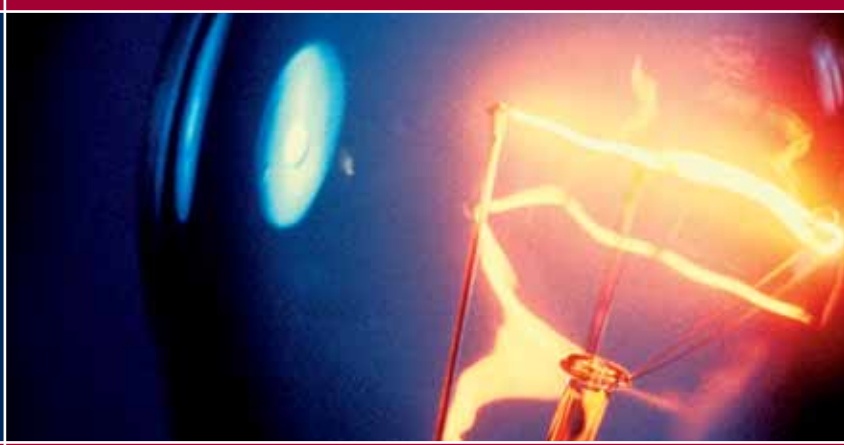
PARP

2013

# (R)ewaluacja
## Poszukiwanie nowych metod oceny efektów



redakcja
Agnieszka Haber
Zuzanna Popis

# (R)ewaluacja
## Poszukiwanie nowych metod oceny efektów

**pod redakcją**
**Agnieszki Haber i Zuzanny Popis**

Polska Agencja Rozwoju Przedsiębiorczości
Warszawa 2013

*(R)ewaluacja. Poszukiwanie nowych metod oceny efektów.*
**pod redakcją Agnieszki Haber i Zuzanny Popis**

Poglądy autorów rozdziałów zawartych w publikacji są wyrazem ich własnych opinii i nie odzwierciedlają oficjalnego stanowiska PARP.

**Współpraca techniczna:** Ewa Kapusta



ISBN: 978-83-7633-272-7

Wydanie pierwsze
Nakład: 1000 egz.

**Tłumaczenie:**
GTC AMG Sp. z o.o.

**Przygotowanie do druku, druk i oprawa:**
Agencja Reklamowo-Wydawnicza A. Grzegorczyk

## Spis treści

**Szanowni Państwo,**

Mamy przyjemność zaprezentować Państwu dziewiąty tom serii wydawniczej PARP poświęconej ewaluacji. Naszym zamierzeniem jest, aby tematyka, jaką prezentujemy Państwu już od blisko siedmiu lat w ramach serii Ewaluacja, odpowiadała na najbardziej aktualne wyzwania związane z realizacją polityk publicznych oraz na oczekiwania wobec badań ewaluacyjnych, jakie z tych wyzwań wynikają.

Inspiracją dla tegorocznego wyboru artykułów były zagadnienia poruszane podczas VIII Konferencji Ewaluacyjnej – *Ewaluacja w systemie polityk publicznych* – konferencji, którą PARP miała przyjemność już po raz ósmy współorganizować z Ministerstwem Rozwoju Regionalnego w listopadzie 2012 roku.

Tym razem, na naszych łamach mamy zaszczyt gościć tak wybitne autorytety ewaluacji jak m.in. Michael Scriven, Michael Q. Patton, Alberto Martini. Nasi autorzy zechcieli podzielić się z nami swoim spojrzeniem zarówno na samą dyscyplinę, trendy jakim ewaluacja podlega, jak również przybliżać konkretne rozwiązania z obszaru metodologii i praktyki badań ewaluacyjnych.

Mam nadzieję, że przekazywana Państwu publikacja spotka się z zainteresowaniem czytelników. Tym bardziej, że nadchodzący rok będzie dla badań ewaluacyjnych szczególnym czasem, w którym będziemy podsumowywać efekty interwencji zrealizowanych w perspektywie finansowej 2007-2013, jak również rozpoczniemy realizację programów zaprojektowanych na lata 2014-2020.

Liczę, że kolejny tom serii Ewaluacja wniesie wkład w debatę środowisk zaangażowanych w prowadzenie ewaluacji i zarządzanie interwencjami publicznymi oraz że rozwiązania w nim zawarte będą stanowić inspirację dla Państwa praktyki zawodowej.

Serdecznie zapraszam do lektury.

**Bożena Lublińska-Kasprzak**
**Prezes Polskiej Agencji Rozwoju Przedsiębiorczości**

*Michael Scriven*

# Ewaluacja jako rewolucyjna dyscyplina

## Wprowadzenie

Czasem warto jest z dystansu przyjrzeć się temu, co często określamy jako całościowy obraz, w którym nasza działalność jest tylko niewielkim komponentem. W naszym przypadku oznacza to spojrzenie na całą dziedzinę ewaluacji (na którą składa się około 20 mniejszych obszarów), a nie tylko na zagadnienia, którymi większość z nas się zajmuje, tzn. na analizę polityki, ewaluację programu czy ewaluację pracowniczą. Powinniśmy to zrobić z trzech powodów: (1) dzięki temu możemy dostrzec związki i rozwiązania, których nie widzimy zajmując się tylko problemami z naszego obszaru zainteresowania. Może to także być dla nas (2) źródłem dumy z tego, czym się zajmujemy, kiedy zaczynamy zdawać sobie sprawę ze znaczenia i wartości dobrze przeprowadzonej ewaluacji, co daje nam siłę do obrony przyjętego przez nas podejścia przed typowymi zarzutami dotyczącymi kosztów, rzekomego braku naukowego charakteru czy braku praktycznej użyteczności. Ponadto, możemy dzięki temu poznać (3) ograniczenia tej dyscypliny i nieco ostrożniej podchodzić do często słyszanych uwag na temat „natury ewaluacji".

Możemy skupić się na poszukiwaniu nowych perspektyw, tzn. albo na perspektywie, którą można określić mianem *perspektywy geograficznej/przestrzennej* (gdzie przestrzeń obrazuje mapę dziedzin wiedzy, tzw. dyscyplin), albo na bardziej popularnej i lepiej znanej *perspektywie historycznej/tymczasowej* – kronice lub historii – którą wszyscy pamiętamy z literatury poświęconej historii idei. Zaprezentuję teraz kilka uwag sformułowanych z punktu widzenia każdej z tych perspektyw i mam nadzieję, że zainspirują one czytelnika do określonych reakcji.

## Zakres ewaluacji

Zacznijmy od panoramicznego spojrzenia na zakres geograficzny tego, co moim zdaniem stanowi ugruntowaną obecnie dyscyplinę ewaluacji (Scriven 1994). Zgodnie z definicjami, jakie można znaleźć w obszerniejszych słownikach, co pozwala zwykle uniknąć niejasności, w dyscyplinie tej wykorzystuje się cały szereg zastosowań terminów **„dobry"** i **„zły"** oraz **„prawidłowy"** i **„nieprawidłowy"**, jak również wszystkie terminy, w definicjach których słowa te się pojawiają. Bardziej precyzyjnie i praktycznie rzecz ujmując, ewaluacja odnosi się do określenia **zalet**, **wartości** (ang. *worth*) i **znaczenia**, czyli trzech terminów o mniej więcej podobnym znaczeniu i odpowiednio do takich terminów jak **jakość, wartość** (ang. *value*) i **ważność.**

Czytelnicy, którzy zdobyli wykształcenie w zakresie nauk społecznych, mogą uznać ten obszar za nieuprawiony, ze względu na to, że mieli kontakt z osobami w dalszym ciągu zanurzonymi w sferze pozytywistycznej lub neopozytywistycznej filozofii nauki, zgodnie z którą nauka nie może obejmować twierdzeń oceniających, ponieważ są one (rzekomo) całkowicie subiektywne, nieprecyzyjne i/lub niepoddające się testom. Pogląd ten jest jednak całkowicie nieprawidłowy i został sformułowany w oparciu o powierzchowną analizę języka ewaluacyjnego. O ile to prawda, że czasem język ewaluacji po prostu wyraża kwestie nierozstrzygalnych – często nawet zjadliwych – sporów o gusta (np. w kłótniach o kwestie polityczne, style

w modzie, restauracje czy sztukę współczesną), większość z nich jest sprawdzalna i często obiektywnie prawdziwa – na przykład, pełnią w nauce rolę polegającą na wyrażaniu (a ostatecznie także określeniu) klasyfikacji dobrych i złych hipotez, teorii, instrumentów, jakości danych oraz ich zastosowania w profesjonalnych badaniach oceny jakości prac zaliczeniowych studentów czy też odpowiedzi udzielanych przez nich w testach. Szczegółowe i profesjonalne działania prowadzone w wielu obszarach, w których ewaluacja jest obiektywnie przydatna, doprowadziły do tego, że współcześnie wykształciło się siedem mniejszych, lecz dobrze czytelnikowi znanych, obszarów wchodzących w skład profesjonalnej ewaluacji: ewaluacja programu, analiza polityki, ewaluacja produktu i pracowników, ewaluacja osiągnięć (np. w lekkoatletyce), ewaluacja propozycji i wreszcie ewaluacja portfolio. Ostatnie prace ujawniły szereg wad, jakie występują w niektórych z tych obszarów, można je jednak naprawić i w niektórych krajach zostało to zrobione.

Nazwałem i pokrótce opisałem dwa nowe elementy do „listy siedmiu" – dodając lub opatrując nazwą dwa kolejne rodzaje ewaluacji, na temat których stopniowo przybywa literatury. Są to:

1. metaewaluacja – ewaluacja ewaluacji (1968);
2. ewaluacja interdyscyplinarna – ewaluacja metodologii w ramach dyscypliny (1980). Ewaluator Chris Coryn, na podstawie konkretnego przykładu prowadzonej na szczeblu federalnym ewaluacji wniosków o dofinansowanie badań, zidentyfikował najważniejsze, lecz możliwe do naprawienia, wady drugiego ze wskazanych rodzajów ewaluacji, których koszty można liczyć w milionach.

Ja także w ubiegłym roku zidentyfikowałem i nazwałem kilka kolejnych mniejszych obszarów ewaluacji, które istnieją od dawna (od ponad dwóch tysiącleci), jednak nie były one traktowane jako istotne elementy tej dyscypliny, choć wkrótce, moim zdaniem, zostaną uznane za ważne obszary:

3. ewaluacja z wykorzystaniem mądrości tłumu – z którą spotykamy się na stronach wielu sklepów internetowych czy recenzujących restauracje, lecz zwykle odrzucamy ten typ ewaluacji jako zniekształcony i często wyraźnie powierzchowny. To tak samo, jakbyśmy narzekali na statystyki przytaczając jako argument fakt, że punkty danych nie tworzą gładkiej krzywej. Tak naprawdę ważne jest, czy określone podejście przynosi nam jakieś korzyści, a nie to, czy pasuje do wcześniejszych modeli zgrabnie zaprezentowanych danych. Zaproponowałem pewne zasady umożliwiające wyciągnięcie przydatnych wniosków ewaluacyjnych z typowych dostępnych w Internecie ewaluacji z wykorzystaniem mądrości tłumu. Moim zdaniem, takie lub podobne zasady mogłyby posłużyć do udoskonalenia „otwartej oceny", eksperymentalnego zastosowania ewaluacji z wykorzystaniem mądrości tłumu na potrzeby recenzowania przesyłanych do publikacji artykułów do czasopism[1], istnieje więc możliwość uzyskania zauważalnych korzyści wynikających z poważnego potraktowania tego zjawiska. Dobrze jest także pamiętać, że ewaluacja z wykorzystaniem mądrości tłumu ingeruje w nasze życie od dawna i w niezwykle głęboki sposób – poprzez wybory polityczne, na przykład wybory prezydenta Stanów Zjednoczonych (oraz chociażby wybory do senatu w Starożytnym Rzymie).
4. ewaluacja meta-dyscyplinarna – ewaluacja całych dyscyplin i potencjalnych dyscyplin, np. niedawna ewaluacja nauk sądowych[2] przeprowadzona przez amerykańską Krajową Radę ds. Badań (ang. *US National Research Council*) na wniosek Kongresu. Sporo już zostało zrobione w trzech z tych czterech ważnych obszarów, co przyniosło ogromne oszczędności i udoskonalenia. Na przykład w ciągu trzech lat od opracowania tezy na temat interdyscyplinarności oceny finansowania badań ze środków krajowych, Coryn został poproszony przez właściwe organy rządowe w Nowej Zelandii, Kanadzie, Szwajcarii i Rosji o udoskonalenie krajowych procedur oceny propozycji badań. Działania takie pochłaniają miliardy dolarów na całym świecie, więc udoskonale-

[1] Dziękuję Mitchowi Feldmanowi za zwrócenie na to uwagi.

[2] Ten rodzaj ewaluacji jest wykorzystywany na szeroką skalę w naukach ścisłych i historii (pomyślmy o bojach stoczonych o uznanie dla psychologii czy zaprzestanie traktowania historii jako źródła anegdot), może jednak znaleźć zastosowanie w każdej dyscyplinie, czy przyszłej dyscyplinie, np. komunikacji, grafice, astrologii, psychoanalizie, parapsychologii.

nia w zakresie metodologii ewaluacji mogłyby przynieść oszczędności w postaci milionów dolarów i uratować przed niesłusznym odrzuceniem setki propozycji ważnych badań i technologii.

To jest właśnie aktualny obraz tradycyjnego zakresu profesjonalnej ewaluacji – przynajmniej z mojej perspektywy! Moim zdaniem, osiągnięcia w tych dziedzinach mogą stanowić dowód, że ewaluacja to z pewnością zawód i faktycznie dyscyplina, tj. zasoby wiedzy i praktyki wymagające długich i trudnych studiów z wiarygodną własną koncepcją, odmienną od innych dyscyplin (zwykle z powodu odrębnych metod, ale także i przedmiotu) i do tego ważna dla ogółu wiedzy i społeczeństwa lub przynajmniej ich pewnej istotnej części.

Jednak ewaluacja to znacznie więcej niż wspominane do tej pory dyscypliny akademickie, gdyż każdy z nas dokonuje jej codziennie, często wykorzystując do tego celu zdobywane przez długie lata umiejętności. Pomyślmy na przykład o ewaluacji produktów żywnościowych przeprowadzanej przez doświadczonego szefa kuchni, który przed wschodem słońca wybiera się na targ, żeby dostać świeże owoce, ryby i warzywa potrzebne do przygotowania posiłków w ciągu dnia. Osoba ta jest jedynie przedstawicielem ogromnej grupy profesjonalnych ewaluatorów poruszających się w środowisku nieakademickim: w tym przypadku kucharz staje się prawie paradygmatem, podczas gdy wysoko wykwalifikowany rzeczoznawca majątkowy czy handlarz diamentów stanowią kolejne przykłady na to, że precyzyjna ocena jest niezbędna, a możliwość weryfikacji jest uznawana z mocy prawa. Handlarz diamentów analizuje cenę oszlifowanego i nieoszlifowanego kamienia biorąc pod uwagę 4 aspekty, tj. kolor, szlif, jasność, karat (wagę) i często wystarczą mu na to sekundy[3] – czasem jest nawet w stanie zaryzykować fortunę i własną przyszłość w oparciu o taką ewaluację. To ciekawy przypadek, ponieważ czasami wszystko odbywa się błyskawicznie; chociaż nabycie profesjonalnych umiejętności wymaga lat nauki pod okiem specjalistów, podobnie jak w przypadku umiejętności analityka polityki i ewaluatora programu, czasem na ich zastosowanie w praktyce wystarczą sekundy. Są to w rzeczywistości – ostatecznie – umiejętności percepcyjne, podobnie jak wiele z tych, jakie nabywają myśliwi, rzeczoznawcy z zakresu technologii drewna czy tropiciele. Tej profesjonalnej wiedzy ewaluacyjnej nie można nauczyć się na uniwersytecie, a ponadto większość zdobywanej przez specjalistów wiedzy jest w znacznej części wiedzą nieakademicką. Jednak jej opanowanie może w takim samym stopniu zająć umysł oraz, co bardzo ważne, w dużej mierze jest to wiedza ewaluacyjna. Kucharz uczy się, które owoce są owocami klimakterycznymi (tj. dojrzewają, zyskują na smaku po zebraniu), u kogo kupić dobre melony, po czym poznać, że gruszki są dojrzałe; ewaluator produktu uczy się jak odróżnić dobre dane wejściowe od złych w przypadku ewaluacji z wykorzystaniem wiedzy tłumu oraz jakie są tego ograniczenia. Rozwój nauki przyczynił się zarówno do znacznego przyrostu wiedzy ewaluacyjnej jak i nieewaluacyjnej, jednak ze względu na błędny pogląd, że wiedza ewaluacyjna nie ma „statusu" naukowego, nie była ona rozwijana w taki sposób, jak na przykład rozwijano statystykę, chociaż na początku traktowana była ona jako nauka w dużej mierze podrzędna w stosunku do „dobrej matematyki".

Podsumowując, można zastanowić się nad umieszczeniem profesjonalnych ewaluatorów w tabeli 2x2, z kolumnami opatrzonymi nagłówkami „profesjonaliści" i „amatorzy" i wierszami oznaczonymi nagłówkami „wnioskowanie" i „postrzeganie". Należy skrupulatnie unikać snobizmu intelektualnego, który przejawia się w twierdzeniu, że to, czym zajmują się ewaluatorzy nieakademiccy jest łatwiejsze i nie tak ważne, jak to, czym zajmujemy się my, ewaluatorzy programu opierający się na wnioskowaniu.

Istnieje jeszcze jedna grupa ewaluatorów zawodowych, o której właściwie nie wspomnieliśmy w naszych rozważaniach o zakresie ewaluacji, a jest to grupa godna szacunku, dlatego też należy o niej pamiętać broniąc naszej dyscypliny przed atakami ludzi, którzy wciąż wierzą w twór w rodzaju „nauki wolnej od oceny". Patrząc na elitarną grupę klasycznych dyscyplin – przedmiotów, jakie studiowali ludzie zdobywający wykształcenie kla-

[3] Sekundy, ale w przypadku opinii *negatywnej*; ponieważ doświadczony szlifierz może potrzebować nawet tygodnia, aby dokładnie oszacować prawdziwą wartość potencjalnie doskonałego kamienia.

syczne w imperiach Greków i Rzymian i w późniejszych stuleciach aż do wieku XX – gdy bliżej się im przyjrzymy, okaże się, że trzy z nich mają w dużej mierze charakter ewaluacyjny, i tylko w jednym przypadku ich wiarygodność była kiedykolwiek poważnie zagrożona. Mam tu na myśli logikę, medycynę, matematykę, etykę i inżynierię (przede wszystkim lądową oraz technologię broni). Zastanówmy się nad każdą z nich z osobna.

Logika w połowie poświęca się ocenie argumentów, podczas gdy etyka w połowie skupia się na ewaluacji działań i postaw (chociaż podstawy tej właśnie nauki *były* kwestionowane) i jest w głównej mierze skoncentrowana na ewaluacji działań dotyczących modeli i projektów – podobnie jak kontrola jakości w produkcji oraz medycyna przy diagnozowaniu chorób, działaniach prozdrowotnych i zapobieganiu chorobom. W tych, w ogromnej mierze ewaluacyjnych dyscyplinach nigdy nie brano poważnie pod uwagę doktryny leżącej u podstaw twierdzenia, że nauka jest wolna od wartości, ponieważ zgoda na pogląd, że sądy wartościujące mogłyby być obiektywnie formułowane, doprowadziłaby do ich niemal całkowitego unicestwienia. A te niekwestionowane trzy dyscypliny były poddawane analizom i z sukcesem uprawiane od tysięcy lat, tak więc ich historia obala sceptyczny pogląd leżący u podstaw stanowiska, że nauka jest wolna od wartości.

Doktrynę nauki wolnej od wartości oparto na nieprzemyślanych zaleceniach sformułowanych przez grupę osób z wykształceniem w dziedzinie fizyki, chemii i biologii jako podstawowy element metodologii nauki, zgodny z tym, jak oni ją rozumieli. Ponieważ dorobek tych trzech nauk na przełomie XIX i XX w., czyli dokładnie wtedy, gdy nauki społeczne walczyły o uznanie, był tak ogromny, zrozumiałe jest, że naukowcy zajmujący się naukami społecznymi wybrali te podstawowe elementy fizyki i innych nauk, które leżały u podstaw sukcesu tych nauk i które miały rekomendację Macha i Koła Wiedeńskiego. Zostali oni jednak wprowadzeni w błąd, a skutki były katastrofalne, nie tylko z punktu widzenia metodologii, tj. utrudniając rozwój nauk społecznych, ale także ze względów etycznych, gdyż etyczne aspekty ludzkiego zachowania i myśli zostały wykluczone jako usankcjonowana domena badań i rozwoju. Oznaczało to, że ogromne kwestie polityki zostały zdominowane przez stronnicze a czasem niedojrzałe systemy wartości przyświecające osobistościom politycznym i partiom sprawującym władzę.

Tak, jak należało przypuszczać, kiedy tylko doktryna nauki wolnej od wartości zyskała zwolenników, ugruntowała się i trudno było ją obalić, pomimo dostarczanych przez nauki z większą tradycją i inne obiekty studiów mocnych dowodów na jej niesłuszność, w tym dowodów z wielu nowych obszarów ewaluacji, np. tego, którym my się zajmujemy, z których wszystkie znalazły zastosowanie w licznych sferach, takich jak zdrowie, oświata, budownictwo, działalność wojskowa uzyskując pozytywne i sprawdzalne wyniki. Zamiast tego, echa tej niefortunnej doktryny wciąż rozbrzmiewają w holach budynków wydziałów nauk społecznych w większości kampusów uniwersyteckich, w najnowszych podręcznikach poświęconych stosowanym naukom społecznym nie wspomina się o ewaluacji, chociaż 90% pytań, na które stosowane nauki społeczne próbują znaleźć odpowiedzi, to pytania ewaluacyjne. Godna podziwu wytrwałość biorąc pod uwagę fakt, że doktryna ta ma tak wiele oczywistych wad! Biorąc pod uwagę fakt, że ewaluacja jest elementem każdej nauki, tym bardziej zmusza to do zastanowienia się nad psychologicznymi powodami tego uporu, co czynimy w dalszej części tekstu. Jednak pierwszą kwestią, jaką musimy się zająć, jest błędna teoria wiedzy naukowej, która wspiera doktrynę nauki wolnej od wartości. Jeżeli w pełni nie przebudujemy filozofii nauki, nie uda nam się doprowadzić do tego, aby ewaluacja zyskała należne jej uznanie[4]. Chciałbym jednak skupić się na analizie rozwoju ewaluacji z perspektywy *historycznej* i wynikających z niej ogromnych implikacjach dla wielu interesujących nas dyscyplin.

Proponowana analiza historyczna, chociaż zbyt krótka, by przeprowadzić dowody we wszystkich aspektach, jakie zaproponuję, może wystarczyć, by zachęcić niektórych czytelników do rewizji ogólnego obrazu relacji – pomiędzy poszczególnymi dyscyplinami i wobec ewaluacji – a szczególnie do zrozumie-

[4] Właściwie mam już prawie gotowe rozwiązanie tego problemu (pod nazwą „pragmatyczna filozofia nauki"), niestety wyjaśnianie jej zajęłoby zbyt wiele miejsca.

nia, co mam na myśli mówiąc o szacunku, na jaki zasługuje ewaluacja. Pozwoli to w pewnym stopniu na stworzenie bezpiecznego fundamentu dla naszej dyscypliny, wraz z dalszymi wskazówkami ułatwiającymi uniknięcia pułapki myślenia, że bycie wolnym od wartości to jeden z atrybutów nauki.

Ewaluacja jest procesem kognitywnym. Większość z nas, realizując zawodowe obowiązki ewaluatorów, postrzega ewaluację jako złożony i świadomy proces wnioskowania; jednak, dla niektórych z nas, zarówno współcześnie, jak i dla żyjących w przeszłości gatunków hominidów, był to, i nadal jest, proces percepcyjny zrutynizowany dzięki latom nauki i praktyki. Dowody archeologiczne wskazują na to, że poważna ewaluacja produktu znalazła swoje zastosowanie już milion lat wstecz, gdy obrabiający krzemień stopniowo doskonalili i rozwijali swoje rzemiosło. Możemy jednak wnioskować z dużą dozą prawdopodobieństwa, że proces ten rozpoczął się na długo przed nastaniem epoki kamienia, chociaż prawie nie zachowały się pozostałości tych przedmiotów; istniały drewniane misy i dzidy, chaty kryte strzechą, sieci rybackie i odzież, do tej pory zachowały się niektóre z tych okazów. Nie ma jednak wątpliwości, że prowadzona była na przykład ewaluacja personelu, nie tylko przed nastaniem ery kamiennej, lecz także wśród naszych przodków żyjących *przed pojawieniem się języka*, przecież wybierali oni przywódców i partnerów, a także nauczycieli, których zadaniem było wyposażenie dzieci w umiejętności polowania, łowienia ryb i zbieractwa. Tam, gdzie istniało nauczanie, u podstaw którego z pewnością leżały demonstrowanie, korygowanie błędów i nagradzanie z wykorzystaniem metod psychologicznych bądź fizycznych – przynajmniej milion lat przed pojawieniem się języka mówionego, musiała istnieć ewaluacja, z uwagi na to, że zarówno nauczyciele, rodzice, jak i przywódcy wolą dobre nauczanie od złego. Ponieważ nauczanie dobre od nauczania złego można odróżnić, przynajmniej w pewnym stopniu, na podstawie wyników osiąganych przez uczniów, a ocena taka wymaga ewaluacji ich osiągnięć, istnienie nauczania oznacza istnienie ewaluacji na dwóch poziomach.

Pojawienie się języka jako narzędzia komunikacji dało możliwość snucia planów i składania propozycji, które podlegały oczywiście ewaluacji, co z pewnością nastąpiło tysiące lat przed tak złożonymi projektami inżynieryjnymi, jak Piramidy czy Wielki Mur Chiński. Tak więc pierwsze hominidy rozwiązywały problemy dotyczące przetrwania już 3,5 miliona lat temu, a jednym z najważniejszych narzędzi, które wykorzystywali był kognitywny proces ewaluacji. Czasem jego wynikiem była wiedza wyraźna, czasem wiedza milcząca, jednak bardzo często była to wiedza ewaluacyjna.

Oczywiście istnieje pokusa, by uznać, że wszystkie te działania ewaluacyjne były bardzo prymitywne, im więcej jednak wiemy o dokonaniach naszych przodków, im bardziej świadomi jesteśmy tego, jak trudno jest nam współczesnym przetrwać w tropikach bez nowoczesnych technologii, co pokazują liczne programy telewizyjne typu „reality show", tym bardziej zdajemy sobie sprawę z trudności, jakie musieli pokonywać nasi przodkowie i z tego, jak wiele udało im się osiągnąć. Wydaje się bezsprzeczne, że tysiące lat przed wykształceniem się nauki Homo sapiens zgromadzili ogromne zasoby okupionej ciężką pracą wiedzy, w dużej części pozawerbalnej, ale w znacznym procencie także werbalnej, oraz, że ogromna część z tej wiedzy była kluczowa dla przetrwania. Co więcej, większość z *tej* wiedzy stanowiła *wiedza ewaluacyjna* na temat tego, jak najlepiej wykonywać określone czynności, czyli jak gotować, co najlepiej jeść, a czego unikać, oraz jak unikać pułapek czyhających przy zdobywaniu pożywienia. Była to także wiedza idiosynkratyczna na temat tego, z którym rybakiem najlepiej wypłynąć, po stronie którego wojownika najlepiej stanąć, czy któremu przywódcy się podporządkować; zawierała ona także wiele generalizacji (co najmniej na poziomie regionów) , np. jeśli chodzi o ogólną charakterystykę dojrzałych mango czy jadowitych węży. Naturalnie należało też opanować rozległą wiedzę nieewaluacyjną, np. o najlepszych drogach do terenów łowieckich, czy najlepszych miejscach do zbieractwa, komu się podporządkować, a kogo unikać. Jednak ostatecznie zdobyta z takim trudem bezcenna wiedza podstawowa, która ułatwiała przetrwanie, była w większości *poddającą się weryfikacji wiedzą ewaluacyjną*. Tak więc spojrzenie na historię naszego gatunku dostarcza dodatkowych argumentów przeciwko twierdzeniu, że taka wiedza wyraża jedynie preferencje i gusta, a zatem, zgodnie z tym, co twierdzili pozytywiści, jest wiedzą w pełni subiektywną. Ponadto, udo-

skonalanie wiedzy ewaluacyjnej i umiejętności – nie odrzucanie ich w całości, lecz doskonalenie – jest jedną z istotnych funkcji nauki (i technologii) i to my, ewaluatorzy, jesteśmy naukowcami, którzy tym właśnie się zajmują walcząc z całej siły o uznanie dla ciężkiej pracy naszych przodków.

## Rewolucyjne paradygmaty ewaluacji

Po tym, jak zagłębiliśmy się już w historię ewaluacji, chciałbym nadać jej pewną strukturę poprzez wyodrębnienie okresów zdominowanych przez poszczególne paradygmaty wiedzy ewaluacyjnej. Następnie chciałbym wykorzystać tę strukturę do zastanowienia się nad przyszłością ewaluacji – albo przynajmniej nad różnymi możliwościami jej rozwoju, zależnie od tego, czy uda mi się przekonać czytelnika do mojej opinii na temat kierunku, w którym powinniśmy zmierzać.

**Paradygmat 1. Od około 3.5 milionów lat temu do ok. 1900 roku**

Dominującą rolę odgrywał **paradygmat zdrowego rozsądku,** tj. pogląd, że wiedza ewaluacyjna istnieje, często ma ogromne znaczenie i poddaje się weryfikacji, przynajmniej na równi z wiedzą nieewaluacyjną. Jest częścią wielkiego drzewa poznania.

**Paradygmat 2. Od ok. 1900 roku do ok. 1950 roku**

**Paradygmat ewaluacji bezużytecznej dla nauki/„niedotykalnej".** Pierwsza wielka rewolucja dotycząca koncepcji i statusu ewaluacji nie była dla niej korzystna: pozytywiści wprowadzili doktrynę nauki wolnej od wartości, którą podchwyciły dobrze zapowiadające się nauki społeczne, a za nimi wiele innych (także naukowcy reprezentujący dyscypliny klasyczne). Ruch ten zignorowały jednak cenione tradycyjne dyscypliny naukowe, takie, jak medycyna, inżynieria, logika, przy czym nie były w tej decyzji odosobnione, ze względu na to, że „praktyczny imperatyw", ogromna praktyczna potrzeba dokonywania ewaluacji produktów, polityk, programów, itp. oznaczała, że siedem wymienionych wcześniej dyscyplin musiało radzić sobie same (pojawiając się i osiągając sukcesy w różnych momentach i w zróżnicowanym tempie). I chociaż na przykład guru psychologii odrzucili naukową wartość ewaluacji, w rzeczywistości istniało i dobrze rozwijało się kilka mniejszych obszarów psychologii, które nie brały tego poglądu pod uwagę, chociaż rzadko kwestionowały go publicznie. Na przykład ewaluacja pracowników jest częścią psychologii przemysłowo-organizacyjnej i posiada własne czasopismo naukowe oraz organizuje specjalistyczne spotkania i konferencje. Obiektywna rzeczywistość dowodzi więc, że psychologia cierpi na schizofrenię w zakresie wartości, choć większa grupa psychologów wycina drzewo wiedzy ewaluacyjnej. Wygląda to podobnie we wszystkich naukach społecznych; żadne z wiodących czasopism naukowych z dziedziny nauk społecznych nie dopuszcza publikacji poświęconych ewaluacji ani artykułów, w których do omówienia tematu użyto terminów ewaluacyjnych, chociaż w dalszym ciągu zamieszczane są treści poświęcone ewaluacji interdyscyplinarnej, np. recenzje książek (jeszcze jeden przejaw schizofrenii). Stanowisko to w dużej mierze podtrzymują, choć nie zawsze w tym samym stopniu, badacze reprezentujący nauki fizyczne i biologiczne.

**Paradygmat 3. Od ok. 1950 roku do chwili obecnej**

Kontrrewolucja rozpoczyna się w połowie stulecia: powraca paradygmat zdrowego rozsądku, chociaż tylko na obrzeżach nauk społecznych, najpierw i z największym nasileniem w badaniach edukacyjnych[5]. Za-

[5] Pewien udział w rozpętaniu kontrrewolucji mieli ewaluatorzy produktu, szczególnie Consumer Reports [niezależna amerykańska organizacja testująca produkty na rynku – tłum.], czasopisma motoryzacyjne i audiofilskie. Systematycznie, rok po roku, publikowali oni dobre (choć nie pozbawione wad) ewaluacje produktów, a prawie każdy, kto krytykował możliwość naukowej ewaluacji, z tych wyników korzystał. Z pewnością musiało to powodować pewne napięcia wśród niektórych sceptyków.

wiera on jednak nowy element, ponieważ **w ewaluacji** programów/personelu/polityki **odchodzi się od akceptacji z zasady na rzecz aktywnego a ostatecznie profesjonalnego statusu**. Wiele osób przeprowadza poważne ewaluacje w obrębie siedmiu dyscyplin, a dzięki temu, że ostrożnie korzystają z różnych dobrze znanych metod naukowych, wyników ich badań nie sposób odrzucić w kontekście zaakceptowania paradygmatu zdrowego rozsądku. Jednak zmiana nie jest w żadnym wypadku doprowadzona do końca: cieszący się prestiżem naukowcy zajmujący się naukami społecznymi wciąż głoszą pogląd, że w „prawdziwej nauce", czy „nauce wysokiej jakości" unika się ewaluacji, lub że istnieją ogromne różnice pomiędzy faktami a wartościami, bądź ewaluacją a badaniami, czy ewaluacją a opisem. Są to trzy fałszywe dychotomie będące dowodem na (często nieuświadomione) sprzyjanie doktrynie nauki wolnej od wartości[6]. Ewaluacje te są w znaczniej mierze niekompletne: trudno w nich znaleźć odniesienia do komponentu wartości, np. nie występuje w nich, lub występuje w niewielkim stopniu, ocena wartości *in situ*; brak też jest zaakceptowanego modelu łączenia wartości z danymi nieewaluacyjnymi, co jest konieczne do wyciągnięcia wniosku ewaluacyjnego.

Jednak ta zmiana skutkująca tym, że zastępy profesjonalnych ewaluatorów prowadzą wartościowe prace w dyscyplinie, która do niedawna była zakazana, to nie tylko przejście od nietykalności do profesjonalizmu – nawet jeśli ma ono miejsce jedynie na obrzeżach nauk społecznych – i pierwsze próby wyeliminowania neopozytywistycznej filozofii nauki. To coś jeszcze bardziej rewolucyjnego, chociaż przez ponad pół wieku ukrytego gdzieś pod powierzchnią; coś, z czego w XXI wieku dopiero zaczynamy zdawać sobie sprawę. To przejście od koncepcji ewaluacji jako dodatku do uznanych dyscyplin naukowych – co określaliśmy mianem *dodatkowego zastosowania nauki* – do koncepcji *ewaluacji jako wiodącej nauki spoza głównego nurtu, zarówno teoretycznej jak i stosowanej*. Często mówi się, że za nauką teoretyczną stoi chęć pogłębiania wiedzy, a za nauką stosowaną chęć rozwiązywania praktycznych problemów. Pogląd ten budzi kontrowersje, ale ogólnie uznaje się, że w jakimś stopniu jest prawdziwy, tzn. wyraża przynajmniej część prawdy. W fazie Paradygmatu 3 stworzono podstawy do traktowania ewaluacji jako dyscypliny obejmującej obie sfery nauki. Dla wielu ewaluatorów, dążenie do dokonywania ewaluacji to czasami po prostu dążenie do odkrywania prawdy o świecie, *w wymiarze ewaluacyjnym*, np. najlepszego i drugiego w kolejności, rzeczywistego bądź możliwego, X.

