ZARZĄDZANIE PUBLICZNE 1(25)/2014, s. 15–28
doi:10.4467/20843968ZP.14.002.2175
www.ejournals.eu/Zarzadzanie-Publiczne

Tomasz Dąbrowski
Uniwersytet Jagielloński w Krakowie

KRYTYCZNA ANALIZA METODY RCT W KONTEKŚCIE OCENY EFEKTYWNOŚCI POLITYK PUBLICZNYCH

Abstract

Critical analysis of research method based on RCT to evaluate effectiveness of public policies

The aim of this article is critical analysis of utilization of research method based on Randomized Controlled Trial to evaluate effectiveness of public policies. Due to the increase in importance of RCT method, resulting in more frequent elaborations presenting the recommended process of its implementation, this article represents an attempt which focuses more on describing the key issues related to utilization of the method and in particular: identity of the experimental group and the controlled group, selection bias, assigning individuals to groups, internal disruptions as well as the problem of utilizing devising the RCT method to evaluate multileveled policies.

Key words: RCT, Randomized Controlled Trial, public policy, public policy evaluation, Evidence-Based Policy

Streszczenie

Celem niniejszego artykułu jest krytyczna analiza wykorzystania metody badań opartych na losowych grupach kontrolnych (*Randomized Controlled Trial* – RCT) do oceny efektywności polityk publicznych. Ze względu na rosnące znaczenie metody RCT, skutkujące coraz liczniej pojawiającymi się opracowaniami prezentującymi rekomendowany proces jej stosowania, w artykule skupiono się na przybliżeniu głównych problemów wiążących się z wykorzystaniem metody, w szczególności: tożsamości grupy eksperymentalnej i kontrolnej, obciążeń selekcyjnych, przydziału jednostek do grup, zakłóceń zewnętrznych oraz problemu wykorzystania RCT do oceny wielopoziomowych polityk.

Słowa kluczowe: RCT, *Randomized Controlled Trial*, polityki publiczne, ocena polityk publicznych, polityki oparte na dowodach

Wprowadzenie

Celem niniejszego artykułu jest krytyczna analiza wykorzystania metody badań opartych na losowych grupach kontrolnych (RCT – *Randomized Controlled Trial* albo *Randomized Comparative Trial*) do oceny efektywności polityk publicznych. W artykule skupiono się na przybliżeniu najważniejszych problemów wiążących się z wykorzystaniem metody, a nie na przedstawieniu procesu jej stosowania. Ze względu na dążenie do przedstawienia najistotniejszych obszarów problemowych świadomie zostały pomięte także szczegółowe informacje statystyczne wiążące się z analizą danych pozyskiwanych w wyniku badań.

Bezpośrednią przyczyną, dla której tak postawiony temat wydaje się istotny, jest rosnąca popularność badań korzystających z metody RCT na gruncie europejskim, co jest związane między innymi z rosnącym znaczeniem koncepcji polityk opartych na dowodach [Cartwright, Hardie, 2012; Head, 2010; Mulgan, 2011]. Ponieważ zaś w kontekście polityk tego typu (*Evidence-Based Policy*) metoda RCT określana jest jako „najlepsza metoda do zbadania czystych efektów oddziaływania projektów na różne grupy interesariuszy" [Pylak, 2009: 75] albo jako „złoty standard" [West i in., 2008: 1362], istnieje potencjalne ryzyko bezrefleksyjnego jej stosowania jako metody podstawowej. Z tych względów istotne wydaje się pytanie badawcze o to, jakie ograniczenia należy brać pod uwagę przy podejmowaniu decyzji o wykorzystywaniu metody RCT do oceny polityk publicznych.

Rosnące znaczenie RCT

O zwiększającym się znaczeniu RCT świadczy przede wszystkim rosnące zainteresowanie metodą w instytucjach publicznych odpowiedzialnych za ewaluację polityk czy też programów[1], będące zapewne konsekwencją rekomendacji instytucji unijnych odnoszących się do kolejnego okresu programowania budżetu UE [Komisja Europejska, 2010: 18–19]. Nie bez znaczenia jest także rosnące zainteresowanie metodą zarówno na gruncie literatury anglojęzycznej [Bloom, 2006; Duflo i in., 2006; Haynes i in., 2012; White, 2013], jak i polskiej (Niżankowski i in., 2002; Olejniczak, 2012; Trzciński, 2009], a także pojawiające się opracowania omawiające tzw. najlepsze praktyki (*best practice*) w zakresie korzystania z metody w prowadzeniu badań ewaluacyjnych [Bandiera i in., 2012; Barrera-Osorio, Linden, 2009].

Na rosnącą popularność RCT można również patrzeć z punktu widzenia pewnej słabości metod ewaluacji stosowanych w praktyce oceny polityk publicznych.

[1] Ekspertyza na temat źródeł danych wykorzystywanych do realizacji badań kontrfaktycznych w ramach ewaluacji EFS, wykonana na zlecenie Departamentu Zarządzania EFS MRR, została ukierunkowana na „wskazanie rodzaju, zakresu oraz źródeł danych (…), służących do konstruowania grup kontrolnych" [Ministerstwo Rozwoju Regionalnego, 2013].

