**Октябрь 2018**

# Основы мониторинга и оценки для доноров и исполнителей

(A primer about monitoring and evaluation, for funders and implementers)

[Выдержка]

*Кэролайн Файнс (Caroline Fiennes)**
*Наталия Киритопулу (Natalia Kiryttopoulou)*
*Марк Максмайстер (Marc Maxmeister)*

*Ответственный автор*: caroline.fiennes@giving-evidence.com

keystone
accountability for social change

GIVING EVIDENCE
Enabling **giving** based on sound **evidence**

*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*

# Введение

Данный документ подготовлен по заказу клиента - финансирующей организации, для которой мониторинг (monitoring), оценка (evaluation) и обучение (learning) являются сравнительно новыми понятиями. Мы публикуем эту работу, потому что считаем (и надеемся), что этот материал будет полезен широкой аудитории фондов/доноров и исполнителей программ.

Документ объясняет, что такое мониторинг и оценка, и чем они отличаются друг от друга. Мы структурировали и представили свои идеи в виде **четырехуровневой рамочной концепции**, которая сортирует и группирует вопросы, имеющие отношение к мониторингу и оценке [Обратите внимание, что (как это видно из дальнейших рассуждений) мониторинг и оценка – это две разные вещи, несмотря на то что они зачастую тесно связаны между собой]:

- *Уровень 1*: параметры гранта: **ресурсный вклад (выраженный, например, в виде суммы гранта)** и мероприятия, выполняемые грантополучателем
- *Уровень 2*: отслеживание **изменений в зоне действий получателя гранта** – например, увеличение количества рабочих мест, изменение доходов у партнерской организации-грантополучателя, количество проведенных мастер-классов
- *Уровень 3*: **оценка влияния грантополучателя**: т.е. выяснение, какие из произошедших изменений являются следствием (возникли в результате) активности партнерской организации-грантополучателя
- *Уровень 4*: **оценка влияния фонда/донора**: т.е. выяснение, какие из произошедших изменений являются следствием (возникли в результате) активности донора

Мы представляем эти четыре уровня в виде лестницы, руководствуясь тем, что вопросы, которые относятся к Уровню1, намного проще, чем на Уровне 2 и т.д. – как с точки зрения видов данных и необходимых методов анализа, так и с позиции их понятийной сложности.


*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*

Для ясности уточним:

1. Если вы можете (или в состоянии) собрать данные по необходимым показателям, будучи просто осведомленными о том, что происходит ***внутри организации, получившей грант*** (например, количество полученных займов, доходы грантополучателя, его управленческие компетенции), то это является признаком Уровня 1.

2. Если показатели требуют знаний об изменениях, происходящих **в зоне действий грантополучателя** и за рамками организации (например, количество рабочих мест, не предполагающих оплату труда за счет грантополучателя; количество построенных домов; поправки, внесенные в законы), то вы находитесь на Уровне 2.

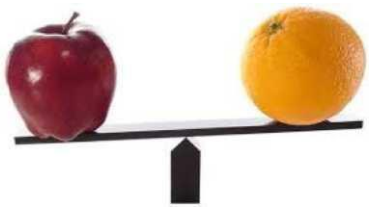
3. Надлежащая работа на Уровнях 3 и 4 неизбежно требует инструментов для сопоставления данных и сравнительного анализа.

Уровни 1 и 2 – это мониторинг, Уровни 3 и 4 – оценка.

Многие доноры/фонды обладают информацией для Уровней 1 и 2, очень немногие располагают убедительными данными для Уровня 3, и практически ничего не могут предложить для Уровня 4.

*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*

# Сравнение должно быть объективным



Представим себе эксперимент, в ходе которого мы выбираем самые крупные семена бобов и включаем их в группу, получающую полив. Впоследствии они всходят лучше, чем более мелкие семена, не получающие воды. Соответственно, мы не можем установить, каким образом на показатель роста влияет (i) размер семян или (ii) режим полива. Иными словами, мы не сможем увидеть разницу между:

i. «эффектом заботы» ('treatment effect'), т.е. режимом полива, и/или

ii. «эффектом селекции» (‘selection effect'), т.е., некоторыми характеристиками выборки (в данном случае это размер), которые определяют включение в группу для «заботы» (обработки в определенном режиме).