Przejawiało się to w wysypie, począwszy od lat 50 XX w., kilkunastu teorii ewaluacji, czego koszty ponosiła ewaluacja stosowana. Ludzie ci chcieli wiedzieć, czym *jest* ewaluacja, w jaki sposób można ją *udoskonalić* i jak *najlepiej o niej pisać*. Natomiast, jeśli chodzi o ewaluację stosowaną, przejawiało się to w części raportu ewaluacyjnego zawierającej rekomendacje, w której ewaluator starał się stworzyć coś lepszego od tego, co było przedmiotem ewaluacji. Dla nauk politycznych, na przykład, miało to następujące konsekwencje. Podczas, gdy przed rokiem 1950 specjalista w zakresie nauk społecznych musiał ograniczyć się do badania rzeczywistego funkcjonowania różnych wcieleń poszczególnych form rządów, po roku 1950 *mógł* już pytać, które z nich funkcjonują lepiej bądź najlepiej, w jakich okolicznościach, zamiast stwierdzić „Cóż, to czysto filozoficzne pytanie", co niezwykle często zdarzało się przed rokiem 1950 i zdarza się także i obecnie – tym, którzy od 1950 r. wciąż nie obudzili się ze snu. Innymi słowy, pytania ewaluacyjne są obecnie uzasadnione *w obszarze* nauk społecznych, choć nikt specjalnie nie pali się, by na nie odpowiadać.

Istnieją zatem ewaluatorzy, którzy przesuwają granice nauki – tworząc przy tym drzewo wiedzy ewaluacyjnej – podobnie, jak istnieją inni naukowcy, którzy tworzą drzewo nieewaluacyjne. Zauważyć należy, że ewaluator w poszukiwaniu prawdy często sięga *dalej* niż naukowcy nieewaluacyjni prowadzący badania przed 1950 rokiem: w momencie, gdy gotowy jest już nieewaluacyjny opis wymiarów i działania przed-

[6] Zdroworozsądkowy pogląd, że te kategorie się zazębiają jest z pewnością słuszny: niezaprzeczalnym faktem jest to, że Einstein był wyjątkowo dobrym fizykiem, że ewaluacja Head Start wymagała przeprowadzenia wielu badań oraz, że użycie w opisie podejrzanego w liście gończym sformułowania „dobrze ubrany" nie jest sprzecznością samą w sobie. Nawet takie kategorie jak „twierdzenia ewaluacyjne" i „twierdzenia empiryczne" wzajemnie się nie wykluczają, ponieważ całkiem możliwe, że faktem empirycznym jest to, że ponad milion Kalifornijczyków jest dwujęzycznych lub „dobrze radzi sobie z obsługą broni z rodziny AR-15".

miotu badań, pozostaje ważne i bardzo ciekawe pytanie o to, na ile *dobrze lub źle* przedmiot ten realizuje zadania, na potrzeby których został lub nie został zaprojektowany. Ponadto, pozostaje także bardziej podstawowe pytanie, w niektórych przypadkach właściwe, dotyczące tego, czy to, co dany przedmiot badań czyni, jest samo w sobie ogólnie dobre czy złe: na przykład, czy można znaleźć uzasadnienie dla wojny (stosowania tortur, aborcji, czy kary śmierci). Prowadzi to do pytania, ważnego dla walidacji wartości w ramach ewaluacji, które stanowi główny przedmiot omawianego poniżej Paradygmatu 6.

**Paradygmat 4. Od ok. 1990 roku do chwili obecnej i w przyszłości**

Zaczyna się krystalizować koncepcja **ewaluacji jako superdyscypliny**, co stanowi podwójny skok, po pierwsze, przejście od statusu zawodu do statusu dyscypliny, dziedziny badań, którą cechują wyraźnie zaznaczone granice i przedmiot badań, zasadność przyjętych metodologii i koncepcji oraz jej znaczenie społeczne i intelektualne.

Ponieważ zagadnienia te były nieustannie omawiane na przestrzeni XX w., wykształciły się pewne szczególne cechy określające rolę ewaluacji na tle innych dyscyplin, co przyczyniło się do drugiego skoku. Jedną z tych cech jest idea ewaluacji jako *transdyscypliny*, tj. jako jednej z niewielu dyscyplin, łączących statystykę z komunikacją, które dostarczają narzędzi innym dyscyplinom, lecz są także dyscyplinami samodzielnymi i autonomicznymi (Scriven 1991). Unikatową cechą ewaluacji jest to, że stanowi ona zasadniczy element *każdej* innej dyscypliny, włącznie z dyscyplinami fizycznymi, jak gimnastyka, balet, trening maratoński, ponieważ każda dyscyplina, z definicji, dysponuje zestawem norm określających jakość danych, zasadność wnioskowania, dopuszczenie do publikacji, istotność dla różnych wyróżnień, itp., przy czym normy te muszą być zgodne z wymaganiami obowiązującymi dla ewaluacji. Sytuację tę można opisać twierdzeniem, że **ewaluacja jest dyscypliną alfa**, tj. na pierwszy rzut oka widać, że jest właścicielem badań nad kontrolą jakości prowadzonych we wszystkich dyscyplinach. Potrzeba, aby rolę tę pełniła właśnie ewaluacja wynika stąd, że: (1) badania poświęcone niepewnej sytuacji recenzowania – kluczowemu mechanizmowi kontroli jakości we wszystkich naukach i wielu innych dyscyplinach – niewątpliwe wskazują, że stanowi ono chwiejną podstawę dla kontroli jakości, nawet przy zastosowaniu najłagodniejszych norm[7]; a (2) głośne skandale, np. w anestezjologii i naukach sądowych dowodzą, że w poszczególnych naukach nie korzysta się z nawet najprostszych systemów kontroli jakości (np. brak kontroli służących wykrywaniu i zapobieganiu oszustwom) czy systemów rozproszenia funduszy. Systemy kontroli jakości wykorzystywane we wszystkich dyscyplinach powinny być po prostu traktowane jak stosowany aspekt ewaluacji. Ewaluatorzy we współpracy z wiodącymi naukowcami podjęli już prace nad udoskonaleniem tych systemów. Krótko mówiąc, ewaluacja to skarbnik posiadający klucze do królestwa dyscyplin, stąd termin „dyscyplina alfa".

Takie podejście jest przez wielu naukowców traktowane jak inwazja na ich terytorium, co można uznać za przynajmniej rewolucyjne, jednak transdyscyplinarna rola ewaluacji oznacza dopełnianie, a nie dominację. Oczywiście, teraz sprawia to wrażenie inwazji na królestwa poszczególnych dyscyplin, jednak tylko dlatego, że nie zajmowały się one do tej pory na poważnie kontrolą jakości. Obecne problemy angażują prymitywną metodologię ewaluacji, o czym wie każdy naukowiec – np. oszustw i recenzowania (w tym otwartego recenzowania) – niemniej jednak, ewaluatorzy dysponują dużą wiedzą na temat trików w procesie ewaluacji i mogą służyć pomocą przy doskonaleniu innych aspektów kontroli jakości obejmujących dyscyplinę, np. kwestię wprowadzenia korzyści społecznych i intelektualnych do rachunku ewaluacji projektów badawczych i ich wyników w postaci nowych teorii. Oczywiście, mogą jedynie zająć się tym, na co mają wpływ; wszelkie rozwiązania będą wymagały współpracy ze strony ekspertów danego przedmiotu.

[7] W stosunku do liczby odrębnych dziedzin badań i biorąc pod uwagę fakt, że proces powszechnej kontroli przebiega wyjątkowo powoli, bardzo niewiele badań poświęca się recenzowaniu.

**Paradygmat 5. 2012 rok do chwili obecnej**

Spojrzenie na współczesną scenę nauk społecznych pozwala z dużą dozą prawdopodobieństwa stwierdzić, że wiele z nich (siedem dyscyplin i nie tylko) rzeczywiście prowadzi ewaluację, nawet jeśli nie jest ona idealna, chociaż ważne badania rzadko takie są. Chciałbym w tym miejscu zaryzykować twierdzenie, że prawie wszystkie badania stosowane w nauce prowadzone są, by znaleźć odpowiedzi na pytania ewaluacyjne – o najlepszy lek, najlepszą formę nauczania matematyki oraz o to, czy program ubezpieczeń społecznych lub szybki pociąg pasażerski nie obciążają zbytnio budżetu państwa, itp. I to właśnie w tej grze ekspertami jest 7 dyscyplin. Dlatego też, w sytuacji, gdy na ośrodki badawcze, w tym najważniejsze uczelnie, wywiera się coraz większą presję, aby uzasadniały wydatki na badania podając praktyczne korzyści, naukowcy prowadzący badania stosowane w dziedzinie nauk społecznych powinni posługiwać się najbardziej odpowiednią metodologią, tj. tą samą, która wykorzystywana jest na potrzeby ewaluacji prowadzonej na obrzeżach nauk społecznych. Miejsce tej metodologii jest w naukach z głównego nurtu, nie tych działających na obrzeżach. Ufam więc, że Paradygmat 5 będzie charakteryzował nauki społeczne w przyszłych latach: tzn. **ewaluacja stanie się dominującym modelem w dyscyplinach naukowych – dyscypliną przykładową – dla nauk stosowanych**.

Jedną z trudności w realizacji tego zalecenia będzie znalezienie wystarczająco mocnych argumentów, aby przekonać liderów badań stosowanych (szczególnie w stosowanych naukach społecznych), że jest to konieczne. Jeżeli to się nie uda, staną się oni pozbawionymi wiarygodności kandydatami do otrzymania funduszy publicznych. Kolejny problem to wyjaśnienie modelu metodologicznego, jaki stosuje się w dobrych badaniach ewaluacyjnych – narzędzi, które wnosi – a w szczególności procesów oceny, klasyfikowania oraz integrowania wartości z danymi nieewaluacyjnymi.

Oczywiście pomysł, że nauki społeczne – i inne dyscypliny stosowane – powinny zmienić stosunek do dyscypliny, którą od dawna uważały za bezwartościową i zacząć ją doceniać, czy nawet szanować, jest rewolucyjny, więc prawdopodobnie zajmie to jakieś 50 lat, czyli mniej więcej tyle samo, przez ile dyscyplina ta była niezasłużenie odrzucana. Trzeba to jednak zrobić, bo, w przeciwnym razie, akademickie nauki społeczne w obecnym kształcie staną się zabytkiem muzealnym.

**Paradygmat 6. Po roku 2012**

Kwestie poruszane w poprzednim akapicie prowadzą nas do ostatniej zmiany paradygmatu. Nie można uniknąć problemu wyraźnego wkomponowania kwestii etycznych w ewaluację, ponieważ stanowią one wartość alfa[8] – czyli, jak wszyscy uważamy (albo przynajmniej gołosłownie twierdzimy), przebijającą wszystkie inne wartości w przypadku konfliktu. W wielu przypadkach w toku prowadzenia ewaluacji nie pojawiają się problemy etyczne, jednak nie zawsze tak jest. A ponieważ ewaluacja zyskała miano dyscypliny alfa, a także dyscypliny przykładowej dla nauk stosowanych, musi przyjąć na siebie[9] rolę **głównego badacza wartości alfa.** Kiedy tylko podejście oparte na roli alfa się upowszechni, uwaga będzie musiała zostać skupiona na potrzebie kontrolowania bezsprzecznie słusznego dążenia do wrażliwości kulturowej, tak, aby nie doszło do relatywizmu etycznego. Ewaluacja musi podejmować się nie tylko oceny wartości etycznych, tak samo, jak wszystkich innych wartości, na których się opiera, ale także uzasadnić konieczność ich priorytetowego traktowania w przypadku pojawienia się konfliktów pomiędzy wartościami. Poprzedni domniemani właściciele etyki oraz ci, którzy się jej wyrzekli, np. religie i naukowcy neopozytywistyczni, uznają ten paradygmat za niedopuszczalne przesunięcie władzy, chociaż od dawna było już wiadomo, że stracili oni swoją pozycję w wyniku

[8] A dokładniej „aksjomat etyczny" dotyczący tego, że pozornie wszystkim przysługują równe prawa. Pozostałe liczne prawdy, które można znaleźć w zaleceniach moralnych każdej wiary wywodzą się od tego twierdzenia, które w takiej czy innej formie można znaleźć w każdym kodeksie wraz z innymi nieewaluacyjnymi założeniami dotyczącymi preferencji kulturalnych, takich jak instytucja własności prywatnej czy monogamia.

[9] Naszkicowałem dowód, który moim zdaniem został potwierdzony w rozdziale poświęconym etyce zawartym w publikacji mojego autorstwa *Primary Philosophy*, (McGraw-Hill 1966).

postępu, jaki dokonał się w teorii gier, teorii ewolucji, psychologii porównawczej, krytyce teologii, kosmologii, metaetyce i logice ewaluacji, tak więc rewolucja ta wydaje się być spóźniona. To, że we wspólnej spuściźnie intelektualnej nadal brak uzasadnienia dla etyki jako wartości alfa stanowi prawdopodobnie największe zaniedbanie intelektualne, które szybko należy nadrobić. Zebranie tych siedmiu komponentów razem jako dowodów będzie wymagało ogromnych multidyscyplinarnych starań, ale są szanse na powodzenie.

## Podsumowanie

Tę krótką wędrówkę po stosunkowo szybko zmieniających się frontach walki w historii idei chciałbym zamknąć następującą refleksją. Podejrzewam, że najpowszechniejszą reakcją na przedstawione w tej pracy podejście będą głosy, iż próbuje się tu nadać przesadnie duże znaczenie stosunkowo nowej dyscyplinie. Moja odpowiedź brzmi tak: to nauka popełniła grzech pychy w tym sporze. Po pierwsze, wadliwa nauka doprowadziła naukowców do odrzucenia ewaluacji pod zarzutem, że wyraża ona jedynie preferencje, co jest absurdalnym wnioskiem, biorąc pod uwagę fakt, że naukowcy zawsze prowadzili ewaluację komponentów swoich nauk oraz to, że w każdym podręczniku z dziedziny archeologii mogli znaleźć owoce dwóch milionów lat ciężkiej pracy poświęconej na gromadzenie wiedzy ewaluacyjnej, co dowodzi, że nie mieli racji. A ponieważ dzisiaj w nauce dzieje się jeszcze gorzej, zmuszeni jesteśmy wykorzystywać ewaluację, aby, pomimo błędów popełnionych w zakresie zapobiegania oszustwom i niedbalstwa, nie ucierpiało dobro nauki. Dlatego też, nie tylko uważam, że oddanie ewaluacji należnej jej pozycji jest naszym obowiązkiem, ale że do powstania tego obowiązku przyczynili się ci, którzy skarżą się teraz na tych, co powiewają sztandarem ewaluacji chcąc przyciągnąć prawdziwych naukowców, osoby, które są naprawdę zainteresowane poszukiwaniem prawdy, a nie tylko utrzymaniem obecnej wysokiej, lecz nieuzasadnionej pozycji.

Wniosek jest taki, że te rewolucyjne działania dają jedyną nadzieję na ratunek przed negatywnymi skutkami wadliwej nauki.

---

**Dr Michael Scriven** ukończył studia licencjackie w dziedzinie matematyki oraz magisterskie w dziedzinie matematyki i filozofii na Uniwersytecie w Melbourne. Posiada tytuł doktora filozofii Uniwersytetu w Oxfordzie. Dr Scriven obecnie pełni funkcję profesora psychologii w Claremont Graduate University oraz starszego adiunkta na Western Michigan University. Jest wykładowcą akademickim od ponad 60 lat. Wykładał na takich uczelniach jak Berkeley, Minnesota, Swarthmore College, Western Michigan w USA, Uniwersytecie w Auckland w Nowej Zelandii, University of Western Australia oraz Stanford i Harvard. Dr Scriven opublikował ponad 450 artykułów naukowych z 11 różnych dyscyplin, z czego około 100 dotyczyło ewaluacji, wydał kilka książek, a także był członkiem ponad 40 redakcji i komitetów redakcyjnych.

## Bibliografia


- Scriven M., *Evaluation as a discipline*, [w:] Elsevier, Summer (red.), *Studies in Educational Evaluation*, 1994, s. 147-166.
- Scriven M., *Evaluation Thesaurus*, Sage, Newbury Park, Ca 1991, wydanie 4.

*Michael Quinn Patton*

# Ewaluacja skoncentrowana na wykorzystaniu

Ewaluacja skoncentrowana na wykorzystaniu, podejście z 35-letnią historią, ma na celu zwiększenie potencjalnej użyteczności i faktycznego wykorzystania badań ewaluacyjnych. Ewaluacja skoncentrowana na wykorzystaniu opiera się na badaniu czynników wpływających na charakter i zakres wykorzystania ewaluacji. Początkowe badania skupiały się na wykorzystaniu dużych ewaluacji prowadzonych w sektorze zdrowia publicznego (Patton 1978). Kolejne badania nad wykorzystaniem badań ewaluacyjnych przeprowadzone w późniejszych latach poszerzyły nasze rozumienie czynników decydujących o użyteczności i doprowadziły do tego, że wykorzystanie zaczęto pojmować jako złożony proces, który powinien opierać się na myśleniu systemowym i w takim myśleniu być zakorzeniony (Patton 2012). Niniejszy artykuł zawiera charakterystykę ewaluacji skoncentrowanej na wykorzystaniu i stanowi wkład w istotną działalność ewaluacyjną Polskiej Agencji Rozwoju Przedsiębiorczości w zakresie upowszechniania głównych nurtów w teorii i praktyce ewaluacji.

Ewaluacja skoncentrowana na wykorzystaniu opiera się na założeniu, że badania ewaluacyjne należy oceniać pod kątem ich użyteczności i faktycznego wykorzystania, w związku z czym ewaluatorzy powinni, planując proces ewaluacji, zwracać uwagę na to, w jaki sposób wszystkie realizowane zadania, od początku do końca procesu, będą wpływały na wykorzystanie badania. Ewaluacja skoncentrowana na wykorzystaniu jest związana z tym, jak ludzie doświadczają procesu ewaluacji i w jaki sposób aplikują wyniki ewaluacji w rzeczywistym świecie. W ewaluacji skoncentrowanej na wykorzystaniu nacisk kładzie się więc na **zamierzone wykorzystanie badań przez docelowych użytkowników.**

## Nacisk na zamierzone wykorzystanie badań przez docelowych użytkowników

W każdej ewaluacji występuje wiele potencjalnych interesariuszy i całe spektrum możliwych sposobów wykorzystania badania. Ewaluacja skoncentrowana na wykorzystaniu wymaga przejścia od elementów ogólnikowych i abstrakcyjnych, tj. możliwych odbiorców i potencjalnych sposobów wykorzystania, do elementów rzeczywistych i konkretnych: faktycznych głównych użytkowników docelowych i ich wyraźnie sprecyzowanego zobowiązania do przyjęcia konkretnych, ściśle określonych sposobów wykorzystania. Ewaluator umożliwia przede wszystkim dokonanie oceny i podjęcie decyzji przez docelowych użytkowników, nie działa natomiast jako odległy, niezależny sędzia. Jako że żadna ewaluacja nie może być pozbawiona wartości, ewaluacja skoncentrowana na wykorzystaniu odpowiada na pytanie o to, na czyich wartościach ewaluacja ma się opierać, poprzez współpracę z jasno określonymi głównymi użytkownikami docelowymi, odpowiadającymi za wykorzystywanie wyników ewaluacji i wdrażanie rekomendacji. Krótko mówiąc, ewaluacja skoncentrowana na wykorzystaniu opiera się na zrozumieniu, że wykorzystanie ewaluacji jest zbyt ważne, aby jedynie założyć lub mieć nadzieję, że zostanie ona wykorzystana. Wykorzystanie musi być zaplanowane i umożliwione.

Ewaluacja skoncentrowana na wykorzystaniu jest kwestią wysoce indywidualną i zależną od sytuacji. Ewaluator nawiązuje relację zawodową z docelowymi użytkownikami, aby pomóc im w ustaleniu, jakiego rodzaju ewaluacji potrzebują. Wymaga to negocjacji, podczas których ewaluator oferuje wachlarz możliwości. Ewaluacja skoncentrowana na wykorzystaniu nie faworyzuje ani nie zależy od żadnego kon-

kretnego zagadnienia, modelu, metody, teorii czy nawet wykorzystania. Stanowi ona raczej proces, którego celem jest pomoc głównym użytkownikom docelowym w wyborze zagadnień, modelu, metod, teorii i sposobów wykorzystania najbardziej odpowiednich w ich konkretnej sytuacji. Tym interaktywnym procesem zachodzącym między ewaluatorem a głównymi użytkownikami docelowymi kieruje umiejętność odpowiedniego reagowania w danej sytuacji. Obecnie mamy do dyspozycji wiele możliwości w tej obszernej dziedzinie, jaką stała się ewaluacja. Rozpatrując szeroki i zróżnicowany wachlarz możliwości w zakresie ewaluacji, można stwierdzić, że ewaluacja skoncentrowana na wykorzystaniu może mieć dowolną funkcję (formatywną, podsumowującą, rozwojową), korzystać z dowolnego rodzaju danych (ilościowych, jakościowych, mieszanych), czy dowolnego modelu (np. naturalistycznego, eksperymentalnego) oraz badać dowolny przedmiot (m.in. procesy, rezultaty, efekty, koszty, stosunek korzyści do kosztów). Ewaluacja skoncentrowana na wykorzystaniu stanowi proces, którego celem jest podjęcie decyzji we wspomnianych kwestiach, we współpracy z określoną grupą głównych użytkowników, poprzez skupienie się na ich zamierzonym wykorzystaniu ewaluacji.

Fundamentem i uzasadnieniem ewaluacji skoncentrowanej na wykorzystaniu jest psychologia wykorzystywania. Zasadniczo z badań nad wykorzystaniem ewaluacji (Patton 2008) wynika, że użytkownicy docelowi są bardziej skłonni wykorzystać badania ewaluacyjne wtedy, gdy rozumieją proces ewaluacji i jego efekty oraz czują się za ten proces odpowiedzialni. Z większym prawdopodobieństwem będą rozumieć i czuć się odpowiedzialni za ten proces, jeżeli będą w nim czynnie uczestniczyć, zaś poprzez czynne angażowanie głównych użytkowników docelowych ewaluator przy okazji szkoli użytkowników w zakresie wykorzystania, przygotowując fundamenty tego użycia i zwiększając docelową użyteczność ewaluacji. Choć siłą napędową ewaluacji skoncentrowanej na wykorzystaniu jest niepewność co do jej użyteczności, ewaluator musi również uwzględnić kwestię dokładności ewaluacji, jej wykonalności, stosowności i zakresu odpowiedzialności (AEA 2012; Yarbrough, Shulha, Hopson i Caruthers 2011).

## Podstawowe definicje

Ewaluacja programu polega na systematycznym gromadzeniu informacji o działaniach, cechach charakterystycznych i wynikach programów w celu sformułowania opinii o programie, zwiększeniu jego efektywności lub podjęciu świadomych decyzji o przyszłych programach. Ewaluacja programu skoncentrowana na wykorzystaniu (w odróżnieniu od ogólnie pojętej ewaluacji programu) to ewaluacja dokonywana na rzecz określonych głównych użytkowników docelowych i wspólnie z nimi, odnosząca się do określonego zamierzonego wykorzystania.

Podana wyżej definicja ogólna zawiera trzy wzajemnie powiązane elementy:

1) systematyczne gromadzenie informacji o
2) potencjalnie szerokim zakresie tematów w odniesieniu do
3) wielu możliwych ocen i sposobów wykorzystania.

Definicja ewaluacji skoncentrowanej na wykorzystaniu zawiera dodatkowo konieczność określenia zamierzonego wykorzystania przez użytkowników docelowych. Kwestia definicji ewaluacji ma duże znaczenie, ponieważ różne podejścia opierają się na różnych definicjach. Zaproponowana powyżej definicja oparta na wykorzystaniu znacząco przeciwstawia się innym podejściom, w których ewaluację określa się jako pomiar stopnia osiągnięcia celu przy zapewnieniu poczucia odpowiedzialności lub podkreśla się stosowanie metodologii nauk społecznych w celu dokonania oceny skuteczności programu.

## Włączenie użytkowników docelowych w proces podejmowania decyzji ewaluacyjnych: czynnik ludzki

W każdej ewaluacji trzeba podjąć wiele decyzji. Koniecznie należy ustalić cel ewaluacji. Zazwyczaj niezbędne jest wyznaczenie konkretnych kryteriów ewaluacyjnych w celu dokonania oceny realizacji programu. Trzeba dokonać wyboru metod i uzgodnić ramy czasowe. Wszystkie wyżej wymienione działania są istotnymi kwestiami w każdej ewaluacji. Pytanie brzmi: Kto podejmie decyzję w tych kwestiach? Odpowiedź skupiająca się na wykorzystaniu brzmi: *główni docelowi użytkownicy ewaluacji.*

Jasne i precyzyjne wskazanie osób mogących skorzystać na ewaluacji jest kwestią tak istotną, że ewaluatorzy wprowadzili specjalne określenie na potencjalnych użytkowników ewaluacji: *interesariusze.* Interesariusze ewaluacji to osoby mające swój interes w wynikach ewaluacji. W każdej ewaluacji występuje wielu potencjalnych interesariuszy: podmioty finansujące program, zespół wdrażający program, administratorzy oraz klienci lub uczestnicy programu. Za interesariuszy można także uznać inne osoby mające bezpośredni lub nawet pośredni interes w skuteczności programu, np. dziennikarzy i przedstawicieli opinii publicznej lub też konkretniej podatników – w przypadku programów publicznych. Interesariuszem jest każdy, kto podejmuje decyzje lub chce uzyskać informacje dotyczące programu. Należy jednak pamiętać, że interesariusze mają zazwyczaj różne, często sprzeczne interesy. Żadna ewaluacja nie jest w stanie dostarczyć odpowiedzi na wszystkie potencjalne pytania równie skutecznie. Oznacza to, że, w celu ukierunkowania ewaluacji, należy poświęcić część procesu na zawężenie zestawu możliwych pytań. W przypadku ewaluacji programu skoncentrowanej na wykorzystaniu, proces ten rozpoczyna się od zawężenia grupy potencjalnych interesariuszy do znacznie mniejszej, ściśle określonej grupy głównych użytkowników docelowych. Ich potrzeby informacyjne, tj. ich zamierzone sposoby wykorzystania, fokusują ewaluację.

Różni ludzie inaczej postrzegają te same rzeczy, mają różne interesy i potrzeby. Stwierdzenie to można uznać za oczywiste. Chodzi jednak o to, że ta oczywistość jest stale ignorowana podczas projektowania badań ewaluacyjnych. Ukierunkowanie ewaluacji na potrzeby informacyjne danej osoby lub grupy określonych, współzależnych i współpracujących osób jest czymś zupełnie innym od tego, co tradycyjnie zaleca się jako „określenie odbiorców" ewaluacji. Odbiorcy to jednostki amorficzne, anonimowe. Nie wystarczy też wskazanie organu lub organizacji jako odbiorcy raportu z ewaluacji. Organizacje są bezosobowym zbiorem hierarchicznie uporządkowanych stanowisk. To ludzie, a nie organizacje, wykorzystują informacje wynikające z ewaluacji, stąd istotne znaczenie czynnika ludzkiego.

Czynnik ludzki to obecność możliwej do określenia osoby lub grupy osób, którym osobiście zależy na ewaluacji i jej wynikach. Z badań nad wykorzystaniem ewaluacji (Patton 2008) wynika, że w przypadku, gdy w ewaluacji aktywnie uczestniczy określona zainteresowana osoba lub grupa, zachodzi większe prawdopodobieństwo, że ewaluacja zostanie wykorzystana. Tam, gdzie czynnik ludzki nie występował, zauważono wyraźny brak wpływu ewaluacji.

Czynnik ludzki jest wyrazem roli przywódczej, zainteresowania, entuzjazmu, determinacji, zaangażowania, asertywności i troski ściśle określonych osób. Są to osoby intensywnie poszukujące informacji w celu dokonywania ocen i zmniejszenia niepewności związanej z podejmowaniem decyzji. Osoby te chcą zwiększyć swoją zdolność przewidywania rezultatów działań w ramach programu, przez co zwiększają swój poziom świadomości jako osób podejmujących decyzje, tworzących polityki, konsumentów, uczestników programu, podmiotów finansujących lub osób odgrywających inne role. Są to główni użytkownicy ewaluacji.

Choć każdy przypadek jest inny, wyraźnie można zaobserwować pewien wzór: tam, gdzie pojawia się czynnik ludzki i pewne osoby przyjmują bezpośrednią, indywidualną odpowiedzialność za przekazanie wyników odpowiednim osobom, ewaluacja przynosi efekty. Tam, gdzie czynnik ludzki nie występuje, zauważa się wyraźny brak wpływu ewaluacji. Wykorzystanie nie zależy od jakiejś konfiguracji abstrakcyjnych czynników, w dużej mierze wykorzystanie jest zależne od prawdziwych, żyjących i przejmujących się istot ludzkich.

„Nic bardziej nie wpływa na wykorzystanie ewaluacji niż czynnik ludzki – interes urzędników w czerpaniu wiedzy z ewaluacji oraz pragnienie ewaluatora, aby zwrócić uwagę na to, co wie" (Cronbach i in. 1980, s. 6).

Znaczenie czynnika ludzkiego w wyjaśnianiu i przewidywaniu wykorzystania ewaluacji prowadzi bezpośrednio do zwrócenia – w ramach ewaluacji skoncentrowanej na wykorzystaniu – szczególnej uwagi na pracę z docelowymi użytkownikami w celu określenia zamierzonego wykorzystania. Ze względu na czynnik ludzki zwracamy się do określonych osób rozumiejących i doceniających ewaluację i takich, którym na niej zależy oraz zwracamy uwagę na to, czym są zainteresowani. Dla branży jest to najważniejszy wniosek w zakresie zwiększania wykorzystania; to wiedza obecnie powszechnie uznawana przez praktykujących ewaluatorów (Cousins i in. 1996; Preskill i Caracelli 1997).

## Skoncentrowanie na użytkowniku

Zasadniczo, ewaluacja skoncentrowana na wykorzystaniu skupia się na użytkowniku (Alkin 1995). Ponieważ żadna ewaluacja nie może służyć wszystkim interesom potencjalnych interesariuszy w równym stopniu, w ewaluacji skoncentrowanej na wykorzystaniu jednoznacznie określa się, czyim interesom ona służy – interesom ściśle określonej grupy głównych użytkowników docelowych.

Zwrócenie się do głównych użytkowników docelowych nie jest tylko i wyłącznie zadaniem akademickim, wykonywanym tylko dla samej świadomości wykonania. Włączenie określonych osób, które mogą i chcą wykorzystywać informacje, umożliwia im określenie kierunku ewaluacji, zaangażowanie się w nią oraz wzięcie za nią odpowiedzialności na wszystkich jej etapach, od zainicjowania badania, przez etap projektowania i gromadzenia danych, aż po raport końcowy i proces rozpowszechniania wyników. Jeżeli decydenci wykazywali niewielkie zainteresowanie badaniem na jego początkowych etapach, raczej nie zainteresują się nagle wykorzystaniem wyników po jego zakończeniu. Nie będą wystarczająco przygotowani do ich wykorzystania.

## Etapy procesu ewaluacji skoncentrowanej na wykorzystaniu

Po pierwsze, określa się grupę docelowych użytkowników. Docelowych użytkowników łączy się ze sobą lub w jakiś sposób organizuje, o ile jest to możliwe (np. tworzy się grupę zadaniową ds. ewaluacji, której członkami są główni interesariusze), aby współpracowali z ewaluatorem i wspólnie podejmowali kluczowe decyzje w sprawie ewaluacji.

Po drugie, ewaluator i docelowi użytkownicy zobowiązują się do zamierzonego wykorzystania ewaluacji oraz ustalają, jaką funkcję ma pełnić ewaluacja, np. formatywną, podsumowującą lub generowania i dostarczania wiedzy. Priorytetyzowanie pytań ewaluacyjnych często obejmuje względne znaczenie skupienia się na osiągnięciu celów, wdrożeniu programu i/lub teorii działania programu (modelu logicznego). Zestaw możliwości ewaluacyjnych jest szeroki, a więc może okazać się konieczne omówienie wielu różnych typów ewaluacji. Ewaluator współpracuje z docelowymi użytkownikami przy ustalaniu priorytetowych sposobów wykorzystania ewaluacji, uwzględniając względy polityczne i etyczne. Poprzez interakcję i uwzględnienie indywidualnych przypadków ewaluator pomaga docelowym użytkownikom w udzieleniu odpowiedzi na pytanie: uwzględniając oczekiwane sposoby użycia, czy warto jest przeprowadzić ewaluację? Do jakiego stopnia i w jaki sposób docelowi użytkownicy zobowiązują się do zamierzonego wykorzystania ewaluacji?

Trzeci, ogólny etap procesu obejmuje decyzje w sprawie metod, pomiaru i projektu ewaluacji. Główni użytkownicy docelowi biorą udział w podejmowaniu decyzji w sprawie metod, aby dokładnie rozumieli

mocne i słabe strony wyników, które będą wykorzystywać. Można tutaj uwzględnić różne opcje: dane jakościowe i ilościowe, model naturalistyczny, eksperymentalny i quasi-eksperymentalny, celowe i probabilistyczne podejścia do doboru próby, większy lub mniejszy nacisk na uogólnienia oraz alternatywne sposoby rozwiązywania problemu potencjalnych zagrożeń dla aktualności, rzetelności i użyteczności. Ściślej mówiąc, dyskusja na tym etapie będzie polegała na zwróceniu uwagi na kwestie metodologicznej odpowiedniości, wiarygodności danych, zrozumiałości, dokładności, zrównoważenia, praktyczności, stosowności i kosztu. Jak zawsze, nadrzędną kwestią jest użyteczność. Czy wyniki uzyskane w efekcie zastosowania tych metod będą użyteczne i faktycznie wykorzystywane?

Po zgromadzeniu danych i uporządkowaniu ich w celu przeanalizowania rozpoczyna się czwarty etap procesu skoncentrowanego na wykorzystaniu. Użytkownicy docelowi biorą czynny i bezpośredni udział w interpretowaniu wniosków, formułowaniu opinii na podstawie danych oraz opracowywaniu rekomendacji. Następnie, w świetle faktycznych wniosków, można przystąpić do sformalizowania konkretnych strategii ich wykorzystania, a ewaluator może umożliwić przejście do faktycznego użycia.

Wreszcie, można podjąć decyzje w sprawie rozpowszechniania raportu z ewaluacji, wykraczając poza wszelkie pierwotne zobowiązania podjęte w ramach planowania zamierzonego wykorzystania. To uwypukla rozróżnienie pomiędzy zamierzonym wykorzystaniem przez użytkowników docelowych (użycie planowane) a bardziej ogólnym rozpowszechnianiem w szerszym gronie odbiorców (w którym mogą pojawić się zarówno sposoby użycia oczekiwane i niezamierzone).

Choć zasadniczo występuje jasny, stopniowy sposób postępowania podczas tworzenia ewaluacji skoncentrowanej na wykorzystaniu, w rzeczywistości rzadko jest to proces prosty czy liniowy. Na przykład, ewaluator może stwierdzić, że istotni stają się nowi użytkownicy, lub że podczas podejmowania decyzji w sprawie metod pojawiają się nowe pytania. Niekoniecznie musi też występować jasne i precyzyjne rozróżnienie między procesami precyzowania pytań ewaluacyjnych a podejmowaniem decyzji w sprawie metod; pytania pomagają w świadomym wyborze metod, a preferencje metodologiczne – w świadomym wyborze pytań.