Ponieważ nadal część interwencji oceniana jest jedynie przez pryzmat osiągnięcia założonych wskaźników monitoringowych [Ferraro, 2009: 77; White, 2013: 30], przyjęcie metody RCT może być postrzegane jako swego rodzaju panaceum na problemy z rzetelną oceną, jako standard, który podniesie jakość dowodów na efektywność polityk, a tym samym – który pozwoli udokumentować zasadność angażowania środków publicznych w określone projekty.

Ogólne założenia i cechy charakterystyczne RCT

RCT jest jedną z metod wykorzystywanych w toku ewaluacji. Cechą charakterystyczną RCT jest wprowadzenie pojęcia grupy kontrolnej wybieranej w sposób losowy w grupie jednostek, których ma dotyczyć dana interwencja [Haynes i in., 2012: 4]. RCT zakłada konieczność wprowadzenia w całej grupie eksperymentalnej określonego bodźca (interwencji), oddziałującego na sytuację tej grupy, przy jednoczesnej stałej kontroli czynników nazywanych zakłóceniami, które na osiągnięte rezultaty mogły mieć wpływ. RCT zatem koncentruje się na określeniu związku przyczynowo-skutkowego między określonym działaniem a osiągniętymi dzięki niemu skutkami, następnie zaś odnosi je do stanu kontrfaktycznego, reprezentowanego przez grupę kontrolną, w której dany bodziec nie występował.

RCT jest więc metodą eksperymentalną, odpowiadającą na pytanie o to, jaki wpływ wywarła interwencja, przez porównanie zmian w grupie, w której tę interwencję wprowadzono (grupa eksperymentalna), z sytuacją grupy kontrolnej, w której badana interwencja nie występowała. Parafrazując White'a [2013: 31] – jest próbą odpowiedzi na pytanie o to, co osiągnięto dzięki wdrożeniu określonej polityki, a co można było osiągnąć bez niej (*with versus without*).

Zasadnicze założenia RCT nie są skomplikowane. Które zatem elementy mogą okazać się problematyczne podczas realizacji badań?

Problem tożsamości grupy eksperymentalnej i kontrolnej

Fundamentem, na którym została zbudowana koncepcja RCT, jest założenie, że między grupą eksperymentalną i kontrolną nie zachodzą istotne różnice – zarówno na poziomie obserwowalnym, jak i nieobserwowalnym. Jedyną istotną zmienną jest wpływ (lub jego brak) określonego bodźca, czyli wpływ prowadzonej interwencji. Mechanizmem mającym zapewnić taki stan jest losowy dobór grupy kontrolnej (*randomization*), reprezentującej stan kontrfaktyczny, a więc „hipotetyczny stan, jaki miałby miejsce w przypadku, gdyby grupa eksperymentalna nie została objęta badanym działaniem" [Trzciński, 2009: 19]. Ponieważ tożsamość obydwu grup ma podstawowe znaczenie, błędy popełniane w tym obszarze skutkować będą zawsze obniżeniem jakości lub całkowitym zdyskredytowaniem uzyskiwanych wyników.

Założeniem jest, że porównywalność obydwu grup zapewnia losowość doboru jednostek. Randomizacja zniweluje potencjalne różnice między obydwiema grupami, ale tylko pod warunkiem, iż jednostki wchodzące w ich skład są tożsame pod względem zdolności selekcyjnej. Oznacza to, że każda jednostka znajdująca się w dowolnej z grup powinna spełniać kryteria pozwalające jej zakwalifikować się do objęcia daną interwencją, a przydział do grupy kontrolnej powinien opierać się wyłącznie na mechanizmie losowym.

Zasadniczy problem pojawia się wówczas, gdy jako kontrolną wskazuje się grupę, której brak objęcia interwencją nie wiązał się z faktem doboru losowego, lecz z odgórnym założeniem, że grupa ta jest tożsama z grupą eksperymentalną. Interesujący przykład konsekwencji takich działań przytacza White [2013: 32], opisując skutek przyjęcia, że grupą kontrolną dla szkół zlokalizowanych na obszarach cechujących się wyższym poziomem ubóstwa będą szkoły z innych obszarów. Obydwie grupy funkcjonowały tymczasem w różnych kontekstach społecznych. Młodzież uczęszczająca do szkół zlokalizowanych na uboższych obszarach częściej wywodziła się z rodzin słabiej wykształconych, o mniejszych ambicjach edukacyjnych i mniejszych możliwościach poświęcenia czasu na naukę. Sytuacja ta skutkowała tym, że nawet w obliczu wsparcia udzielonego szkołom na obszarach uboższych osiągane przez nie wyniki były niższe niż w pozostałych szkołach. Na podstawie takiego porównania nie można jednak wyciągnąć wniosku o słabym wpływie interwencji. Na wynikach badania zaciążył bowiem niewłaściwy dobór grup, brak porównania jednostek tożsamych.

Z analogicznym błędem będziemy mieć do czynienia także wówczas, gdy wpływ interwencji polegającej na wsparciu grupy osób długotrwale bezrobotnych w znalezieniu pracy porównamy do szeroko rozumianej grupy ogółu osób bezrobotnych, ale niekoniecznie trwale. Sytuacja tych dwóch grup może być różna – zarówno na poziomie motywacji, statusu, jak i kwalifikacji tych osób.