В процессе отбора людей для оценки каких-либо программ или для получения информации, всегда существует опасность «эффекта селекции». Ниже приводим несколько иллюстраций.
Fitted For Work (FFW – «Готова к работе») – это австралийская благотворительная организация, которая помогает женщинам устроиться на работу. FFW демонстрирует свою эффективность, сообщая о том, что «75% женщин, получивших от организации помощь по подбору гардероба и прохождению собеседований, находят рабочие места в течение трех месяцев… в сравнении с данными о том, что… среди женщин, которые полагаются на услуги австралийских федеральных служб занятости, в течение трех месяцев трудоустраиваются только 48%».

Данное сравнение некорректно, и ничего не говорит об эффекте, произведенном работой FFW. Потому что женщины, получающие поддержку от FFW, отличаются от тех, кто за этой помощью не обратился, (по меньшей мере) по двум критериям:

- Во-первых, они находят информацию о FFW и решаются обратиться за помощью. Вполне возможно, что женщины, действующие подобным образом, активнее взаимодействуют с окружением и более мотивированы к переменам по сравнению с теми, кто этого не делает. Так может проявляться «эффект селекции» среди женщин, получающих услуги FFW.

- Во-вторых, безусловно, женщины, которые приходят в FFW, получают от FFW определенную поддержку. А это уже «эффект заботы».

Приведенный пример сопоставления данных не разъясняет, какая часть изменения возникла вследствие «эффекта селекции», и что именно произошло благодаря «эффекту заботы», т.е. в результате поддержки от FFW.

Все это не означает, что программы FFW «не работают». Скорее, приведенные данные вообще не говорят о том, работают программы или нет.

Это не просто теория. По итогам сравнительного анализа общедоступных данных микрозаймы для жителей малообеспеченных сел на северо-востоке Таиланда тоже производили впечатление положительного эффекта. Но в этой аналитике не была учтена погрешность, возникшая при формировании выборки людей, получивших займы. Тщательное исследование - которое скорректировало «погрешность выборки» (selection bias) и прояснило, как чувствовали бы себя эти люди без каких-либо вмешательств – показало, что эффект займов незначителен. Они никак не повлияли на размеры семейных сбережений, затрат на образование либо на объем времени, которое люди посвящали трудовой деятельности. Среди обратившихся изначально были наиболее мотивированные сельские жители, которые хотели получить займы, получили их, и в любых ситуациях «переиграли» бы своих односельчан. В данном случае «эффект селекции» замаскировал тот факт, что «эффект заботы» (в виде займов), по сути, не состоялся.

Бывают ситуации, когда «эффект селекции» настолько силен, что заслоняет собой тот факт, что «забота» на самом деле приносит вред. Люди от лечения умирают. Бен Голдакр (Ben Goldacre), британский врач и автор научных публикаций, приводит такой пример:

> *«Мы привыкли считать, что заместительная гормональная терапия (ЗГТ) почти вдвое снижает риск инфарктов - в частности, потому, что такой вывод был сделан по итогам небольшого тестирования и крупного наблюдательного исследования. Эти изыскания имели*

*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*

*некоторые ограничения. В рамках маленького тестирования рассматривались только «косвенные исходы» (surrogate outcomes) – кардиомаркеры, имеющие отношение к сердечно-сосудистым заболеваниям, но не к фактической ситуации с инфарктами. А наблюдательное исследование столкнулось с тем, что женщины, которым доктора назначили ЗГТ, на момент начала терапии были более здоровы. Однако в тот момент исследователи озвучили наши лучшие ожидания, и это то, что нередко срабатывает лучше всего.*

*Когда для изучения сердечно-сосудистых заболеваний было проведено масштабное рандомизированное исследование, то оказалось, что ЗГТ увеличивает риск инфарктов на 29%».*

# Рандомизация как способ получения объективных сравнений

Американский инвестиционный банк Goldman Sachs реализует программу «10,000 Women» («10 000 женщин»), которая поддерживает женщин-предпринимателей. На первый взгляд, «70% опрошенных выпускниц программы увеличили доходы, и 50% создали новые рабочие места». Банк Goldman Sachs, безусловно, весьма впечатлен этими цифрами, потому что они публикуются на разворотах рекламных изданий. Но стоит ли нам поддаваться подобным эмоциям?

