годов был издан классический учебник по оцениванию П. Росси, одного из учеников П. Лазарсфельда [13].

4. Расширяется число методов оценочных исследований и новых концептуализаций оценивания. После нескольких десятилетий преобладания стандартизованных тестов в оценивании развивается методология эксперимента; появляются такие методы оценивания, как кейс-стади [14], смешанные методы оценивания (т.е. использование и количественных, и качественных методов) и др. Помимо заложенного в тестировании представления об оценивании как сравнении результатов какого-либо теста в школе с неким «объективным», заданным заранее уровнем, начал активно развиваться подход к оцениванию как к сравнению разных образовательных программ, учебных планов и др. (т.е. происходило развитие взглядов Райса и Тайлера на оценивание), к изучению ожидаемых и непредусмотренных последствий социальной программы, а также ее целей и др.

5. Появляются стандарты качества работ в оценивании – создаются специальные организации, которые занимаются оценкой исследований (например, *National Study Committee on Evaluation*), разрабатываются стандарты для такой оценки [1]. Появляются метаисследования – работы, в которых анализируются достоинства и недостатки разных методов оценивания.

Рассмотрим подробнее два последних пункта – развитие методологии оценивания и метаисследований в этой области.

Пересмотр существовавших методов оценивания

В работе [1] выделяются две основных реформы 1950–60-х годов, которые стимулировали поиск новых методов в оценивании образования – Закон об образовании в целях национальной обороны, 1958 г. (*National Defense Education Act*), и Закон о начальном и среднем образовании, 1965 г. (*Elementary and Secondary Education Act, ESEA*).

Первый закон был выпущен после запуска СССР спутника, когда правительство США начало обращать особое внимание на состояние образования. Среди прочего вследствие этого появились новые образовательные программы и учебные планы по математике, иностранным языкам, а также национальные программы развития учебных планов, особенно в области естественных наук, расширились программы тестирования в школьных округах. Подход Тайлера использовался для определения целей новых учебных планов и оценки степени реализации этих целей. Были созданы новые стандартизованные тесты, которые лучше отражали содержание новых учебных планов. Кроме этого, для оценки новых учебных планов проводились эксперименты [1].

Л. Кронбах в своей работе [4] критически рассмотрел существовавшие методы оценивания эффективности разных учебных планов. Он отметил, что тестирование делает акцент на точности измерений, но на самом деле важна еще и валидность – не ясно, почему измерение фактологических знаний (с помощью тестов) важнее, чем измерение общих навыков и знаний. Линдквист и Тайлер работали над тестами для изучения общих навыков, но эти тесты распространения не получили. Помимо этого, посредством тестирования нельзя изучить отдельные аспекты работы учебных планов.

Критические аргументы Кронбаха в адрес экспериментов заключались в следующем: в условиях, когда группы плохо выровнены, результаты эксперимента уже заранее можно считать недействительными (Кронбах, естественно, описывал существовавшие на тот момент исследования; в дальнейшем была развита методология и рандомизированных, и нерандомизированных экспериментов). Кроме этого, не ясно, эффект какого воздействия изучается, так как сравниваются разные по множеству показателей учебные курсы, и как именно происходит воздействие.

По мнению Кронбаха, вместо экспериментов и тестов должны быть применены другие методы – тщательное изучение отдельных случаев; использование, помимо тестов, еще и интервью с уче-

никами, и эссе. Иначе нельзя получить информацию о том, что в учебном плане необходимо изменить, и понять происходящие образовательные процессы – а в этом, по мнению Кронбаха, и состоит основная задача оценивания.

Второй закон – о начальном и среднем образовании – был издан в рамках программы «война с бедностью», которая предусматривала реформы, направленные на выравнивание и повышение возможностей для всех граждан – в медицине, других социальных и образовательных услугах[1]. В законе об образовании подчеркивалась необходимость обеспечить всем гражданам, в том числе детям из неблагополучных семей, равный доступ к образованию, а школам стать более подотчетными; улучшить академическую успеваемость детей из неблагополучных семей. Первая глава этого закона требовала от каждой школы, которая получает финансирование по программе ESEA, проводить ежегодную оценку учащихся с использованием стандартизированных тестов. Это требование (оценка результатов на пути достижения целей и стандартизированный тест) – отражает взгляд на оценочные исследования, который существовал на тот момент.