## Negocjacje w sprawie ewaluacji w celu dostosowania jej do konkretnych sytuacji

Ewaluacja skoncentrowana na wykorzystaniu obejmuje negocjacje między ewaluatorem a docelowymi użytkownikami, prowadzone przez cały czas trwania procesu ewaluacji. Jest to być może najbardziej widoczne na etapie projektowania ewaluacji. Projekt danej ewaluacji zależy od zaangażowanych osób i ich sytuacji. *Ewaluacja sytuacyjna* to coś w rodzaju etyki sytuacyjnej (Fletcher 1966), sytuacyjnego przywództwa (Hersey 1985) lub sytuacyjnego uczenia się: „działanie zależy od konkretnej sytuacji, w której ma miejsce" (Anderson i in. 1996, s. 5). Standardy i zasady ewaluacji zapewniają ogólny kierunek, podstawy zasad etycznych oraz zobowiązanie do przestrzegania kompetencji i uczciwości zawodowej, lecz nie istnieją reguły bezwzględne, których ewaluator może przestrzegać, jeśli chce wiedzieć, co robić w przypadku określonych użytkowników w danej sytuacji. To dlatego Newcomer i Wholey (1989) w swoim podsumowaniu wiedzy o strategiach ewaluacji na rzecz budowania wysokowydajnych programów stwierdzili: „Przed dokonaniem ewaluacji ewaluatorzy i menedżerowie programu powinni współpracować w celu określenia idealnego produktu końcowego" (s. 202). Oznacza to negocjowanie zamierzonego i oczekiwanego wykorzystania ewaluacji.

Każda sytuacja ewaluacyjna jest inna. Skuteczna ewaluacja (czyli taka, która jest użyteczna, praktyczna, zgodna z zasadami etyki i dokładna) wynika z określonych cech i warunków danej sytuacji: połączenia osób, polityki, historii, kontekstu, zasobów, ograniczeń, wartości, potrzeb, interesów i przypadku. Mimo raczej oczywistego, niemal banalnego i zasadniczo zdroworozsądkowego charakteru tego spo-

strzeżenia, nie jest ono wcale oczywiste dla większości interesariuszy, którym zależy na tym, aby ewaluacja została przeprowadzona prawidłowo. Jedynym wspólnym zastrzeżeniem interesariuszy w kwestii czynnego udziału w etapie projektowania ewaluacji jest to, że nie mają oni wiedzy niezbędnej do robienia tego w sposób „właściwy". Panuje uporczywe przekonanie, że istnieje tylko jeden właściwy sposób działania. Właściwym sposobem, z perspektywy skoncentrowanej na wykorzystaniu, jest dokonanie tego w taki sposób, aby ewaluacja miała znaczenie i była użyteczna dla określonych ewaluatorów oraz zaangażowanych użytkowników docelowych, zaś znalezienie tego sposobu wymaga interakcji, negocjacji i analizy sytuacyjnej.

Charakter interakcji konsultacyjnych między ewaluatorami skoncentrowanymi na wykorzystaniu oraz użytkownikami docelowymi określają słowa „aktywny", „reakcyjny", „interakcyjny", „adaptacyjny". Te pojęcia mają być zarówno opisem, jak i zaleceniem. Opisują one sposób, w jaki faktycznie przebiega proces podejmowania decyzji w świecie rzeczywistym. Niemniej stanowią też zalecenie mające na celu uwrażliwienie ewaluatorów, aby działali w sposób świadomy i celowy, reagowali odpowiednio do sytuacji i dostosowywali się do niej, w celu zwiększenia swojej skuteczności przy współpracy z głównymi użytkownikami docelowymi.

Ewaluatorzy skoncentrowani przede wszystkim na wykorzystaniu działają w sposób świadomy i celowy identyfikując użytkowników docelowych i dobierając użyteczne pytania. Uważnie słuchają użytkowników docelowych i reagują na to, czego dowiadują się o danej sytuacji, w której ma miejsce ewaluacja. Biorą czynny udział w dwukierunkowym procesie negocjowania. Dostosowują się i w razie konieczności zmieniają pytania ewaluacyjne i projekt ewaluacji w miarę uzyskiwania coraz szerszej wiedzy o danej sytuacji i zmieniających się warunków. Ewaluatorzy aktywni, wchodzący w reakcję, interakcję i dostosowujący się do danej sytuacji nie narzucają z góry określonych modeli. Nie postępują za każdym razem tak samo. Uważnie obserwują każdą indywidualną sytuację i rzeczywiście reagują na zachowanie użytkowników docelowych każdej nowej ewaluacji.

Bycie aktywnym, wchodzenie w reakcję i interakcję oraz dostosowywanie do okoliczności to postawy charakteryzujące wszystkie etapy interakcji między ewaluatorem a użytkownikiem, począwszy od wskazania głównych użytkowników docelowych, przez wybór odpowiednich pytań i metod, aż po analizę wyników. Wszystkie etapy wymagają procesów współpracy opartych na działaniu, reakcji i adaptacji, podczas rozpatrywania dostępnych możliwości przez ewaluatorów i użytkowników docelowych. Zestaw możliwości obejmuje szeroki zakres metod, składników ewaluacji (od „mdłych" po „pikantne") oraz wielu różnych ról ewaluatorów: partnera, trenera, osoby prowadzącej grupę, specjalisty od spraw technicznych, polityka, analityka organizacyjnego, współpracownika, eksperta zewnętrznego, metodologa, pośrednika informacji, osoby odpowiedzialnej za komunikację, agenta zmiany, dyplomaty, osoby rozwiązującej problemy i kreatywnego konsultanta. Role, jakie odgrywa ewaluator w danej sytuacji zależą od jego celu, specyficznych okoliczności, w jakich się on znajduje, *oraz jego własnej wiedzy, umiejętności, stylu, wartości i zasad etycznych*.

Bycie aktywnym, wchodzenie w reakcję i interakcję oraz dostosowywanie się do sytuacji jest wyraźnym przejawem istotnego znaczenia indywidualnego doświadczenia ewaluatora, jego ukierunkowania i wkładu poprzez umiejscowienie na pierwszej pozycji w tym trójkącie konsultacji elementu aktywności. Odpowiednie reagowanie w danej sytuacji nie oznacza „turlania się" i „udawania trupa" (bierności) w świetle interesów lub postrzeganych potrzeb interesariuszy. Ponieważ w ewaluacji skoncentrowanej na wykorzystaniu ewaluator nie narzuca jednostronnie żadnego kierunku ani zestawu metod pracy nad programem, tak samo interesariusze nie narzucają swoich początkowych upodobań w sposób jednostronny czy dogmatyczny. Uzgodnienie ostatecznego modelu ewaluacji jest procesem podlegającym negocjacji, umożliwiającym uwzględnienie wartości i możliwości zarówno ewaluatora, jak i użytkowników docelowych.

Ewaluator skoncentrowany na wykorzystaniu, będąc aktywnym, wchodząc w reakcję i interakcję oraz dostosowując się do danej sytuacji, jest jedną z wielu stron przy stole negocjacyjnym. Czasem w procesie negocjacji mogą występować dysonanse; czasem przebiega on harmonijne. Bez względu na pojawiające się okoliczności, ewaluator skoncentrowany na wykorzystaniu nie działa sam.

## Wykorzystanie procesu

Większość dyskusji o wykorzystaniu ewaluacji skupia się na wykorzystaniu wniosków. Jednak zaangażowanie w procesy ewaluacyjne może być użyteczne niezależnie od wniosków płynących z ewaluacji. Procesy rozumowania stanowią główną „siłę pociągową" ewaluacji: na nich spoczywa cały ciężar. Jeżeli w wyniku zaangażowania w ewaluację główni użytkownicy docelowi nauczą się rozumować podobnie jak ewaluator i działać zgodnie z wartościami ewaluacji, wówczas rezultaty ewaluacji obejmują coś więcej niż tylko wnioski. Była ona użyteczna w zakresie przekraczającym same wnioski, ponieważ zwiększyła zdolność uczestników do wykorzystywania logiki i racjonalnego rozumowania charakterystycznych dla ewaluacji. „Wykorzystanie procesu" odnosi się wówczas do wykorzystywania logiki, racjonalnego rozumowania i przestrzegania wartości leżących u podstaw profesjonalnej ewaluacji.

Osoby wyszkolone w zakresie metod badawczych i ewaluacyjnych mogą łatwo uznać logikę leżącą u podstaw tych metod za pewnik. Podobnie jak osoby żyjące na co dzień w danej kulturze, sposób myślenia osób funkcjonujących w kulturze badawczej wydaje się dla nich naturalny i łatwy. Jednak dla praktyków, decydentów i osób tworzących politykę logika ta może być trudna do uchwycenia i dość nienaturalna. Myślenie w kategoriach tego, co jest jasne, określone, konkretne i możliwe do zaobserwowania nie przychodzi łatwo osobom, które z łatwością funkcjonują wśród pojęć ogólnych, uogólnień i niezweryfikowanych przekonań, a nawet na nich opierają swoje działanie. Nauczenie się patrzenia na świat w sposób, w jaki widzi go ewaluator, ma często długotrwały wpływ na osoby biorące udział w ewaluacji – wpływ, który może być większy i bardziej długotrwały niż wnioski płynące z samej ewaluacji.

Wykorzystanie procesu odnosi się do – i jest widoczne w formie – indywidualnych zmian w myśleniu i zachowaniu oraz programowych lub organizacyjnych zmian procedur i kultury, które mają miejsce wśród osób zaangażowanych w ewaluację w wyniku procesu uczenia się zachodzącego w procesie ewaluacji. Dowodem wykorzystania procesu może być m.in. następujące stwierdzenie po przeprowadzeniu ewaluacji: „Wpływ na nasz program nie wynikał tak bardzo z wniosków, ale z przejścia przez proces myślowy, jakiego ewaluacja wymagała".

Ewaluacja może przynosić, i często przynosi, takiego rodzaju skutki. Tym, co odróżnia ewaluację skoncentrowaną na wykorzystaniu jest to, że proces czynnego angażowania użytkowników docelowych zwiększa tego rodzaju wpływ ewaluacji. Ponadto, można sprawić, że możliwość i chęć uczenia się z procesów ewaluacyjnych oraz wynikających z nich wniosków będzie zamierzona i celowa. Innymi słowy, zamiast traktować wykorzystanie procesu jako nieformalne odgałęzienie, należy stwierdzić, że wyraźne uprzednie zwrócenie uwagi na potencjalne skutki logiki i procesów ewaluacyjnych może zwiększyć te skutki i przekształcić je w planowany cel podjęcia się ewaluacji. W ten sposób zwiększa się ogólna użyteczność ewaluacji.

Podstawą wykorzystania procesu jest praca z użytkownikami docelowymi mająca im pomóc w myśleniu o potencjalnych i pożądanych skutkach sposobu przeprowadzenia ewaluacji. Pytania o to, kto będzie zaangażowany, nabierają innego znaczenia przy uwzględnieniu, że osoby zaangażowane w sposób najbardziej bezpośredni będą nie tylko odgrywać kluczową rolę w ustalaniu treści ewaluacji – a co za tym idzie przedmiocie wniosków – ale także będą to osoby najbardziej wyeksponowane na logikę i procesy ewaluacji. Stopień wewnętrznego zaangażowania i odpowiedzialności będzie wpływał na charakter i stopień wpływu na kulturę programu.

Wpływ wywierany jest także na sposób myślenia o kosztach i korzyściach i obliczania ich przez podmioty finansujące i użytkowników ewaluacji. Stosunek kosztów do korzyści zmienia się po obu stronach równania, gdy ewaluacja przynosi nie tylko wnioski, ale także służy długoterminowym potrzebom programu jak np. rozwój pracowników i proces uczenia się organizacji.

Wyróżnia się sześć głównych typów wykorzystania procesu:

1) zakorzenienie w programie lub kulturze organizacji myślenia ewaluacyjnego;
2) pogłębienie wzajemnego porozumienia, zwłaszcza w kwestii wyników;
3) wspieranie i wzmacnianie programu poprzez ewaluację skupioną na interwencji;
4) zwiększenie zaangażowania uczestników, poczucia odpowiedzialności i własnej determinacji (ewaluacja uczestnicząca i wzmacniająca pozycję);
5) skutki pomiarowe, czyli „to, co jest mierzone, jest realizowane";
6) rozwój programu lub organizacji (Patton 2008).

Przykład wykorzystania procesu podają m.in. Cousins i Earl (1995), którzy opowiadali się za podejściem opartym na uczestnictwie i współpracy głównie w celu zwiększenia wykorzystania wniosków. Jednak wykraczają oni poza zwiększone wykorzystanie wniosków, kiedy omawiają sposób, w jaki zaangażowanie w ewaluację może pomóc w stworzeniu organizacji uczącej się. Postrzeganie ewaluacji uczestniczącej jako sposobu na stworzenie kultury organizacji opartej na ciągłym uczeniu się stało się ostatnio istotnym tematem w literaturze, łącząc ewaluację z „organizacjami uczącymi się" (np. King 1995; Sonnichsen 1993).

Ewaluacja skoncentrowana na wykorzystaniu jest automatycznie oparta na uczestnictwie i współpracy poprzez aktywne angażowanie głównych użytkowników docelowych we wszystkie aspekty ewaluacji, co stanowi strategię zwiększania wykorzystania wniosków. Zwrócenie dodatkowo uwagi na wykorzystanie procesu odzwierciedla to, w jaki sposób uczestnictwo i współpraca mogą prowadzić do ciągłego, długoterminowego zobowiązania się do stosowania logiki ewaluacyjnej i budowania kultury uczenia się w ramach danego programu lub organizacji. Uwypuklenie tego rodzaju wykorzystania procesu poszerza spektrum potencjalnych sposobów wykorzystania ewaluacji. To, na ile takie wykorzystanie procesu powinno być istotne w danej ewaluacji jest kwestią negocjacji z użytkownikami docelowymi. Praktyczną konsekwencją zdecydowanego podkreślenia tworzenia kultury uczenia się jako części procesu będzie zwrócenie w ramach ewaluacji uwagi na logikę i umiejętności ewaluacyjne i szkolenia w tym zakresie.

## Podstawowe założenia ewaluacji skoncentrowanej na wykorzystaniu

Podsumowując, podstawowe założenia ewaluacji skoncentrowanej na wykorzystaniu są następujące:

1) Siłą napędową ewaluacji powinno być zobowiązanie użytkowników docelowych do zamierzonego wykorzystania. Na każdym etapie podejmowania decyzji – niezależnie od tego, czy decyzja dotyczy celu, przedmiotu, modelu, metod, pomiaru, analizy czy raportowania – ewaluator zadaje użytkownikom następujące pytanie: „W jaki sposób wpłynęłoby to na Państwa wykorzystanie tej ewaluacji?"

2) Planowanie strategii w zakresie wykorzystania trwa od samego początku ewaluacji. Nie można zainteresować się wykorzystaniem dopiero na końcu procesu ewaluacji. Pod koniec ewaluacji potencjał wykorzystania jest już w dużej mierze określony. Od momentu, w którym interesariusze i ewaluatorzy zaczynają wchodzić w interakcje i tworzą koncepcje ewaluacji, podejmowane są decyzje, które w dużej mierze będą miały wpływ na wykorzystanie.

3) Na wykorzystanie znaczny wpływ ma czynnik ludzki. Czynnik ludzki odnosi się do wniosku badawczego, zgodnie z którym osobiste interesy i zaangażowanie osób biorących udział w ewaluacji określają sposób wykorzystania badania. W ten sposób badania ewaluacyjne powinny być w szczególności zorientowane na użytkownika – na interesy i potrzeby informacyjne konkretnych, możliwych do zidentyfikowania osób, a nie jakiejś nieokreślonej, biernej grupy odbiorców.

4) Uważna i przemyślana analiza interesariuszy powinna pomóc w świadomej identyfikacji głównych użytkowników docelowych, przy uwzględnieniu wielu różnych interesów występujących w przypadku każdego programu, a więc i ewaluacji. Zespół, uczestnicy programu, kierownictwo, urzędnicy państwowi, podmioty finansujące oraz liderzy społeczności – wszyscy oni są zainteresowani ewaluacją, ale stopień i charakter ich interesów jest zróżnicowany. Określanie głównych użytkowników docelowych i sposobów wykorzystania ewaluacji wiąże się z wrażliwością polityczną i wydawaniem etycznych sądów.

5) Badania ewaluacyjne muszą być w jakiś sposób ukierunkowane. Najbardziej użytecznym sposobem jest skoncentrowanie się na zamierzonym wykorzystaniu przez użytkowników docelowych. Ograniczone zasoby i czas nie pozwalają na to, aby jakiekolwiek pojedyncze badanie ewaluacyjne dostarczyło odpowiedzi na wszystkie pytania lub zajęło się wszystkimi możliwymi kwestiami. Ponieważ żadna ewaluacja nie może służyć wszystkim interesom potencjalnych interesariuszy w równym stopniu, interesariusze reprezentujący różne okręgi wyborcze powinni wspólnie negocjować to, jakie kwestie i pytania powinny mieć znaczenie priorytetowe.

6) Skoncentrowanie się na zamierzonym wykorzystaniu wymaga dokonania świadomych i przemyślanych wyborów. Cele ewaluacji są różne i obejmują: ocenę merytoryczną i wartościującą (ewaluacja podsumowująca), usprawnienie programów (wykorzystanie użytkowe) i generowanie wiedzy (wykorzystanie koncepcyjne). Z biegiem czasu, w miarę „dojrzewania" programu, główne potrzeby informacyjne i sposoby wykorzystania ewaluacji mogą ulegać zmianom i ewolucji.

7) Użyteczne badania ewaluacyjne muszą być zaprojektowane tak, aby były dostosowane do indywidualnej sytuacji. Podejścia oparte na standaryzowanych koncepcjach nie zdają egzaminu. Względną wartość konkretnego kierunku wykorzystania można ocenić wyłącznie w kontekście konkretnego programu i interesów użytkowników docelowych. Czynniki sytuacyjne mają wpływ na sposób wykorzystania. Czynniki te obejmują zmienne charakterystyczne dla danej społeczności, cechy organizacji, charakter ewaluacji, wiarygodność ewaluatora, względy polityczne i ograniczenia w zakresie zasobów. Przeprowadzając ewaluację skoncentrowaną na wykorzystaniu, aktywny, wchodzący w reakcję i dostosowujący się do indywidualnej sytuacji ewaluator współpracuje z użytkownikami docelowymi, aby ocenić, w jaki sposób różne czynniki i warunki mogą wpływać na potencjał wykorzystania.

8) Zobowiązanie użytkowników docelowych do danego wykorzystania można wspierać i wzmacniać poprzez aktywne włączenie ich w proces podejmowania istotnych decyzji w sprawie ewaluacji. Zaangażowanie zwiększa znaczenie, stopień zrozumienia i poczucie odpowiedzialności za ewaluację, a to wszystko umożliwia świadome i odpowiednie wykorzystanie.

9) Celem jest uczestnictwo na wysokim poziomie, a nie uczestnictwo „w dużej ilości". Czas interakcji w grupie może mieć odwrotny wpływ na jakość procesu. Ewaluatorzy przeprowadzający ewaluację skoncentrowaną na wykorzystaniu muszą być zdolnymi animatorami grup.

10) Wysoki poziom zaangażowania użytkowników docelowych przekłada się na wysoką jakość i użyteczność ewaluacji. Wielu badaczy martwi się, że będą musieli poświęcić rygor metodologiczny, jeśli w podejmowaniu decyzji w sprawie metod będą uczestniczyć osoby niebędące naukowcami. Niemniej decydentom zależy na danych użytecznych i rzetelnych. Rzetelność i użyteczność są od siebie wzajemnie zależne. Eliminowanie zagrożeń dla użyteczności jest tak samo ważne, jak eliminowanie zagrożeń dla rzetelności. Zdolne osoby przeprowadzające ewaluację mogą pomóc osobom niebędącym naukowcami w zrozumieniu kwestii metodologicznych, aby mogły one same ocenić kompromisy, jakie mają miejsce podczas dokonywania wyboru między silnymi i słabymi stronami różnych podejść i metod.

11) Ewaluatorzy mają słuszny interes w ewaluacji, ponieważ ich wiarygodność i uczciwość są zawsze narażone na ryzyko, stąd właśnie prawo ewaluatorów do bycia aktywnym, reagowania i dostosowywania do sytuacji. Ewaluatorzy aktywnie przedstawiają użytkownikom docelowym ich własne, najlepsze opinie na temat odpowiedniego ukierunkowania i metod ewaluacji; wchodzą w reakcję poprzez uważne słucha-

nie i uwzględnianie wątpliwości innych osób; dostosowują się do sytuacji poprzez znajdywanie sposobów projektowania ewaluacji, które uwzględniają różne interesy, w tym ich własny interes, przy jednoczesnym przestrzeganiu wysokich standardów praktyki zawodowej. Wiarygodność i uczciwość ewaluatorów są czynnikami wpływającymi na wykorzystanie, a zarazem fundamentem zawodu. W tym względzie ewaluatorzy powinni stosować się do standardów i zasad zawodowych.

12) Ewaluatorzy zaangażowani w zwiększanie wykorzystania są odpowiedzialni za szkolenie użytkowników w zakresie procesów ewaluacji i wykorzystywania informacji. Szkolenie interesariuszy w zakresie metod i procesów ewaluacji stanowi wkład w zarówno krótkoterminowe, jak i długoterminowe sposoby wykorzystania ewaluacji. Rozszerzenie wiedzy decydentów o ewaluacji może przyczynić się do zwiększenia wykorzystania ewaluacji w czasie. Każda indywidualna ewaluacja dostarcza zatem możliwości szkolenia użytkowników ewaluacji i zwiększania możliwości organizacyjnych zakresie wykorzystania – co określa się jako „wykorzystanie procesu" – przy jednoczesnym wykorzystaniu procesu ewaluacji do wspierania długoterminowego programu i rozwoju organizacji.

13) Wykorzystanie jest czymś innym niż raportowanie i rozpowszechnianie wyników. Sporządzanie raportu i rozpowszechnianie wyników może być sposobem na umożliwienie wykorzystania, lecz tych działań nie należy mylić z zamierzonym wykorzystaniem takim, jak podejmowanie decyzji, usprawnianie programów, zmiana myślenia i generowanie wiedzy.

14) Poważnie zwrócenie uwagi na wykorzystanie pociąga za sobą koszty finansowe i czasowe, które wcale nie są nieważne. Korzyści tych kosztów przekładają się na zwiększone wykorzystanie. Koszty te powinny być jasno określone w ofertach i budżetach badania ewaluacyjnego, tak aby nie dochodziło do zaniedbania użycia na skutek braku zasobów.

## Kwestie związane z przeprowadzaniem ewaluacji skoncentrowanej na wykorzystaniu

Podczas przeprowadzania ewaluacji skoncentrowanej na wykorzystaniu występuje szereg określonych kwestii.

### Reakcje użytkownika i jakość techniczna

Reagowanie i aktywne angażowanie głównych użytkowników docelowych w ewaluację nie powinno oznaczać poświęceń w kwestii jakości technicznej. Punktem wyjścia jest uznanie, że standardy jakości technicznej są różne w przypadku różnych użytkowników i różnych sytuacji. Nie chodzi tutaj o przestrzeganie jakichś bezwzględnych standardów jakości technicznej, ale raczej o dopilnowanie, aby metody i środki były odpowiednio dopasowane do potrzeb w zakresie rzetelności i wiarygodności danego celu ewaluacji i określonych użytkowników docelowych.

Jennifer Greene (1990) dogłębnie przeanalizowała debatę o „jakości technicznej wobec reakcji użytkowników". Stwierdziła, że panuje powszechna zgoda co do tego, że oba te elementy są istotne, lecz występują rozbieżne opinie dotyczące względnego znaczenia każdego z nich. Jennifer Greene stwierdziła, że w debacie chodzi tak naprawdę o to, na ile należy uznać i rozpatrywać kwestię związku ewaluacji z polityką: „Ewaluatorzy powinni przyjąć, że napięcia i konflikty w praktyce badań ewaluacyjnych są czymś wręcz nieuniknionym, że wymogi wynikające z większości – jeśli nie wszystkich – definicji reagowania i jakości technicznej (nie wspominając o wykonalności i stosowności) będą w sposób charakterystyczny odzwierciedlały konkurujące ze sobą kierunki polityki i wartości danego środowiska" (s. 273). Następnie autorka zaleciła, aby ewaluatorzy „wyjaśnili zasady i wartości" stanowiące podstawę decyzji w sprawie celu, odbiorców, modelu i metod. Jej zalecenie jest spójne z ewaluacją skoncentrowaną na wykorzystaniu.

**Rotacja użytkowników – słaby punkt ewaluacji skoncentrowanej na wykorzystaniu**

Najsłabszym punktem ewaluacji skoncentrowanej na wykorzystaniu jest rotacja głównych użytkowników docelowych. Proces tak bardzo zależy od aktywnego zaangażowania użytkowników docelowych, że utrata użytkowników w trakcie na skutek zmian miejsca pracy, reorganizacji, nowego podziału funkcji i wyborów politycznych może negatywnie wpłynąć na ostateczne wykorzystanie ewaluacji. Nowi użytkownicy, włączani do ewaluacji na późnym etapie procesu, rzadko kierują się tymi samymi celami, co osoby, które brały udział w procesie na początku. Najlepszym antidotum jest współpraca z grupą zadaniową złożoną z wielu użytkowników docelowych, po to, aby utrata jednego lub dwóch nie miała aż tak negatywnego wpływu. Mimo to w przypadku znacznej rotacji głównych użytkowników docelowych może być konieczne ponowne ożywienie procesu poprzez wynegocjowanie na nowo zobowiązań w sprawie modelu i wykorzystania z nowymi osobami włączonymi do badania.

Przy dokonywaniu wyboru odpowiednich interesariuszy, skłanianiu ich do poświęcenia swego czasu i uwagi ewaluacji, stawianiu czoła dynamice politycznej, budowaniu wiarygodności i przeprowadzaniu ewaluacji w sposób zgodny z zasadami etyki pojawia się wiele wyzwań. Wszystkie te wyzwania krążą wokół relacji między ewaluatorem a użytkownikami docelowymi. Gdy nowi użytkownicy docelowi zastępują tych, którzy odchodzą, konieczne jest nawiązanie nowych relacji. To może oznaczać opóźnienia względem pierwotnych harmonogramów, lecz opóźnienia te opłacają się w kontekście końcowego wykorzystania poprzez wspieranie podstaw porozumienia i relacji będących podstawą ewaluacji skoncentrowanej na wykorzystaniu.

**Rozwijanie potencjału na potrzeby wykorzystania ewaluacji**

Podobnie jak studenci potrzebują doświadczenia i praktyki, aby nauczyć się prowadzenia ewaluacji, programy i organizacje potrzebują doświadczenia i praktyki, aby stać się biegłymi w wykorzystywaniu ewaluacji w celu usprawnienia programu i procesu uczenia się w organizacji. Dziedzina ewaluacji zwraca coraz większą uwagę na sposoby wbudowywania potencjału ewaluacyjnego w programy i organizacje (Kuzmin 2005; Patton 1994). Otwartość na ewaluację wzrasta, gdy organizacje mają pozytywne doświadczenia w jej zakresie i uczą się wyciągać lekcje z tych doświadczeń. Powszechnym problemem przy wprowadzaniu ewaluacji do organizacji jest zbyt intensywne działanie (wysiłki na dużą skalę i uniwersalne uprawnienia), zanim potencjał jest wystarczający dla zarządzania użyteczną ewaluacją. Zdolność ta obejmuje uświadomienie pracownikom i osobom odpowiedzialnym za kwestie administracyjne, na czym polegają logika i wartości ewaluacji, opracowanie właściwych dla danej organizacji procesów włączania ewaluacji w proces planowania i realizacji programu oraz powiązanie ewaluacji z najnowszą wiedzą dotyczącą uczenia się organizacji (Sonnichsen 2000; Preskill i Torres 1998).

Ćwierć wieku badań nad „gotowością do ewaluacji" (Preskill i Torres 2000; Seiden 2000; Mayer 1975) doprowadziło do wniosku, że określenie wartości ewaluacji i uczenia się są warunkami koniecznymi wykorzystania ewaluacji. Określenie wartości ewaluacji nie jest czymś oczywistym. Nie odbywa się też w sposób naturalny. Zaangażowanie użytkowników w ewaluację jest zazwyczaj delikatną kwestią, nierzadko kapryśną, a więc należy ją pielęgnować jak roślinę, która ma potencjał do ogromnego wzrostu, ale tylko pod warunkiem, że będziemy o nią odpowiednio dbać, pielęgnować i odżywiać. W ewaluacji skoncentrowanej na wykorzystaniu takie „odżywianie" ma kluczowe znaczenie, nie tylko na potrzeby zwiększania wykorzystania danej ewaluacji, ale także rozwijania potencjału (wykorzystania procesu) wykorzystywania przyszłych ewaluacji.

**Zmienne role ewaluatora związane ze zmiennymi celami ewaluacji**

Różne cele ewaluacji wymagają zróżnicowania ról ewaluatora. Trzy typy odzwierciedlają historyczny rozwój ewaluacji na gruncie trzech różnych tradycji:

(1) badań socjologicznych;
(2) pragmatycznej praktyki terenowej, zwłaszcza ewaluatorów wewnętrznych i konsultantów;
(3) kontroli programów oraz kontroli finansowych.

Gdy badanie ewaluacyjne ma na celu uzyskanie możliwej do generalizacji wiedzy o związkach przyczynowo-skutkowych zachodzących między działaniem a rezultatami programu, wymagane jest rygorystyczne stosowanie metod socjologicznych, a podstawą powinna być tu rola ewaluatora jako eksperta w dziedzinie metodologii. Gdy nacisk kładzie się na ustalenie ogólnych zasług lub wartości programu, centralne miejsce zajmuje rola ewaluatora jako osoby oceniającej. Jeżeli badanie ewaluacyjne zlecono w wyniku wątpliwości społeczeństwa w kwestii odpowiedzialności, dla decydentów i społeczeństwa będzie widoczna rola ewaluatora jako niezależnego audytora, inspektora lub kontrolera. Gdy głównym celem jest usprawnienie programu, ewaluator pełni funkcję doradczą i współpracuje z zespołem realizującym program. Będąc członkiem zespołu projektowego, ewaluator programów rozwojowych będzie odgrywał rolę konsultanta. Jeśli ewaluacja służy celom związanym ze sprawiedliwością społeczną, ewaluator staje się agentem zmiany (Patton 2010).

*W ewaluacji skoncentrowanej na wykorzystaniu ewaluator zawsze pełni funkcję negocjatora* – negocjującego z głównymi użytkownikami docelowymi w sprawie innych ról, które ma odgrywać. Poza tym, wszystkie role są jasno określone, dopuszczalne są także wszystkie metody. Wybór roli wynika z i zależy od zamierzonego wykorzystania przez użytkowników docelowych.

Proszę wziąć pod uwagę na przykład międzynarodową ewaluację pomocy żywnościowej dla mieszkańców obszarów wiejskich w okresie dotkliwej suszy. Na potrzeby rozliczania i przeglądu polityki główni użytkownicy docelowi są członkami komitetów nadzorujących program w międzynarodowych instytucjach finansujących. W czasie międzynarodowego kryzysu żywnościowego program będzie bardzo widoczny, kosztowny, i prawdopodobnie kontrowersyjny, zwłaszcza dlatego, że specjalne grupy interesów często nie są zgodne co do sposobu rozdzielania żywności i kwestii, które grupy potrzebujących powinny mieć w tym względzie pierwszeństwo. W takich warunkach wiarygodność i użyteczność ewaluacji będzie w dużym stopniu zależeć od niezależności ewaluatorów, ich ideologicznej neutralności, wiedzy metodologicznej i mądrości politycznej.

Proszę teraz zderzyć taką międzynarodową ewaluację wykonania zobowiązań, z rolą ewaluatora we wspieraniu zwiększania wpływu przywództwa na małych obszarach wiejskich. Program działa w kilku lokalnych wspólnotach. Głównymi użytkownikami docelowymi są nauczyciele szkolni, miejscowi urzędnicy oraz miejscowi pracownicy służby zdrowia, którzy pomagali w przygotowaniu programu przy wsparciu ze strony zagranicznego darczyńcy. Ewaluacja skupia się na usprawnieniu programu w celu zwiększenia satysfakcji uczestników i wspierania pożądanego wzrostu poziomu wiedzy oraz zmian w zachowaniu uczestników. W takich warunkach wykorzystanie ewaluacji będzie w dużym stopniu zależało od relacji ewaluatora z lokalnymi członkami zespołu realizującego program. Ewaluator będzie musiał zbudować silną relację opartą na wzajemnym zaufaniu i szacunku, aby skutecznie umożliwić zespołowi podejmowanie decyzji w sprawie priorytetów ewaluacji i metod gromadzenia danych. Wówczas ewaluator przeprowadzi ich przez proces osiągania konsensusu, w miarę interpretowania wyników i uzgadniania zmian.

Te kontrastujące ze sobą przykłady odzwierciedlają zakres kontekstów, w których prowadzona jest ewaluacja programu. Rola ewaluatora w konkretnym badaniu będzie zależała od dopasowania jego/jej roli do kontekstu i celu ewaluacji, będących przedmiotem negocjacji z głównymi użytkownikami docelowymi. Dzieje się tak przede wszystkim wówczas, gdy ewaluator skoncentrowany na wykorzystaniu i główni użytkownicy docelowi postanawiają w sposób jednoznaczny skupić się na wykorzystaniu procesu. Wykorzystanie procesu wykracza poza tradycyjne skupienie się na wnioskach i raportach jako podstawowych nośnikach wpływu ewaluacji. Każda ewaluacja może przynosić i często przynosi skutki w sposób niezamierzony lub jako następstwo wykorzystania jej wyników. Tym, co odróżnia ewaluację skoncentrowaną

na wykorzystaniu, jest to, że możliwość i chęć uczenia się z procesów ewaluacyjnych oraz wniosków może być zamierzona i celowa – (użytkownicy docelowi mają możliwość rozważenia wprowadzenia jej od samego początku procesu). Innymi słowy, zamiast traktować wykorzystanie procesu jako nieformalny efekt domina, należy stwierdzić, że wyraźne uprzednie zwrócenie uwagi na potencjalny wpływ logiki i procesów ewaluacyjnych może zwiększyć ten wpływ i przekształcić go w planowany cel podjęcia się ewaluacji. W ten sposób zwiększa się ogólną użyteczność ewaluacji i wzmacnia jej przyszły potencjał.

Niemniej ewaluator skoncentrowany na wykorzystaniu, który przedstawia użytkownikom docelowym możliwości wykraczające poza wąskie i tradycyjne wykorzystanie wyników, ma obowiązek ujawnić i omówić zastrzeżenia dotyczące takiego podejścia. Gdy ewaluatorzy badają nowe, innowacyjne możliwości, muszą oni jasno zaznaczyć, że nieuczciwość, korupcja, zniekształcanie danych i sprzedawanie danych są niedopuszczalne. Gdy główni użytkownicy docelowi chcą i potrzebują niezależnej, podsumowującej ewaluacji, powinni taką otrzymać. Gdy chcą, aby ewaluator działał niezależnie przynosząc wyniki skupiające się na usprawnianiu na potrzeby ewaluacji formatywnej, należy im to zapewnić. Nie są to jednak już jedyne możliwości w zestawie możliwych sposobów wykorzystania ewaluacji. Dziś już wykorzystuje się nowe podejścia oparte na uczestnictwie, współpracy, zorientowane na interwencję i rozwój. W ewaluacji skoncentrowanej na wykorzystaniu nowym wyzwaniem jest praca z głównymi użytkownikami docelowymi w celu zrozumienia, kiedy takie podejścia są odpowiednie, i pomaganie użytkownikom docelowym w podejmowaniu świadomych decyzji co do ich stosowności, w odniesieniu do konkretnej ewaluacji.