Сейчас вы увидите всю безнадежность ситуации. С одной стороны, это не те данные, которые говорят о том, что было «до» и «после»: это только то, что стало «после». Речь не идет о том, что «до программы увеличение доходов наблюдалось у 30% женщин, а теперь их доля возросла до 70%».

С другой стороны, информация о способах контроля и сравнения тоже отсутствует. Нам не говорят о том, что «70% выпускниц программы увеличили доходы, в то время как среди представителей других видов бизнеса аналогичный показатель (за тот же период времени) составил всего 20%».
Есть еще и третья большая проблема, которая тоже достаточно распространена. Давайте на мгновение представим, чего могли бы добиться упомянутые женщины в обычных условиях.

Нетрудно догадаться, что женщины особого типа, для которых участие в программе Goldman Sachs - подходящее занятие - это очень целеустремленные люди, добивающиеся успеха практически при любых обстоятельствах. То есть, данная программа может привлекать и отбирать необычайно предприимчивых людей. Это значит, что результаты, которые широко публикует Goldman, могут просто являться следствием «**эффекта селекции**».

Чтобы «установить причинно-следственные связи», т.е. разглядеть эффект программы, исследователю потребовалось бы сформировать две группы женщин-предпринимателей, которые в полном смысле идентичны. Одну группу следовало бы внедрить в программу «10,000 Women» и посмотреть, насколько улучшится их карьера по сравнению с теми, кто не участвует (эта вторая «контрольная» группа предназначена для того, чтобы показать, чего достигли бы эти женщины в любом случае).

Весьма досадно, что мы не можем создавать группы подобным образом, потому что люди не приходят подходящими парами: все они имеют странности, особый опыт, мнения и черты характера, которые могут повлиять на их эффективность (указывая также на возможное воздействие со стороны других инициатив). Тем не менее, если исследователь возьмет достаточно большую группу женщин, в которой все отвечают условиям программы, и разделит ее произвольным (или «случайным») образом (‘randomly’), то появится основание ожидать, что все эти странности и индивидуальные особенности равномерно распределятся между этими двумя группами. «Случайность» устраняет «погрешность выборки», создавая условия, при которых участие в программе становится единственным отличием между группами. Она изолирует эффект программы, и благодаря этому сравнение результатов двух групп позволяет демонстрировать эффект программы. Вуаля.