Для достижения целей закона об образовании школьным округам и местным образовательным комитетам выделялись денежные средства для расширения программ, адаптированных для детей с ограниченным доступом к образованию, найма специального персонала для обучения детей по специальным компенсаторным программам (например, для улучшения навыков чтения) и покупки специальных учебных материалов, строительства дополнительных зданий и улучшения здоровья детей.

Школьные округа сразу обнаружили, что существующие инструменты и стратегии не подходят для решения поставленных законом задач. Существовавшие стандартизованные тесты были

[1] В 2001 г. в этот закон были внесены поправки, и он получил название *No Child Left Behind*.

предназначены для того, чтобы построить рейтинг учащихся по определенному показателю; эти тесты не годились для определения нужд и оценки достижений детей с ограниченным доступом к образованию, отстававших от программы школьного обучения. Более того, эти тесты были нечувствительны к разнице между школами и программами [1]. Была и еще одна проблема: для обеспечения коммерческой выгоды от проведения тестирования содержание стандартизированных тестов должно было отражать нужды большинства школьных округов и игнорировать потребности отдельных округов. И наконец, учителям не хватало информации о нуждах детей с ограниченным доступом к образованию, поэтому учителя не могли корректировать задачи, поставленные разработчиками программы [1].

Оценка результатов этой программы показала ее неэффективность – компенсаторные программы по чтению не оказывали положительного воздействия на детей (сравнивались участвовавшие в программе дети с какими-либо ограничениями в доступе к образованию с детьми без таких ограничений, не участвовавшими в программе; однако сами авторы исследования отмечают, что это сравнение оказалось неудачным, так как эти группы не были выровнены по другим параметрам). Данные для оценки собирались с помощью опросов учителей, директоров школ и школьных округов; также использовалась предоставленная школами статистика – данные об учениках.

Выявленные недостатки в существовавших методах оценивания привели к развитию новых методов – кейс-стади, смешанных методов (количественных и качественных), а также к развитию методологии эксперимента.

Развитие методологии эксперимента и основные исследования с использованием экспериментальных планов

В 1960-е годы появляются очень важные работы Кэмпбелла [15], а также Кэмпбелла и Стэнли [16], посвященные методологии экс-

периментальных исследований в психологии и педагогике. Кэмпбелл отмечает специфику эксперимента[1] в социальных науках – экспериментатор не может полностью владеть ситуацией (организовать экспериментальное воздействие, ограничить действие других переменных; не всегда можно провести случайный отбор в контрольную и экспериментальную группы[2], предварительное тестирование и т.д.), так как эксперименты в этих науках полевые, а не лабораторные. Следовательно, появляются дополнительные угрозы валидности вывода, основанного на эксперименте. Важность этих работ состоит в том, что в них показано, что, несмотря на эти трудности, социальным ученым не стоит отказываться от этого метода. Кэмпбелл и Стенли описали способы обеспечения контроля в случаях, когда невозможна рандомизация, и для обозначения этих методов ввели термин «квазиэксперимент» – это эмпирические исследования, которые, как и подлинный эксперимент, направлены на оценку причинно-следственной связи, но в этих исследованиях не полностью контролируется порядок экспериментального воздействия [12, с. 107] (т.е. эксперимент без рандомизации).

В эти же годы в исследованиях образования появились первые полевые эксперименты со случайным распределением на контрольную и экспериментальную группы (*randomized field trials*, RFT), а также квазиэксперименты [1]. Крупный квазиэксперимент в образовании 1960-х годов – оценка эффекта проекта *Head Start*

[1] Эксперимент в определении Кэмпбелла – это та часть исследования, которая «заключается в том, что исследователь осуществляет манипулирование переменными и наблюдает эффекты, производимые этим воздействием на другие переменные». Эксперимент – средство проверки каузальных гипотез [12, с. 39].

[2] Случайное распределение по группам означает отсутствие систематических различий между группами, которые бы препятствовали выводу о том, что именно экспериментальное воздействие обусловило различия между группами; нет никаких скрытых факторов, которые бы повлияли на различия между группами, кроме как воздействие в ходе эксперимента.