**Polityczne podstawy ewaluacji skoncentrowanej na wykorzystaniu**

Ewaluacja skoncentrowana na wykorzystaniu wymaga sprytnego wyczucia politycznego przy identyfikacji zarówno zamierzonego wykorzystania, jak i użytkowników docelowych, ponieważ model i wykorzystanie ewaluacji zawsze występują w określonym kontekście politycznym. Oto niektóre wnioski płynące z praktyki:

1) Nie wszystkie informacje są użyteczne. Aby informacja posiadała jakąś moc, musi ona być istotna i przedstawiona w formie zrozumiałej dla użytkowników. Socjolog badający organizacje, Michael Crozier, zaobserwował następujące zjawisko: „Ludziom i organizacjom zależy tylko na tym, co mogą uznać za coś, co ma na nich wpływ, oraz nad czym mogą mieć kontrolę" (1964, s. 158).

2) Nie wszystkie osoby są użytkownikami informacji. Poszczególne osoby różnią się między sobą pod względem umiejętności wykorzystania informacji i procesów ewaluacyjnych. Różnice te pogłębia ich zróżnicowanie społeczne, poziom wykształcenia i doświadczenie. W praktyce ewaluacji oznacza to, że informacja ma największą moc w rękach osób, które wiedzą, jak ją wykorzystać i są otwarte na wykorzystywanie jej. Wyzwaniem w kwestii wykorzystania jest dopasowanie, czyli przekazanie właściwych informacji właściwym ludziom. A co z osobami, które nie są skłonne do wykorzystywania informacji, osobami, które mają ostrożny, obojętny, czy nawet wrogi stosunek do ewaluacji? Ewaluator skoncentrowany na wykorzystaniu szuka możliwości i strategii pozyskiwania i szkolenia użytkowników informacji. Tak więc wyzwanie związane ze zwiększaniem wykorzystania składa się z dwóch elementów: a) znalezienie i zaangażowanie osób, które są z natury użytkownikami informacji oraz b) przeszkolenie tych, które nie mają takich predyspozycji.

3) Informacja ukierunkowana na wykorzystanie z większym prawdopodobieństwem trafia do celu. Trudno jest z góry przewidzieć, jakie decyzje zostaną podjęte co do tego, które informacje będą najbardziej wartościowe. Ewaluacja skoncentrowana na wykorzystaniu ma na celu zwiększenie prawdopodobieństwa zebrania odpowiednich i istotnych informacji poprzez skupienie się na realnych kwestiach, w realnych ramach czasowych, mając na uwadze realne decyzje. W ten sposób ewaluacja skoncentrowana na wykorzystaniu ma na celu wypełnienie luki pomiędzy potencjalnym i faktycznym wykorzystaniem, pomiędzy wiedzą a działaniem. Ukierunkowanie ewaluacji na zamierzone wykorzystanie przez użytkowników docelowych zwiększa szanse na osiągnięcie celu.

4) Tylko informacje wiarygodne mają rzeczywistą moc. Alkin i in. (1979) stwierdzili, że charakterystyczne cechy zarówno ewaluacji, jak i ewaluatora, wpływają na wykorzystanie, zaś jedną z ich najważniejszych cech jest wiarygodność. Eleanor Chelimsky, jedna z najbardziej doświadczonych i skutecznych ewaluatorów w branży, we współpracy z Kongresem podkreśliła tę kwestię: „Niezależnie od tego, czy problem tkwi w uczciwości, równowadze, jakości metodologicznej lub rzetelności, żadne działanie na rzecz zapewnienia wiarygodności nie idzie na marne. Wspomnienie słabej jakości utrzymuje się długo w pamięci...". (Chelimsky 1987, s. 14). Im bardziej upolityczniony jest kontekst, w którym dokonywana jest ewaluacja i im bardziej widoczna jest ona w tym upolitycznionym środowisku, tym ważniejsza dla wiarygodności będzie niezależna ocena jakości ewaluacji mająca na celu ustalenie wiarygodności. Chodzi tutaj o formę dopasowania ewaluacji skoncentrowanej na wykorzystaniu, w której zabezpieczenia wiarygodności ewaluacji mają pomóc w przewidywaniu i eliminowaniu określonych politycznych przejawów ingerencji w danym środowisku politycznym.

O ile jest to możliwe i wykonalne, można zorganizować grupę zadaniową ds. ewaluacji, aby podejmowała główne decyzje dotyczące ukierunkowania, metod i celu ewaluacji. Grupa zadaniowa jest narzędziem aktywnego angażowania kluczowych zainteresowanych stron w ewaluację. Ponadto, same procesy związane z podejmowaniem decyzji w sprawie ewaluacji zazwyczaj zwiększają zaangażowanie interesariuszy w wykorzystanie rezultatów, zwiększając jednocześnie ich wiedzę o ewaluacji, ich umiejętność przeprowadzania badań ewaluacyjnych oraz ich zdolność do interpretowania wniosków. Grupa zadaniowa umożliwia ewaluatorowi dzielenie się odpowiedzialnością za podejmowanie decyzji poprzez zapewnienie forum dla politycznych i praktycznych perspektyw, które pochodzą od tych interesariuszy, którzy docelowo będą zaangażowani w wykorzystanie ewaluacji.

**Ewaluatorzy skoncentrowani na wykorzystaniu potrzebują specjalnych umiejętności**

Aby podsycać wykorzystanie ewaluacji i utrzymać ją z dala od destrukcyjnych procesów grupowych lub polityki władzy, ewaluator skoncentrowany na wykorzystaniu musi wykazywać się polityczną mądrością, umiejętnością kierowania grupą, odczytywania odpowiedniej wewnętrznej dynamiki organizacji oraz musi być przyjazną dla użytkownika osobą przekazującą informacje (Patton 2008; Torres i in. 1996). To wyraźnie pokazuje, że ewaluatorzy skoncentrowani na wykorzystaniu potrzebują nie tylko umiejętności technicznych i metodologicznych, ale także umiejętności z zakresu procesów grupowych oraz politycznej przenikliwości – czegoś, co czasem określa się jako „umiejętności interpersonalne ewaluatorów" (Ghere i in. 2006).

**Niewłaściwe wykorzystanie ewaluacji**

Ewaluacja skoncentrowana na wykorzystaniu ma na celu umożliwienie odpowiedniego wykorzystania wniosków i procesów ewaluacyjnych, a więc ewaluatorzy skoncentrowani na wykorzystaniu muszą także uwzględniać kwestię niewłaściwego wykorzystania. Procesy i wyniki ewaluacji mogą być błędnie interpretowane i niewłaściwie wykorzystane podczas poszukiwań politycznej przewagi. Alkin i Coyle (1988) dokonali istotnego rozróżnienia pomiędzy niewłaściwą ewaluacją, w której ewaluator osiąga słabe wyniki lub nie przestrzega norm i zasad, a niewłaściwym wykorzystaniem, w którym użytkownicy manipulują ewaluacją w sposób zniekształcający wyniki lub badanie. King (1982) twierdził, że zamierzone niewykorzystanie źle przeprowadzonych badań należy postrzegać jako słuszne i odpowiedzialne. Oto kilka kwestii ogólnych dotyczących niewłaściwego wykorzystania.

W miarę coraz szerszego wykorzystania może także dochodzić do niewłaściwego wykorzystania, tak więc ewaluatorzy skoncentrowani na wykorzystaniu muszą zachować czujność, po to, aby ich działania nie kierowały uwagi na odwrotne skutki ewaluacji. Gdy ludzie ignorują badania ewaluacyjne, ignorują ich potencjalne sposoby wykorzystania, ale także nadużycia. Gdy ewaluatorzy z sukcesem skupiają większą

uwagę na danych z ewaluacji i zwiększają faktyczne wykorzystanie, może wystąpić w związku z tym wzrost nadużyć, często w ramach tej samej ewaluacji. Donald T. Campbell przedstawił taką samą prognozę podczas formułowania „pewnego zniechęcającego prawa, które – jak się wydaje – zaczyna się pojawiać: im więcej jakiś wskaźnik społeczny jest stosowany przy podejmowaniu decyzji społecznych, tym bardziej jest on narażony na większą presję w formie korupcji" (1988, s. 306).

Współpraca z wieloma użytkownikami, którzy rozumieją i doceniają ewaluację jest jednym z najlepszych środków zapobiegających niewłaściwemu wykorzystaniu. Sojusznicy w wykorzystaniu są jednocześnie sojusznikami przeciwko niewłaściwemu wykorzystaniu. Rzeczywiście niewłaściwe wykorzystanie można ograniczyć poprzez działania na rzecz przyjęcia przez użytkowników docelowych odpowiedzialności za ewaluację w takim stopniu, że stają się mistrzami właściwego użycia, strażnikami chroniącymi przed niewłaściwym wykorzystaniem i obrońcami wiarygodności ewaluacji, gdy pojawia się jej niewłaściwe wykorzystanie.

Kontrolowanie niewłaściwego wykorzystania jest czasem niezależne od ewaluatora, lecz tym, co leży zawsze w zakresie bezpośredniej odpowiedzialności ewaluatora jest niewłaściwa ewaluacja: uchybienia ewaluatora, które prowadzą niniejszą dyskusję na poziom etyki ewaluacji.

**Etyka podejścia opartego na skoncentrowaniu na użytkowniku**

Czasem zachodzi obawa, że przeprowadzając ewaluację skoncentrowaną na wykorzystaniu ewaluator może zostać przeciągnięty na stronę interesariuszy. W jaki sposób ewaluatorzy mają zachować swoją uczciwość, jeśli nawiązali bliskie stosunki z interesariuszami? W jaki sposób ewaluator ma uwzględnić politykę, nie stając się przy tym politycznym narzędziem chroniącym interes tylko jednej strony?

Charakter relacji występujących między ewaluatorami i osobami, z którymi ewaluatorzy współpracują, jest złożony. Z jednej strony, ewaluatorzy pragną utrzymać uprzejmy dystans wobec osób, które badają, w celu zapewnienia ochrony obiektywności i zminimalizowania osobistej i politycznej stronniczości. Z drugiej strony, perspektywa związków międzyludzkich pokazuje, że bliski kontakt interpersonalny jest warunkiem koniecznym w budowaniu wzajemnego zaufania. Ewaluatorzy znajdują się więc między przysłowiowym „młotem a kowadłem": zbyt mocne zbliżenie się do decydentów może nadwerężyć naukową wiarygodność; natomiast zachowanie dystansu może zaszkodzić wykorzystaniu.

Jednym ze sposobów rozwiązywania problemów związanym z „przeciąganiem na stronę" jest skupienie się na empirycznych podstawach ewaluacji. Empiryczna podstawa ewaluacji obejmuje jasne określanie założeń i wartości, weryfikację zasadności założeń oraz dokładne zbadanie programu, w celu uzyskania wiedzy o tym, co się dokładnie dzieje. Uczciwość ewaluacji zależy od jej empirycznej orientacji, tj. jej zaangażowania w systematyczne gromadzenie wiarygodnych danych oraz sprawozdawczość. Podobnie uczciwość procesu grupowego ewaluacji zależy od pomagania uczestnikom w przyjęciu perspektywy empirycznej. Musi pojawić się zaangażowanie, aby naprawdę dowiedzieć się, co się dzieje, przynajmniej w stopniu, w jakim jest to możliwe, uwzględniając ograniczenia metod badawczych i skąpe zasoby. Pojawienie się takiego zaangażowania wiąże się z nauczaniem i kierowaniem. Mądry ewaluator będzie monitorował empiryczne ukierunkowanie użytkowników docelowych i sposób aktywno-reaktywno-adaptacyjny, w ramach reagowania na konkretne sytuacje, będzie podejmował odpowiednie kroki mające na celu utrzymanie ewaluacji na właściwej ścieżce empirycznej i użytecznej.

Ewaluatorzy stawiają czoła różnym sytuacjom wymagającym silnego zakorzenienia w zasadach etycznych, które mogą wymagać odwagi. Poza ogólną wrażliwością etyczną, etyka ewaluatorów skoncentrowanych na wykorzystaniu może być zakwestionowana w odniesieniu do dwóch zasadniczych aspektów ewaluacji skoncentrowanej na wykorzystaniu:

1) ograniczaniu zaangażowania interesariuszy do głównych użytkowników docelowych; oraz
2) ścisłej współpracy z tymi użytkownikami.

Etyka ograniczania i ukierunkowywania zaangażowania interesariuszy dotyczy tego, kto ma dostęp do władzy, jaką jest wiedza wynikająca z ewaluacji. Etyka budowania bliskich relacji dotyczy uczciwości, neutralności i możliwości korupcji ewaluatora. Oba te problemy opierają się na zasadniczym pytaniu etycznym: komu służy ewaluacja i ewaluator?

Po pierwsze, ewaluatorzy muszą świadomie i celowo określić swoje własne podstawy moralne i dokładnie przeanalizować to, czyje interesy są reprezentowane w zadawanych pytaniach oraz kto będzie miał dostęp do wyników. Aktywny element bycia aktywno-reaktywno-interaktywno-adaptacyjnym skłania ewaluatorów do przedstawiania swoich własnych wątpliwości, problemów i wartości podczas negocjacji w sprawie ewaluacji. Ewaluator jest również interesariuszem – nie głównym interesariuszem, ale w każdej ewaluacji reputacja ewaluatora, jego wiarygodność i przekonania mogą być wystawione na próbę. Ewaluator skoncentrowany na wykorzystaniu nie przyjmuje pasywnej postawy w akceptacji tego, czego na początku chce użytkownik docelowy. Proces aktywności, reakcji i adaptacji obejmuje zobowiązanie ze strony ewaluatora do reprezentowania standardów i zasad zawodowych oraz jego poczucie moralności i uczciwości, przy jednoczesnym poszanowaniu przekonań i wątpliwości głównych użytkowników.

Druga kwestia dotyczy tego, w jaki sposób interesy poszczególnych grup interesariuszy są reprezentowane w procesie skoncentrowanym na wykorzystaniu. Preferowanym rozwiązaniem jest działanie na rzecz skłonienia uczestników grup, aby reprezentowali sami siebie w procesie negocjowania ewaluacji. Jak wspomniano wcześniej, ewaluacja skoncentrowana na wykorzystaniu angażuje rzeczywiste osoby, a nie tylko ogólnych, abstrakcyjnych odbiorców. W związku z tym, jeżeli chodzi o interesy osób będących w niekorzystnej sytuacji, należy znaleźć sposoby wysłuchania i zaangażowania ich bezpośrednio, nie wystarczy samo ich reprezentowanie, w sposób potencjalnie protekcjonalny przez osoby uprzywilejowane. To, czy i w jaki sposób należy tego dokonać, może być częścią zadań ewaluatora w trakcie interakcji polegających na byciu aktywnym, reagowaniu, wchodzeniu w reakcje i dostosowywaniu się do sytuacji.

Inne obawy dotyczące ewaluacji skoncentrowanej na wykorzystaniu mają ci, którzy martwią się, że różne role dostępne dla ewaluatorów skoncentrowanych na wykorzystaniu mogą szkodzić temu, co niektórzy uznają za podstawowy (lub jedyny) cel ewaluacji – dostarczenie obiektywnych ocen zalet lub wartości. Jeśli ewaluatorzy przyjmują role wykraczające poza dokonywanie ocen zalet lub wartości, np. tworzenie organizacji uczących się lub umożliwianie dokonywania ocen użytkownikom docelowym, czy to wprowadza zamęt w sprawie, czym jest ewaluacja?

Michael Scriven na przykład twierdzi, że ewaluatorzy nie służą żadnym określonym osobom. Służą prawdzie. Jego zdaniem prawda może być ofiarą, gdy ewaluatorzy nawiązują bliskie relacje zawodowe z zespołem realizującym program. Scriven przestrzega ewaluatorów, aby skrupulatnie chronili swoją niezależność. Angażowanie użytkowników docelowych może tylko zagrozić osłabieniem trafnych ocen, których ewaluator musi dokonywać. Scriven twierdzi, że ewaluatorzy muszą być zdolni do radzenia sobie z samotnością, jaka może towarzyszyć niezależności, i chronić się przed spoufalaniem, tendencją do bycia przeciągniętym na którąś stronę i opowiadaniem się za programem będącym przedmiotem ewaluacji (1991a, s. 182). Spoufalanie się prowadzi do „kazirodczych stosunków", w których ewaluator „idzie do łóżka" z programem będącym przedmiotem ewaluacji (s. 192). Każdą porażkę w wydaniu obiektywnego sądu Michael Scriven potępia jako „zniesienie zawodowej odpowiedzialności ewaluatora..." (1991, s. 32). Wyśmiewa on to, co szyderczo nazywa „milszym, delikatniejszym podejściem" do ewaluacji (s. 39). Jego wątpliwości wynikają z tego, czego doświadczył – oporu klientów ewaluacji przed negatywnymi wynikami oraz psychologiczne trudności, z jaką ewaluatorzy przekazują niekorzystne informacje zwrotne. W ten sposób uczula ewaluatorów, aby byli bezkompromisowi w przekazywaniu negatywnych wyników. „Głównym powodem, dla którego ewaluatorzy unikają negatywnych wniosków, jest to, że nie mają odwagi..." (s. 42).

Moje doświadczenia jako ewaluatora skoncentrowanego na wykorzystaniu są inne od doświadczeń Scrivena, więc osobiście dochodzę do innych wniosków. Staram się pracować z klientami, którym zależy na informa-

cjach wysokiej jakości, umożliwiających im usprawnienie programów. Są to ludzie o wysokim poziomie kompetencji i uczciwości, potrafiący wykorzystać i zrównoważyć zarówno pozytywne, jak i negatywne informacje, aby móc podejmować świadome decyzje. Za swój obowiązek uważam współpracę z nimi w taki sposób, aby mogli otrzymać wyniki zarówno pozytywne, jak i negatywne, oraz wykorzystać je do swoich zamierzonych celów. Nie wyczuwam ich oporu. Raczej sądzę, że bardzo chcą uzyskać informacje wysokiej jakości, dzięki którym będą mogli rozwijać programy, na które poświęcili swoją energię. Staram się dokonywać ocen, jeżeli przyjąłem taką rolę w toku naszych negocjacji, w sposób zapewniający, że zostanę wysłuchany; pracuję z użytkownikami docelowymi, aby umożliwić im dojście do własnych wniosków. Często oni sami są dla siebie bardziej surowi niż ja.

Ze swojego doświadczenia wiem, że przekazanie negatywnych informacji zwrotnych nie tyle wymaga aż tak dużo odwagi, co umiejętności. Nie uważam też, że klienci ewaluacji muszą być niezwykle oświeceni, aby móc usłyszeć i wykorzystać negatywne informacje zwrotne, jeżeli poprzez swoje umiejętne działanie ewaluator zbudował podstawę dla takich informacji zwrotnych, a więc są one pożądane dla dobra długofalowej efektywności. Zaangażowane zespoły realizujące programy nie chcą tracić czasu na coś, co nie działa.

## Podsumowanie

Zasadniczy przedmiot ewaluacji skoncentrowanej na wykorzystaniu (współpraca z głównymi użytkownikami docelowymi na rzecz zapewnienia zamierzonego wykorzystania ewaluacji) stał się punktem centralnym w praktyce większości zawodowych ewaluatorów. Cousins i jego współpracownicy zbadali grupę 564 ewaluatorów oraz 68 praktyków z list członkowskich zawodowych towarzystw ewaluacyjnych ze Stanów Zjednoczonych i Kanady. Kwestionariusz obejmował listę możliwych przekonań, z którymi respondenci mogli się zgodzić lub nie zgodzić. Największa zgoda panowała co do stwierdzenia: „Ewaluatorzy powinni formułować rekomendacje z badania". Pozycją wywołującą następny najwyższy poziom zgodności (71%) było stwierdzenie: „Główną funkcją ewaluatora jest maksymalizacja zamierzonego wykorzystania danych wynikających z ewaluacji przez użytkowników docelowych" (Cousins i in. 1996 s. 215). Preskill i Caracelli (1997) przedstawili podobne wyniki z badania przeprowadzonego w 1996 r. na członkach Amerykańskiego Towarzystwa Ewaluacyjnego. W ten sposób w ciągu 35 lat, od ukazania się pierwszego wydania *Utilization-focused evaluation* (Patton 1978) jej podstawowe założenia przeszły od kontrowersyjnej idei (por. Alkin 1990) do dominującej filozofii ewaluacji.

---

**Michael Quinn Patton** jest konsultantem w zakresie ewaluacji oraz rozwoju organizacji, byłym Prezesem Amerykańskiego Towarzystwa Ewaluacyjnego, autorem pięciu publikacji dotyczących tematyki ewaluacji oraz współautorem szóstej. Dr Patton posiada dyplom socjologii (licencjat) Uniwersytetu w Cincinnati oraz socjologii wsi (magister) Uniwersytetu Wisconsin. Również na tym uniwersytecie zdobył tytuł doktora socjologii. Dr Patton przez 18 lat był wykładowcą na Uniwersytecie Minnesoty. Pełnił wówczas przez 5 lat funkcję Dyrektora Centrum Badań Społecznych Minnesoty – Minnesota Center for Social Research. Obecnie Dr Patton prowadzi prywatną firmę doradczą Utilization-Focused Information and Training oraz wykłada w Union Institute Graduate School.

## Bibliografia


- AEA (American Evaluation Association), *The program evaluation standards,* 2012, http://www.eval.org/evaluationdocuments/progeval.html
- *Alkin M., Lessons Learned About Evaluation Use, Prezentacja panelowa, Międzynarodowa Konferencja Ewaluacyjna, Amerykańskie Towarzystwo Ewaluacyjne, Vancouver 2 listopada 1995.*

- *Alkin M., Debates on Evaluation, SAGE Publications, Newbury Park 1990.*
- *Alkin M. i Karin C., Thoughts on Evaluation Misutilization, "Studies in Educational Evaluation" 1988, vol. 14.*
- *Alkin M. C., Daillak R. i White P. Using Evaluations: Does Evaluation Make a Difference, SAGE Publications, Newbury Park 1979.*
- *Anderson J., Reder L. i Simon H., Situated Learning and Education, "Educational Researcher" 1996, vol. 25, s. 4: 5-21.*
- *Campbell, D. T., Methodology and Epistemology for Social Science. [w:] Overman E.S. (red.), Wybrane opracowania, University of Chicago Press, Chicago 1988.*
- *Chelimsky E., The Politics of Programme Evaluation. S. 5-22 [w:] Cordray D.S., Bloom H.S. i Light R.J. (red.), Evaluation Practice in Review. New Directions for Programme Evaluation, Jossey-Bass, San Francisco 1987, Nr 34letni.*
- *Cousins J. B., Donohue J. i Bloom G., Collaborative Evaluation in North America: Evaluators Self-reported Opinions, Practices and Consequences, "Evaluation Practice", 1996, vol. 17, s. 3: 207-226.*
- *Cousins J., Earl B. i Earl L. M. (red.), Participatory Evaluation in Education: Studies in evaluation use and organizational learning, Falmer Press, Londyn 1995.*
- *Cronbach Lee J. i in., Toward Reform of Programme Evaluation, Jossey-Bass, San Francisco 1980.*
- *Crozier M., The Bureaucratic Phenomenon, University of Chicago Press, Chicago 1964.*
- *Fletcher J. Situation Ethics: The New Morality, Westminster John Knox, Londyn 1966.*
- *Ghere G., King J. A., Stevahn L. i Minnema J., A Professional Development Unit for Reflecting on Programme Evaluation Competencies, "American Journal of Evaluation" 2006, vol. 27(1), s. 108-123.*
- *Greene J. C., Technical Quality Versus User Responsiveness in Evaluation Practice," Evaluation and Programme Planning" 1990, vol. 13 (3), s. 267-74.*
- *Hersey P., Situational Leader, Center for Leadership, North Carolina 1985.*
- *Joint Committee on Standards for Educational Evaluation, The Programme Evaluation Standard, Sage, Thousand Oaks, Ca 1994.*
- *King J.A., Studying the Local Use of Evaluation: A Discussion of Theoretical Issues and an Empirical Study, "Studies in Educational Evaluation" 1982, vol. 8, s. 175-183.*
- *King J. A., Involving Practitioners in Evaluation Studies: How Viable is Collaborative Evaluation in Schools, [w:] Cousins J. Earl B. i L. (red.), Participatory Evaluation in Education: Studies in evaluation use and organizational learning, Falmer Press, Londyn 1995b, s. 86-102.*
- *Kuzmin A., Exploration of Factors That Affect the Use of Evaluation Training in Evaluation Capacity Development, Doctoral dissertation, Union Institute and University, Cincinnati, Ohio 2005.*
- *Newcomer K. E. i Wholey J. S., Conclusion: Evaluation Strategies for Building High-Performance Programmes, [w:] Wholey J.S. i Newcomer K. E. (red.), Improving Government Performance: Evaluation Strategies for Strengthening Public Agencies and Programmes, Jossey-Bass, San Francisco 1989, s. 195-208.*
- *Patton M. Q., Essentials of Utilization-Focused Evaluation, Sage Publications, Thousand Oaks, Ca 2012.*
- *Patton M. Q., Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use, Guilford Press, Nowy Jork 2010.*
- *Patton M. Q., Utilization-Focused Evaluation, Sage Publications, Thousand Oaks, Ca 2008, wydanie 4.*
- *Patton M. Q., Developmental Evaluation, "Evaluation Practice" 1994, vol. 15,3 (październik), s. 311-320.*
- *Patton M. Q., Utilization-Focused Evaluation, Sage, Beverly Hills, Ca 1978.*
- *Preskill H. i Caracelli V., Current and Developing Conceptions of Evaluation Use: Evaluation Use TIG Survey Results, "Evaluation Practice" 1997, vol. 18(3), s. 209-225.*
- *Preskill H. i Torres R., The Readiness for Organizational Learning and Evaluation Instrument, Developmental Studies Center, Oakland, Ca 2000.*
- *Preskill H. i Torres R., Evaluative Inquiry for Learning in Organizations, Sage Publications, Thousand Oaks, Ca 1998.*
- *Scriven M., Beyond Formative and Summative Evaluation, [w: ] McLaughlin M.W. i Phillips D.C. (red.), Evaluation and Education: At Quarter Century, 90th Yearbook of the National Society for the Study of Education, University of Chicago Press, Chicago 1991, s.18-64.*
- *Scriven M., Evaluation Thesaurus, Sage, Newbury Park, Ca 1991a, wydanie 4.*
- *Seiden K., Development and Validation of the `Organizational Readiness for Evaluation' Survey Instrument, Niepublikowana rozprawa doktorska, University of Minnesota 2000.*
- *Shadish W. R., Jr., Newman D. L., Scheirer M. A., Wye C., Guiding Principles for Evaluators, "New Directions for Programme Evaluation", Jossey-Bass, San Francisco 1995, vol. 66.*
- *Sonnichsen R. C., High Impact Internal Evaluation, Sage Publications, Thousand Oaks, Ca 2000.*
- *Sonnichsen R. C., Can Governments Learn? [w:] Leeuw F., Rist R., Sonnichsen R. (red.) Comparative Perspectives on Evaluation and Organizational Learning, Transaction, New Brunswick, N.J. 1993.*
- *Torres R., Preskill H. i Piontek M. E., Evaluation Strategies for Communicating and Reporting: Enhancing Learning in Organizations, Sage, Thousand Oaks, Ca 1996.*
- *Yarbrough D. B., Shulha L. M., Hopson R. K. i Caruthers F. A., The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users, Sage, Thousand Oaks, Ca 2011, wydanie 3.*

*Michael Quinn Patton*

# Przyszłe trendy w ewaluacji

Ewaluacja jako profesja przeszła ogromne zmiany w ciągu ostatniego ćwierćwiecza. Początki zawodu ewaluatora sięgają 1975 roku, kiedy opublikowano pierwszy „Podręcznik badań ewaluacyjnych" (ang. *Handbook of Evaluation Research*). W tym samym roku w Stanach Zjednoczonych powstała Sieć Ewaluacyjna (ang. *Evaluation Network*) i Towarzystwo Badań Ewaluacyjnych (ang. *Evaluation Research Society*) – w 1984 roku połączono je w Amerykańskie Towarzystwo Ewaluacyjne (ang. *American Evaluation Association*). W tym okresie nie ukazywały się żadne specjalistyczne czasopisma poświęcone ewaluacji, nie istniały odpowiednie instytucje szkoleniowe, nie zdefiniowano standardów ewaluacji, a podręczników poświęconych tematowi ewaluacji opublikowano zaledwie kilka. Od tamtego czasu sytuacja zmieniła się diametralnie: globalna profesja ewaluacji tworzy obecnie bogatą i zróżnicowaną mozaikę. W niniejszym tekście przedstawię główne tendencje, które będą moim zdaniem kształtować ewaluację w następnym ćwierćwieczu.

## 1. Międzynarodowa i międzykulturowa ekspansja ewaluacji: globalizacja i różnorodność

Żaden trend nie wywarł większego wpływu na rozwój ewaluacji w ostatniej dekadzie niż jej nieustannie rosnący, globalny zasięg. W 1995 roku ewaluatorzy z 61 krajów zgromadzili się na pierwszej prawdziwie międzynarodowej konferencji ewaluacyjnej, która odbyła się w Vancouver w Kanadzie. Dziesięć lat później, na drugą międzynarodową konferencję w Toronto ściągnęło 2330 ewaluatorów z całego świata. W latach 90. powstało również Europejskie Towarzystwo Ewaluacyjne, założone w 1994 roku w Hadze, oraz Afrykańskie Towarzystwo Ewaluacyjne, utworzone w 1999 roku w Nairobi. W 2012 roku to ostanie zorganizowało szóstą, ogólnoafrykańską konferencję w Ghanie. Obecnie istnieje ponad 60 krajowych stowarzyszeń ewaluacyjnych na całym świecie, m.in. w Polsce, Rosji, Japonii, Malezji, Mongolii, Brazylii, Kolumbii, Peru, RPA, Zimbabwe, Nigerii, Nowej Zelandii i na Sri Lance – te kraje to zaledwie kilka przykładów z długiej listy. W 2003 roku w Limie (Peru) odbyło się inauguracyjne posiedzenie nowej Międzynarodowej Organizacji Współpracy Ewaluacyjnej (ang. *International Organization for Cooperation in Evaluation*, IOCE), mającej charakter inicjatywy sieciowej i patronackiej, wspierającej krajowe i regionalne stowarzyszenia ewaluacyjne działające na całym świecie. W 2002 roku w Pekinie utworzono Międzynarodowe Stowarzyszenie Ewaluacji Programów Rozwojowych (ang. *International Development Evaluation Association,* IDEAS), aby wspierać ewaluatorów zajmujących się przede wszystkim zagadnieniami związanymi z krajami rozwijającymi się, a jego pierwsza konferencja (organizowana raz na dwa lata) miała miejsce w New Delhi w 2005 roku. Sieć Monitorowania, Ewaluacji i Systematyzacji Ameryki Łacińskiej i Karaibów (ang. *Network for Monitoring, Evaluation, and Systematization of Latin America and the Caribbean,* ReLAC) powstała w 2005 roku w Peru.

Za pośrednictwem swojego Międzynarodowego programu szkoleń w zakresie ewaluacji programów rozwojowych (ang. *International Program for Development Evaluation Training,* IPDET), Bank Światowy organizuje na Uniwersytecie Carleton w Ottawie (Kanada) miesięczne szkolenia dla uczestników z krajów rozwijających się. Instytucje międzynarodowe opracowały kompleksowe wytyczne dotyczące przeprowadzania ewaluacji. Poszczególne stowarzyszenia krajowe dokonały przeglądu i przyjęły Standardy Ewaluacji opra-

cowane przez Wspólny Komitet ds. Standardów Ewaluacji (ang. *Joint Committe on Standards for Evaluation*), dostosowując je do własnego kontekstu społeczno-politycznego, a zarazem podkreślając, że ewaluacje należy oceniać przez pryzmat ich użyteczności, wykonalności, prawidłowości i dokładności.

Rządy na całym świecie tworzą nowe systemy monitoringu i ewaluacji, mające na celu budowanie zarządzania opartego na wynikach i pomiar skuteczności działań służących wspieraniu rozwoju. Instytucje międzynarodowe zaczęły stosować ewaluację również dla oceny pełnego zakresu działań rozwojowych realizowanych w krajach rozwijających się. Większość dużych organizacji międzynarodowych posiada własne jednostki ewaluacyjne, a także wytyczne, protokoły, konferencje, szkolenia, strony internetowe i specjalistów do spraw zasobów.

Ten wspólny wysiłek podejmowany na skalę globalną doprowadził do opracowania strategii i podejść, które mogą być udostępniane na całym świecie. Globalizacja ewaluacji wspiera zatem naszą międzynarodową współpracę służącą pogłębianiu wiedzy na temat czynników zwiększających skuteczność programów i wykorzystanie ewaluacji. Perspektywa międzynarodowa stanowi również wyzwanie dla „zachodnich" definicji i założeń kulturowych dotyczących sposobu przeprowadzania i oceny jakości ewaluacji. Biorąc pod uwagę, że standardy ewaluacji są tłumaczone na różne języki, krajowe stowarzyszenia wzbogacają poszczególne wersje o własne niuanse kulturowe i dostosowują praktyki do lokalnego kontekstu i uwarunkowań politycznych, społecznych, organizacyjnych, ekonomicznych i kulturowych. Uważam, że ta tendencja polegająca na adaptacji kulturowej i politycznej będzie kształtować ewaluację również w przyszłości.

## 2. Ewaluacja jako „transdyscyplina" i zawód

Filozof i teoretyk ewaluacji Michael Scriven scharakteryzował ją jako *transdyscyplinę*, ponieważ każda dyscyplina, profesja i dziedzina korzysta z pewnej formy ewaluacji, czego najbardziej oczywistym przykładem jest prawdopodobnie ewaluacja dokonań studentów uczestniczących w różnych programach naukowych i kursach, a także periodyki naukowe, w których nowe badania są poddawane ewaluacji przez innych naukowców z danej dziedziny zanim zostaje podjęta decyzja o tym, czy dany artykuł zasługuje na publikację. Ewaluacja służy innym dyscyplinom, nawet jeżeli jest dyscypliną samą w sobie – stąd jej nowy, transdyscyplinarny status. Statystyka, logika i ewaluacja są przykładami transdyscyplin, ponieważ ich metody, właściwy im sposób myślenia i bazy wiedzy są wykorzystywane w innych dziedzinach badań, np. w edukacji, ochronie zdrowia, opiece społecznej, inżynierii, badaniach środowiskowych, itd. Spodziewam się, że ewaluacja będzie w coraz większym stopniu uznawana za dziedzinę transdyscyplinarną, co będzie miało wpływ zarówno na sposób jej przeprowadzania (przy udziale zespołów interdyscyplinarnych), jak i na sposób prowadzenia badań dotyczących samej ewaluacji.

Jedną z ważnych implikacji tej tendencji jest uznanie, że ewaluacja posiada własną bazę wiedzy na temat czynników, które wpływają na skuteczność programu i sposobów przeprowadzania użytecznych ewaluacji. Zbyt często tych ostatnich dokonują ekonomiści i inni przedstawiciele nauk społecznych, którym brakuje odpowiedniej wiedzy ewaluacyjnej lub doświadczenia. Wielu spośród nich nie wie, że istnieją standardy jakości ewaluacji. Nie znają nowych rozwiązań, które pojawiły się w zakresie metod i modeli ewaluacji. Mam nadzieję, że ewaluacja zostanie w końcu uznana zarówno za zawód, jak i transdyscyplinarną dziedzinę wiedzy, oraz że ci, którzy zlecają i finansują ewaluacje, będą pamiętać o tym, by w zespołach dokonujących ewaluacji na całym świecie znaleźli się wykwalifikowani profesjonalni ewaluatorzy.