Na marginesie należy wspomnieć o decyzji dotyczącej kategorii jednostek, które będą podstawą przeprowadzenia randomizacji (*unit of randomization*). W największym uproszczeniu – podczas oceny możliwe jest analizowanie pojedynczych podmiotów (np. osób, instytucji) lub ich określonej grupy (*cluster randomization*). Podział ten będzie implikować kwestię zastosowania metod analitycznych w czasie badania, determinując konieczność skupienia się na pomiarze efektów na poziomie jednostek lub całych grup [Jadad, 1998: 8]. Wybór podziału w oczywisty sposób będzie zależeć od celów danej interwencji, oczekiwanych rezultatów, przyjętych uprzednio założeń dotyczących wskaźników, jak również od możliwości finansowych i organizacyjnych mających wpływ na zdolność do gromadzenia danych.

Mając na uwadze powyższe, należy podkreślić, że RCT nie oznacza prostego porównania dwóch grup między sobą, lecz przede wszystkim ma na celu zdefiniowanie jednego zbioru jednostek spełniających kryteria wymagane do objęcia ich określoną interwencją, a dopiero następnie podzielenie tych jednostek na grupę eksperymentalną i kontrolną.

Obciążenia selekcyjne

Specyficznym przykładem problemu z zapewnieniem tożsamości grup są obciążenia selekcyjne (*selection bias*). Występują one wówczas, gdy na etapie przydziału jednostek do grupy kontrolnej i eksperymentalnej popełniono błędy skutkujące niemożliwością rzetelnego ich porównania.

Jako teoretyczny przykład może służyć odniesienie się do praktyki powiatowych urzędów pracy, w których spotyka się mechanizm punktowy (rangowy) w procesie selekcji. Skutkuje on stworzeniem listy rankingowej osób zainteresowanych określonym wsparciem. Ponieważ miejsce na liście decyduje o możliwości korzystania ze wsparcia, oznacza to, że do udziału w określonej interwencji zapraszane są jednostki najwyżej ocenione, więc spełniające na najwyższym poziomie wymagania dotyczące ich statusu, wykształcenia, doświadczenia itp. Ich porównanie z pozostałymi osobami, które zgłosiły akces, ale zostały ocenione niżej, musiałoby budzić wątpliwości. Wprawdzie spełniają one podstawowe kryteria selekcyjne, ale jednocześnie proces kwalifikacyjny jednoznacznie wykazał, że z punktu widzenia zdefiniowanych kryteriów oceny jednostki te prezentują się gorzej niż osoby wybrane do objęcia wsparciem.

Innym przykładem mogą być te interwencje, w których udział we wsparciu jest oparty na zasadzie autoselekcji (*self-selection*), choćby w formie dobrowolnych zgłoszeń limitowanych jedynie liczbą miejsc. Gdy sytuacja taka prowadzi do skompletowania grupy osób bezrobotnych mających wziąć udział w szkoleniu, a następnie do porównania wyników osiąganych przez tę grupę z grupą pozostałych osób bezrobotnych zarejestrowanych w danym urzędzie – możliwe jest, że niedostrzeżony zostanie problem silniejszej motywacji osób, które same się zgłosiły do udziału w szkoleniu. Autoselekcja sprawia bowiem, że do korzystania z danego wsparcia zgłaszają się przede wszystkim osoby z określonym, często pozytywnym do niego nastawieniem [White, 2013: 31], które dzięki wyższej motywacji być może potrafiłyby osiągnąć założony cel interwencji (np. znaleźć pracę) także i bez niej. Ich porównanie z grupą osób, które nie zgłosiły swojego udziału, nie będzie więc właściwe.

Powyższe przykłady pokazują, że brak losowego doboru może mieć znaczenie dla tożsamości obydwu grup ujętych w badaniu, lecz w kontekście pozytywnym (osoby wybrane do objęcia wsparciem w pewnym stopniu dają większą gwarancję osiągnięcia określonych rezultatów). Możliwy jest jednak przykład odwrotny – gdy osoby korzystające ze wsparcia są nim potencjalnie mniej zainteresowane niż osoby, które z niego nie korzystają. Podobnie jak w przykładach przytoczonych wyżej, sytuację taką można prześledzić również na tle wsparcia w zakresie przekwalifikowania się oferowanego osobom bezrobotnym. Otóż udział w interwencjach o tym charakterze, w przypadku części osób bezrobotnych, obwarowany jest istotnymi konsekwencjami prawnymi (utrata statusu osoby bezrobotnej w sytuacji odmowy udziału we wsparciu oferowanym przez

urząd pracy[2]). Oznacza to, że udział we wsparciu osób zagrożonych utratą statusu bezrobotnego może się wiązać ze swego rodzaju przymusem sprawiającym, iż realnie nie zawsze będą one zainteresowane osiągnięciem jakichkolwiek rezultatów, lecz samym faktem ukończenia określonego szkolenia, do udziału w którym zostały zaproszone przez urząd. Może to oznaczać, że pomimo oferowanego wsparcia ich sytuacja może nie być lepsza niż pozostałych osób bezrobotnych, które mają jednak silniejszą motywację lub predyspozycje pozwalające im znaleźć pracę. Porównanie tych grup metodą RCT nie wydaje się zatem zasadne.

Należy przy tym wskazać, że obciążenia selekcyjne mogą mieć charakter jawny (*overt bias*) lub ukryty (*hidden bias*). Pierwsze z nich są możliwe do zmierzenia, kontrolowania, drugie – niemożliwe albo niezwykle trudne do zidentyfikowania i zwymiarowania. O ile w przypadku obciążeń jawnych możliwe jest ich usunięcie lub zniwelowanie, o tyle w przypadku obciążeń ukrytych w zasadzie możliwe jest jedynie ich bierne akceptowanie lub próba kontrolowania stopnia niepewności z nimi związanego [Rosenbaum, 2005: 1].