(«скачок на старте») – программы Министерства здравоохранения и социального обеспечения США, направленной на предоставление образования и других социальных услуг детям из малообеспеченных семей. В 1968–1969 гг. оценка этой программы проводилась *Westinghouse learning corporation* (консультационный центр, который занимался тестированием в образовании). Оценивались результаты летних и круглогодичных программ проекта *Head Start*, направленных на улучшение здоровья, физической формы, развитие таких качеств, как уверенность в себе, самодисциплина, любознательность и др. Основной вопрос исследования состоял в том, каково воздействие программы *Head Start* на интеллектуальное и психологическое развитие детей дошкольного возраста, и сохраняется ли достигнутый эффект до первых классов школы.

Для ответа на этот вопрос сравнивались группы детей, попавших под действие программы, и не участвовавших в ней. Результаты показали, что летняя программа не оказывала сильного воздействия на развитие дошкольников, и эффект не сохранялся до начала школы; круглогодичная программа давала незначительный эффект [17]. Были выявлены методологические проблемы: контрольная группа была плохо подобрана (дети в этой группе оказались не из таких же необеспеченных семей, поэтому незначительные различия между контрольной и экспериментальной группами могли быть обусловлены просто тем, что у членов контрольной группы изначально был более высокий «старт», поэтому эффекты программы, вероятно, были систематически недооценены). Во-вторых, в дизайне исследования отмечались проблемы внутренней валидности – разные параметры окружения детей в семье и в школе не были проконтролированы, т.е. эффект программы нельзя отделить от влияния других факторов [18]. Наконец, в измерении развития детей использовались стандартизованные тесты, тогда как дети, участвовавшие в программе, принадлежали к очень разным социокультурным группам [17].

Другой эксперимент 1960-х годов – уже со случайным распределением на группы – изучение эффекта программы *HighScope Perry Preschool* («высокая планка»). Метод обучения *HighScope,* так же как и в «прогрессивных» школах-лабораториях, основывался на принципе *learning by doing*. Проект *Perry Preschool* был предназначен для детей из афроамериканских семей с низким доходом, которые показывали невысокий уровень IQ и с высокой вероятностью могли после поступления в школу учиться с низкой успеваемостью. В рамках проекта дети были случайным образом поделены на контрольную и экспериментальную группы. Детей в последней группе обучали по методу *HighScope*. В программе участвовали дошкольники 3-5 лет.

Эффект программы замеряли каждый год, пока детям было 4–11 лет, а затем, когда им исполнилось 14, 15, 19 и 27 лет [19]. Среди детей из экспериментальной группы в последующие после эксперимента годы уровень распространенности преступлений и подростковой беременности оказался более низким, чем в контрольной группе; улучшалась их успеваемость, они стали более успешны в трудоустройстве, получали более высокий доход, а также достигали других социальных и академических успехов [19].

Другие эксперименты 1960-х годов были посвящены изучению эффекта группировки школьников по способностям. Было всего 27 таких исследований, обзор которых дается в статье [20]. Среди этих экспериментов в двух применялось случайное деление на контрольную и экспериментальную группы, в пяти – выравнивание контрольной и экспериментальной групп, в девяти – анализ корреляций между оценками учеников перед выделением групп по способностям и после. В этих исследованиях не было обнаружено никакого эффекта от деления студентов на группы в зависимости от способностей.

В начале 1970-х годов были осуществлены несколько крупных исследований с рандомизацией. Один из них – эксперимент по изучению обучающих эффектов после первого года существования

передачи «Улица Сезам» С. Болла и Дж. Богатца (S. Ball, G. Bogatz). Предполагалось сравнить знания по нескольким темам, которые рассматривались в сериях «Улицы Сезам» первого сезона (знание алфавита, цифр, частей тела, геометрических фигур, навыки классификации и др.), у двух сформированных случайным образом групп дошкольников – тех, кого просили смотреть программу и тех, кого не просили.

Но оказалось, что дети в контрольной группе смотрели не меньше серий, чем дети в экспериментальной группе, что не позволило изучить эффект передачи сравнением двух групп (на следующий год в контрольную группу включили тех детей, в чьих семьях не было возможности смотреть передачу). В итоге использовался другой способ анализа данных: сравнивались 4 группы детей в зависимости от частоты просмотра (количества просмотренных серий в неделю). Дети, которые смотрели передачу дольше, в итоговом тестировании чаще умели распознавать геометрические фигуры, части тела и т.д. Тест также показал, что те темы, которым уделялось больше экранного времени, были усвоены детьми лучше, чем темы, которые получали меньше экранного времени. Обучающий эффект передачи не зависел от того, смотрели ли дети передачу в домашней обстановке или в классе. Эти результаты повторились и во втором исследовании [21].