## 3. Wzrost zainteresowania polityków odpowiedzialnością, wskaźnikami efektywności i transparentnością

Pomiar efektywności jest obecnie *de rigueur* w polityce na całym świecie. Znajduje to odzwierciedlenie w coraz większej uwadze poświęcanej wartościom docelowym wskaźników efektywności, benchmarkom i „kamieniom milowym" w Milenijnych Celach Rozwoju i międzynarodowych traktatach, takich jak protokół z Kioto w zakresie emisji gazów cieplarnianych. Wskaźniki efektywności stały się tak ważne i powszechnie stosowane, że stanowią obecnie stały element ustawodawstwa, zarządzania i międzynarodowych umów dotyczących monitoringu. Bieżący monitoring wskaźników i ich porównywanie z ustalonymi wartościami docelowymi można nazwać pomiarem efektywności lub monitorowaniem efektywności. Służy to trzem głównym celom: (1) ocenie wpływu polityk realizowanych przez rząd na usługi publiczne, (2) identyfikacji sprawnie funkcjonujących instytucji i urzędników państwowych, a także tych, które nie realizują w pełni swojego potencjału, oraz (3) odpowiedzialności publicznej. Rządy i organizacje międzynarodowe zarazem monitorują usługi publiczne i są monitorowane w oparciu o wskaźniki efektywności. To sprawia, że polityczne znaczenie monitoringu jest ogromne.

Dobrze przeprowadzony monitoring efektywności jest przydatny w szerszych ramach monitorowania i ewaluacji. Przeprowadzony nieprawidłowo może okazać się bardzo kosztowny i nie tylko nieskuteczny, ale również szkodliwy czy wręcz destrukcyjny. W związku z tym, w przyszłości potrzeba będzie bardziej zaawansowanych systemów monitoringu efektywności, które będą uwzględniały ich nieuniknione ograniczenia. Potencjalny pozytywny wkład monitoringu efektywności odzwierciedla często powtarzane stwierdzenie, że „co zostaje zmierzone, zostaje również wykonane". Odpowiednio opracowane wskaźniki koncentrują uwagę na priorytetowych rezultatach i zapewniają odpowiedzialność za ich osiągnięcie. Minusem wskaźników efektywności jest to, że pomiar nieodpowiedniego działania oznacza, że to nieodpowiednie działanie zostaje zrealizowane.

Potrzebna jest edukacja i szkolenia w zakresie właściwego stosowania i interpretacji wskaźników efektywności. Szczególny nacisk trzeba będzie położyć na znaczenie niezależnej kontroli i transparentności jako gwarantów odpowiedzialności publicznej, dyscypliny metodologicznej i sprawiedliwego traktowania monitorowanych osób i/ lub instytucji.

Obawy o niewłaściwe wykorzystanie wskaźników efektywności wynikają z prawa Campbella, sformułowanego przez Donalda T. Campbella, jednego z najwybitniejszych pionierów ewaluacji: *Im bardziej dany ilościowy wskaźnik społeczny zostaje wykorzystany w społecznym procesie decyzyjnym, tym bardziej będzie on przedmiotem wypaczającej go presji, i tym silniej będzie zakłócać i zaburzać procesy społeczne, które ma monitorować*[1]. Rozważmy następujący przykład: policjanci w Nowym Orleanie manipulowali statystykami dotyczącymi przestępczości lokalnej, aby sprawić wrażenie, że spadek przestępczości jest rezultatem polityki polegającej na przyznawaniu nagród oficerom odpowiedzialnym za okręgi z najniższym współczynnikiem przestępczości. W następstwie tego skandalu pięciu policjantów zostało zwolnionych.

Oczekuje się, że rządy i politycy będą wyznaczać cele i relacjonować postępy, co ma stanowić podstawę ich odpowiedzialności publicznej. Przydatność wskaźników efektywności zależy od ich wiarygodności, odpowiedniości, ważności, transparentności i sensowności – oraz od właściwego i sprawiedliwego procesu ich interpretacji. Wskaźniki efektywności są jednym z bardzo szerokiego zestawu narzędzi ewaluacyjnych, który obejmuje szeroką gamę metodologii, technik gromadzenia danych, środków i modeli. Biorąc pod uwagę szybkie rozpowszechnianie podejść do monitoringu efektywności, istnieje niebezpieczeństwo, że wiele osób uzna pomiar efektywności za wystarczający, lub wręcz za zamiennik ewaluacji. Niemniej jednak, pomiar efek-

[1] Ang. *"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor".*

tywności zaledwie pokazuje trendy i kierunki. Wskaźniki mówią nam, czy coś wzrasta, maleje czy pozostaje bez zmian. Ewaluacja pozwala przejść na wyższy poziom, pytając, dlaczego zmiany wskaźników przyjmują dany kierunek, w jaki sposób zmiany wskaźników związane są z konkretnymi interwencjami, co napędza zmiany wskaźników i jakimi wartościami powinniśmy się kierować interpretując wskaźniki przy dokonywaniu osądów. Pomiar efektywności skoncentrowany na wykorzystaniu dodaje do powyższego znaczenie precyzyjnego określenia głównych zamierzonych użytkowników i zakładanych sposobów wykorzystania wskaźników efektywności. Biorąc pod uwagę rosnące znaczenie pomiaru efektywności w sektorze publicznym na całym świecie, ewaluatorzy, decydenci i ogół społeczeństwa powinni zrozumieć zarówno zalety, jak i ograniczenia kluczowych wskaźników efektywności (ang. *Key Performance Indicators* – KPI).

## 4. Budowanie potencjału ewaluacyjnego i rozwój umiejętności

Dowiedzieliśmy się już, że organizacje, programy i projekty wymagają odpowiednich zasobów, wiedzy i poziomu rozwoju organizacyjnego w celu skutecznego zarządzania i dokonywania ewaluacji. Zwłaszcza w ostatnim dziesięcioleciu lepiej zrozumieliśmy znaczenie zwiększania potencjału ewaluacyjnego w organizacjach poprzez zaszczepienie myślenia ewaluacyjnego w kulturze organizacyjnej. Instytucje mają coraz większą zdolność do generowania, przechowywania i pobierania ogromnych ilości informacji i danych. Problem polega na tym, by wiedzieć jak korzystać ze wszystkich tych informacji. Potencjał technologiczny w zakresie gromadzenia i komputeryzacji informacji znacznie przekracza możliwości większości organizacji w zakresie przetwarzania i zrozumienia ich.

Osoby stojące na czele organizacji nieustannie zmuszone są podejmować decyzje dotyczące tego, co warto wiedzieć, co można pominąć i jak przekładać wyniki na działania i decyzje. Oznacza to, że w przyszłości coraz bardziej będziemy polegać na ewaluatorach nie tylko w zakresie dokonywania ewaluacji, ale również budowania trwałego potencjału ewaluacyjnego organizacji. Ewaluatorzy będą musieli dołożyć starań by zbudować i utrzymać zainteresowanie ewaluacją. Identyfikacja zamierzonych użytkowników polega częściowo na selekcji, a częściowo na wychowaniu. Ci potencjalni użytkownicy, którzy nisko oceniają ewaluację lub nie interesują się nią, mieli być może złe wcześniejsze doświadczenia, lub po prostu nie zastanowili się nad korzyściami płynącymi z ewaluacji. Częścią ewaluacji będzie zatem podtrzymywanie zainteresowania i zapewnianie, aby użytkownicy zobowiązali się do wykorzystania wyników. Nawet ci, którzy od początku cenią ewaluację, potrzebują szkoleń i wsparcia, aby skutecznie wykorzystywać informacje.

Poza wiedzą metodologiczną, ewaluatorzy muszą również rozwijać swój potencjał i umiejętności. Okazuje się, że aby skutecznie umożliwiać wykorzystanie ewaluacji, potrzebują umiejętności budowania relacji, wspierania współpracy grupowej, zarządzania konfliktami, umiejętności „chodzenia po politycznej linie" i efektywnej komunikacji interpersonalnej. Umiejętności techniczne i wiedza z zakresu nauk społecznych nie wystarczą, aby ewaluacje rzeczywiście były wykorzystywane – niezbędne są umiejętności interpersonalne. Niezależnie od ideałów racjonalności w procesach decyzyjnych w nowoczesnych organizacjach, dynamika osobista i polityczna wpływa na to, co rzeczywiście się dzieje. Ewaluatorzy pozbawieni doświadczenia i umiejętności w zakresie relacji międzyludzkich i polityki szybko zauważą, że wyniki ich pracy są ignorowane lub, co gorsza, niewłaściwie wykorzystywane.

## 5. Wykorzystanie procesów będzie coraz lepiej rozumiane i doceniane

W rozdziale dotyczącym ewaluacji skoncentrowanej na wykorzystaniu omówiono wykorzystanie procesu i jego znaczenie dla ewaluacji. Uważam, że to znaczenie wzrośnie. *Wykorzystanie procesu* koncen-

truje się na uczeniu się i na potencjale budowanym dzięki udziałowi w ewaluacji. Wpływ podejść uczestniczących i opartych na współpracy wykracza poza wykorzystanie wyników. Uczestnicy ewaluacji uczą się myśleć bardziej krytycznie. Uczą się, w jaki sposób formułować pytania, interpretować dane, określać priorytety, wyjaśniać modele interwencji i skupić się na rezultatach. Uczą się korzystania z logiki i rozumowania ewaluacyjnego. W ten sposób, wykorzystanie procesu buduje potencjał trwałego zaangażowania i wykorzystania ewaluacji.

Rozumowanie w kategoriach tego, co jest jasne, szczegółowe, konkretne i obserwowalne nie przychodzi łatwo tym, dla których dwuznaczności, ogólniki i niesprawdzone przekonania stanowią podstawę działania. Stanowią oni większość. Praktykujący logikę ewaluacyjną stanowią bardzo nieliczną mniejszość. Dobrą wiadomością jest jednak to, że rozumowanie ewaluacyjne staje się niezwykle cenne dla wszystkich, którzy kiedykolwiek go spróbowali. To te osoby tworzą popyt na usługi ewaluacyjne.

Wykorzystanie procesu różni się od korzystania z wyników zawartych w raporcie ewaluacyjnym. Można tę zależność porównać do różnicy pomiędzy nauką tego, jak skutecznie się uczyć, a zdobywaniem konkretnej wiedzy na dany temat. Nauka myślenia ewaluacyjnego oznacza naukę tego, jak się uczyć oraz jak krytycznie myśleć, a ci, którzy angażują się w ewaluację, uczą się poprzez działanie. Ułatwienie myślenia ewaluacyjnego otwiera nowe możliwości dla oddziaływania, które organizacje i pracodawcy cenią, ponieważ zdolność angażowania się w tego rodzaju myślenie może przynieść trwalszą wartość niż pewien określony zbiór wyników. Ma to szczególne znaczenie dla organizacji, które są powszechnie nazywane „organizacjami uczącymi się" (ang. *learning organizations*). Uczenie się patrzenia na świat oczyma ewaluatora często ma trwały wpływ na tych, którzy uczestniczą w ewaluacji – wpływ, który może być większy i trwalszy niż wyniki tej samej ewaluacji. Te ostatnie mają bardzo krótki „czas połowicznego rozpadu" (ang. „*half life*"); bardzo szybko się starzeją, co wynika z szybkiego tempa zmian zachodzących na świecie. Konkretne wyniki zazwyczaj są „odpowiednie" w ograniczonym zakresie. W przeciwieństwie do nich, uczenie się ewaluacyjnego myślenia i działania może mieć długotrwały wpływ. W związku z tym, doświadczenie związane z angażowaniem się w ewaluację, dla faktycznie zaangażowanych interesariuszy, może ukształtować ich sposób myślenia, otwartość na „testowanie" rzeczywistości i sposób, w jaki postrzegają to, co robią. Spodziewam się zatem, że w przyszłości wykorzystanie procesu będzie jeszcze istotniejszym elementem budowy potencjału w ramach ewaluacji.

## 6. Metodologiczne debaty na temat rygoru

W przeszłości ewaluacji prowadzono intensywne dyskusje na temat tego, co składa się na rygor metodologiczny, w szczególności dyskusje na temat wartości i wiarygodności metod jakościowych w porównaniu z metodami ilościowymi. Dyskusje te przybierały różne formy; adwersarze mniej lub bardziej zaciekle spierali się ze sobą. Debata ponownie przybiera na sile, tym razem skupiając się na tym, czy randomizowane eksperymenty kontrolowane stanowią złoty standard w ewaluacji wpływu.

Zasadności odpowiednio zastosowanych metod eksperymentalnych i pomiarów ilościowych nigdy nie poddawano w wątpliwość. Jednak w latach 90. powszechna estyma dla metod jakościowych znacznie wzrosła. Dzisiaj szczególnie cenione są metody mieszane. O ile w środowisku panuje konsensus co do tego, że ewaluatorzy muszą znać i stosować różne metody, aby móc reagować na niuanse poszczególnych pytań ewaluacyjnych i indywidualnych potrzeb interesariuszy, o tyle kwestia, co stanowi *metodologiczny złoty standard* jest przedmiotem żywej polemiki. Z jednej strony, istnieje zgoda co do tego, że rygorystyczność należy oceniać z perspektywy stosowania odpowiednich metod dla konkretnego celu i pytania ewaluacyjnego. O ile to możliwe, cenne jest stosowanie wielu metod – zarówno ilościowych, jak i jakościowych. Panuje jednak również powszechne przekonanie, że jedno pytanie jest ważniejsze niż pozostałe (pytanie

o związek przyczynowy), oraz że jedna z metod (randomizowane próby kontrolne) jest lepsza od pozostałych. Tu właśnie pojawia się kwestia złotego standardu.

Debata toczy się nie tylko wśród metodologów ewaluacji. Dotyczy ona również praktyków, podobnie zresztą jak użytkowników – decydentów, pracowników i menedżerów programów oraz ich sponsorów. Wszystko może jednak zatonąć w dyskusji na temat tego, czy wyniki w formie statystyk eksperymentów („twarde" dane) są ważniejsze i bardziej naukowe niż wyniki quasi-eksperymentów i jakościowych studiów przypadków („miękkie" dane). Kto zechciałby przeprowadzić (lub finansować) drugorzędną ewaluację jeśli istnieje ogólnie przyjęty złoty standard? Jakie naprawdę są mocne i słabe strony różnych metod, w tym eksperymentów (które, jak się okazuje, mają również swoje minusy)? Co oznacza dostosowanie metody do zadanego pytania?

Jeśli ewaluatorzy mają zaangażować użytkowników w podejmowanie decyzji na temat metod, zarówno ewaluatorzy, jak i użytkownicy muszą zrozumieć debatę na temat metod, a ewaluatorzy muszą ułatwiać wybór metod odpowiednich dla danego celu ewaluacyjnego. Oznacza to informowanie głównych interesariuszy na temat dostępnych, uzasadnionych opcji, potencjalnych korzyści stosowania wielu metod oraz mocnych i słabych stron poszczególnych podejść.

Zarówno Amerykańskie Towarzystwo Ewaluacyjne, jak i Europejskie Towarzystwo Ewaluacyjne wspierają w swojej polityce eklektyzm metodologiczny i dostosowanie metod ewaluacji do jej charakteru i potrzeb informacyjnych głównych użytkowników, którym wyniki ewaluacji mają służyć. Według przyjętych przez te dwie organizacje deklaracji, złoty standard powinna stanowić stosowność metodologiczna.

Problemem w tym, że rzekoma wyższość ilościowego/eksperymentalnego podejścia uniemożliwia poważne rozważenie alternatywnych metod, a w konsekwencji miliony dolarów przekazuje się na ewaluacje prowadzone metodami eksperymentalnymi. Metody te mają swoje zalety, ale również kilka poważnych wad. Pochwała złotego standardu oznacza, że zlecający i przeprowadzający ewaluacje muszą wyjść od pytania „W jaki sposób możemy zastosować metody eksperymentalne w tej ewaluacji?", zamiast pytać „Jakie metody będą odpowiednie, biorąc pod uwagę sytuację i potrzeby informacyjne?". To prestiż metody określa pytanie i projekt ewaluacji, a nie względy użyteczności, wykonalności, prawidłowości i dokładności.

W ramach złotego standardu, ewaluację wpływu o wysokiej jakości definiuje się jako testowanie hipotez, sformułowanych w sposób dedukcyjny, poprzez losowe przypisywanie uczestników programu do grupy objętej działaniem programu i kontrolnej oraz ilościowy pomiar rezultatów. Już *z definicji*, żadne inne opcje nie są warte poważnego rozważenia.

Istnieją jednak alternatywy. Aby ocenić procesy, rezultaty oraz wpływ programu, eksperymenty można zastąpić innymi metodami. W ostatnim ćwierćwieczu, alternatywy te zostały wykorzystane przez ewaluatorów, którzy stwierdzili, że dominującemu paradygmatowi nie udało się odpowiedzieć na zadane pytania, a nawet właściwie ich zadać. Debata o tym, czy metody eksperymentalne stanowią metodologiczny złoty standard kręci się, częściowo, wokół tego, jaki poziom i rodzaj dowodów jest niezbędny do określenia czy dana interwencja jest skuteczna. Pozwolę sobie zilustrować to przykładem z mojej książki, *Utilization-Focused Evaluation* (Patton 2008, Rozdział 12).

Rozważmy wyzwanie polegające na eliminacji robaków jelitowych u dzieci – ogromny problem w krajach rozwijających się. Załóżmy, że chcemy zewaluować interwencję, polegającą na tym, że dzieciom w wieku szkolnym cierpiącym na biegunkę podawane są leki na odrobaczenie, co ma na celu zwiększenie frekwencji szkolnej i wyników. W celu przypisania interwencji do pożądanego rezultatu, zwolennicy randomizowanych prób kontrolnych nalegaliby na zastosowanie takiego modelu ewaluacji, w którym uczniowie cierpiący na biegunkę zostają losowo podzieleni na grupę objętą interwencją (czyli tych, którym podaje się lek) i grupę kontrolną (tych, którzy nie otrzymują leku). Następnie porównaliby frekwencję szkolną i wyniki testów członków obu grup. Jeśli po miesiącu frekwencja wśród dzieci przyjmujących lek byłaby wyższa (na

statystycznie istotnym poziomie) w porównaniu z grupą kontrolną, poprawę wyników możnaby przypisać przeprowadzonej interwencji (podawaniu leku).

Zwolennicy badań jakościowych kwestionują zasadność uczestnictwa grupy kontrolnej w tym przypadku. Załóżmy, że przeprowadzi się rozmowy z uczniami, rodzicami, nauczycielami i miejscowymi pracownikami służby zdrowia na temat przyczyn niskiej frekwencji szkolnej i gorszych wyników testów. Niezależnie od siebie, każda z tych grup stwierdza, że to biegunka jest główną przyczyną słabej frekwencji szkolnej i gorszych wyników. Gromadzenie danych od różnych grup (uczniów, rodziców, nauczycieli, pracowników służby zdrowia) nazywa się *triangulacją*, co jest sposobem sprawdzania zgodności danych pochodzących z różnych źródeł. Po przeprowadzeniu podstawowych wywiadów, uczniowie otrzymują lek odrobaczający. U pacjentów przyjmujących lek obserwuje się wzrost frekwencji szkolnej i polepszenie wyników, a w przeprowadzonych później rozmowach uczniowie, rodzice, nauczyciele i pracownicy służby zdrowia niezależnie od siebie potwierdzają, że zmiany są wynikiem przyjmowania leku odrobaczającego i mniejszą częstotliwością występowania biegunki. Czy są to wiarygodne, przekonujące dowody?

Ci, którzy uważają taki model ewaluacji za wystarczający twierdzą, że wyniki charakteryzuje zarówno racjonalność, jak i empiryczność, a wysoki koszt związany z dodaniem grupy kontrolnej nie jest niezbędny do ustalenia związku przyczynowego. Uznaliby również za nieetyczne pozbawienie uczniów cierpiących na biegunkę dostępu do leku, kiedy jego stosowanie samo w sobie jest korzystne. Zwolennicy randomizowanych prób kontrolnych stwierdziliby, że bez grupy kontrolnej na rezultaty mogłyby wpłynąć inne, nieznane czynniki, i że tylko istnienie sytuacji kontrfaktycznej (udział grupy kontrolnej) pozwala ustalić rzeczywisty wpływ interwencji.

Jak pokazuje powyższy przykład, ewaluatorzy i metodologowie znajdujący się po różnych stronach tej debaty mają inne spojrzenie na to, co jest w prawdziwym świecie wystarczającym dowodem na to, że dany rezultat możemy przypisać danemu działaniu. Nie jest to po prostu akademicka debata. Chodzi o miliony dolarów przeznaczane na ewaluację, której wyniki mają wpływ na sposób wydatkowania miliardów dolarów na międzynarodową pomoc rozwojową.

W 2008 roku, główni fundatorzy zainteresowani ewaluacją powołali do życia Międzynarodową Inicjatywę na rzecz Oceny Wpływu (ang. *International Initiative for Impact Evaluation, 3ie*) . Misją 3ie jest „wspieranie dążeń do dobrobytu poprzez zachęcanie do tworzenia i wykorzystania wyników rygorystycznych ewaluacji wpływu w decyzjach politycznych, które mają na celu doskonalenie programów rozwoju społecznego i gospodarczego w krajach o niskich i średnich dochodach". Przyszłość ewaluacji będzie obejmować ożywioną, międzynarodową debatę na temat definicji „rygorystyczności".

## 7. Myślenie systemowe i nauka o złożoności jako ramy ewaluacji

Ostatnią tendencją, którą dostrzegam, jest położenie większej wagi na myślenie systemowe w ewaluacji i jego wykorzystanie w większym stopniu. Ewaluacja została silnie uzależniona od liniowych modeli logicznych, które pozwalają na konceptualizację interwencji. W ostatniej dekadzie przeprowadzenie ewaluacji często obejmowało konceptualizację i testowanie modelu logicznego lub teorii zmiany danego programu. Biorąc pod uwagę, że ewaluatorzy zaangażowali się w pracę z realizatorami programów, aby w sposób bardziej przejrzysty określić model lub teorię danego programu, oczywiste stało się, że nie tylko działaniem końcowym, ale również początkowym. Oznacza to, że tradycyjne modele planowania obejmują pewien szereg etapów, pośród których planowanie jest pierwszym, po czym następuje realizacja programu, a następnie ewaluacja, co znaczy, że jest ona ostatnim elementem działania. Jednak aby opracować plan lub projekt programu, który mógłby rzeczywiście zostać poddany ewaluacji, niezbędny jest udział ewaluatorów oraz myślenie ewaluacyjne od samego początku. Myślenie ewaluacyjne staje się częścią

procesu projektowania programu, obejmującego w szczególności konceptualizację modelu logicznego programu lub teorii zmiany i zadanie następującego pytania: W jaki sposób program doprowadzi do osiągnięcia pożądanych rezultatów? Ten sposób działania jest przykładem *wykorzystania procesu*, w którym ewaluacja ma wpływ na program zupełnie niezależnie od ustaleń dotyczących jego skuteczności. Sam proces konceptualizacji teorii zmiany może wpłynąć na sposób jego realizacji, jego zrozumienie, sposób, w jaki się o nim mówi i doskonali go. Jak wspomniano wcześniej, proces myślenia ewaluacyjnego przynosi takie właśnie rezultaty.

Dla ewaluatorów ma to ogromne znaczenie. Oznacza, że muszą oni (1) wykazać się przenikliwością w procesie konceptualizacji programu i teorii zmiany oraz (2) sprawnie współpracować z osobami zaangażowanymi w realizację programu, decydentami i podmiotami finansującymi, co ma ułatwić werbalizację ukrytych teorii zmian. Biorąc pod uwagę znaczenie tych zadań, ogromne znaczenie ma to, jakie ramy teorii zmiany może zaoferować ewaluator. Myślenie systemowe jest jedną z takich ram – spodziewam się, że będzie ono coraz bardziej cenione i częściej wykorzystywane w procesie ewaluacji.

Liniowe modele logiczne prowadzą do konstruowania modeli i schematów, w których wkłady przekładają się na działania, działania na produkty, a produkty na rezultaty. Z drugiej strony, patrząc na program z perspektywy systemów skupiamy się na współzależnych konfiguracjach czynników, które prowadzą do rezultatów, a nie na prostym modelu przyczynowo-skutkowym. Ramy systemowe opierają się na kilku podstawowych założeniach:

a. Całość jest większa niż suma poszczególnych części.
b. Części są od siebie wzajemnie zależne, przez co zmiana zachodząca w jednej ma wpływ na wszystkie pozostałe i na ich wzajemne relacje.
c. Model opiera się na wzajemnych relacjach.
d. Systemy składają się z podsystemów i funkcjonują w jeszcze większych systemach.

W 2006 roku, Amerykańskie Towarzystwo Ewaluacyjne opublikowało pierwszą w swojej historii monografię: antologię tekstów zatytułowaną *Systems Concepts in Evaluation* pod redakcją Boba Williamsa i Iraja Imana. W monografii tej przedstawiono szeroką gamę rozwiązań systemowych i pokazano różnorodność podejść w ramach systemów. Komentując tę różnorodność, redaktorzy tomu napisali:

*Ci spośród Was, którzy poszukują spójności w tym, co uznajemy za istotne dla ewaluacji systemy koncepcyjne, powinni podczas lektury niniejszej publikacji poszukiwać raczej wzorców, a nie definicji. My dostrzegamy trzy takie wzorce:*

*1. Perspektywy. Wykorzystanie koncepcji systemów zakłada, że ludzie skorzystają na umiejętności spojrzenia na świat inaczej. Dla praktyków systemów, ta motywacja jest oczywista, celowa i ma zasadnicze znaczenie dla ich podejścia. Jednakże samo globalne podejście i umiejętność dostrzegania szerokiego kontekstu lub odkrywanie wzajemnych powiązań nie oznacza „systemowego" podejścia do danego problemu. Tym, co czyni je systemowym jest sposób patrzenia na kontekst – węższy lub szerszy – i odkrywanie wzajemnych powiązań pomiędzy elementami. „System" jest tyleż samo „wyobrażeniem" świata rzeczywistego, co jego fizycznym opisem.*

*2. Granice. Od granic zależy nasze postrzeganie systemów. Definiują, kto i co leży w granicach danego badania i poza nimi. Granice wyznaczają i wskazują na istotne różnice (np. co jest „w", a co „poza"). Określają, kto lub co skorzysta z danego badania, a także kto i co na nim ucierpi. Mówiąc o granicach, mówimy przede wszystkim o wartości, ponieważ stanowią one ocenę wartości. Określanie granic jest elementarną częścią funkcjonowania, badania i myślenia systemowego.*

*3. Poplątane systemy. Dostrzegamy systemy w ramach systemów, systemy nakładające się na inne systemy i systemy splątane z innymi systemami. Nierozsądne byłoby zatem skupienie się na jednej wizji lub definicji systemu nie podejmując wysiłku zbadania jego relacji z innym. Gdzie kończy się jeden, a zaczyna drugi system? Czy nakładają się one na siebie? Kto ma największe szanse doświadczyć lub odczuć wpływ wzajemnego przenikania się systemów? Jakie systemy istnieją w ramach innych systemów i dokąd prowadzą? Osoba myśląca systemowo*

*zawsze spogląda wewnątrz, na zewnątrz, poza i pomiędzy łatwo identyfikowalne granice systemów, a następnie poddaje je krytycznej ocenie i, jeśli to konieczne, zmienia ich początkowe granice (Williams i Iman 2006, s. 6).*

Ewaluacja jako zawód i transdyscyplina dopiero zaczyna dostrzegać i uwzględniać implikacje teorii systemów. Uważam, że w przyszłości ewaluatorzy będą coraz intensywniej wykorzystywać koncepcje i myślenie systemowe.

## Podsumowanie

Wydaje się, że ewaluacja ma zapewnioną przyszłość. Szybki rozwój tej dziedziny sprawił, że stała się ona dynamiczną i żywą profesją. W powyższym, krótkim przeglądzie wskazałem na siedem tendencji, które będą towarzyszyć międzynarodowemu rozwojowi ewaluacji. Spodziewam się, że będziemy obserwować następujące zjawiska:

1. zwiększoną międzynarodową i międzykulturową ekspansję ewaluacji, jej globalizację i rosnącą różnorodność;
2. rosnące uznanie dla ewaluacji jako transdyscypliny i zawodu;
3. wzrost zainteresowania odpowiedzialnością, wskaźnikami efektywności i transparentnością;
4. większy nacisk na budowanie potencjału i rozwój umiejętności;
5. większe zrozumienie i uznanie dla wykorzystania procesu;
6. kontynuację debaty na temat tego, co stanowi dyscyplinę metodologiczną;
7. szersze wykorzystanie myślenia systemowego i nauki o złożoności jako ram ewaluacji.

---

**Michael Quinn Patton** jest konsultantem w zakresie ewaluacji oraz rozwoju organizacji, byłym Prezesem Amerykańskiego Towarzystwa Ewaluacyjnego, autorem pięciu publikacji dotyczących tematyki ewaluacji oraz współautorem szóstej. Dr. Patton posiada dyplom socjologii (licencjat) Uniwersytetu w Cincinnati oraz socjologii wsi (magister) Uniwersytetu Wisconsin. Również na tym uniwersytecie zdobył tytuł doktora socjologii. Dr. Patton przez 18 lat był wykładowcą na Uniwersytecie Minnesoty. Pełnił wówczas przez 5 lat funkcję Dyrektora Centrum Badań Społecznych Minnesoty – *Minnesota Center for Social Research*. Obecnie Dr. Patton prowadzi prywatną firmę doradczą Utilization-Focused Information and Training oraz wykłada w Union Institute Graduate School.

## Bibliografia:

- *Patton M. Q., Utilization-Focused Evaluation, Sage Publications, Thousand Oaks, Ca 2008, wydanie 4.*
- *Struening E. L., Guttentag M. (red.), Handbook of Evaluation Research, Sage, Beverly Hills, Ca 1975.*
- *Williams B., Iman I, Systems Concepts in Evaluation: An Expert Anthology, Monografia Amerykańskiego Towarzystwa Ewaluacyjnego, EdgePres, Point Reynes 2006.*

*Michael Wiseman*

# Droga do nagrody: spojrzenie na ewaluację opartą na podejściu kontrfaktycznym z dwóch perspektyw

Polskie Ministerstwo Rozwoju Regionalnego jest powszechnie uznawane za instytucję wyjątkowo aktywnie wykorzystującą fundusze Unii Europejskiej w celu promowania polityki spójności. Taka reputacja wynika częściowo z wysiłków ministerstwa związanych z ewaluacją programów. Z uwagi na coraz większe środki finansowe do wykorzystania oraz inne zmiany, zapotrzebowanie na ewaluację będzie rosło. Obawy te nie ograniczają się do systemów finansowanych w ramach funduszy strukturalnych i Funduszu Spójności. Istotnie, mało prawdopodobne jest, aby instytucja, która na ogół nie zajmuje się ewaluacją, mogła zaprojektować lub przeprowadzić skuteczne oceny skutków działań krajowych, które wspiera Unia Europejska. Dlatego też postępy w zakresie poprawy ewaluacji programów finansowanych ze środków Unii Europejskiej są zależne od postępów w zakresie rozwoju kultury ewaluacji i odpowiednich umiejętności w ramach instytucji zarządzających oraz wzajemnie z nimi powiązane. Konieczne jest również, aby przywódcy polityczni w większym stopniu doceniali rolę ewaluacji.

Niniejszy tekst zawiera dwa spojrzenia na rozwój ewaluacji. Jedno z nich rozpoczyna się od instytucji zarządzających. Drugie obejmuje zaangażowanie transnarodowe. Stoję na stanowisku, że niektóre problemy z ewaluacją prowadzą do tego, że otrzymujemy zbyt mało w kontekście na przykład Unii Europejskiej lub też dowolnego innego podmiotu politycznego, w którego budżecie przewiduje się środki na ewaluację. Na zakończenie przedstawiam sugestię dotyczącą sposobu zwrócenia uwagi na to, co jest konieczne do promowania lepszej ewaluacji w kontekście zdecentralizowanego procesu podejmowania decyzji.

Moje przemyślenia wywodzą się z trzech źródeł. Jednym z nich jest mój były mentor, Aaron Wildavsky, który lata temu w Berkeley uczył mnie, (wówczas) młodego profesora ekonomii, jak ważne jest myślenie o zarządzaniu programem jako o kluczowym elemencie analizy polityki (zob. Wildavsky 1987). Profesor Wildavsky uważał, że głównym problemem zarządzania publicznego jest stworzenie kultury instytucji, która stale wspiera wysiłki na rzecz ulepszeń i wynagradza pracowników, którzy przyczyniają się do osiągnięcia tego celu. W takiej kulturze kładzie się ogromny nacisk na ewaluację. Drugie źródło stanowi wieloletnie doświadczenie współpracy z amerykańskimi stanami i obserwowanie porażek tego, co jeden z sędziów Sądu Najwyższego określił (w 1932 r.) słynnym terminem „laboratoria demokracji". Stany Zjednoczone i inne państwa mogą się wiele nauczyć z europejskich doświadczeń z ewaluacją opartą na współpracy. Trzecie źródło to ostatnie 2 lata spędzone na współpracy z analitykami z biura Dyrekcji Generalnej ds. Zatrudnienia, Spraw Społecznych i Włączenia Społecznego (DG Employment) nad opracowaniem wytycznych w sprawie ewaluacji oddziaływania opartej na podejściu kontrfaktycznym (ang. *Counterfactual Impact Evaluation* – CIE) działań subsydiowanych w ramach Europejskiego Funduszu Społecznego. Właśnie w tym trzecim kontekście dowiedziałem się o wysiłkach Polski i innych państw członkowskich UE związanych z ewaluacją i zacząłem je doceniać.

## Ewaluacja „w domu"

Moją argumentację rozpoczynam od ewaluacji przeprowadzanej jako standard przez instytucje zarządzające w stylu Wildavsky'ego niezależnie od źródła finansowania.

## Podstawy Counterfactual Impact Evaluation

Dążenie do poprawy polega na poszukiwaniu zmian w działalności instytucji, które są opłacalne, tzn. korzyści przeważają nad kosztami. Ewaluacja oddziaływania oparta na podejściu kontrfaktycznym (CIE) jest podstawą analizy kosztów i korzyści. W ramach CIE ocenia się konsekwencje wprowadzenia programu lub zmiany dla rezultatów będących przedmiotem zainteresowania w porównaniu z tym, co osiąga się, stosując rozwiązanie alternatywne. W przypadku zarządzania, alternatywą jest zazwyczaj dotychczasowy scenariusz postępowania, ale może to być także inna strategia programu, która ma umożliwić osiągnięcie równoważnych celów.

*Oddziaływanie* jest to niefortunne określenie na różnicę między rezultatami zarejestrowanymi dla osób objętych wprowadzonym lub zmodyfikowanym programem a rezultatami przewidywanymi dla scenariusza alternatywnego. Określenie e*fekt* byłoby lepsze. Analiza korzyści i kosztów poszukuje zmiany, która nie byłaby zbyt kosztowna do wprowadzenia, ale nie można szacować korzyści i kosztów bez oszacowania tego, co by się wydarzyło w przypadku braku inicjatywy.

Taką prognozę, tj. *sytuację kontrfaktyczną*, można opracować na wiele sposobów. To, co określa się mianem „spójności wewnętrznej" ewaluacji, zależy od wiarygodności konstrukcji sytuacji kontrfaktycznej. W wielu sytuacjach najbardziej wiarygodną sytuację kontrfaktyczną tworzy się przez dobór losowy potencjalnych uczestników programu do grup „objętej oddziaływaniem bodźca" i „kontrolnej". Osoby w grupie objętej oddziaływaniem bodźca mają możliwość zaangażowania się w nową działalność, zaś osoby z grupy kontrolnej – nie. Jeżeli bowiem dobór losowy jest przeprowadzony prawidłowo, nie istnieją różnice systematyczne między grupą objętą oddziaływaniem bodźca a grupą kontrolną i wówczas rezultat dla grupy kontrolnej staje się rzetelną prognozą tego, co wydarzyłoby się w przypadku alternatywnym.