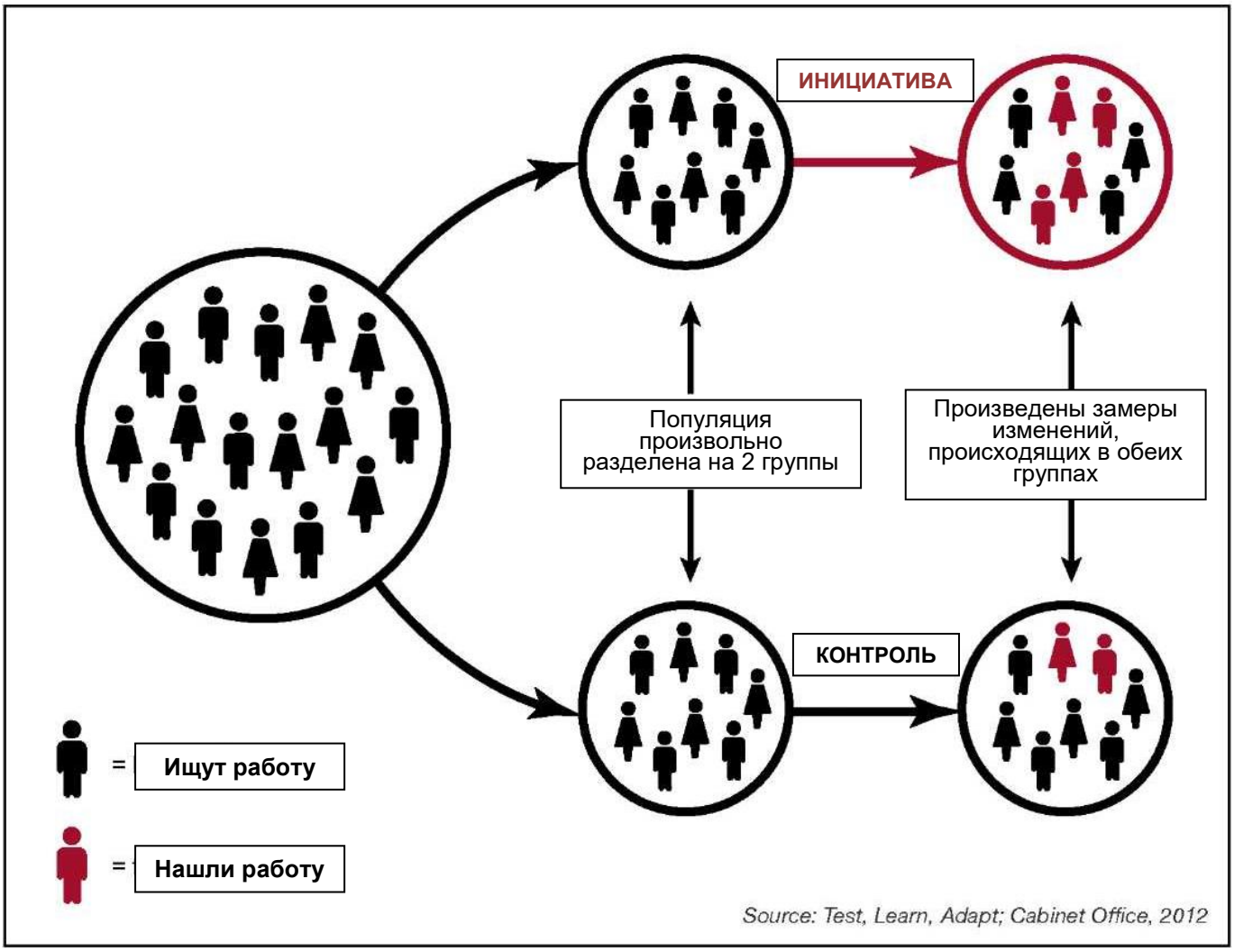
Эксперимент, о котором мы только что рассказали – это **рандимизированное контролируемое исследование** (РКИ/RCT - randomized controlled trial). Если РКИ проводится правильно, то оно

*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*

изолирует программу от всех других возможных инициатив и тем самым показывает, что могло бы произойти при условии отсутствия программы. Поэтому РКИ часто называют «золотым стандартом» оценки (предназначенным для конкретных случаев изучения эффектов), который изначально был разработан для тестирования медицинских препаратов, а теперь используется для получения надежных и объективных выводов повсеместно.

Например, РКИ (проведенные исследовательской организацией Innovations for Poverty Action/«Инновации для борьбы с бедностью» при поддержке J-PAL - Лаборатории по изучению проблем бедности им. Абдула Латифа Джамиля, США) доказали полезность раздачи чечевицы в качестве поощрения для родителей за доставку детей в пункты вакцинации в аграрных регионах Индии, и показали условную экономическую эффективность различных программ очистки воды в Кении. IPA и J-PAL также используют РКИ для тестирования программ улучшения сексуального здоровья, снижения коррупции, и даже в ходе восстановления мирной жизни после конфликтов. Они применяют надежный научный метод (являющийся великим интеллектуальным достижением нашей эпохи) для решения самых острых социальных проблем современности. РКИ использовались для изучения эффектов деятельности по профилактике детского неблагополучия в США (этой практикой сегодня активно интересуется Великобритания) и для снижения детской смертности в Уганде.