Problem przydziału jednostek do grupy kontrolnej i eksperymentalnej

Jak wspomniano wyżej, przyporządkowanie jednostek do grupy kontrolnej i eksperymentalnej jest głównym elementem metody RCT. Jest to jednocześnie etap, w którym potencjalna możliwość popełnienia błędu może obciążyć wyniki badania. Działania te wymagają znajomości zasad doboru próby badawczej, których omówienie, jak wskazano we wstępie, nie jest przedmiotem niniejszego opracowania.

Niemniej jednak przydział poszczególnych jednostek z punktu widzenia metod statystycznych jest tylko swego rodzaju następstwem udanej rekrutacji (zakwalifikowania) jednostek do eksperymentu. Możliwość przeprowadzenia losowania opartego na właściwej próbie zależy bowiem od charakteru dostępnych jednostek mających potencjalną zdolność korzystania ze wsparcia. Brak ich rzeczywistej dostępności sprawi, że poprawny pod względem doboru próby podział natrafi zapewne na trudności. Tym samym skuteczność etapu randomizacji w dużym stopniu będzie zależeć od rzeczywistych możliwości pozyskania odpowiednich jednostek do badania.

Jako uzupełnienie należy także za Bloomem [2006: 2] zwrócić uwagę na sposób informowania jednostek uczestniczących w eksperymencie (lub asygnowanych do grupy kontrolnej) o samym badaniu. Zakres przekazywanych im informacji również może mieć bowiem wpływ na potencjalne wyniki.

[2] Por. art. 33 ust. 3 Ustawy z dnia 20 kwietnia 2004 r. o promocji zatrudnienia i instytucjach rynku pracy: „Starosta (…) pozbawia statusu bezrobotnego, który (…) odmówił bez uzasadnionej przyczyny (…) udziału w szkoleniu, stażu, przygotowaniu zawodowym w miejscu pracy (…)" [*Ustawa*, 2004]. Ustawa określa także kryteria, jakie musi spełnić osoba bezrobotna, by pozbawianie jej statusu bezrobotnego mogło dojść do skutku.

Szczególnym przypadkiem zakłóceń, które mogą być następstwem przekazania badanym jednostkom szczegółowych informacji o samym eksperymencie, jak i wynikającym z obecności badaczy w jego trakcie, może być efekt Hawthorne, który przejawia się w osiąganiu lepszych wyników zarówno przez grupę eksperymentalną, jak i kontrolną, przy czym bodźcem do zmiany zachowań w grupie kontrolnej jest sama świadomość uczestniczenia w badaniu [Duflo i in., 2006: 68–69; Weber, 2002].

Problem małych grup

Na dodatkową uwagę zasługuje kwestia adekwatności metody w sytuacji występowania relatywnie małych grup objętych określonym wsparciem.

Problem ten odwołuje się do zwiększania się ryzyka obciążeń selekcyjnych wraz ze zmniejszaniem się grupy, która może być objęta określoną interwencją. RCT zakłada bowiem, że poszczególne jednostki zostaną zakwalifikowane do grupy eksperymentalnej bądź losowej na podstawie pewnej uśrednionej charakterystyki (*similar average characteristic*), co jednak sprawia, iż metoda RCT będzie bardziej właściwa dla relatywnie dużych grup docelowych (*large-n*). Mniejsze liczebnie grupy (*small-n*) oznaczają z punktu widzenia uśrednienia ich charakterystyk dodatkowe utrudnienie na gruncie chociażby statystycznym.

Przykładowo, chcąc ocenić politykę transportową aglomeracji śląskiej [Jackiewicz i in., 2010], trudno byłoby jako metodę przyjąć RCT. Brakuje bowiem dla niej obszaru, który byłby punktem odniesienia. Jest to rzecz jasna skrajny przypadek, przytoczony jednak celowo dla pokazania, że w niektórych interwencjach nie będzie możliwe zastosowanie metody RCT. Niemniej problemy mogą pojawić się także w grupach pozornie łatwiejszych do zdiagnozowania na podstawie uśrednionych charakterystyk. White [2013: 35] przytacza przykład badań w zakresie zdrowia publicznego. Przy badaniach odnoszących się do zdrowia kobiet dopiero $n = 2000$ pozwalało na rzetelne porównanie poszczególnych grup. Przyjęcie mniejszych prób prowadziło do uzyskania danych o niskiej wiarygodności.

Oznacza to, że problem liczebności zarówno grupy eksperymentalnej, jak i kontrolnej nie jest oczywisty. Odnosi się on w zasadzie do pytania o dobór próby wystraczającej do tego, by móc prowadzić prawidłowe wnioskowanie o rzeczywistym wpływie danej interwencji. W literaturze wskazuje się przy tym, że w trakcie badania interwencji o spodziewanych, łatwo obserwowanych, znaczących efektach (*large effect size*), grupy te mogą być relatywnie mniejsze niż w przypadku interwencji o spodziewanych efektach, słabiej dostrzegalnych, o mniejszych różnicach – *small effect size* [Bloom, 2006: 4].