Niezależnie jednak od tego, czy przeprowadza się formalną ocenę, czy też nie, dobrzy menedżerowie programów zawsze myślą o sytuacjach kontrfaktycznych: „Załóżmy, że zmienimy A na B. Czy różnica w rezultatach będzie warta zachodu?". Rozważanie sytuacji kontrfaktycznej wymaga dwóch prognoz: jednej dotyczącej tego, co zdarzy się w przypadku braku zmiany, tj. przy utrzymaniu operacji A, i drugiej dotyczącej konsekwencji zmiany. W niektórych okolicznościach z innych prób przeprowadzania działania B można uzyskać wystarczająco dużo informacji, aby można było z wystarczającą wiarygodnością przewidzieć konsekwencje – oraz ocenić spodziewane korzyści i koszty – aby poprzeć to działanie. W takim przypadku mówi się, że informacje z innych źródeł mają *spójność zewnętrzną*. Jeżeli dowody są podejrzane lub niepewne, powinno się przeprowadzić eksperyment lub przynajmniej przygotować plany oceny skutków po wystąpieniu zmiany. W ten sposób wracamy do CIE.

Należy wspomnieć o trzech aspektach CIE przeprowadzanej przez dobry (znów: w rozumieniu Wildavsky'ego) zarząd. Pierwszy z nich dotyczy procesu, drugi – spójności zewnętrznej, zaś trzeci – kosztów.

Spójrzmy jeszcze raz na stwierdzenie: „Załóżmy, że zmienimy A na B". Składa się ono z trzech części: części dotyczącej zmiany, części dotyczącej różnicy i tego, czy zmiana jest warta zachodu. Pierwsza jest zmiana: innowacja zmienia wkład lub proces. Zarządzający są zainteresowani zmianami, jakie powoduje innowacja w charakterze usług lub innych działań należących do obowiązków instytucji. Innowacje z reguły określa się jako idealny, nowy model robienia czegoś.

Taka zmiana w sposobie wykonywania czynności jest domeną analizy procesu. Analiza procesu zazwyczaj obejmuje nie jedną, a *dwie* sytuacje kontrfaktyczne. Pierwszą sytuację kontrfaktyczną stanowi pro-

ces określony w planie lub ideał rozważanej zmiany, tj. model. Druga sytuacja kontrfaktyczna obejmuje działania podejmowane na potrzeby kontroli, często scenariusz dotychczasowego postępowania. W ewaluacji procesu analizuje się obydwie sytuacje – jak bliska osiągnięcia zamierzonego celu jest instytucja wdrażająca i jak duża jest różnica między procesem objętym oddziaływaniem bodźca a sytuacją kontrfaktyczną.

Analizę procesu nazywa się czasami „monitoringiem" i w istocie może być to właściwe określenie, ponieważ ocena procesu bez monitoringu jest praktycznie niemożliwa. Z moich doświadczeń wynika jednak, że to, co nazywa się monitoringiem, nie ma nic wspólnego z ewaluacją procesu, ponieważ żadna z sytuacji kontrfaktycznych dotyczących procesu – idealna lub kontrolna – nie jest dobrze zidentyfikowana. Uniemożliwia to pomiar. Ewaluacja procesu wymaga dokonania w jakiś sposób pomiaru różnicy między tym, co dzieje się w grupie objętej oddziaływaniem bodźca i grupie kontrolnej w wyniku innowacji oraz dokonania w jakiś sposób pomiaru różnicy między tym, co dzieje się w grupie objętej oddziaływaniem bodźca a zamiarem zarządzających.

Analiza procesu jest kluczowa dla dobrego zarządzania, jest niezbędnym działaniem poprzedzającym ewaluację oddziaływania. Jest także źródłem strony „kosztów" w ocenie korzyści i kosztów – w stwierdzeniu „załóżmy, że zmienimy" jest to część „warte zachodu". Taki związek z oddziaływaniem pojawia się, ponieważ analiza procesu jest kluczowa do potwierdzenia tego, co dokładnie wydarzyło się w następstwie innowacji. Jeżeli wiemy, co naprawdę wydarzyło się w procesie, mamy wskazówki dotyczące tego, co spowodowało obserwowany efekt. Jeżeli nie osiągniemy żadnego efektu, analiza procesu może nam powiedzieć, czy błąd tkwił w teorii (pomysł był zły) czy w realizacji (nie udało się zrealizować programu w taki sposób, jak zamierzyli planujący).

Pozwolę sobie tutaj na dygresję na temat „modeli logicznych". Model logiczny jest teorią, która łączy wkład z produktem i wyjaśnia, dlaczego oczekuje się, że pewne zmiany w działaniu będą miały określony wpływ na rezultaty będące przedmiotem zainteresowania. Modele logiczne stanowią kluczową część planowania innowacji i mogą zwrócić uwagę na zasadnicze cechy, które należy monitorować w ramach ewaluacji procesu. W połączeniu z dowodami na skuteczność różnych elementów w zakładanym łańcuchu związku między programem i rezultatem, modele logiczne mogą uzasadniać zmiany w polityce; jeżeli dowody są dostatecznie mocne, przeprowadzenie dodatkowej oceny skutków może nie być konieczne. Niezależnie jednak od tego, jak przekonująca jest logika, monitorowanie procesu jest kluczowe do zrozumienia – i oceny – wdrożenia. Chociaż jasne przedstawienie logiki zmiany może być bardzo ważne, modele logiczne nie zastąpią kontrfaktycznej ewaluacji procesu lub oddziaływania.

Wracając jednak do dobrego menedżera według Wildavsky'ego: stwierdziłem, że pierwszą cechą ewaluacji programu w stylu dobrego menedżera jest zwrócenie uwagi na proces. Druga cecha dotyczy spójności zewnętrznej. Spójność zewnętrzna nie dotyczy wiarygodności sytuacji kontrfaktycznej (jest to spójność *wewnętrzna*), ale znaczenia rezultatów jednej ewaluacji dla sytuacji poza tą, w której przeprowadzono ewaluację oddziaływania. Czy można sobie wyobrazić, że podobna innowacja, wprowadzona w innym czasie lub miejscu, miałaby podobny wpływ? Według mojego słownika *sztuka* wymaga umiejętności „nabytych dzięki doświadczeniu, badaniu lub obserwacji". Uważam, że ocena spójności zewnętrznej jest sztuką. Krajowe instytucje mają dobrą pozycję, aby, w kontekście wspólnego środowiska kulturalnego, prawnego i gospodarczego, ocenić znaczenie tego, czego nauczono się w jednym miejscu, w odniesieniu do innego miejsca. Gdy pozostałe warunki są podobne, bardziej pewnie czujemy się, jeśli ewaluacja została przeprowadzona „blisko", przy czym „blisko" oznacza „w miejscu podobnym do tego". Dobrzy menedżerowie potrafią ocenić bliskość, czyli co oznacza słowo „podobny".

Po trzecie, ewaluacja jest kosztowna. Nie słyszałem o żadnej instytucji zarządzającej z nieograniczonym budżetem. Nie można po prostu przetestować wszystkich możliwych innowacji ani zmierzyć wpływu wszystkich cech realizowanych programów w tym samym momencie. Ponadto w każdej zwyczajnej insty-

tucji zwielokrotnienie liczby ewaluacji zmniejsza uwagę poświęcaną poszczególnym wysiłkom i zwiększa ryzyko niepowodzenia. W związku z tym wybór działań ewaluacyjnych jest ważny i wiąże się z kompromisami. Takie jest życie.

## Koncentracja na nagrodzie

Biorąc pod uwagę powyższe, załóżmy, że kluczowi członkowie parlamentu krajowego studiowali Wildavsky'ego i postanowili stworzyć zachęty do ewaluacji bieżących programów oraz tworzenia i testowania pomysłów na ulepszenia. Zachętą jest nagroda za wyróżniający się plan ewaluacji. Kiedy należy wręczyć nagrodę? Czego powinien szukać komitet przyznający nagrodę CIE?

Najważniejszy wniosek jest następujący: interesowi publicznemu najlepiej służy przyznawanie nagród na podstawie *planu*. Nie oznacza to, że realizacja nie jest ważna, w istocie jest ona bowiem kluczowa, a instytucja zorientowania na ulepszenia będzie postrzegać ewaluację ewaluacji jako część kultury ulepszeń. Podstawę osiągnięć stanowią jednak plany. Odkładanie rozważenia przeprowadzenia ewaluacji aż do momentu wprowadzenia innowacji drastycznie zmniejsza prawdopodobieństwo wyciągnięcia wniosków.

Wynikiem konsultacji powinny być same kryteria. Pomaga to zapewnić poczucie zaangażowania w konkurs; uczestnicy stają się interesariuszami podejmowanych wysiłków. Niezależnie od tego, jak się je wyrazi, prawdopodobne jest, że kryteria przyznawania nagród będą obejmowały większość spośród następujących elementów. Podane tutaj elementy dotyczą przypadku, w którym wniosek zawiera wezwanie do innowacyjności. Nie należy jednak zapominać, że w pewnych przypadkach innowacją może być zamknięcie programu zamiast modyfikowania go lub rozpoczynania.

- Logika interwencji

  Czy teoria leżąca u podstaw interwencji, zmiana mająca być przedmiotem ewaluacji ma sens? Czy model przyczynowo-skutkowy jest poparty innymi ocenami?
- Potencjał

  Czy istnieje dobry powód, by wierzyć, że korzyści z ewaluacji przewyższą koszty?
- Metodologia ewaluacji

  Czy plan jest wykonalny? Czy rezultaty będą miały spójność wewnętrzną? Czy prognoza rezultatów w przypadku braku innowacji jest wiarygodna?
- Analiza procesu

  Czy plan ewaluacji obejmuje porównanie wyników innowacji zarówno do modelu, jak i do procesu, których doświadczyła grupa kontrolna?
- Użyteczność zewnętrzna

  Czy wyniki ewaluacji będą miały wartość jako wkład w przyszłe decyzje, w tym w dalsze stosowanie interwencji?

Spójność zewnętrzna jest tylko jednym z elementów tego, co decyduje o użyteczności ewaluacji. Ewaluacja jest użyteczna, gdy rezultaty mają potencjał ulepszenia późniejszego procesu podejmowania decyzji w sektorze publicznym.

Trudno jest przypisać wagi tym elementom, gdyż nie są one od siebie niezależne. Na przykład „potencjał" zależy zarówno od jakości metodologii, jak i od użyteczności zewnętrznej wiedzy, jaką ma wygenerować ewaluacja. Spójność zewnętrzna rezultatów ewaluacji interwencji, której brakuje spójności logicznej, jest zależna od tego, czy w trakcie samej ewaluacji uwidoczni się jej logika, której nie oczekiwano. Jest to ryzykowne przedsięwzięcie! Powyższa lista sugeruje jednak kolejność elementów oceny. Ostatnią kwestią do rozważenia powinna być użyteczność zewnętrzna: czy to, co projekt ma przynieść, faktycznie okaże się użyteczne?

Załóżmy, że Ministerstwo Rozwoju Regionalnego, albo nawet jeszcze szerszy zbiór instytucji rządowych, ma przeprowadzić konkurs wśród swoich jednostek lub na przykład wśród swoich pracowników na pomysły na innowacje/ewaluacje – po wspólnym opracowaniu kryteriów. Kto wie, jaki może być wynik? Należy zauważyć, że w tym konkursie poprzeczka ustawiona jest wysoko. Uczestnicy nie tylko muszą mieć interesujący pomysł, ale muszą przynajmniej przedstawić szkic planu realizacji oraz ewaluacji.

Zastosowanie przez rząd nagród jako zachęt nie jest wcale przesadą. W 1714 r. brytyjski parlament zaproponował słynną nagrodę *Longitude Prize* za odkrycie praktycznego sposobu pomiaru długości geograficznej. Nagrody są kluczowym elementem „Strategii dla Amerykańskich Innowacji"[1] administracji Obamy (*Strategy for American Innovation*, zob. www.challenge.gov). Nowością w tej propozycji nagród jest to, że dotyczy ona planowania ewaluacji.

## Poważne problemy

Rozważania na temat użyteczności zewnętrznej prowadzą nas do ewaluacji w kontekście UE. W większości obszarów funduszy strukturalnych i Funduszu Spójności zwiększa się nacisk na ewaluację oddziaływania, zwłaszcza, że trwa okres planowania perspektywy finansowej 2014–2020. Mnożą się „wytyczne", ponieważ różne Dyrekcje Generalne usiłują wykazać, że fundusze przynoszą efekty nie tylko w postaci transferu zasobów. Co się zmienia, gdy zmienimy skalę i zaczniemy myśleć o ewaluacji na poziomie różnych państw członkowskich?

Pojawiają się dwa duże problemy dotyczące zachęt.

Pierwszy problem z zachętami pojawia się, ponieważ niektórzy uważają, że środki finansowe przeznaczone na spełnianie wymogów funduszy związanych z ewaluacją stanowią deadweight i nie są niczym więcej niż biletem, który trzeba nabyć, aby zdobyć i utrzymać dostęp do funduszy. W świecie Wildavsky'ego wzywanie do przeprowadzenia ewaluacji byłoby zbędne, ponieważ instytucje otrzymujące wsparcie już by ją przeprowadzały, a środki otrzymane z Brukseli traktowanoby tak samo jak te zebrane w formie podatków od dobrych obywateli, powiedzmy, Nowego Tomyśla. W rzeczywistości, jak stwierdził Wildavsky, utworzenie i utrzymanie kultury ulepszeń i ewaluacji nie jest łatwe, instytucje często nie radzą sobie z nią, a politycy rzadko ją wspierają. Nawet jeśli istnieje taka wola, instytucje zarządzające, nieposiadające wieloletniego doświadczenia w zakresie ewaluacji, mogą zwyczajnie nie mieć wystarczających zdolności do przeprowadzania ich we właściwy sposób. Rozwijanie takiej zdolności może być trudnym przedsięwzięciem zarówno z przyczyn politycznych, jak i taktycznych. Trudno robić to z Warszawy, nie mówiąc już o Brukseli (czy Waszyngtonie).

Drugi problem dotyczący zachęt jest bardziej skomplikowany. Obejmuje efekty rozlania, tj. korzyści uzyskanych przez jedną instytucję lub państwo w wyniku działań innej instytucji lub państwa. Ekonomiści już dawno uznali, że efekty rozlania mogą doprowadzić do „nieprawidłowości" w funkcjonowaniu rynku w takim sensie, że gdy efekty rozlania są pozytywne, jest ich zbyt mało, a gdy są negatywne – zbyt dużo. Ewaluacja może być użytecznym sposobem regulacji i uzasadnienia ex post dla dotacji. Użyteczność zewnętrzna ewaluacji może być jednak źródłem największych korzyści z ewaluacji. Jeżeli ewaluacja ma przekonującą spójność zewnętrzną, a wyciągnięte wnioski są użyteczne, wówczas taka wiedza daje efekt rozlania. Jeżeli przy opracowywaniu planów ewaluacji nie uwzględnia się użyteczności zewnętrznej, korzyści będą zaniżone. Jeżeli kosztami obciążona jest wyłącznie instytucja przeprowadzająca ewaluację, niektórych ewaluacji będzie się unikać albo będzie się je przeprowadzać na nieoptymalną skalę, nawet jeśli z punktu widzenia wartości zdobytej w ten sposób wiedzy, należy je podjąć.

[1] Biuro Zarządzania i Budżetu Białego Domu (*Office of Management and Budget*) przedstawia użyteczny opis pod adresem http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-11.pdf

Innymi słowy, niektóre ewaluacje mogą mieć więcej „interesariuszy", niż identyfikują władze lokalne. Klasycznym rozwiązaniem problemu efektów zewnętrznych jest ich „internalizacja" poprzez przydzielenie ewaluacji do tego szczebla zarządzania, który obejmie wszystkich interesariuszy. Ewaluację można przeprowadzać na poziomie lokalnym, ale jeżeli zaistnieją efekty rozlania, uzasadniona jest dotacja na przeprowadzenie ewaluacji, zebrana od całej społeczności beneficjentów. Celem dotacji jest zachęcenie ewaluatorów do uznania korzyści z wytwarzania wiedzy, która przynosi zarówno korzyści zewnętrzne, jak i użyteczność zewnętrzną. Podobnie, jak większość innych problemów w obszarze polityki publicznej, projektowanie takich dotacji nie jest łatwe. Dlatego właśnie Bóg dał nam ekonomistów (gdybyście mieli Państwo jakieś wątpliwości).

Efekty zewnętrzne są z pewnością obecne w ewaluacjach realizacji polityki krajowej na szczeblu lokalnym i przynajmniej z zasady efekty rozlania takich ewaluacji mogą zostać zinternalizowane przez rząd krajowy, który ponosi część lub całość kosztów ewaluacji lokalnych. Instutycjom UE brakuje jednak (często z uzasadnionych powodów) wielu instrumentów polityki dostępnych rządowi krajowemu. Niemniej, jeżeli nie podejmuje się wysiłków mających na celu promowanie zwracania uwagi na korzyści zewnętrzne z projektowania i ewaluacji inicjatyw finansowanych z Funduszy, obywatele UE ogółem stracą na tym. Potrzeba czegoś więcej niż wytycznych. Należy zastanowić się nad nagrodami i klubami.

## „Jestem z Brukseli i jestem tutaj, aby cię nagrodzić"

Zastanówmy się najpierw nad nagrodami.

Załóżmy, że Dyrekcja Generalna ds. Polityki Regionalnej i Miejskiej (*Directorate General for Regional and Urban Policy*) ogłasza konkurs na plany ewaluacji działań finansowanych z funduszy DG REGIO. Abstrahując od samych nagród, w jaki sposób należy zmienić listę kryteriów, jeżeli ma być ona stosowana w kontekście międzynarodowym?

Ponownie kwestia ta powinna zostać rozstrzygnięta w drodze decyzji zbiorowej. Na pierwszy rzut oka wydaje się, że można by zastosować tę samą listę z potencjalną zmianą akcentów. Ważne jest, aby skoncentrować się jeszcze bardziej na spójności i użyteczności zewnętrznej. Warto zmienić punkt widzenia z perspektywy ewaluatora na perspektywę konsumenta. Załóżmy, że jesteśmy w państwie członkowskim A i obserwujemy ewaluację przeprowadzoną w państwie członkowskim B. Jak zmienia się perspektywa w stosunku do sposobu analizowania ewaluacji przeprowadzonej w naszym kraju?

Z pewnością, niektóre aspekty w ogóle się nie zmieniają. Spójność wewnętrzna pozostaje kluczowym elementem programu ewaluacji i pragniemy poznać metodologię. To, co może okazać się ważniejsze, to szczegóły procesu. Chcemy wiedzieć, jakie były rezultaty, nie w teorii, ale w praktyce. Jesteśmy zainteresowani tym, co można nazwać „funkcją produkcji" w odniesieniu do rezultatów. Ponadto z pewnością pragniemy dowiedzieć się, ile faktycznie osiągnięto z tego, co zamierzano.

Oddziaływanie bodźca jest to jednak tylko jedna strona medalu. Ważne jest, aby zrozumieć sytuację kontrolną. Co wydarzyło się w sytuacji kontrolnej? W jakim stopniu można to porównać do sytuacji bazowej w naszym kraju? To właśnie różnica w sytuacji kontrolnej jest najbardziej problematyczna dla spójności ponad granicami. Jaką stanowi to dla nas różnicę, jeżeli innowacja wprowadzona w jakimś innym kraju zwiększyła utrzymanie zatrudnienia w grupie X, jeżeli to, co zdarzyłoby się w grupie X przy braku innowacji w tym innym kraju, całkowicie różni się od polityki stosowanej w naszym kraju? Z pewnością, sztuka określania użyteczności zewnętrznej wymaga danych szczegółowych zarówno na temat procesu, jak i oddziaływania. Na poziomie UE plany przyznawania nagród za ewaluację powinny przewidywać gromadzenie tych informacji.

To, czego potrzeba, aby zwiększyć użyteczność zewnętrzną oraz zrównoważyć korzyści z wysiłku włożonego w zgromadzenie i przedstawienie takich danych z kosztami, prawdopodobnie różni się w zależności od rodzaju wprowadzanej innowacji. Chodzi mi o to, że, w przypadku omawiania ewaluacji w śro-

dowisku krajowym, charakter i możliwość uogólnienia doświadczeń grupy kontrolnej są często czytelne i decydent może poczynić odpowiednie założenia, zastanawiając się, czy wyniki uzyskane w jednym miejscu/czasie mają zastosowanie do innego miejsca i czasu. W przypadku rozszerzenia poza granice danego państwa zwiększa się ryzyko związane z takim założeniem.

Nagroda może służyć do zwrócenia uwagi na konieczność zapewnienia szczegółów procesu oraz na zyski z prawidłowego przeprowadzenia ewaluacji dla osób postronnych, ale nie jest jasne, czy zainteresowanie konkurencją będzie wystarczające, aby przynieść faktyczne zwiększenie bazy dowodów na potrzeby polityki w dłuższym okresie, co w naszym przypadku oznacza lata 2014–2020. Potrzeba czegoś więcej. Po raz kolejny pomocna może być zmiana perspektywy.

## Ogólnoeuropejski klub ewaluacyjny

Do tej pory dyskusja była ustrukturyzowana w sposób pionowy, począwszy od ewaluacji na szczeblu państw, po ewaluacje innowacji finansowanych na szczeblu UE. Wróćmy do poziomu horyzontalnego, tj. stosunków między państwem A a państwem B. Stwierdziłem, że nieuznawanie korzyści dla państwa B płynących z wiedzy na temat oddziaływania innowacji pochodzącej z ewaluacji sprawia, że państwo A niedostatecznie inwestuje w ewaluacje, i na odwrót. Ponadto państwo A nie tylko niedostatecznie inwestuje, ale prawdopodobnie nie gromadzi ani nie przekazuje tych rodzajów informacji, które umożliwiłyby rezultatom „podróżowanie" przez granicę.

Wydaje się, że wymiana przyniosłaby tutaj korzyści. Załóżmy, że nasze dwa państwa utworzyły związek ewaluacyjny – „klub", jeżeli Państwo wolą. Członkostwo w tym klubie wiąże się z zobowiązaniem i korzyścią. Zobowiązanie polega na uznaniu interesów drugiego członka, zarówno przy określaniu, co zostanie poddane ewaluacji, jak i rodzajów gromadzonej szczegółowej wiedzy. Korzyścią jest to, że partner robi to samo. Najłatwiejszym sposobem na zapewnienie uwzględnienia takich interesów partnera-interesariusza jest bezpośrednie ich włączenie. Czy brzmi to jak swego rodzaju „otwarta metoda koordynacji" między A i B (Heidenreich iZeitlin 2009? Owszem!

Klub ewaluacyjny A&B przynosi dodatkowy pozytywny skutek. Jak wspomniano wcześniej, krajowe zdolności do przeprowadzania ewaluacji są ograniczone, niezależnie od tego, jak wielki jest entuzjazm dla zarządzania w stylu Wildavsky'ego. Dzięki koordynacji, państwa mogą się skupić na jednym zestawie kwestii do rozwiązania i polegać na tym, że partner zajmie się pozostałymi. Aby jednak osiągnąć takie korzyści, zaangażowanie w ewaluację musi być poważne po obu stronach wymiany. Otwarta koordynacja i częsta komunikacja mają zasadnicze znaczenie. Trafna wydaje się tutaj uwaga prezydenta Reagana na temat traktatu o redukcji uzbrojenia strategicznego START. Parafrazując, „ufaj ewaluacji swojego partnera, ale sprawdzaj".

Droga do ogólnoeuropejskiego klubu ewaluacyjnego jest jeszcze bez wątpienia długa. Z kolei dążenie do koordynacji poziomej w ewaluacji programów między państwami Unii pozostaje nieco w tyle za obecnym wyzwaniem związanym z rozwijaniem kultury ewaluacji w ramach instytucji zarządzających. Niemniej taki lider, jak Polska, powinien zacząć myśleć w ten sposób, być może przez tworzenie dwustronnych partnerstw koordynujących ewaluacje w poszczególnych obszarach polityki. Takie podejście oddolne mogłoby być bardzo użytecznym uzupełnieniem wysiłków Brukseli związanych z promowaniem CIE.

## Podsumowanie

Obecnie poświęca się dużo uwagi promowaniu opartych na podejściu kontrfaktycznym ewaluacji inicjatyw państw członkowskich finansowanych z funduszy strukturalnych i Funduszu Spójności. Chociaż ewaluacja jest kluczowa dla dobrego zarządzania, ważnym celem długoterminowym jest rozwijanie

i wspieranie w ramach instytucji zarządzających ogólnej kultury, która stale promuje wysiłki na rzecz ulepszeń i nagradza tych pracowników, którzy przyczyniają się do osiągnięcia tego celu. Dążenie do niego jest po części kwestią zachęt. Jednym z podejść do zachęt jest ustanowienie nagród za plany oceny inicjatyw, które uwzględniają znaczenie zarówno analizy procesu i oddziaływania, jak i roli ewaluacji we wspólnym europejskim wysiłku związanym z budowaniem bazy dowodów na potrzeby polityki spójności i rozwoju regionalnego. W tym celu użyteczne może być postrzeganie instytucji zarządzających jako swoistego klubu, w którym wkład stanowi dostarczenie spójnych zewnętrznie i użytecznych ocen innowacji. Rygoryzm tych ewaluacji można zapewnić przez ustanowienie wymiany wewnątrzunijnej zgodnie z pierwotnie zaproponowanymi założeniami dla otwartej metody koordynacji. Państwa najbardziej zaangażowane w rozwój ewaluacji, np. Polska, mogą przejąć inicjatywę.

Autor skorzystał na dyskusjach z Veronicą Gaffey, Herthą Schönhofer oraz Alberto Martinim. Żadne z nich nie powinno być pociągnięte do odpowiedzialności.

---

**Prof. Michael Wiseman** jest profesorem polityk publicznych, administracji i ekonomii na George Washington University w Waszyngtonie. W trakcie swojej kariery naukowej przez 18 lat był profesorem ekonomii na Uniwersytecie w Berkeley oraz profesorem spraw publicznych na Uniwersytecie Madison w Wisconsin. Praca naukowa na trzech różnych wydziałach George Washington University odzwierciedla zainteresowanie prof. Wisemana politykami publicznymi, zarówno z punktu widzenia ich tworzenia, jak i zarządzania nimi (w tym ewaluacji). W ostatnich latach prof. Wiseman pełnił funkcję konsultanta ds. ewaluacji dla instytucji, takich jak: Biuro Świadczeń Socjalnych ds. Polityk Emerytalnych i Integracji Osób Niepełnosprawnych (US *Social Security Administration's Office of Retirement and Disability Policy*), Biuro Pomocy Rodzinom w Wydziale Dzieci i Rodzin Ministerstwa Zdrowia i Usług Społecznych (*Office of Family Assistance in the Administration for Children and Families of the US Department of Health and Human Services*) Centrum Badań Ekonomicznych Ministerstwa Rolnictwa (*Economic Research Service*), Komisja Europejska, OECD, brytyjski Departament Pracy i Emerytur oraz firmy specjalizujące się w eksperymentach w dziedzinie polityki społecznej oraz w ewaluacji. Jest autorem szeregu opracowań w dziedzinie rozwoju regionalnego, polityki społecznej, zarządzania publicznego oraz ewaluacji programów.

## Bibliografia


- Heidenreich M., Zeitlin J., *Changing European Employment and Welfare Regimes: The influence of the open method of coordination on national reforms*, Routledge, Nowy Jork 2009.
- Rossi P. H., Lipsey M. W. i Freeman H. E., *Evaluation: A Systematic Approach*, SAGE Publications, Thousand Oaks, Ca 2004, wydanie siódme.
- Wildavsky A., *Speaking Truth to Power: The Art and Craft of Policy Analysis*, Transaction Publishers, Piscataway, N.J. 1987.

*Alberto Martini*

# Różne oblicza randomizowanych prób kontrolnych

## Wprowadzenie

*Co działa?* Skuteczność polityki społecznej, edukacyjnej, przemysłowej, a także polityki rynku pracy jest prawie zawsze niepewna. Głównym wyzwaniem dla ograniczenia takiej niepewności jest „przypisanie": Czy zmiany obserwowane w czasie są skutkiem interwencji, czy wystąpiłyby również *bez niej*? Czy zaobserwowane różnice w rezultatach pomiędzy uczestnikami programu i nieuczestniczącymi w nim są *spowodowane* interwencją, czy wystąpiłyby także w sytuacji, *gdyby jej nie było*? Aby odpowiedzieć na te pytania, trzeba odtworzyć sytuację kontrfaktyczną – tj. *co by się stało w przypadku braku interwencji*. Wpływ interwencji na dany rezultat jest różnicą między obserwowanym rezultatem a hipotetyczną sytuacją kontrfaktyczną. Trudność uzyskania wartości takiego hipotetycznego rezultatu stanowi *podstawowy problem wnioskowania przyczynowego* (Holland 1986).

Rzecz jasna, zrozumienie, *co działa, a co nie*, nie jest jedyną informacją potrzebną do kształtowania polityki. Przydatne jest również ustalenie, *dla kogo*, *w jakich kontekstach* i, wreszcie, *dlaczego* interwencja działa lub nie. Przynajmniej dwa różne rodzaje dowodów są potrzebne do kształtowania lepszych polityk publicznych: jeden dotyczy przede wszystkim *kwantyfikacji efektów* i obejmuje pytanie „dla kogo", a drugi wiąże się z *wyjaśnieniem efektów (lub ich braku)*. W niniejszym tekście zajmiemy się metodami poświęconymi kwantyfikacji efektów interwencji dla niektórych wcześniej ustalonych obszarów zainteresowania.

Kwantyfikacja efektów, mimo że jest tylko pierwszym krokiem, sama w sobie stanowi wyzwanie. W ostatnim ćwierćwieczu poczyniono jednak ważne postępy w dążeniu do ustalenia, *co działa* w programach rządowych. Zasadniczo efekty uzyskuje się poprzez porównanie rezultatów jednostek (osób, społeczności, szkół, przedsiębiorstw, miast), które uczestniczą w programie, z tymi, które w nim nie uczestniczyły. Jednak różnica ta sama z siebie nie ujawnia prawdziwego wpływu interwencji na rezultat – nie może być *interpretowana* w sensie przyczynowym. Luka między obserwowaną zmianą i prawdziwym wpływem to tzw. *obciążenie selekcyjne*, które jest spowodowane istniejącymi przed rozpoczęciem interwencji różnicami między beneficjentami i niebeneficjentami. *Idealna* strategia uniknięcia istniejącej już różnicy między tymi dwiema grupami polega na losowym wyborze, kto staje się beneficjentem. Generuje to *statystycznie identyczną* grupę porównawczą, która może być wykorzystana do obliczenia wiarygodnych szacunków wpływu programu.

## Randomizowana próba kontrolna

Randomizowana próba kontrola (ang. *Randomized Control Trial* – RCT) jest badaniem mającym na celu oszacowanie wpływu interwencji na jeden lub kilka wskaźników rezultatów w odniesieniu do zbioru jednostek (takich, jak osoby indywidualne, rodziny, szkoły lub firmy). W eksperymencie randomizowanym ewaluator może manipulować tym, kto trafia do „grupy objętej oddziaływaniem bodźca", a kto trafia do „grupy kontrolnej". Randomizacja gwarantuje, że (średnio) przed interwencją grupa objęta oddziaływaniem bodźca i grupa kontrolna są zasadniczo identyczne i dlatego osiągnęłyby bardzo podobne rezultaty w przypadku braku działania bodźca. Dlatego też różnicę w rezultacie w obu grupach można *z przekonaniem* przypisać działaniu polityki.

Spojrzenie przez ten optymistyczny pryzmat sprawia, że RCT otrzymuje tytuł *złotego standardu* ewaluacji wpływu. Z drugiej strony jedni nie przywiązują żadnej wagi do dowodów dostarczonych przez RCT (Pawson and Tilley 1997), inni natomiast kładą na jednej szali potencjał poznawczy, na drugiej zaś liczne praktyczne utrudnienia w pomyślnej realizacji RCT oraz ich efektywne wykorzystanie w procesie kształtowania polityki. Na przykład Berk (2002) mówi o „standardzie brązowym" (ang. *bronze standard*), natomiast Bell (2012) omawia 15 różnych ograniczeń RCT, przytaczając dla każdego z nich praktyczne sposoby na przezwyciężenie problemu. W niniejszym tekście przypomnimy najczęściej przywoływane ograniczenia, a następnie skupimy się na kwestii niepodporządkowania się:

- W wyniku randomizacji otrzymujemy szacunki wpływu, które same są wewnętrznie spójne, ale mogą być trudne do generalizowania, a taka generalizacja jest kluczem do uzyskania użytecznych dowodów.
- Eksperymenty są *czasochłonne* i *wymagają ścisłego monitorowania*, aby zagwarantować, że są skutecznie przeprowadzone.
- Możliwość odmowy objęcia działaniem bodźca może rodzić wrażliwe politycznie *wątpliwości natury etycznej*. Może to ograniczać szanse na przyjęcie podejścia eksperymentalnego oraz zwiększać prawdopodobieństwo, że osoby odpowiedzialne za realizację programu nie będą zainteresowane współpracą.
- Randomizacja wymaga *wczesnego zaangażowania* ewaluatora oraz pewnego stopnia *stabilności otoczenia*, w którym odbywa się eksperyment.
- Randomizacja wymaga, aby interwencja była *stosunkowo prosta*, podczas gdy polityki socjalne są tradycyjnie złożone, gdyż dotyczą problemów wielowymiarowych/wielopoziomowych: chociaż złożoność jest główną przeszkodą w ewaluacji i bardziej ogólnie w uzyskiwaniu wiedzy, w przypadku randomizacji konflikt pomiędzy metodami i takimi okolicznościami jest szczególnie widoczny.

## Kwestia niedoskonałego podporządkowania się i jej konsekwencje

W klasycznym eksperymencie badacz ma pełną kontrolę nad tym, które podmioty zostaną poddane interwencji, i wszystkie podmioty *podporządkowują się* przydzielonemu im statusowi. RCT w takiej czystej postaci zdarzają się rzadko. W praktyce trudno jest dopilnować, aby wszyscy wskazani do objęcia działaniem bodźca zostali nim faktycznie objęci, a wybrani do grupy kontrolnej – nie. Takie niepodporządkowanie się przybiera dwie różne formy: *no-shows* (osoby przypisane do grupy objętej oddziaływaniem bodźca, które rezygnują przed zakończeniem, a czasami nawet jeszcze przed rozpoczęciem działania) oraz *cross-overs* (osoby przypisane do grupy kontrolnej, które mimo wszystko zostają objęte oddziaływaniem bodźca). Niepodporządkowanie się jest problemem, ponieważ podważa sam powód, dla którego w ogóle wprowadzono randomizację.

Podejmowanie kwestii niepodporządkowania się w kontekście eksperymentów społecznych doprowadziło jednak do istotnych zmian w *zrozumieniu, jaką wiedzę można uzyskać w wyniku randomizacji*, oraz uwidoczniło konsekwencje *heterogeniczności* wpływu. Jeżeli założyć, że wpływ *nie* różni się dla poszczególnych jednostek, wówczas niepodporządkowanie się przestaje być problemem – niezależnie od tego, z jakiej podgrupy uzyskamy szacunki wpływu, można je generalizować na całą populację. Obecnie nikt nie twierdziłby na poważnie, że oddziaływanie jest homogeniczne dla poszczególnych jednostek. Tabela 1 (na końcu tekstu) ilustruje wspomniane różne strategie.