В экспериментальных исследованиях 1970-х годов был распространен вопрос о влиянии численности учеников в классе на эффективность обучения. В статье [22] приводится метаанализ 78 исследований, в которых изучался эффект размера класса. При этом использовались различные методы – случайное деление на группы, выравнивание групп, панельное исследование одной группы (одни и те же ученики сначала обучались вместе в одном классе, затем их делили на небольшие группы), неконтролируемое деление на группы. Варьировались и изучаемые предметы, и количество учебных часов, которые учителя проводили с детьми в рамках исследования, и число детей в «больших» и «маленьких»

классах и др. Общий результат таков: обучение более успешно в «маленьких» классах. Масштаб эффекта «маленького класса» различался от исследования к исследованию.

Метаисследования

Метаисследования, или дискуссия о преимуществах разных представлений об оценочном исследовании, о разных моделях объекта изучения начались в 1960-х годах с упомянутой ранее критики Кронбахом наиболее распространенных методов и господствовавшего определения оценивания как сравнения результатов разных программ. После этого обсуждались и другие вопросы.

В частности, было проблематизировано представление о школе как о «черном ящике» или о «фабрике», которое лежит в основе большинства оценочных исследований. Такой взгляд на школу, вероятно, остается неизменным со времен первых тестирований и первых исследований школ. Это представление получило распространение в 1960–1970-х годах, когда приступили к исследованию производственных функций в образовании. После доклада Дж.С. Коулмана [23], где говорилось об отсутствии связи между ресурсами школы и успехами учеников, начали изучать связь между конкретными характеристиками школ и результатами учеников или выпускников школ. Как отмечают, например, Р. Уэйсс и М. Рейн [24], такая модель школы предполагает изучение воздействия программы или реформы на какие-то показатели успеваемости учеников (оценки за школьные экзамены, результаты независимых тестирований и т.д.), при этом из внимания упускаются процессы, которые приводят к наблюдаемым результатам.

Эксперимент также подвергается критике из-за того, что с его помощью можно измерить лишь эффект программы, но не сами изменения, происходящие в ходе воздействия на экспериментальную группу и вызывающие определенный результат. В работе [25] автор рассматривает ограничения экспериментального подхода и

взгляда на школу как на «черный ящик» и предлагает альтернативы, основанные на этнографии и феноменологии.

Однако оценке с помощью экспериментов могут быть подвергнуты не только результаты какой-либо социальной программы. П. Росси выделяет несколько типов оценочных исследований, в том числе и оценивание процессов работы какой-то программы, а не только исходного плана реализации программы, ее итоговых результатов, эффективности и соотношения выгод и издержек [13]. Д. Кэмпбелл отмечает, что в изучении процессов, приводящих к тому или иному результату реформы или социальной программы, также могут применяться экспериментальные планы [26].

На текущем этапе развития оценочных исследований эксперименты и квазиэксперименты (с предшествующим или последующим выравниванием групп, например, с помощью регрессионного анализа либо с применением специальных методов «условной» рандомизации, скажем, экспериментального плана с «разрывом регрессии» (*regression discontinuity*)[1]) считаются основным методом для установления каузальных отношений.

[1] Этот квазиэкспериментальный план, разработанный в 1960 г. Д. Тислуайтом и Д. Кэмпбеллом [27] применяется в ситуации, когда случайный отбор невозможен. Решение о том, включать ли единицу исследования в экспериментальную или контрольную группы, принимается в зависимости от того, как значение определенной переменной-критерия (зависимой переменной) для этой единицы соотносится с установленным пороговым значением (например, экспериментальному воздействию подвергаются только школы, где средняя успеваемость учеников ниже определенной величины [28]). Проводится предварительный замер, выделяются две группы, затем экспериментальная группа подвергается воздействию, и проводится еще один замер (пост-тест). После этого строится регрессионная модель, описывающая поведение изучаемой переменной в экспериментальной группе, а также регрессионная модель для предсказания изменений рассматриваемой зависимой переменной в экспериментальной группе в случае, если бы никакого воздействия не было. Величина «разрыва» (отсюда название метода) между двумя регрессионными кривыми в точке, выбранной в качестве порогового значения, и принимается за размер эффекта от экспериментального воздействия. Если при проведении рандомизированного эксперимента

Эксперименты и квазиэксперименты используются, в частности, при изучении эффектов совместного обучения (*peer-effects*) [29], влияния специальных программ для детей из неблагополучных семей на снижение уровня отчислений из школ [30] и др.