Problemowi temu osobną publikację poświęcili Phillips i White [2012]. Konkluzje z niej płynące ukierunkowane są jednak w stronę rekomendowania wykorzystania w takiej sytuacji metod dedukcyjnych, opartych na analizach związków przyczynowych, a więc na wykorzystaniu metod jakościowych lub

mieszanych. Stosowanie zatem wyłącznie ujęć statystycznych (zasadniczo właściwych istocie RCT) przy małych grupach może stanowić istotne ograniczenie, wymagające szczególnej uwagi.

Problem zakłóceń zewnętrznych

Z punktu widzenia rzetelności procesu badawczego dużym utrudnieniem są zakłócenia (*confounding*) występujące podczas jego realizacji [Niżankowski i in., 2002: 35], które mogą występować zarówno w grupie eksperymentalnej, jak i kontrolnej. Teoretycznie losowy wybór grupy kontrolnej powinien zniwelować to zagrożenie, w praktyce jednak czynniki zakłócające (*confounding factors*), szczególnie w obszarze polityk publicznych, są niezwykle trudne do zidentyfikowania i oszacowania, także dlatego, że mogą zaistnieć w obydwu grupach w różnych konfiguracjach (np. gdy wystąpią zakłócenia, które sprawiają, że grupa kontrolna może w polityce osiągnąć określone rezultaty pomimo braku wsparcia, a grupa eksperymentalna natrafi na zakłócenia, które mimo podjętej interwencji nie pozwolą jej na osiągnięcie założonego efektu). W kontekście występowania zakłóceń mówi się wprawdzie o ocenie efektów netto polityki publicznej (*net effect*) i efektach brutto (*gross effect*) uwzględniających potencjalne zakłócenia, nie zmienia to jednak zasadniczej konkluzji, że ich zidentyfikowanie oraz kontrolowanie jest w wielu wypadkach problematyczne.

Problematyka zakłóceń zewnętrznych jest szczególnie widoczna w przypadku polityk środowiskowych, w odniesieniu do których wpływ na realizację danej polityki mogą mieć takie czynniki, jak sytuacja ekonomiczna, zdolności inwestycyjne firm w regionie, nastawienie lokalnych społeczności czy choćby warunki pogodowe [Ferraro, 2009: 78]. Ponieważ programy ochrony środowiska odnoszą się często do określonych obszarów geograficznych, wystąpienie zakłóceń (przykładowo niezaplanowanych warunków pogodowych, kryzysu ekonomicznego zmniejszającego zapotrzebowanie na surowce naturalne) na obszarze traktowanym jako miejsce wdrażania polityki lub na obszarze kontrolnym może istotnie wpłynąć na wyniki oceny. Ponieważ zakłócenia mogą mieć charakter ukryty, trudno obserwowalny, zawsze będą one stanowić potencjalny czynnik obciążający badanie.

Problem wielopoziomowych polityk

O ile omówione wyżej problemy dotyczą sytuacji, w której ocena stopnia osiągnięcia celów wydaje się możliwa, jeśli tylko da się zidentyfikować odpowiednie grupy, o tyle zagadnienie stosowania RCT w analizie wielopłaszczyznowych, wielopoziomowych projektów wiąże się z szerszym pytaniem o to, czy w ogóle metoda ta może być zastosowana.

Dylemat dotyczy wykorzystywania RCT przy ocenie złożonych, międzysektorowych polityk, w których sama struktura celów jest na tyle rozbudowana, że ich odniesienie do jednej grupy docelowej jest wykluczone lub też zdefiniowanie związku przyczynowo-skutkowego – wyjątkowo skomplikowane przez wielopoziomowość wzajemnie zależnych od siebie celów.

Przykładem może być obszar rozwoju zasobów ludzkich będący częścią polityki spójności. Ze względu na jego złożoność realizacja jednego procesu badawczego jest niemożliwa. Dlatego ocenia się oddzielnie wpływ interwencji na beneficjentów ostatecznych, rezultaty poszczególnych działań czy same instytucje wdrażające pod kątem jakości ich działań [Bienias, Sudak, 2008: 44–45]. Przy tak złożonych politykach jedyną możliwością wykorzystania RCT jest ocena poszczególnych efektów cząstkowych polityki, a więc rezygnacja z prowadzenia jednego, kompleksowego badania opartego na RCT przy skupieniu się na wybranych, definiowanych jej aspektach. Należy przy tym dodać, że całokształt badań ewaluacyjnych na poziomie złożonych polityk powinien być mierzony za pomocą triangulacji metod badawczych [Olejniczak, 2012: 47].

Problem czasu przystąpienia do wykonywania RCT

Dodatkowym problemem jest to, w którym momencie wdrażania określonej polityki możliwe jest zastosowanie RCT. Nie jest to pytanie bezpodstawne, gdyż wpisuje się ono w szerszy kontekst dyskusji o metodach ewaluacji *ex-post* [Haber, 2007a]. Biorąc jednak pod uwagę sam charakter RCT i jego zasadniczą cechę, jaką jest świadomy wybór grupy kontrolnej, a nie porównanie sytuacji grupy uczestniczącej w projekcie z sytuacją grupy poza projektem, wydaje się, że zastosowanie RCT w kontekście ewaluacji *ex-post* nie jest możliwe. Wykorzystanie RCT do uzasadniania efektów interwencji już zakończonych przeczyć będzie jej eksperymentalnemu charakterowi.