**Пример РКИ (программа возврата к трудовой деятельности)**




*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*

# Популяция, вмешательство, сравнение, изменение

Все клинические исследования можно рассматривать как PICO – тщательное изучение популяции (**P**opulation) и вмешательства (**I**ntervention), сопоставление данных (**C**omparison) и анализ результатов (**O**utcome), т.е. как изучение изменений, которые происходят в какой-либо популяции вследствие вмешательства/лечения и определяются через сопоставление с результатами сравнительного лечения. В примере, приведенном Беном Голдакром:

- Вмешательство: заместительная гормональная терапия
- Популяция (целевая аудитория): женщины*
- Сравнение: отсутствие заместительной гормональной терапии
- Изменения: сердечные приступы/инфаркты

| P | I | C | O |
|---|---|---|---|
| Популяция<br>Пациент<br>Проблема | Вмешательство или экспозиция | Сравнение | Изменение |
| Кто является пациентами?<br>В чем заключается проблема? | Что мы делаем?<br>Каким воздействиям подвергаем пациентов? | С чем мы сравниваем наше вмешательство? | Что происходит?<br>Каков результат/изменение? |

*Первичные исследования привели к неверным выводам – к сожалению, это так – потому что они не различали сравнительно здоровых и менее здоровых женщин, относящихся к другой популяции.

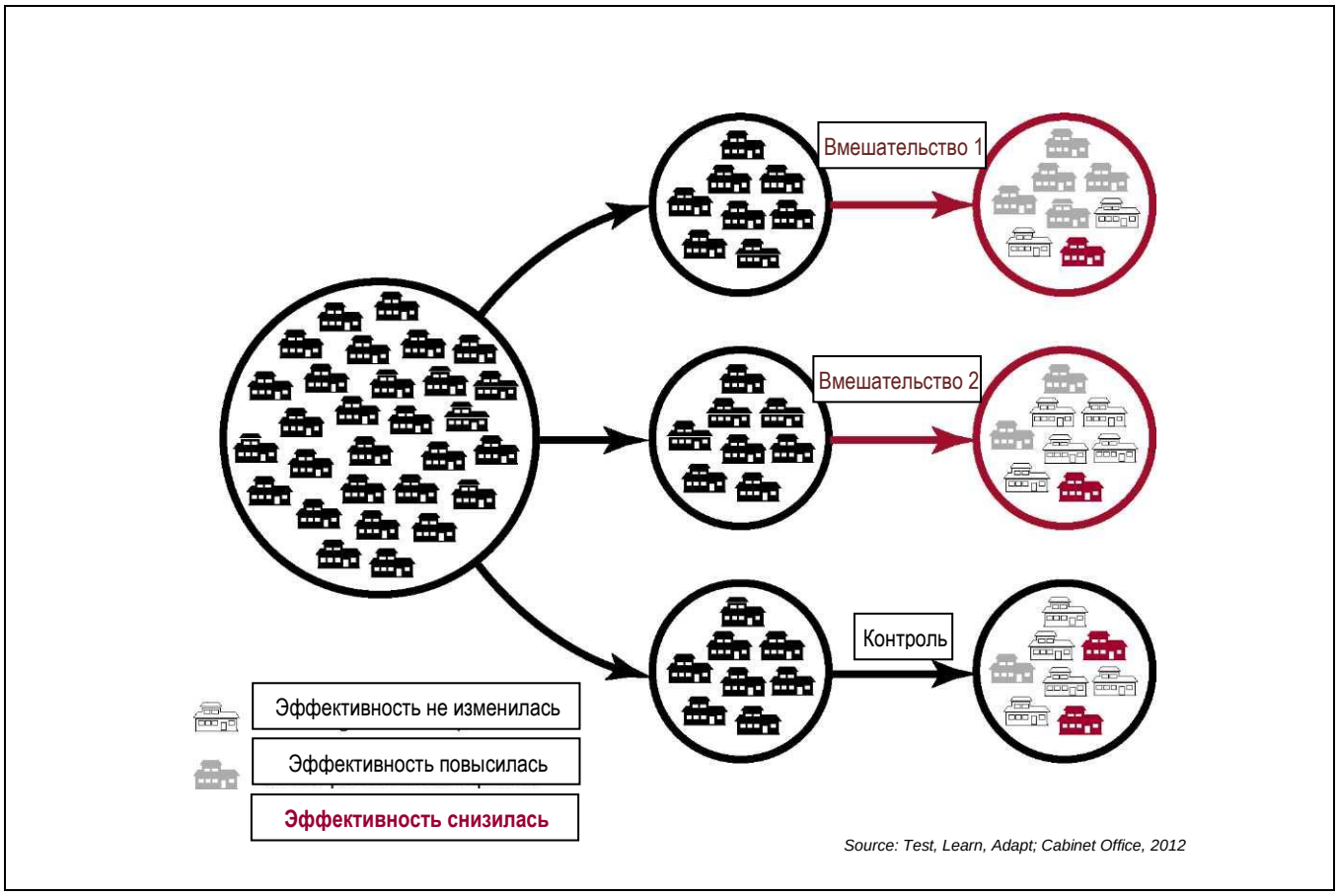
Мы вновь обращаемся к этим выводам, потому что они показывают, что важность сравнительной методики нельзя недооценивать. PICO преподают всем медикам. *Данное вмешательство/лечение конкретной группы людей генерирует данное изменение по сравнению с чем?*