## Oferta zamiast objęcia oddziaływaniem bodźca: rozwiązanie ITT (ang. *Intention To Treat*)

Jeżeli chodzi o kwestię heterogeniczności i wynikająca z niej niemożliwość zignorowania niepodporządkowania się, istnieją dwa odmienne sposoby definiowania efektów dostarczonych w wyniku eksperymentu: efekt *oferty objęcia oddziaływaniem bodźca* oraz efekt *faktycznego objęcia oddziaływaniem bodźca*. Należy pamiętać, że w obu przypadkach efekt dotyczy danej zmiennej wynikowej będącej przedmiotem zainteresowania oraz jest to efekt uśredniony. Ponieważ jednak oddziaływanie jest heterogeniczne, średnie obliczone dla różnych podzbiorów populacji również będą zróżnicowane.

Udział w programach społecznych jest najczęściej dobrowolny. W przypadku udziału dobrowolnego przedmiotem zainteresowania może być pomiar efektu samej oferty udziału w programie, a nie faktycznego otrzymania bodźca. *Niepodporządkowanie się jest wliczone w rezultat, tak więc zakres niepodporządkowania się staje się nieistotny.* Podejście to określa się mianem **analizy przeznaczenia do objęcia działaniem bodźca** (ITT). ITT mierzy średni wpływ *zaoferowania* programu. Efekty ITT programu mierzy się różnicą między średnią zmiennej wynikowej dla próby z grupy objętej oddziaływaniem bodźca (jednostki, które pierwotnie przypisano do otrzymania bodźca) oraz dla grupy kontrolnej (jednostki, które pierwotnie przypisano do nieotrzymania bodźca). Przedstawione zostało to w wierszu A w Tabeli 1.

Oszacowanie efektu w postaci ITT stanowi po prostu różnicę między średnim rezultatem obserwowanym wśród podmiotów, które *przypisano* do otrzymania bodźca, a tym obserwowanym wśród podmiotów nieprzypisanych. Analizę ITT opisuje się zwykle stwierdzeniem *„raz zrandomizowana, zawsze analizowana"*. Analiza obejmuje bowiem wszystkie randomizowane jednostki w grupach, do których zostały losowo przypisane, niezależnie od tego, czy faktycznie otrzymały one bodziec (co często jest trudne, jeśli nie niemożliwe do zaobserwowania), niezależnie od ich zgodności z kryteriami początkowymi, od odchyleń w protokole, wycofania się oraz wszelkich innych zdarzeń, jakie następują po randomizacji.

Jedną z przyczyn, która uzasadnia taką praktykę, jest fakt, że zjawisko niepodporządkowania się występujące w trakcie RCT może zaistnieć wówczas, gdy objęcie bodźcem jest oferowane całej populacji. ITT jest standardowym podejściem w badaniach klinicznych. W analizie ITT można uniknąć przeszacowania efektywności interwencji, ale można ją krytykować za większą podatność na błędy typu II (wykazanie braku efektu tam, gdzie on w praktyce wystąpił). Kolejną przyczyną, która może uzasadniać ITT, jest fakt, że w niektórych sytuacjach podporządkowania nie da się zaobserwować. W badaniach klinicznych, w których pacjenci przyjmują leki w domu, bez żadnego monitorowania, nie można obserwować podporządkowania się. Stąd analiza ITT jest jedynym realnie wykonalnym podejściem. Trzecią kwestią są koszty. Jeżeli koszty dla wszystkich uczestników badania są ponoszone z góry, branie pod uwagę kwestii podporządkowania się nie jest tak istotne. Rozważmy jednak przypadek, w którym koszty są ponoszone tylko w odniesieniu do jednostek, które faktycznie otrzymują bodziec. Analiza ITT przestaje być adekwatna i przedmiotem większego zainteresowania staje się wpływ otrzymania bodźca. Przyjrzyjmy się programowi, w ramach którego oferuje się fizjoterapię i tylko niektórzy pacjenci korzystają z tej oferty. Koszty są ponoszone wyłącznie w odniesieniu do faktycznych uczestników, dlatego chcielibyśmy poznać wpływ programu właśnie na nich.

Prawdą jest, że w programach dobrowolnych można jedynie oferować, a nie narzucać objęcie oddziaływaniem bodźca. Celem polityki jest jednak często skierowanie programu do tych, którzy mogą z niego skorzystać w większym stopniu, stąd zainteresowanie wykroczeniem poza ITT przy podejmowaniu decyzji dotyczących takich programów. Szacując efekt faktycznego otrzymania bodźca, trzeba zmierzyć się z kwestią niepodporządkowania się, nie można jej po prostu ignorować. Idealny scenariusz całkowitego podporządkowania się zapewniłby szacunek średniego efektu objęcia bodźcem (ang. *Average Treatment Effect,*

ATE), tj. średniego efektu, który otrzymujemy w wyniku objęcia wszystkich kwalifikujących się jednostek oddziaływaniem bodźca. Oczywiste jest, że w przypadku całkowitego podporządkowania się, szacunki ITT i ATE są zbieżne i są różnicę między średnimi rezultatami dla jednostek z grupy przypisanej do objęcia oddziaływaniem bodźca oraz dla jednostek nieprzypisanych do tej grupy. Jest to zilustrowane w wierszu B w Tabeli 1. Kiedy opuścimy jednak wyimaginowany świat, w którym podporządkowanie się jest całkowite, jak również zrezygnujemy z wygodnego rozwiązania, jakim jest ITT, sytuacja staje się nieco bardziej skomplikowana.

## Zawężenie perspektywy: efekt objęcia bodźcem dla jednostek uczestniczących

W przypadku, gdy jedyny typ niepodporządkowania się dotyczy jednostek przypisanych do grupy objętej oddziaływaniem bodźca, które jednak nie wzięły udziału w interwencji (tzw. *no-shows*), można obliczyć średni efekt objęcia bodźcem dla jednostek uczestniczących (ang. T*reatment Effect for the Treated,* TOT). TOT obejmuje średni zysk z udziału w programie dla tych, którzy faktycznie są nim objęci. TOT jest zwykle przedmiotem zainteresowania decydentów, którzy chcą wiedzieć, co mogą osiągnąć poprzez pełne wdrożenie swoich pomysłów, i nie są zainteresowani „rozwodnionym" efektem, jak w przypadku ITT. TOT stanowi różnicę średniej zmiennej wynikowej dla grupy objętej oddziaływaniem bodźca i grupy kontrolnej, *podzieloną przez prawdopodobieństwo objęcia oddziaływaniem bodźca* w grupie nim objętej. Prosty wzór zawarty w wierszu C w Tabeli 1 ma bardzo intuicyjne wyjaśnienie.

Podstawowym założeniem dla tego dostosowania jest fakt, że *efekt objęcia bodźcem wynosi zero dla jednostek no-show*. Weźmy na przykład eksperyment, w którym niektórzy członkowie grupy objętej oddziaływaniem bodźca nie otrzymują bodźca (stają się jednostkami *no-show*), ale żaden z członków grupy kontrolnej nie otrzymuje bodźca (nie ma żadnych jednostek *cross-over*). Jeżeli *jednostki no-show nie odczuwają żadnych efektów* interwencji ani randomizacji jako takiej, ITT równa się średniej ważonej TOT dla odbiorców bodźca i zero dla jednostek *no-show*, przy czym wagi są równe wskaźnikowi poddania oddziaływaniu bodźca i dopełnienie równa się zero (Bloom 1985):

(1) ITT = [poddanie oddziaływaniu bodźca]*TOT + [1 – poddanie oddziaływaniu bodźca]*ZERO.

Stąd:

(2) TOT = ITT/poddanie oddziaływaniu bodźca.

Podejście to *nie* wymaga, aby jednostki *no-show* były podobne do jednostek objętych wpływem bodźca. Wymaga ono jedynie, aby jednostki *no-show* nie odczuwały żadnego efektu bodźca czy randomizacji. Ponadto, ze względu na heterogeniczność oddziaływania bodźca, efekt objęcia bodźcem dla jednostek objętych oddziaływaniem bodźca *dotyczy wyłącznie odbiorców bodźca* i nie podlega generalizacji na wszystkie jednostki kwalifikujące się do objęcia oddziaływaniem bodźca.

Założenie, że efekt wynosi zawsze zero dla jednostek *no-show*, ma kluczowe znaczenie. Założenie takie jest w sposób banalny naruszone w sytuacji, gdy program najpierw testuje się na zasadzie dobrowolnego udziału, a później staje się on „obowiązkowy" (np. poprzez nałożenie sankcji lub zapewnienie dodatkowych zachęt). Średni efekt dla potencjalnych jednostek *no-show*, które zostały „zmuszone" do udziału może stać się większy niż zero, ale najprawdopodobniej będzie niższy od wartości TOT. Przy takich, stosunkowo przekonujących, założeniach szacunki ITT i TOT uzyskane w wyniku analizy wpływu dobrowolnego programu stanowią odpowiednio *dolną granicę* i *górną granicę* średnich efektów, jakie uzyskujemy w przypadku,

gdy objęcie oddziaływaniem bodźca staje się obowiązkowe. Można to wywnioskować z obserwacji, że nowi uczestnicy „z obowiązku" raczej nie odczują większych efektów niż pierwotni „dobrowolni" uczestnicy. Z drugiej strony, mało prawdopodobne jest, aby udział przyniósł szkodę nowym uczestnikom „z obowiązku", tak więc ITT dla uczestników dobrowolnych stanowi dolną granicę średniego efektu dla wszystkich objętych oddziaływaniem bodźca.

## Pójście na całość: dopuszczenie jednostek *cross-over*

Dodajmy teraz drugą formę niepodporządkowania się, czyli członków grupy kontrolnej, którzy otrzymują bodziec, tzw. jednostki *cross-over*. Sprostanie tej sytuacji wymaga bardziej złożonych ram analitycznych i przyjęcia dodatkowych założeń. Ramy te, po raz pierwszy opracowane w publikacji Angrist, Imbens i Rubin (1996), opierają się na czterech konceptualnych podgrupach, które ze względu na randomizację obejmują taki sam odsetek grupy objętej oddziaływaniem bodźca i grupy kontrolnej – takie jest przynajmniej założenie. Analityczne szczegóły tej metody wykraczają poza zakres niniejszego przeglądu, dlatego w celu uzyskania przystępnego wyjaśnienia logiki leżącej u podstaw tej metody zalecamy zapoznanie się z pozycją Bloom (2006).

Na bardziej intuicyjnym poziomie jednostki *cross-over* dodatkowo „rozwadniają" kontrast między grupą objętą oddziaływaniem bodźca a grupą kontrolną. Potrzebne jest większe dostosowanie niż w przypadku samych jednostek *no-show*. Należy odjąć od wskaźnika poddania oddziaływaniu bodźca odsetek jednostek, które przezwyciężyły wykluczenia, aby wziąć udział w interwencji. Średni efekt objęcia bodźcem uzyskuje się zatem ponownie, dzieląc różnicę w wyniku dla grupy objętej oddziaływaniem bodźca i grupy kontrolnej, tym razem przez wskaźnik poddania oddziaływaniu bodźca pomniejszony o odsetek jednostek z grupy kontrolnej, które otrzymują bodziec. W ten sposób otrzymujemy wzór podany w wierszu D w Tabeli 1.

Ten estymator został nazwany przez jego twórców LATE (ang. *Local Average Treatment Effect*, tj. lokalny średni efekt objęcia bodźcem). Termin „lokalny" odnosi się do faktu, że ten szacowany efekt ma zastosowanie wyłącznie do podzbioru kwalifikujących się jednostek, które podporządkowują się przypisanemu im statusowi. Podporządkowujących się można traktować jako grupę jednostek, które faktycznie trzymają się protokołu badania; poddadzą się oddziaływaniu bodźca lub nie, w zależności od tego, czy zostaną przypisane do grupy objętej oddziaływaniem bodźca czy do grupy kontrolnej. Z punktu widzenia decydenta jednostki podporządkowujące się stanowią interesującą podgrupę populacji, ponieważ jako jedyne faktycznie odczuwają skutki istnienia oferty. Należy pamiętać, że nie wszystkie podmioty w próbie będą jednostkami podporządkowującymi się: niektórzy *zawsze poddadzą się* oddziaływaniu bodźca, nawet jeśli nie powinni, zaś inni *nigdy nie poddadzą się* oddziaływaniu bodźca, nawet jeśli powinni.

Takie ograniczenie nie eliminuje znaczenia tego szacunku dla polityki. Zmusza ono po prostu do zauważenia, że kiedy mamy coraz mniej kompletne dane, uzyskujemy rezultaty, które mają zastosowanie do coraz mniejszych podgrup kwalifikującej się populacji. W przypadku LATE konieczne jest dostosowanie pod kątem podporządkowania się, w którym uwzględnia się zarówno jednostki *no-show* (w praktyce jednostki *show-up*) i jednostki *cross-over*.

(3) LATE = ITT / wskaźnik podporządkowania się,

gdzie:

(4) Wskaźnik podporządkowania się = odsetek podmiotów, które zostały poddane oddziaływaniu bodźca w grupie objętej oddziaływaniem bodźca – odsetek podmiotów, które zostały poddane oddziaływaniu bodźca w grupie kontrolnej.

W słynnym badaniu Agrista (1984) dotyczącym wpływu służby wojskowej w Wietnamie na zarobki wykorzystano wprowadzone w 1970 r. losowanie kolejności poboru do wojska i wyjazdu do Wietnamu. Spośród mężczyzn urodzonych w 1950 r. tylko 35% wylosowanych faktycznie odbyło służbę, a spośród tych niewylosowanych 19% zgłosiło się. W ramach ITT pomiar wpływu losowania kolejności poboru na roczne zarobki wynosił zaledwie 638 USD, co jest niewielką kwotą nawet po kursie dolara z 1985 r. Jednak taki mały efekt „rozwadnia" bardzo niski poziom podporządkowania się wynikowi losowania. Jeżeli podzielić 638 USD przez 16% (= 35% – 19%), szacunek LATE średniej utraty zarobków *z powodu* służby w Wietnamie daje znacznie wyższą kwotę 3880 USD, która jest dużą sumą po kursie z 1985 r.

Ci, którzy doświadczyli negatywnego wpływu w wysokości prawie 4000 USD, stanowili jednak zaledwie 16% populacji mężczyzn urodzonych w 1950 r. – byli to ci, którzy pojechali do Wietnamu, ponieważ zostali wylosowani, ale w przeciwnym razie nie zrobiliby tego. Fakt, że mamy do czynienia zaledwie z 16% populacji, nie znaczy, że ten szacunek ma mniejsze znaczenie czy mniej istotne konsekwencje. Wręcz przeciwnie – są to *właśnie* te osoby, którym należy się rekompensata; nie ci wylosowani, którzy uniknęli poboru, nie ci, którzy sami się zgłosili, mimo że nie zostali wylosowani, ani też nie ci, którzy nie zostali wylosowani i byli szczęśliwi z tego powodu. Podsumowując, estymator LATE ma zastosowanie do podgrup populacji, które mogą być małe pod względem liczebności, ale które mogą się wydawać ogromne z punktu widzenia polityki.

## Ostatni element – randomizowanie zachęty

Świadomość, że „nie można nikogo zmuszać, nie można nikogo wykluczać" w wielu sytuacjach doprowadziła do wypracowania podejścia, w którym uzyskuje się średnie efekty objęcia bodźcem dla podgrupy populacji poprzez losowe zachęcanie docelowej populacji do udziału (Bradlow 1998). *Podejście oparte na zachęcie* jest szczególnym przypadkiem podejścia eksperymentalnego, które można zastosować w sytuacjach małej kontroli nad podporządkowaniem się jednostek. Kluczowa idea jest następująca: zamiast losowego oferowania interwencji, randomizuje się *zachętę* do udziału i objęcia bodźcem. Dzięki randomizacji zachęty i szczegółowemu śledzeniu rezultatów dla wszystkich jednostek, zarówno tych, które otrzymały zachętę, jak i tych, które jej nie otrzymały, możliwe jest uzyskanie wiarygodnych szacunków dla zachęty i samej interwencji. Warunek jest tylko jeden, tj. aby zachęta zwiększała prawdopodobieństwo, że jednostki zrobią to, do czego się je zachęca. Taka jest podstawa wzoru dla efektów objęcia bodźcem (wzór jest zasadniczo identyczny z (3), poza tym że

$$(5) \quad \text{LATE}_{\text{z zachętą}} = \text{ITT} / \text{podporządkowanie się}_{\text{z zachętą}}$$

Podporządkowania się z zachętą = odsetek jednostek objętych oddziaływaniem bodźca w grupie, która otrzymała zachętę/odsetek jednostek objętych oddziaływaniem bodźca w grupie, która nie otrzymała zachęty.

Zachęta jest tylko zachętą. Na przykład firmy otrzymujące zachętę do ubiegania się o dotację mogą nie złożyć wniosku. A inne firmy, które nie otrzymają żadnej zachęty, mogą mimo to otrzymać dofinansowanie dzięki informacjom uzyskanym z innych źródeł. Ponieważ sama zachęta jest randomizowana, porównanie między grupami, które otrzymały zachętę, i tymi, które jej nie otrzymały, będzie wolne od wszelkich obciążeń związanych z autoselekcją, jeżeli zachętę zapewniono zgodnie z planem. Jednym z kluczowych wymogów uzyskania rozsądnych szacunków jest skuteczność zachęty w nakłanianiu osób do udziału. Ogólnie rzecz ujmując, jest to uzależnione od rodzaju zachęty, ponieważ niektóre zachęty będą stosunkowo skuteczne i będą miały duży wpływ na poddawanie się oddziaływaniu bodźca, natomiast inne będą znacznie mniej skuteczne lub zupełnie nieskuteczne.

Istnieje kilka kluczowych założeń, które należy poczynić. Zachęta nie może przynosić odwrotnego skutku, tzn. nie może zmniejszać prawdopodobieństwa, że podmioty otrzymają bodziec. Jest to często rozsądne założenie, ale trzeba je starannie przemyśleć w każdym indywidualnym przypadku. *Zachęta nie ma bezpośredniego wpływu na rezultaty, z wyjątkiem tego, że zwiększa prawdopodobieństwo otrzymania bodźca.* Dlatego właśnie zachęta powinna być jak najprostsza. Na przykład jeśli zachęta przybiera formę szkolenia, które może przynieść własne rezultaty *poza* zachęceniem do poddania się oddziaływaniu bodźca, wówczas to założenie zostałoby naruszone.

Zachęta może przybrać formę informacji dodatkowych w stosunku do jakichkolwiek informacji, które są już elementem wdrażania programu, i ukierunkowanych na jednostki. Na przykład w kontekście polityki antynikotynowej ewaluator może zaplanować telewizyjną i radiową kampanię reklamową; *zachęta* mogłaby przybrać formę dodatkowych bezpośrednich wiadomości reklamowych wysłanych do losowo wybranej próby jednostek.

## Wnioski

Stosowanie RCT znacznie się rozpowszechniło w ostatnich latach, w związku z czym RCT *są powszechniej stosowane* i *lepiej rozumiane*. Jednym z problemów, z jakimi sobie poradzono jest niepodporządkowanie się, a więc fakt, że po losowym przypisaniu do jednej z grup niektóre osoby znajdują powody, aby nie podporządkować się: niektóre z tych przypisanych do grupy objętej oddziaływaniem bodźca rezygnują przed zakończeniem, a czasami nawet jeszcze przed rozpoczęciem działania, natomiast innym przypisanym do grupy kontrolnej mimo wszystko udaje się otrzymać bodziec.

Okazało się, że zamiast być nierozwiązywalnym problemem, kwestia niepodporządkowania się doprowadziła do lepszego zrozumienia tego, co można uzyskać za pomocą RCT, zwłaszcza, jeśli dopuści się heterogeniczność oddziaływania, która oznacza, że dla różnych podgrup populacji średnie efekty mogą być różne. Im większy zakres niepodporządkowania się, tym mniejsza podgrupa populacji, dla której można uzyskać korzystny szacunek oddziaływania. Nie należy jednak bagatelizować znaczenia tych mniej ogólnych szacunków dla polityki, ponieważ decydenci w dalszym ciągu są zainteresowani tymi, którzy zmieniają swoją decyzję o udziale z powodu wprowadzenia polityki w życie. Ci, którzy pozostają w miejscu, w którym byliby nawet w przypadku braku polityki, dostarczają bardzo mało informacji na temat jej wpływu.

**Tabela 1.** Różne rodzaje niepodporządkowania się w randomizowanych próbach kontrolnych

| | Który efekt? | Który rodzaj niepodporządkowania się jest uwzględniony? | Który estymator? | W jaki sposób uzyskuje się szacunki? | Której podgrupy populacji dotyczy szacowany średni efekt? |
|---|---|---|---|---|---|
| **A** | Efekt **otrzymania oferty** objęcia oddziaływaniem bodźca | **Żaden**<br>W tym przypadku uważa się, że **nie jest istotne**, która część tych, którym zaoferowano objęcie oddziaływaniem bodźca, faktycznie została nim objęta | **ITT**<br>Przeznaczenie do objęcia działaniem bodźca | $ITT = \bar{Y}_{oferta} - \bar{Y}_{brak\ oferty}$ | Średni **efekt oferty** odnosi się do **wszystkich kwalifikujących się jednostek** |
| **B** | Efekt faktycznego **otrzymania bodźca** | **Wszyscy, którym zaproponowano objęcie oddziaływaniem bodźca, zostają nim objęci, ale tylko oni**<br>*Prawdopodobieństwo (objęty/oferta) = 1*<br>*Prawdopodobieństwo (objęty/BRAK oferty) = 0* | **ATE**<br>Średni efekt objęcia bodźcem | $ATE = \bar{Y}_{oferta} - \bar{Y}_{brak\ oferty}$<br>$ATE = ITT$<br>przy perfekcyjnym podporządkowaniu się | Średni **efekt objęcia bodźcem** odnosi się do **wszystkich kwalifikujących się jednostek** |
| **C** | | **Tylko część** osób którym zaproponowano objęcie oddziaływaniem bodźca, **zostaje nim objęta,** natomiast te, którym nie zaproponowano, nie mogą zostać nim objęte<br>*Prawdopodobieństwo (objęty/oferta) < 1*<br>*Prawdopodobieństwo (objęty/BRAK oferty) = 0* | **TOT**<br>Efekt objęcia bodźcem dla jednostek uczestniczących | $TOT = \frac{\bar{Y}_{oferta} - \bar{Y}_{brak\ oferty}}{Prob\ (objęty \mid oferta)}$<br>Zawsze tak jest<br>TOT > ITT | Średni **efekt objęcia bodźcem** odnosi się do **faktycznych uczestników programu** |
| **D** | | **Część** tych, którzy otrzymali ofertę, nie korzysta z niej, natomiast część tych, którzy nie otrzymali oferty, mimo wszystko bierze udział<br>*Prawdopodobieństwo (objęty/oferta) < 1*<br>*Prawdopodobieństwo (objęty/BRAK oferty) > 0* | **LATE**<br>Lokalny średni efekt objęcia bodźcem | $LATE = \frac{\bar{Y}_{oferta} - \bar{Y}_{brak\ oferty}}{Prob\ (objęty \mid oferta)}$ | Średni efekt objęcia bodźcem odnosi się tylko do tych, których **nakłoniono** do udziału w programie **za pomocą oferty albo zachęty**, ale którzy w przeciwnym razie nie zrobiliby tego |
| **E** | Efekt **otrzymania bodźca** w wyniku zachęty | Część tych, którzy otrzymali zachętę, poddaje się działaniu bodźca, natomiast część tych, którzy nie otrzymali zachęty, zostaje objęta oddziaływaniem bodźca<br>*Prawdopodobieństwo (objęty/zachęta) < 1*<br>*Prawdopodobieństwo (objęty/BRAK zachęty > 0* | | $LATE = \frac{\bar{Y}_{zachęta} - \bar{Y}_{brak\ zachęty}}{P(objęty \mid zachęta) - p(objęty \mid zachęta)}$ | |

**Alberto Martini** – Absolwent Wydziału Prawa Uniwersytetu w Turynie (1980), Doktor Ekonomii Uniwersytetu Wisconsin-Madison (1988). W latach 1988-1993 w Mathematica Policy Research – firmie badawczej specjalizującej się w ewaluacji polityk społecznych i wspierających zdrowie. W latach 1993-1998 – Starszy

badacz w Urban Institute, gdzie zajmował się kwestiami związanymi z dobrobytem oraz projektowaniem modeli mikrosymulacyjnych. Od 1998 r. – Profesor statystyki i ewaluacji polityk na Uniwersytecie Piemonte Orientale w Turynie. W latach 2001-2002 pełnił funkcję Prezesa Włoskiego Towarzystwa Ewaluacyjnego. Od 2007 r. członek Consiglio Italiano delle Scienze Sociali (Włoskiej Rady Nauk Społecznych), gdzie współprzewodniczy Komisji ds. Ewaluacji Oddziaływania.

## Bibliografia


- *Angrist J.D., Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence form Social Security Administrative Records, „American Economic Review" 1990, vol. 80, s. 313-336.*
- *Angrist J.D., Imbens G.W., Rubin D.B. , Identification of Causal Effects Using Instrumental Variables, "Journal of the American Statistical Association" 1996, vol. 91(434), s. 444-455.*
- *Bradlow E., Encouragement Designs: An Approach to Self-Selected Samples in an Experimental Design, "Marketing Letters" 1998, vol. 9 (4), s. 383-391(9).*
- *Bell S., Peck L., Obstacles to and Limitations of Social Experiments: 15 False Alarms, opracowanie przedstawione na Międzynarodowej Konferencji poświęconej eksperymentom terenowym w ewaluacji polityki, Nürburg, Niemcy, 22 maja 2012 r.*
- *Berk R.A., Randomized experiments as the bronze standard, dokument 2005080201, Wydział Statystyki, UCLA, Los Angeles 2005.*
- *Bloom H., Accounting for No-Shows in Experimental Evaluation Designs, „Evaluation Review" 1984,vol. 8(2), s. 225-246.*
- *Bloom H., The Core Analytics of Randomized Experiments for Social Research, MDRC Working Papers on Research Methodology, Nowy Jork, NY 2006.*
- *Pawson R.,Tilley N., Realistic Evaluation, Sage Publications, Thousand Oaks, Ca 1997.*

*Philip Davies*

# Stosowanie metod mieszanych w ewaluacji na potrzeby kształtowania polityk publicznych

Ewaluacja polityk publicznych ma na celu znalezienie wysokiej jakości dowodów teoretycznych i empirycznych dotyczących szeregu pytań, na które napotykają osoby odpowiedzialne za projektowanie polityk i instytucje świadczące usługi publiczne. Oto wspomniane pytania:

- W jaki sposób ma funkcjonować polityka, projekt lub program?
- Jakie istnieją dowody na skuteczność proponowanej polityki, proponowanego projektu lub programu?
- Jakie istnieją dowody na skuteczne wdrażanie i realizację proponowanej polityki, proponowanego projektu lub programu?
- Dla kogo proponowana polityka, proponowany projekt lub program są skuteczne lub nieskuteczne?
- Jakie są doświadczenia i spostrzeżenia obywateli związane z polityką, projektem lub programem?
- Jakie są koszty, opłacalność i stosunek kosztów do korzyści polityki, projektu lub programu?

Aby pomóc odpowiedzieć na te pytania, ewaluacja musi przedstawić teorię zmiany, określająca mechanizmy, za pomocą których można osiągnąć różne rodzaje skuteczności. Wymaga to pełnego zakresu metod badawczych (Rossi, Freeman i Lipsey 1999, s. 20), które są wykorzystywane w ewaluacji bez traktowania w sposób uprzywilejowany żadnej metody ani żadnego typu ewaluacji. Zasadą przewodnią ewaluacji polityki powinna być kwestia: „jakie pytanie, lub problem, wymaga odpowiedzi?", a nie preferencje ewaluatorów dotyczące konkretnego typu ewaluacji lub metody dochodzenia. Ta druga kwestia ma znaczenie jedynie w przypadku, gdy pierwsza kwestia jest jasna.

W niniejszym tekście siedem pytań, które pojawiają się w kontekście kształtowania polityki rozpatrywane jest pod kątem kwestii, co ewaluacja polityki może zaoferować, generując odpowiednie dowody. Te siedem pytań to:

1. W jaki sposób polityka ma funkcjonować pod kątem osiągania pożądanych wyników?
2. Co już wiadomo o polityce lub o problemie, który ma rozwiązać?
3. Jaki jest charakter i rozmiar problemu?
4. Jakie inicjatywy polityczne są skuteczne?
5. W jaki sposób sprawić, aby polityka działała?
6. Jakie są koszty, opłacalność i stosunek kosztów do korzyści różnych wariantów polityki?
7. Jakie są konsekwencje etyczne różnych wariantów polityki?

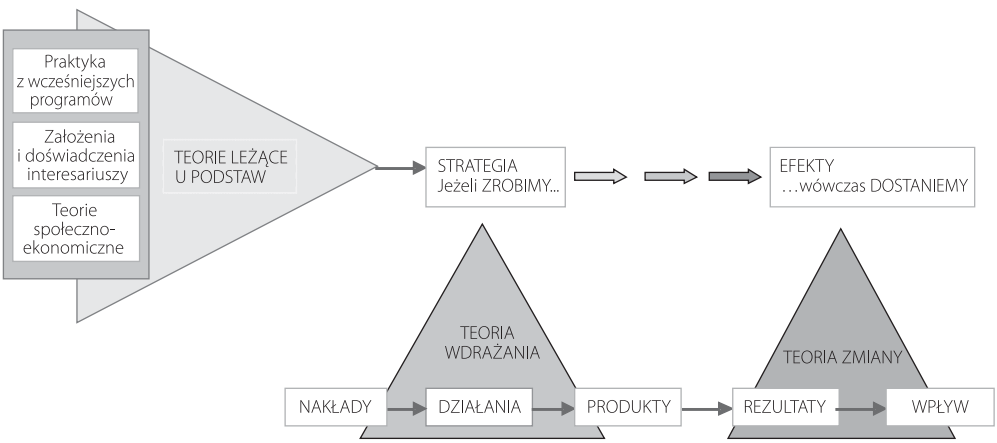
## 1. W jaki sposób polityka ma funkcjonować pod kątem osiągania pożądanych wyników?

Minęły ponad cztery dekady, odkąd Carol Weiss przekonywała, że ewaluacja powinna być oparta na dobrej teorii. Od tego czasu znaczenie podejść ewaluacyjnych opartych na teorii ewoluowało w ważny samoistny paradygmat, jak również stało się niezbędną częścią składową ewaluacji wykorzystujących szereg różnych podejść i metod. Ewaluatorzy, tacy jak Bickman (1987), Chen i Rossi (1983), Chen (1994, 2004, 2005), Petrosino i in. (2000), Pawson (2002), Patton (2008) i White (2011), zwrócili uwagę na potrzebę

zidentyfikowania mechanizmów i założeń, na których opierają się programy i które mogą odpowiadać za powodzenie lub niepowodzenie w osiągnięciu celów. Wspomniane mechanizmy i założenia zapewniają połączenia między nakładami, produktami i rezultatami programu lub interwencji, umożliwiając tym samym przygotowanie łańcucha przyczynowego. To podejście jest bardziej wyszukane niż podejście, które koncentruje się jedynie na nakładach i rezultatach, ponieważ pomaga ustalić, w jaki sposób, dlaczego i w jakich warunkach program może zrealizować zamierzone cele. Teoria zmiany zwraca również uwagę na czynniki kontekstowe, takie jak położenie geograficzne, kultura, religia, pochodzenie etniczne oraz otoczenie polityczne, w którym programy i interwencje są wdrażane.

Teoria zmiany jest zatem czymś więcej niż *jedynie* podejściem teoretycznym. Może mieć praktyczny wpływ na politykę i usługi publiczne, ponieważ może pozwolić ewaluatorom, osobom odpowiedzialnym za tworzenie polityk i innym interesariuszom określić, co powinno istnieć, aby interwencja zakończyła się powodzeniem (Rys. 1). Olejniczak (2009) pokazuje, że teoria programu może zapewnić teorię wdrażania, jak również teorię zmiany, z których pierwsza określa nakłady, działania i produkty, które należy zidentyfikować w celu wywołania rezultatów, które będą stanowiły wymaganą zmianę.

**Rys. 1.** Teoria wdrażania i teoria zmiany




Ustalenie teorii zmiany nie zawsze jest jednak proste. Wiele interwencji w ramach polityk publicznych jest złożonych lub ma różne subkomponenty i funkcjonuje na wielu poziomach (wspólnotowym, instytucjonalnym, osobistej sieci, rodzinnym i indywidualnym). Wdrażanie tych inicjatyw łączy się często z różnymi kontekstami (geograficznym, regionalnym, gospodarczym, etnicznym, kulturowym, religijnym itd.) i z wieloma interesariuszami (na poziomie krajowym, regionalnym, samorządu terytorialnego, organizacjami pozarządowymi, sektorem prywatnym, trzecim sektorem, interesami miast/wsi itd.), tym samym wymaga różnych mechanizmów i założeń. Tam, gdzie występuje taka złożoność i różnorodność, mogą istnieć liczne, jeśli nie sprzeczne, teorie zmiany. Connell i Kubisch podkreślili, że „liczne teorie zmiany mogą funkcjonować jednocześnie w ramach jednej [interwencji] i że różni interesariusze mogą działać w ramach różnych teorii zmian, które mogą nawet być konkurencyjne." (Connell i Kubisch 1998, s. 7).

Metody stosowane przy przygotowywaniu teorii zmiany obejmują analizę logiczną, analizę operacji, konsultacje z interesariuszami z wykorzystaniem techniki delfickiej i grupy nominalnej, pogłębione wywiady i grupy fokusowe z udziałem kluczowych interesariuszy, obserwację/obserwację uczestniczącą i etnografię. Może zaistnieć potrzeba uzupełnienia tych jakościowych metod opracowywania teorii zmiany dowodami z systematycznych przeglądów istniejącej literatury naukowej oraz danymi ilościowymi z kwestionariuszy, spisów i źródeł administracyjnych. White (2009) zaproponował metodę analizy służącą przygotowywaniu i testowaniu teorii

zmiany opartą na określeniu łańcucha przyczynowego, zrozumieniu kontekstu, w którym będzie wdrażana polityka, przewidywaniu różnorodności populacji, kontekstów i doświadczeń oraz rygorystycznej ewaluacji wpływu z wykorzystaniem odpowiednich metod kontrfaktycznych i mieszanych.

## 2. Co już wiadomo o polityce lub o problemie, który ma rozwiązać?

Podczas przygotowywania polityki i jej ewaluacji należy wykorzystać to, co już wiadomo na temat kwestii, której dotyczy dana polityka, na podstawie istniejących dostępnych dowodów. Na przestrzeni ostatnich kilku dekad rozwinęły się różne metody syntezy badawczej, które obejmują metaanalizę statystyczną, metaanalizę narracyjną, szybkie oceny dowodów, mapy dowodów, mapy braków i jakościową syntezę/metaetnografię. Metaanaliza statystyczna i metaanaliza narracyjna są dwoma rodzajami systematycznego przeglądu, które „starają się odkryć zgodności i wyjaśniają różnorodność badań, które wydają się podobne" (Cooper i Hedges 1994, s. 4). W systematycznych przeglądach odróżnia się również badania o wysokiej jakości od badań o niższej jakości, ustalając jasne i przejrzyste kryteria wewnętrznej i zewnętrznej spójności i jakości raportowania. Badania, które spełniają te kryteria włącza się do systematycznego przeglądu, podczas gdy badania, które ich nie spełniają są z niego wyłączone lub przyznaje im się niższą ocenę jakości. Szybkie oceny jakości korzystają z zasad i procedur systematycznych przeglądów, ale przeprowadzają przegląd i procesy analityczne w krótszym czasie, w celu wypełnienia napiętych harmonogramów planowania i tworzenia polityk. W konsekwencji szybkie oceny jakości są zazwyczaj mniej kompletne niż pełne, systematyczne przeglądy i mogą bardziej powierzchownie traktować obciążenia zawarte w dostępnych dowodach.