Существенно улучшить качество информации, получаемой в оценочных исследованиях, позволило появление в 1970-х годах статистических техник интегрирования результатов экспериментов и квазиэкспериментов, проведенных для изучения сходных взаимосвязей, или метаанализа (термин введен Дж. Глассом[1]). Единицами изучения в метаанализе являются отдельные исследования; различные характеристики исследований кодируются, а затем полученная база данных изучается различными статистическими методами. С помощью метаанализа можно получить более валидную оценку влияния какой-либо переменной (величины эффекта), так как агрегирование результатов множества исследований, изучающих один и тот же вопрос, позволяет скорректировать разного рода ошибки, из-за которых возникают различия в полученных результатах (ошибки выборки, ошибки измерения и др. [31]).

Подводя итог, можно сказать, что за время, прошедшее с середины XIX в., в США произошли значительные сдвиги в представлении об оценивании и его основных методах. Изначально изменения в оценочных исследованиях были практически неот-

---

предполагается, что контрольная и экспериментальная группы эквивалентны до введения воздействия, и поэтому разницу можно приписать воздействию, то в случае дизайна «разрыв регрессии» эквивалентность групп не важна; валидность вывода об экспериментальном эффекте зависит от качества регрессионных моделей. Кроме того, при выборе порогового значения необходимо проверять, не может ли возникнуть какой-либо третьей переменной, которая будет по-разному влиять на группы «по обе стороны» этого порогового значения и тем самым создавать «ложный разрыв» [32].

[1] См., например: [13; 22].

делимы от развития показателей в образовании (появления количественных оценок, письменных экзаменов, стандартизованных тестов), а оценивание в образовании, как внешняя по отношению к школе процедура, почти ничем не отличалось от оценивания внутри самих школ – для того чтобы изучить работу школы или учителей, школьные оценки учеников лишь обобщались и сравнивались с заранее определенными критериями. С развитием и профессионализацией этой области представление об оценивании как о сравнении результатов работы (учеников, учителей, школ, учебных планов и т.д.) с некими «объективными» значениями было практически замещено концептуализацией оценивания как сопоставления предполагаемых и фактических результатов различных программ в образовании и изучения воздействия программ. Выделились и специфические процедуры оценивания, не совпадающие с внутришкольным оцениванием учеников. Помимо стандартизованных тестов начали активно использоваться эксперименты (с развитым математическим аппаратом для преодоления смещений при неслучайном распределении на группы, разнообразными экспериментальными планами для повышения надежности и валидности измерений, а также процедуры метаанализа).

ЛИТЕРАТУРА

1. *Maddaus G.F., Stufflebeam D.L., Kellaghan T*. Program Evaluation: a Historical Overview // Evaluation Models: Viewpoints on Educational and Human Services Evaluation / Ed. by D.L. Stufflebeam, G.F. Madaus, T. Kellaghan. Boston; Dordrecht; L.: Kluwer academic publishers, 2002.

2. *Maddaus G.F., O'Dwyer L.M*. A Short History of Performance Assessment: Lessons Learned // Phi Delta Kappan. 1999. Vol. 80. No. 9.

3. *Graham P.A*. Joseph Mayer Rice as a Founder of the Progressive Education Movement // Journal of Educational Measurement. 1966. Vol. 3. No. 2.

4. *Cronbach L. J.* Course Improvement through Evaluation // Evaluation Models: Viewpoints on Educational and Human Services Evaluation / Ed. by G.F. Madaus, D.L. Stufflebeam, T. Kellaghan. N.Y.: Kluwer Academic Publishers, 2002.

5. Standards and Tests for the Measurement of the Efficiency of Schools and School Systems. Part I: National Society for the Study of Education Fifteenth Yearbook. Chicago: Univ. of Chicago Press, 1916.

6. *Ballou F. A.* Work of the Department of Educational Investigation and Measurement, Boston, Massachusetts // Standards and Tests for the Measurement of the Efficiency of Schools and School Systems. Part I: National Society for the Study of Education Fifteenth Yearbook. Chicago: Univ. of Chicago Press, 1916.

7. *Coleman J.* Output-driven Schools: Principles of Design // Redesigning American Education / Ed. by J.S. Coleman, B. Schneider, S. Plank, K.S. Schiller, R. Shouse, H. Wang, S.-A. Lee. Boulder: Westview Press, 1997.