Problem zdefiniowania efektów i wskaźników

Z punktu widzenia rzetelności badania także ustalenie skutków, które powinny nastąpić dzięki danej interwencji, oraz dobór wskaźników ich pomiaru nastręczają problemów praktycznych. Przed przystąpieniem do badania powinno się bowiem określić, czy pomiar określonych efektów wprost jest możliwy, czy też konieczne jest zdefiniowanie pewnych pośrednich, zastępczych wskaźników pozwalających na wnioskowanie o osiąganiu tych efektów, które są niemożliwe do pomiaru wprost lub które następują po stosunkowo długim okresie (*surrogate outcome*).

Przykładowo w odniesieniu do przestępczości można wykazać, że łatwiej jest wykorzystywać wskaźnik ponownych skazań (*reconviction rates*) niż wskaźnik

ponownego popełnienia przestępstwa (*reoffending rates*), ponieważ w drugim przypadku, z przyczyn oczywistych, trudno jest definitywnie określić liczbę przypadków powrotu do przestępczości, która nie została wykryta [Haynes i in., 2012: 23]. Jakość doboru ewentualnych wskaźników pośrednich (zastępczych) rzutować będzie na ostateczny kształt wniosków z badania.

Oczywiście można postulować każdorazowe odwoływanie się wyłącznie do efektów bezpośrednich, ale – jak pokazuje powyższy przykład – w jednych przypadkach może to być niewykonalne z przyczyn obiektywnych, w innych zaś koszty pozyskania danych mogą być na tyle duże, że nie będzie to po prostu opłacalne. W projektowaniu badania nie bez znaczenia jest bowiem rozdźwięk pomiędzy dążeniem do uzyskania szczegółowych, wyczerpujących danych a kosztami ich pozyskania. Na przykład przy ocenie efektów interwencji szkoleniowych zalecana jest czteropoziomowa ocena wyników, w tym między innymi ocena stopnia aplikacji nowych umiejętności w miejscu pracy [Kirkpatrick, 2006: 40––46]. Jest to jednak metoda kosztochłonna, wymagająca dużego zaangażowania organizacji zatrudniającej poszczególne osoby, dlatego nie należy oczekiwać jej szerokiego stosowania nawet pomimo dużych zalet. Wnioski na temat efektów szkoleń będą raczej oparte na zmianie wiedzy i umiejętności uczestników, gdyż te dają się stosunkowo prosto mierzyć. Oznacza to pewne uproszczenie wnioskowania na podstawie pośrednich dowodów, argument finansowo-organizacyjny może tu jednak okazać się przeważający.

Problem trudności wyjaśniania istoty zmiany przez badanie RCT

Metody, narzędzia, wskaźniki i czas pomiaru, jak napisano wyżej, powinny być przedmiotem uzgodnień na etapach poprzedzających wdrożenie określonej interwencji. Oczywiście czas niezbędny do oceny efektów będzie mocno skorelowany z typem interwencji i będzie obejmować w jednym przypadku kilka tygodni, w innym wiele lat. Najistotniejszy w tym zakresie jest jednak nie tyle sam fakt dokonywania pomiaru, a później prowadzenia porównań między grupą eksperymentalną i kontrolną, ile definiowanie wniosków z tych analiz, a w zasadzie – próba wyjaśnienia zmiany (lub jej braku).

Formułowanie wniosków pozwalających wyjaśnić ewentualne zmiany jest działaniem, które w pewnych sytuacjach może wykraczać poza standardowe czynności wykonywane w ramach RCT. Jak pisze Olejniczak, „(…) dobrze wykonane RCT dają co prawda mocne dowody na temat relacji przyczynowo-skutkowej między interwencją a zmianą w konkretnym otoczeniu, jednak nie mogą wyjaśnić tej zmiany" [2012: 48]. Niezależnie od tak postawionej tezy RCT pozwala w niektórych przypadkach na wyjaśnienie zmian wynikłych wskutek interwencji. Zależeć to będzie jednak o tego, jakie wskaźniki brane będą pod uwagę w czasie pomiaru, w szczególności, czy poza generalną próbą oceny stopnia osiągnięcia efektów polityki monitorowano także inne aspekty wdrożenia

(osiągnięcie celów pośrednich, rezultaty i produkty poszczególnych działań podejmowanych w toku realizacji polityki).

Warto przy tym podkreślić, że w wyjaśnieniu zmiany mogą pomóc także metody mieszane, głównie przez dopuszczenie do jakościowej oceny rezultatów badania prowadzonego metodą RCT. Nawet jeśli przyjęcie takiej możliwości traktowane będzie jako odstępstwo od „czystej" metody, to jednak skutkować zapewne będzie lepszym rozumieniem zmiany i efektów samej interwencji, co jest zasadniczym celem stosowania tej metody. Problem zastosowania metod mieszanych szerzej omawiają Spillane, Dorner i in. [2010]. W kontekście analizy danych interesujący jest także artykuł Hubera [2011], odnoszący się do analizy danych w sytuacji posiadania tylko częściowych rezultatów pomiaru.