Сейчас мы подошли к тому, что **некоторые заявления о том, что вмешательства всегда сравниваются с «ничем» (с «отсутствием лечения»), не соответствуют действительности.** Социальные программы, так же как и врачи, редко ориентируются на сопоставление действия с бездействием. Намного чаще сопоставляются Программа А и Программа В, либо комплекс услуг и частичная услуга (например, масштабное финансирование и поддержка небольшими суммами). Фактически, опасность использования «нулевых сравнений» ('null comparisons') была подробно рассмотрена в редакционной статье медицинского журнала с необычайно пламенным заголовком «Кровь на наших руках: зло некорректных сравнений» ('Blood on Our Hands: See The Evil In Inappropriate Comparators'), посвященной медицинским исследованиям, которые практикуют такой метод.

РКИ, сравнивающие программы друг с другом, работают примерно так:

*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*

**Рандомизированное контролируемое исследование двух образовательных инициатив**




# Любой оценке нужна хорошая линейка

Все строгие оценки опираются на сведения об эффективности – такие, например, как предлагаемые уровни образования, количество сердечных приступов, успеваемость, количество всходов (в случае с семенами бобов), количество трудоустроенных женщин, доходы от бизнеса. Оценщики должны выяснить, как эти данные выглядели «до», «после» и, желательно, в период реализации программы во всех группах, вовлеченных в исследование. Они измеряют эффективность по конкретной шкале.

Эти измерительные шкалы мы назовем «линейками».

Линейка не является методом оценки, и, сама по себе, не «устанавливает причинно-следственные связи».

Хорошие линейки (измерители) необходимы, но ни в коей мере не достаточны для проведения оценки. Линейки (измерительные шкалы) могут использоваться для измерения ресурсных вложений (inputs) и/или результатов/изменений (outcomes). Это мониторинг. Оценка отличается от него изучением сравнительных групп.

**У многих доноров есть линейки, но отсутствуют инструменты оценки (т.е. инструментарий для Уровней 1-2, и ничего для Уровней 3-4). Это значит, что у этих доноров нет системы оценки.**

Обратите внимание, что линейки отличаются по своему качеству. Так, некоторые шкалы, предназначенные для измерения, к примеру, «уверенности в себе», ненадежны, потому что вопросы, которые они задают, не исключают разного толкования разными людьми: их «степень взаимного соответствия индивидуальных оценок» (inter-rater reliability) слишком низкая. [Аналогией может послужить неградуированная линейка, с помощью которой невозможно определить длину объекта.] Сегодня существует много надежных и проверенных «линеек» (шкал для измерения), применяемых к широкому спектру явлений. Изобретение собственной линейки – это, как правило, не лучшая идея.

*Материал подготовлен АНО «Эволюция и Филантропия» в рамках проекта «Стандарт 2.0: комплексная поддержка СО НКО на пути повышения доказательности практик в сфере детства», который реализуется с использованием гранта Президента Российской Федерации на развитие гражданского общества, предоставленного Фондом президентских грантов и при стратегической поддержке Фонда Тимченко.*