8. *Lindquist E. F*. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton-Mifflin, 1953.

9. *Kridel C. A., Bullough R. V.* Stories of the Eight-year Study: Reexamining Secondary Education in America. Albany: State Univ. of New York Press, 2007.

10. *Aikin W.* The Story of the Eight-year Study. N.Y.: Harper, 1942 [on-line]. URL: http://www.8yearstudy.org/.

11. *Dewey J.* Experience and Education. N.Y.: Touchstone, 1938.

12. *Кэмпбелл Д.* Модели экспериментов в социальной психологии и прикладных исследованиях. М: Прогресс, 1980.

13. *Rossi P.H., Freeman H.E., Wright S.R*. Evaluation: A Systematic Approach. Beverly Hills: Sage Publications, 1999.

14. *Stake R.E*. The Case Study Method in Social Inquiry // Educational Researcher. 1978. Vol. 7. No. 2.

15. *Campbell D.T.* Reforms as Experiments // American psychologist. 1969. No. 24.

16. *Campbell D.T., Stanley J.* Experimental and Quasi-experimental Designs for Research. Boston: Houghton Mifflin Company, 1963.

17. *Grimmett S., Garrett A. M*. A Review of Evaluations of Project Head Start // The Journal of Negro Education, 1989. Vol. 58. No. 1.

18. *McGroder S.M*. Head Start: What Do We Know About What Works? // Office of the Assistant Secretary for Planning and Evaluation, report, 1990.

19. *Parks G*. The High/Scope Perry Preschool Project // U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention. Juvenile Justice Bulletin, 2000.

20. *Slavin R. E*. Ability Grouping in the Middle Grades: Achievement Effects and Alternatives // Elementary School Journal. 1993. Vol. 93. No. 5.

21. *Palmer E.L., Fisch S.M*. The Beginning of Sesame Street Research // «G» is for Growing: Thirty Years of Research on Children and Sesame Street / Ed. by S.M. Fisch, R.T. Truglio. Mahwah, NJ: Lawrence Erlbaum Associates, 2001.

22. *Glass G., Smith M.N.* Meta-analysis of Research on Class Size Achievement // Educational Evaluation and Policy Analysis. 1979. No. 1.

23. *Coleman J., Campbell E., Hobson C., McPartland J., Mood A., Weinfeld F., York R.* Equality of Educational Opportunity. Washington D.C.: U.S. Government Printing Office, 1966.

24. *Weiss R.S., Rein M.* The Evaluation of Broad-aim Programs: A Cautionary Case and a Moral // Annals of the American Academy of Political and Social Science, 1969. Vol. 385.

25. *Guba E*. Toward a Methodology of Naturalistic Inquiry in Educational Evaluation. L.A.: Center for the Study of Evaluation, 1978.

26. *Campbell D.T.* Considering the Case Against Experimental Evaluations of Social Innovations // Administrative Science Quarterly. 1970. Vol. 15. No. 1.

27. *Thistlewaite D., Campbell D.* Regression-discontinuity Analysis: An Alternative to the Ex Post Facto Experiment // Journal of Educational Psychology. 1960. Vol. 51. P. 309–317.

28. *Lavy V.* Performance Pay and Teachers' Effort, Productivity, and Grading Ethics // American Economic Review. 2009. Vol. 99. No. 5.

29. *Angrist J. D., Lang K.* How Important Are Classroom Peer Effects? Evidence from Boston's METCO Program // NBER Working Paper No. 9263, 2002.

30. *Dynarski M., Gleason P.* How Can We Help? What Have We Learned from Evaluations of Federal Dropout-Prevention Program // Mathematica Policy Research, 1999 [on-line] URL: http://www.mathematica-mpr.com/PDFs/Howhelp.pdf.

31. *Hunter J.E., Schmidt F.L.* Methods of Meta-analysis. Correcting Error and Bias in Research Findings. Thousand Oaks; L.; New Delhi: Sage Publications, 2004.

32. *Trochim W. M. K.* Research Methods Knowledge Database: The Regression-discontinuity Design [on-line]. URL: http://www.socialresearchmethods.net/kb/quasird.php.

33. *Glass G*. Integrating Findings: the Meta-analysis of Research // Review of Research in Education. 1977. Vol. 5.