Problemy etyczne

Uzupełnieniem rozważań dotyczących wykorzystywania metody RCT są podnoszone w literaturze wątpliwości natury etycznej. Ogólny zarzut sprowadza się do pytania, czy w odniesieniu do tych interwencji, które dotykają obszarów wrażliwych społecznie (np. polityk w obszarze zdrowia, przeciwdziałania ubóstwu), etyczne jest ograniczanie dostępu wybranym drogą losowania osobom. Zagadnienie to ma szerszy kontekst – może dotyczyć nie tylko sytuacji braku możliwości dostępu do określonych usług określonym grupom (uznanym za kontrolne), ale także przykładowo ograniczenia konkurencyjności przez uniemożliwienie korzystania ze wsparcia oferowanego w programach realizujących polityki pobudzania wzrostu gospodarczego tylko wybranym podmiotom (grupom eksperymentalnym). Warto przy tym zwrócić uwagę, że te wątpliwości mogą budzić zastrzeżenia na gruncie nie tylko etycznym, ale także prawnym, na przykład gdy istnieje ryzyko naruszenia zasady równego dostępu do świadczeń opieki zdrowotnej [Lach, 2011: 16].

Rozważania na temat dylematów etycznych ograniczają się w zasadzie do pokazania, że próby udowodnienia, czy określone działania są efektywne, nie mogą być odbierane jako nieetyczne z samej zasady, gdyż ich celem jest dostarczenie w konsekwencji rozwiązań o sprawdzonej jakości. Takie działania są zresztą podejmowane wówczas, gdy w ramach określonych polityk wprowadza się programy pilotażowe, które także mają charakter limitujący dostęp [Haynes i in., 2012: 16–17].Trudno tym samym oczekiwać powszechnej zgody co do ich zasadności, należy jednak zawsze brać je pod uwagę w projektowaniu procesu badawczego.

Podsumowanie

Założenia teoretyczne sprawiają, że metoda RCT jest atrakcyjnym i interesującym rozwiązaniem pozwalającym na ocenę polityk publicznych, przede

wszystkim ze względu na jej potencjał dostarczania jednoznacznych wyników definiujących wpływ określonej interwencji. Niezależnie, czy jej rolę postrzegać będziemy przede wszystkim w kategoriach falsyfikacyjnych, czy konfirmacyjnych [Górniak i in., 2007: 132], postulat przyjęcia RCT jako „złotego standardu", który powinien znaleźć swe zastosowanie w procesach oceny efektywności polityk publicznych, wydaje się jednak niemożliwy do zrealizowania. I także – niekonieczny.

Bez względu bowiem na to, czy jest się do RCT nastawionym entuzjastycznie [Haynes i in., 2012], czy też prezentuje się bardziej krytyczne stanowisko [Brass i in., 2006], należy na RCT patrzeć jako na jedną z wielu metod, które prowadzą do lepszego rozumienia tego, czy poszczególne polityki publiczne wywierają założony wpływ. Skupianie się zaś na jednej metodzie, zgodnie z wyrażonym przez Kaplana „prawem jednego narzędzia" [2004: 28], skutkować będzie ograniczoną możliwością percepcji wpływu, jaki wywierają polityki publiczne.

Tym samym trzeba przyjąć, że w pluralizmie metod leży największy potencjał skutecznej oceny. Postulat pluralistycznego ujęcia znajduje z kolei swe odzwierciedlenie zarówno w publikacjach *stricte* dotyczących problematyki praktycznego wykorzystania RCT [West i in., 2008: 1363], jak i w oficjalnych dokumentach Komisji Europejskiej [Hübner, 2008] czy Ministerstwa Rozwoju Regionalnego[3], które koncentrują się na zagadnieniu celu prowadzonych ewaluacji, nie wskazując jednoznacznie metod jej realizacji lub traktując ją jako jedną z wielu.

RCT zatem powinno być tylko jedną z metod, o dużej wartości dowodowej, przy czym jej wykorzystanie musi każdorazowo uwzględniać omówione wyżej ograniczenia. Analiza danych uzyskanych w wyniku RCT powinna zaś dopuszczać wykorzystanie metod mieszanych w miejsce praktykowanej wyłącznej analizy ilościowej.

Literatura

Bandiera O., Burgess R., Das N.C., Selim Gulesci S., Rasul I., Shams R. Sulaiman M. (2012), *Asset Transfer Programme for the Ultra Poor: A Randomized Control Trial Evaluation*, BRAC Centre, Bangladesh.

Barrera-Osorio F., Linden L.L. (2009), *The Use and Misuse of Computers in Education: Evidence from a Randomized Controlled Trial of a Language Arts Program*, http://www.povertyactionlab.org/publication/use-and-misuse-computers-education-evidence-randomized-controlled-trial-language-arts-pr (dostęp: 10.05.2013).

Bienias S., Sudak S. (red.) (2008), *Proces ewaluacji polityki spójności w Polsce*, Ministerstwo Rozwoju Regionalnego, Warszawa.

Bloom H.S. (2006), *The Core Analytics of Randomized Experiments for Social Research*, http://www.mdrc.org/sites/default/files/full_533.pdf (dostęp: 09.05.2013).

[3] W rozdziale poświęconym wyborowi metod ewaluacji najkrócej kwestię adekwatności doboru metod opisano następująco: „Złota reguła: nie ma złotych reguł" [Pylak, 2009: 62].

Brass C.T., Nuñez-Neto B., Williams E.D. (2006), *Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCTs) and Related Issues*, Congressional Research Service Library of Congress, Washington.

Cartwright N., Hardie J. (2012), *Evidence-Based Policy: A Practical Guide to Doing It Better*, Oxford Univerity Press, Oxford.

Duflo E., Glennerster R., Kremer M. (2006), *Using Randomization in Development Economics Research: A Toolkit*, National Bureau of Economic Research, Cambridge.

Ferraro P.J. (2009), *Counterfactual Thinking and Impact Evaluation in Environmental Policy* „New Directions for Evaluation", special issue „Environmental program and policy evaluation", M. Birnbaum, P. Mickwitz (eds.), (122), 75–84, Wiley Periodicals.

Górniak J., Worek B., Krupnik S. (2007), *Zastosowanie podejścia badań jakościowych w ewaluacji ex-post* [w:] A. Haber (red.), *Ewaluacja ex-post. Teoria i praktyka badawcza*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa.

Haber A. (red.) (2007), *Ewaluacja ex-post. Teoria i praktyka badawcza*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa.

Haynes L., Service O., Goldacre B., Torgerson D. (2012), *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*, The Cabinet Office Behavioural Insights Team, London.

Head B. (2010), *Evidence-based Policy: Principles and Requirements*, „Strengthening Evidence-Based Policy in the Australian Federation", Vol. 1: Proceeding, Roundtable Proceedings, Productivity Commission, Melbourne.

Huber M. (2011), *Identification of Average Treatment Effects in Social Experiments under Alternative Forms of Attrition*, „Journal of Educational and Behavioral Statistics", Vol. 37(3), 443–474.

Hübner D. (2008), *Foreword* [w:] *EVALSED: The Resource for the Evaluation of Socio-Economic Development*, Office for Official Publications of the European Communities, Luxemburg.

Jackiewicz J., Czech P., Barcik J. (2010), *Polityka transportowa na przykładzie aglomeracji śląskiej*, „Zeszyty Naukowe Politechniki Śląskiej", Seria Transport, z. 69, 53–62, Wydawnictwo Politechniki Śląskiej, Gliwice.

Jadad A.R. (1998), *Randomised Controlled Trials. A Users' Guide*, BMJ Books, London.

Kaplan A. (2004), *The Conduct of Inquiry: Methodology for Behavioral Science*, Transaction Publishers, New Brunswick.

Kirkpatrick D.L. (2006), *Evaluating Training Programs the Four Levels*, Berrett-Koehler, San Francisco.

Komisja Europejska (2010), *Ewaluacja – jakie stosować metody?*, „Panorama Inforegio. Ewaluacja polityki spójności. Kierunki i wyniki", Vol. 33, 18–19, Bruksela.

Lach D.E. (2011), *Zasada równego dostępu do świadczeń opieki zdrowotnej*, Lex Wolters Kluwer Business, Warszawa.

Ministerstwo Rozwoju Regionalnego (2013), *Raport z wykonania ekspertyzy na temat źródeł danych wykorzystywanych do realizacji badań kontrfaktycznych w ramach ewaluacji EFS*, Warszawa.

Mulgan G. (2011), *Foreword* [w:] A. Shillabee, T.F. Buss, D.M. Rousseau (ed.), *Evidence-based Public Management: Practices, Issues, and Prospects*, Shape Inc., New York.

Niżankowski R., Bała M., Broda M., Dubiel B., Hetnał M., Kawalec P., Łanda K., Plisko R., Podmokły A., Wcisło J., Wójtowicz E. (2002), *Analiza efektywności*, Uniwersyteckie Wydawnictwo Medyczne VESALIUS, Kraków.

Rosenbaum P.R. (2005), *Observational Study* [w:] *Encyclopedia of Statistics in Behavioral Science*, John Wiley & Sons, Hoboken N.J.

Olejniczak K. (red.) (2012), *Organizacje uczące się. Model dla administracji publicznej*, Wydawnictwo Naukowe Scholar, Warszawa.

Pylak K. (2009), *Podręcznik ewaluacji efektów projektów infrastrukturalnych*, Ministerstwo Rozwoju Regionalnego, Warszawa.

Spillane J.P., Pareja A.S., Dorner L., Barnes C., May H., Huff J., Camburn E. (2010), *Mixing Methods in Randomized Controlled Trials (RCTs): Validation, Contextualization, Triangulation, and Control*, Springer Science Business Media, http://www.sesp.northwestern.edu/docs/publications/8689892294c2cf93a26468.pdf (dostęp: 9.05.2013).

Trzciński R. (2009), *Wykorzystanie techniki* propensity score matching *w badaniach ewaluacyjnych*, Polska Agencja Rozwoju Przedsiębiorczości, Warszawa.

Ustawa (2004), Ustawa z dnia 20 kwietnia 2004 r. o promocji zatrudnienia i instytucjach rynku pracy, t.j. Dz.U. 2008 Nr 69, poz. 415 z późn. zm.

Weber A. (2002), *The Hawthorne Works*, http://www.assemblymag.com/articles/88188-the-hawthorne-works (dostęp: 16.05.2013).

West S.G., Duan N., Pequegnat W., Gaist P., Des Jarlais D.C., Holtgrave D., Szapocznik J., Fishbein M., Rapkin B., Clatts M., Mullen P.D. (2008), *Alternatives to the Randomized Controlled Trial*, „American Journal of Public Health", Vol. 98. No. 8, 1359–1366.

White H. (2013), *An Introduction to the Use of Randomised Control Trials to Evaluate Development interventions*, „Journal of Development Effectiveness", 5(1), 30–49.

White H., Phillips D. (2012), *Addressing Attribution of Cause and Effect in Small n Impact Evaluations*: *Towards an Integrated Framework*, International Initiative for Impact Evaluation, New Delhi.