Systematyczne przeglądy podejmowano głównie w celu określenia skuteczności interwencji, odpowiadając na pytanie „co działa". Można również stosować je w celu zidentyfikowania, jakiego rodzaju dowody istnieją w tym temacie i w celu ustalenia charakteru i zakresu przedmiotowej kwestii. Dlatego systematyczne przeglądy dostarczają często mapy dostępnych dowodów (Gough i in.) i wskazują, gdzie występują braki w bazie dowodów (Bhavsar i in.). Metody systematycznego przeglądu były rozwijane również w celu dokonania syntezy badań jakościowych (Britten i Campbell, Snilstveit), w tym badań opartych na wywiadach indywidualnych i grupowych, badań etnograficznych opartych na obserwacji uczestniczącej. Synteza jakościowa zapewnia dowody na to, w jaki sposób ludzie postrzegają interwencje i ich doświadczają lub mówiąc bardziej ogólnie, w jaki sposób ludzie rozumieją i interpretują świat, na jakim żyją. Traktowane łącznie omawiane różne metody syntezy badawczej pozwalają ewaluatorowi nie tylko ustalić, co wiadomo na dany temat od początku prowadzenia ewaluacji, ale także ustalić mocne strony, ważność, rzetelność i wiarygodność tych dowodów.

Systematyczne przeglądy można stosować w celu zagwarantowania, że analiza teorii zmiany opiera się na dowodach empirycznych i w celu ustalenia, jak daleko w łańcuchu przyczynowym znajdują się dowody na poparcie, lub zakwestionowanie, teoretycznych założeń i hipotez, leżących u podstaw inicjatywy politycznej. Systematyczne przeglądy mogą zatem odegrać znaczącą rolę na wczesnych etapach ewaluacji, jak również w trakcie gromadzenia dowodów i budowania bazy dowodów w danym temacie.

## 3. Jaki jest charakter i rozmiar problemu?

Na ewaluacje polityk ma często wpływ to, w jaki sposób przedmiotowe kwestie są ujęte w polityce i procesach politycznych. Na przykład w Wielkiej Brytanii kolejne rządy ujmowały problemy spożywania alkoholu w kontekście jednorazowego spożycia alkoholu w nadmiernych ilościach i zachowań niezgodnych z ogólnymi normami społecznymi wśród młodzieży. Tak właśnie przedstawiano sedno problemu alkoho-

lowego w Wielkiej Brytanii. W rezultacie przygotowywano polityki i szukano dowodów, w celu zmniejszenia spożywania alkoholu i zachowań niezgodnych z ogólnymi normami społecznymi przez młodzież. Nie ma wielu wątpliwości w kwestii, że nadmierne spożywanie alkoholu i zachowania w stanie nietrzeźwości wśród młodzieży są powszechne w niektórych okresach (np. podczas weekendów) i że prowadzą do zakłóceń porządku publicznego oraz do poważnych chorób i urazów. Bardziej szczegółowa analiza oficjalnych statystyk i badań w dziedzinie zdrowia wskazuje jednak, że problemy alkoholowe w Wielkiej Brytanii są szersze i głębsze niż sugeruje częściowa analiza. W ostatnich dwóch dekadach w szczególności znacząco wzrosło występowanie chorób wątroby będących skutkiem spożywania alkoholu, problemów kardiologicznych, wypadków i urazów oraz pewnych aspektów demencji, a wspomniane przypadki wzrostu miały miejsce wśród osób w wieku średnim lub osób starszych, które spożywają alkohol. Zapotrzebowanie na świadczenia zdrowotne będące odpowiedzią na te warunki nałożyło jeszcze dalsze ograniczenia na i tak już napięte budżety instytucji zajmujących się ochroną zdrowia.

Charakter i rozmiar problemów alkoholowych w Wielkiej Brytanii są zatem szersze i bardziej złożone niż sugerują niektóre dyskusje dotyczące polityk i wymagają odpowiedzi w postaci polityki, która odpowiada na problem ilości alkoholu spożywanego przez ludzi w *każdym* wieku i w *szeregu* środowisk (np. dom, restauracje, lokale posiadające koncesje, imprezy sportowe, przestrzenie publiczne itd.). To z kolei może wymagać teorii zmiany i interwencji mających na celu ograniczenie ogólnej dostępności alkoholu, w tym podniesienia akcyzy na napoje alkoholowe, ograniczenia czasu i miejsc, w których można nabywać lub spożywać alkohol oraz ograniczenia sposobów, w jakie supermarkety i inne punkty sprzedaży alkoholu wprowadzają do obrotu napoje alkoholowe. Ten szerszy zakres inicjatyw w ramach polityki może być sprzeczny z przekonaniami i ideologiami politycznymi, które promują wolne rynki i ograniczoną rolę państwa w zakresie wpływania na zachowanie obywateli.

Zdefiniowanie charakteru i wielkości oraz dynamiki danego problemu jest zatem ważną częścią procesu ewaluacji polityki. Można opierać się na korzystaniu z oficjalnych statystyk w postaci danych ze spisów, danych pochodzących z badań ankietowych i danych administracyjnych, jak również na dowodach jakościowych z pogłębionych wywiadów, grup fokusowych i badań etnograficznych. Dane pochodzące z kwestionariuszy, dane administracyjne i dane ze spisów gromadzi zazwyczaj rząd lub krajowy urząd statystyczny. Dane ze spisów są zazwyczaj zbierane co dziesięć lat, a ich zaletą jest objęcie prawie 100% całej populacji. W konsekwencji, ustalenia spisu są zwykle wiarygodne na poziomie małego obszaru (na poziomie dzielnicy, ulicy, a nawet gospodarstwa domowego). Spisy są jednak bardzo drogie i w związku z faktem, że przeprowadza się je co dziesięć lat, dezaktualizują się z upływem czasu. Chociaż dane ze spisu są zasadniczo danymi na poziomie makroekonomicznym (tj. zagregowanymi na poziomie krajowym lub regionalnym), niektóre kraje wydają również próbki zanonimizowanych rekordów, w skład których wchodzą dane mikroekonomiczne na poziomie indywidualnym lub gospodarstwa domowego. W Wielkiej Brytanii, na przykład, w spisie z 2001 r. próbki zanonimizowanych rekordów na poziomie jednostki obejmują 3% populacji (ok. 1,75 mln przypadków), podczas gdy próbki zanonimizowanych rekordów na poziomie gospodarstwa domowego obejmują 1% populacji (ok. 200 tys. gospodarstw domowych i ok. 500 tys. ich członków).

Dane pochodzące z badań ankietowych mogą być przekrojowe lub wzdłużne i mogą zawierać informacje dotyczące statusu indywidualnych osób (wiek, płeć, zatrudnienie, dochód, warunki mieszkaniowe itd.) lub ich postaw, przekonań i perspektyw. *General Household Survey/General Lifestyle Survey*, którego wersja jest prowadzona w większości krajów, jest przykładem badania statusu jednostek, podczas gdy *British Social Attitudes Survey* (Park i in. 2012) jest badaniem postaw, przekonań i perspektyw. Dane pochodzące z badań kwestionariuszowych są zwykle wyraźnie skupione na konkretnej tematyce, chociaż stosuje się je również dla celów administracyjnych, i zazwyczaj dostarczają danych na temat szerokiej gamy zmiennych, jak również ważnych właściwości statystycznych (np. informacje dotyczące metody dobierania prób i błędu próby). Badania ankietowe są na ogół drogie, w szczególności w przypadku, gdy wymagane są duże

próby w celu zapewnienia reprezentatywności i odpowiedniej mocy statystycznej. Ustalenia tych badań są często mniej wiarygodne na poziomie małego obszaru.

Dane administracyjne obejmują dane zbierane w pierwszej kolejności dla celów rządowych i administracyjnych na poziomie krajowym, lokalnym i wielonarodowym. Obejmują dane dotyczące takich tematów, jak ubezpieczenia społeczne, uiszczanie i rozkład podatku dochodowego, zgłaszane przestępstwa, osiągnięcia uczniów i studentów (osiągnięcia szkół i uczelni), wykorzystanie szpitali, zasoby mieszkaniowe i zajmowanie lokali itd. W związku z faktem, że dane administracyjne gromadzi się rutynowo w celach funkcjonalnych, ich gromadzenie nie stanowi zasadniczo obciążenia dla populacji docelowej. Zbiera się je regularnie, zazwyczaj w sposób spójny (chociaż zmiana definicji „bezrobocia", „choroby", „przestępstwa" itd. podważa tę zaletę danych administracyjnych) i zazwyczaj obejmują one prawie 100% populacji będącej przedmiotem zainteresowania. W niektórych przypadkach zestawy danych administracyjnych można łączyć, zapewniając tym samym wartościową analizę międzysektorową dla wielu zmiennych. Dane administracyjne mogą również być wiarygodne na poziomie małego obszaru. Wady danych administracyjnych obejmują niespójności w zakresie definicji i praktyk rejestracyjnych, niepełne zestawy danych, brakujące lub powielone dane oraz brak ciągłości danych na przestrzeni czasu. Również, w związku z faktem, że dane administracyjne gromadzi się dla celów administracyjnych i funkcjonalnych, dane te mogą być nieprzydatne z punktu widzenia zakresu i przedmiotu zainteresowania ewaluacji.

Dane jakościowe odpowiadają na pytania ewaluacyjne, na które odpowiedzi nie możemy uzyskać z danych ilościowych. Różne doświadczenia osób z różnych grup społecznych, kulturowych i etnicznych dotyczące inicjatyw politycznych wymagają zwykle danych z pogłębionych wywiadów indywidualnych, grupowych oraz obserwacji uczestniczącej. Te same metody pomagają również określić konkretne czynniki, okoliczności i konteksty, w których polityka może mieć różnorodne efekty (w tym być skuteczna czy nieskuteczna). Takie dane są ważne dla określenia, w jaki sposób, dlaczego i w jakich warunkach inicjatywa polityczna jest skuteczna/nieskuteczna.

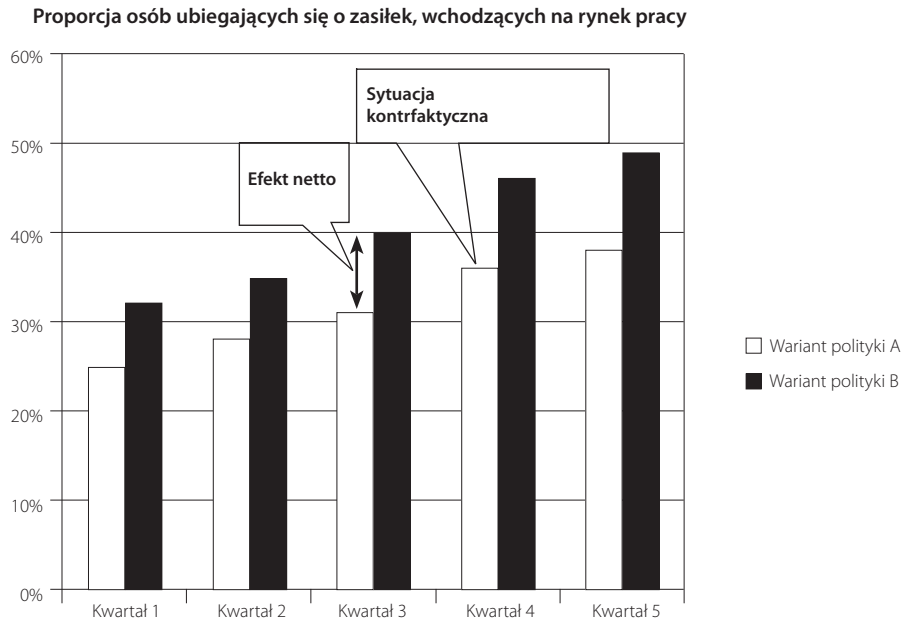
## 4. Jakie inicjatywy polityczne są skuteczne?

W ramach ewaluacji polityki istnieją co najmniej dwa odrębne pojęcia skuteczności. Pierwsze ma swoje źródło w zarządzaniu ukierunkowanym na wyniki i typach ewaluacji nastawionych na osiąganie celów i definiuje skuteczność z punktu widzenia osiągnięcia zamierzonych celów. Według House'a „osiągnięcie celów oznacza spojrzenie na cele programu, a następnie zgromadzenie dowodów odnoszących się do osiągnięcia tych celów" (House 1980, s. 26). Podejście to jest często popularne w instytucjach rządowych, gdzie świadczenie usług publicznych jest monitorowane w celu ustalenia, czy założone wskaźniki celu zostały osiągnięte. Ewaluacja osiągnięcia celów wpisuje się w demokratyczne pojęcia odpowiedzialności i kontroli finansów publicznych. Verdung zasugerował, że „argument demokratyczny oparty na pojęciu prymatu parlamentarnego łańcucha kontroli, a w konsekwencji na perspektywie demokratycznej jest nie do odparcia" (Verdung 2009, s. 41).

Ewaluacja osiągnięcia celów jest raczej niewłaściwym określeniem. Jest to zasadniczo *monitoring* osiągnięcia celów i wskazuje po prostu, czy osiągnięto pewne wyniki lub rezultaty. Nie informuje, *dlaczego* osiągnięto te wyniki lub rezultaty ani, dokładniej, czy te wyniki i rezultaty *są spowodowane* ewaluowaną inicjatywą polityczną. Oznacza to, że nie może *przypisać* żadnych zaobserwowanych zmian wyników lub rezultatów ewaluowanej interwencji. W tym przypadku odpowiednie jest drugie rozumienie skuteczności i wymaga mierzenia wpływu interwencji w porównaniu z wpływem jednego alternatywnego wariantu polityki lub większej ich liczby, w tym niepodejmowania żadnych działań. Wspomniane alternatywne warianty polityki nazywa się *kontrfaktycznymi*.

Rysunek 2 ilustruje *efekt netto* inicjatywy politycznej w porównaniu ze stanem kontrfaktycznym. W tym przypadku wariant polityki A (białe słupki) jest istniejącą polityką (czasami nazywaną „dotychczasowym scenariuszem postępowania") w zakresie „od zasiłku do zatrudnienia", a wariant polityki B (czarne słupki) jest nową, ocenianą polityką. Dane na Rysunku 2 wskazują, że w ramach wariantu polityki B proporcja osób ubiegających się o zasiłek, wchodzących na rynek pracy między kwartałem 1 i kwartałem 5, wzrosła z 32% do 49%. Proporcja osób ubiegających się o zasiłek wchodzących na rynek pracy w ramach wariantu polityki A również rosła w każdym kwartale, z 25% do 38%. *Efekt netto*, wariantu polityki B w porównaniu z wariantem polityki A, waha się zatem od 7% w kwartale 1 do 11% w kwartale 5. Zakładając, że warunki, w których wdraża się warianty polityki A i B są porównywalne, można stwierdzić, że a) wariant polityki B jest bardziej skuteczny niż wariant polityki A i b) że wyższą skuteczność wariantu polityki B można przypisać tej polityce. Rodzi to pytanie: w jaki sposób znaleźć taką sytuację kontrfaktyczną, która zapewni prawdziwą porównywalność scenariusza alternatywnego z ocenianą polityką?

**Rys. 2.** Efekt netto interwencji i sytuacji kontrfaktycznej




W *Magenta Book* starannie podsumowano możliwości znalezienia sytuacji kontrfaktycznej: „metody oparte na wykorzystaniu randomizowanych prób kontrolnych są na ogół uważane za najodpowiedniejszy sposób określania sytuacji kontrfaktycznej dla danej polityki, programu lub projektu, chociaż starannie kontrolowane badania z próbami dopasowanymi według cech i niektóre formy modelowania statystycznego również zapewniają przybliżenie sytuacji kontrfaktycznej" (HMT 2007, s. I:5). Alternatywami dla randomizacji w celu określenia sytuacji kontrfaktycznej są analiza nieciągłości w równaniu regresji (ang. *regression discontinuity design*), metoda *propensity score matching* i metoda różnicy w różnicach (ang. *difference-in-difference*).

W (ostatecznym) podsumowaniu przy pomocy randomizowanych prób kontrolnych można uzyskać porównywalną sytuację kontrfaktyczną dzięki faktowi, że poprzez przydzielanie osób lub jednostek (szkoły, szpitale, urzędy pracy) lub całych społeczności do grupy objętej interwencją (grupa eksperymentalna) lub grupy nieobjętej interwencją (grupa kontrolna) w sposób losowy (Rysunek 3a), wszystkie pozostałe czynniki (przeszkody), które mogą wpłynąć na rezultaty będą równo rozdystrybuowane w grupie eksperymental-

nej i kontrolnej. Jest tak jedynie w przypadku spełnienia pewnych warunków, m.in. wystarczająco dużego rozmiaru prób, aby nadać eksperymentowi odpowiednią moc statystyczną i w przypadku, gdy próba jest starannie administrowana tak, aby uniknąć „zarażenia" (gdy interwencja oddziaływuje także na członków grupy kontrolnej) lub „skażenia" (gdy członkowie grupy objętej interwencją lub kontrolnej mają dostęp do innej interwencji, która może również wpływać na rezultaty, będące przedmiotem zainteresowania).

Analiza nieciągłości w równaniu regresji przydziela jednostki do grupy objętej interwencją lub porównawczej na podstawie wyraźnie określonego wskaźnika lub parametru ze znaną wartością progową kwalifikowalności, taką jak dochód, umiejętność czytania lub zatwierdzona miara ubóstwa. Na Rysunku 3b dzieci przydzielono do programu nauki czytania na podstawie ich punktacji na zatwierdzonej skali ubóstwa wahającej się od 0 (skrajne ubóstwo) do 100 (brak ubóstwa), a wartość progowa wynosiła =/<50. Na wykresie znajdującym się po lewej stronie Rysunku 3b punktacja po interwencji dla całej próby pokazuje ciągłą linię regresji, wskazując tym samym na brak różnicy rezultatów między grupą objętą interwencją i grupą porównawczą, a zatem na brak wpływu programu nauki czytania. Natomiast wykres znajdujący się po prawej stronie Rysunku 3b pokazuje zmianę punktacji po interwencji podmiotów nią objętych i brak ciągłości linii regresji. Pokazuje to, że interwencja w postaci programu nauki czytania była skuteczna. Porównywalność grupy objętej interwencją i porównawczej jest wyraźnie największa najbliżej wartości progowej. Większa jest także wiarygodność szacowanej wartości i kierunku wpływu. Dane znajdujące się dalej od wartości progowej są w mniejszym stopniu porównywalne, a zatem wiarygodność szacowanej wartości i kierunku wpływu jest niższa.

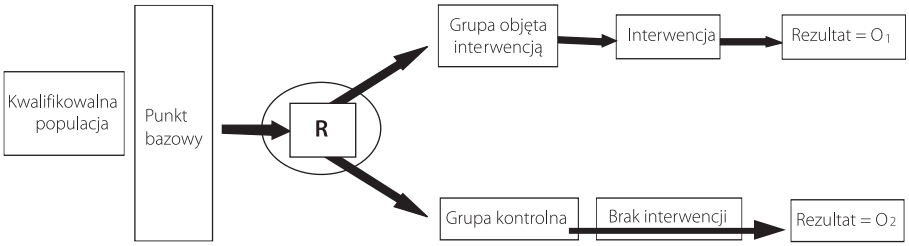
Innym sposobem, w jaki można znaleźć solidne przybliżenie sytuacji kontrfaktycznej jest jak najbliższe połączenie jednostek próby z populacji, która jest przedmiotem interwencji z jednostkami próby z populacji, która nie jest przedmiotem interwencji (zob. Rysunek 3c). Generuje to z kolei pytanie, jakie zmienne należy zastosować w celu połączenia grupy eksperymentalnej i porównawczej. W większości przypadków liczba zmiennych (zmienne towarzyszące) wpływa na rezultaty, które są przedmiotem zainteresowania ewaluacji. Zamiast łączyć grupy na podstawie tylko jednej zmiennej, wprowadzając tym samym obciążenie selekcyjne, lub na podstawie sekwencji zmiennych (co prawda ograniczyłoby to obciążenie selekcyjne, ale również pulę potencjalnych możliwości połączeń), metoda propensity score matching pozwala na łączenie grup na podstawie prawdopodobieństwa (skłonności do) posiadania wszystkich cech istotnych dla rezultatu. Im bardziej pokrywają się charakterystyki grupy objętej interwencją i grupy porównawczej, co nazywa się wspólnym wsparciem, tym większa jest porównywalność obydwu prób. Ograniczenie propensity score matching polega na tym, że jednostki można łączyć jedynie na podstawie cech obserwowanych, co nie pozwala uchwycić wpływu czynników nieobserwowanych, takich jak motywacja.

Na Rysunku 3c pokazano również, że istnieją dwa sposoby oszacowania różnicy rezultatów grupy eksperymentalnej i porównawczej. Jednym jest różnica rezultatów grup po zakończeniu interwencji (O1-O2 na Rysunku 3c). Drugim jest porównanie różnic rezultatów grupy eksperymentalnej między punktem bazowym i zakończeniem interwencji w porównaniu z różnicą rezultatów grupy porównawczej między punktem bazowym i zakończeniem interwencji. Drugi sposób nazywany jest analizą różnicy w różnicach i stanowi również odrębny typ ewaluacji (Rysunek 3d).

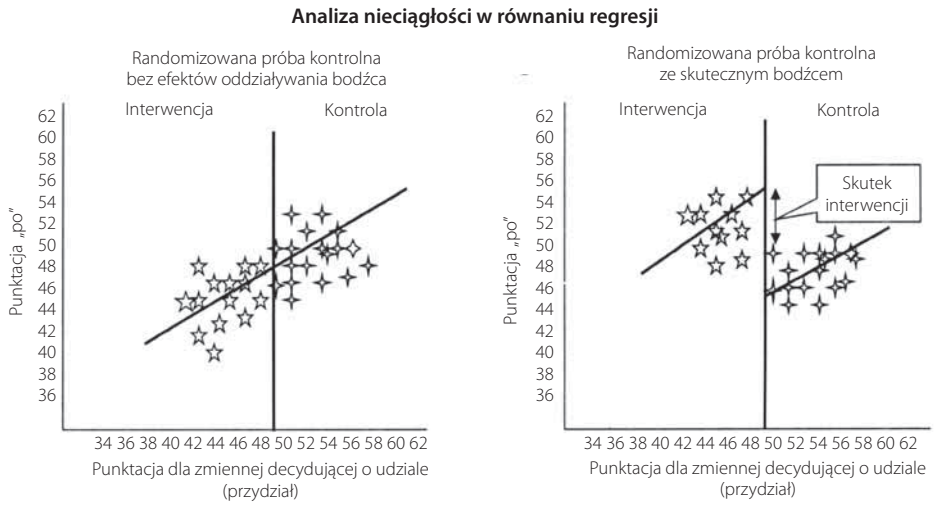
Szacunek wielkości wpływu oparty na różnicy w różnicach jest niekiedy jedynym wariantem ewaluacji, w przypadku, gdy nie można przydzielić osób lub jednostek do grupy eksperymentalnej i porównawczej i gdy niemożliwe jest łączenie obydwu grup. W takich przypadkach rejestruje się różnicę między obydwiema grupami w zakresie zmiennej (zmiennych) będącej (będących) przedmiotem zainteresowania na poziomie punktu bazowego, jak i różnicę między tą samą zmienną (tymi samymi zmiennymi) na poziomie „po". Jeżeli interwencja nie miała żadnego wpływu, wówczas różnica ta powinna mieć taki sam rozmiar na poziomie „po", jak na poziomie punktu bazowego (różnica między linią kropkowaną i niższą linią ciągłą na Rysunku 3d). Jeżeli jednak różnica w grupie eksperymentalnej jest większa na poziomie „po" (różnica między linią kropkowaną i wyższą linią ciągłą na Rysunku 3d), stanowi to skutek interwencji.

**Rys. 3.** Różna podejścia ewaluacyjne do ustalania sytuacji kontrfaktycznej

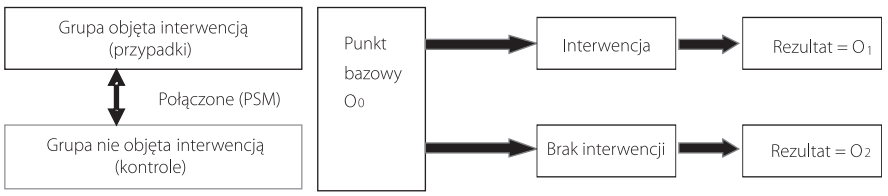
**Rys. 3a.** Analiza randomizowanej próby kontrolnej



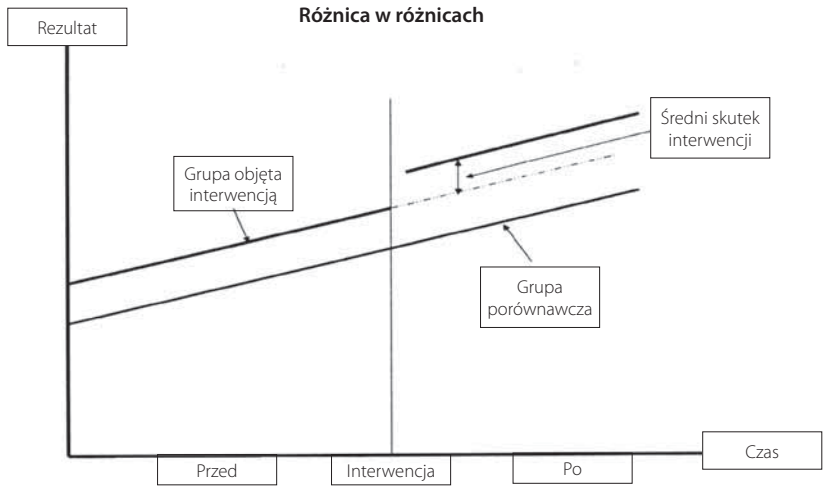

**Rys. 3b**




**Rys. 3c**



**Rys. 3d**




Ten krótki zarys niektórych sposobów znajdowania przybliżenia sytuacji kontrfaktycznej nie oddaje sprawiedliwości szczegółom i złożonościom różnych typów ewaluacji. Bardziej szczegółowy opis różnych metod można znaleźć w publikacjach Campbella i Russo (1998), Ravalliona (1999, 2005), Shadisha, Cooka i Campbella (2002).

## 5. W jaki sposób sprawić, aby polityka działała?

Dla celów kształtowania polityki wiedza, w jaki sposób należy skutecznie wdrażać polityki, aby osiągnąć potencjalne rezultaty jest równie ważna, jak wiedza na temat jej skuteczności w eksperymentalnych warunkach. Ewaluacja koncentruje się w tym przypadku na warunkach, w których można osiągnąć powodzenie we wdrażaniu i realizacji. Ewaluacja skutecznego wdrażania wymaga powrotu do teorii zmiany leżącej u podstaw polityki w celu określenia wkładów, mechanizmów, produktów i zasobów oczekiwanych dla osiągnięcia zamierzonych rezultatów. Te elementy tworzące teorię zmiany polityki można następnie zbadać w celu określenia, czy są właściwe i funkcjonują w przewidziany sposób. Należy również poddać ewaluacji rolę czynników kontekstowych, kulturowych i społeczno-demograficznych jako zmiennych pośredniczących. O'Connor zasugerował, że skuteczne wdrażanie wymaga przygotowania planu realizacji (Rysunek 4), który odzwierciedla tę analizę teorii zmiany.

**Rys. 4.** Przygotowanie planu realizacji


Źródło: O'Connor 2008

Model na Rysunku 4 pokazuje, że ewaluacja wdrażania wymaga analizy teorii zmiany, zrozumienia skuteczności interwencji („co działa") oraz analizy kluczowych uczestników procesu wdrażania. To z kolei opiera się na ilościowych i jakościowych metodach analizy i wymaga dość szczegółowego monitorowania odpowiednich wkładów, mechanizmów, produktów i rezultatów. W przypadku, gdy wskazane zostaną uchybienia w którymkolwiek z tych elementów składających się na proces wdrażania, wymagana jest bardziej pogłębiona ewaluacja. Może ona przyjąć postać pogłębionych wywiadów i grup fokusowych z udziałem kluczowego personelu łańcucha wdrażania, obserwacji kluczowych działań i etnograficznej analizy kontekstów, w których odbywa się proces wdrażania. Metody konsultacyjne, takie, jak technika delficka i grupy nominalnej oraz analiza przypadków krytycznych, mogą również być wymagane w celu ustalenia, w jakich przypadkach występują konsensus i nieporozumienia wśród kluczowego personelu w kwestii, dlaczego nie są osiągane przewidywane wyniki i rezultaty oraz które mechanizmy w procesie realizacji działają, a które nie. O'Connor (2008) sugeruje, że badanie terenowe i analiza wykonywana na potrzeby ewaluacji wdrażania powinny obejmować:

- koncentrację na kluczowych osobach w łańcuchu realizacji,
- zadbanie, aby dane spełniały swoją rolę,
- generowanie jasnych zagadnień i wyraźnych hipotez,
- wykorzystanie tych zagadnień i hipotez w procesie prowadzenia ewaluacji,
- zapewnienie braku defensywnego podejścia ankietowanych,
- utrzymywanie koncentracji wywiadów,
- utrzymywanie ścisłego zakresu i precyzyjnej koncentracji,
- zapewnienie odpowiedniej jasności komunikatów w końcowej fazie ewaluacji,
- wbudowanie etapu działań następczych w proces oceny.

W przypadku istnienia dowodów nieskutecznej realizacji, ewaluację można skoncentrować na tym, co okazało się skuteczne w innych porównywalnych obszarach oraz analizie tego, jakie wkłady, mechanizmy, wyniki i zasoby były wymagane w celu osiągnięcia powodzenia we wdrażaniu i realizacji polityki. Ewaluacja wdrażania pokazuje wzajemne powiązania różnych typów ewaluacji i konieczność stosowania różnych metod ewaluacyjnych w celu odpowiedniego odniesienia się do różnych rodzajów problemów, które pojawiają się w łańcuchu realizacji.

## 6. Jakie są koszt, opłacalność i stosunek kosztów do korzyści różnych wariantów polityki?

Dla celów kształtowania polityki ważne jest, aby wiedzieć nie tylko, które interwencje w ramach polityki są skuteczne i jakie są skuteczne mechanizmy wdrażania, ale również, jakie są względne koszty i korzyści różnych wariantów polityki. W *Zielonej Księdze* Ministerstwa Finansów Wielkiej Brytanii (*The UK Treasury Green Book*) podsumowano zwięźle znaczenie oceny ekonomicznej dla celów kształtowania polityki, stwierdzając że „nie należy przyjmować żadnej polityki, żadnego programu ani projektu bez wcześniejszej odpowiedzi na następujące pytania: 1) czy istnieją lepsze sposoby osiągnięcia tego celu?, 2) czy istnieją lepsze zastosowania dla tych zasobów?" (HMT 2003, s. 1). Pierwsze z tych dwóch pytań wymaga analizy *opłacalności*, w której „porównuje się koszty alternatywnych sposobów osiągania takich samych lub podobnych wyników" (HMT 2003, s. 8), podczas gdy drugie pytanie wymaga *analizy kosztów i korzyści* i szacuje najbardziej optymalny stosunek efektów do kosztów różnych wariantów polityki. Analiza kosztów i korzyści może powodować przeniesienie środków z jednego obszaru polityki, nawet całego departamentu administracji rządowej do innej polityki, innego sektora lub departamentu (np. z wydatków na obronność na edukację lub opiekę zdrowotną). Trzecim rodzajem oceny ekonomicznej jest *analiza kosztów i użyteczności*, która jest „formą oceny ekonomicznej, w której rezultaty alternatywnych procedur lub

programów wyraża się w postaci jednej opartej na użyteczności jednostki miary" (Robinson 1993, s. 859). Dla ekonomistów użyteczność odnosi się do subiektywnych doświadczeń ludzi w styczności z polityką, programem lub projektem, które mogą być różne od bardziej obiektywnych miar rezultatu.

Ocena ekonomiczna i ewaluacja mają za zadanie określić ilościowo „w postaci pieniężnej możliwie wyczerpująco koszty i korzyści propozycji, w tym pozycje, dla których rynek nie zapewnia satysfakcjonującej miary wartości gospodarczej" (HMT 2003, s. 8). Niektóre wartości pieniężne mogą wynikać z działalności rynkowej, jak np. koszty rynku pracy (wynagrodzenia) lub z cen płaconych za towary i usługi na faktycznie istniejących rynkach. Inne wartości pieniężne należy wydedukować, obserwując wybory dokonywane przez ludzi na rynkach pokrewnych lub hipotetycznych. Osiąga się to, szacując skłonność ludzi do płacenia za towary lub usługi na rynku symulowanym lub w hipotetycznej sytuacji. To, ile ludzie są skłonni zapłacić za towar lub za usługę można oszacować na podstawie ich rzeczywistego zachowania na rynku realnym lub symulowanym (tj. ich *ujawnionych* preferencji) lub na podstawie informacji, jakie przekazują na temat tego, ile są gotowi zapłacić (tj. ich *deklarowanych* preferencji).

Podczas szacowania kosztów i korzyści należy wziąć pod uwagę, kto poniesie koszty lub odniesie korzyści. Obejmuje to osoby korzystające z towaru lub usługi, rząd, ogół społeczeństwa (lub dobro publiczne) oraz przyszłe pokolenia. Koszty dla gospodarki mierzy się pod postacią kosztów jednorazowych lub też kosztów ustanowienia oraz kosztów ponoszonych regularnie, jak koszty utrzymania. Korzyści mierzy się również pod kątem tego, czy są odnoszone przejściowo. Koszty, które będą ponoszone w przyszłości muszą być dostosowane do prawdopodobnej inflacji i zmian stawek podatkowych, a korzyści, które zostaną uzyskane w przyszłości muszą zostać pomniejszone o oczekiwaną stopę inflacji (znaną jako stopa dyskontowa). Szacunki kosztów i korzyści polityki, programu lub projektu są zwykle poddawane analizie wrażliwości, w której obliczając różne wartości kosztów i korzyści bierze się pod uwagę założenia dotyczące ryzyka, tendencyjności optymistycznej (przeszacowanie czasu realizacji i niedoszacowanie opóźnień i innych przeszkód) i ogólnych warunków ekonomicznych. Teoria zmiany może pomóc w tej analizie wrażliwości, określając jasno i wyraźnie od początku przygotowywania polityki, jakie są wkłady, działania, mechanizmy, osoby i zasoby, które mogą być potrzebne, jeżeli mają zostać osiągnięte przewidywane wyniki i rezultaty. Pokazuje to ponownie, jak różne zagadnienia, etapy i metody ewaluacji polityki mogą być ze sobą blisko powiązane.

## 7. Jakie są konsekwencje etyczne różnych wariantów polityki?

Zagadnienie to nie jest rutynowo poruszane podczas ewaluacji polityki, ale jest coraz częściej podnoszone przez niektórych analityków i osoby odpowiedzialne za projektowanie polityk. Problem polega na tym, że projektowanie polityki i świadczenie usług publicznych zwykle wymaga kompromisów i wyborów między interesami różnych grup społecznych. Jeżeli, na przykład, tworzenie placówek specjalnej opieki nad wcześniakami odbywa się kosztem usług w zakresie pomocy socjalnej i zdrowotnej dla osób w bardzo zaawansowanym wieku, pokazuje to, że osoby odpowiedzialne za projektowanie polityki, i społeczności, które reprezentują, cenią wczesny etap życia bardziej niż życie osób starszych. Projektowanie polityki w zakresie opieki zdrowotnej obejmuje często dyskusje dotyczące tego, czy osoby, które angażują się z wyboru w ryzykowne zachowania, takie jak palenie tytoniu lub uprawianie sportów ekstremalnych, należy wyłączyć z niektórych rodzajów interwencji, jeżeli zachorują lub odniosą urazy z powodu takich zachowań. W pewnym stopniu decyzje dotyczące takich wyborów i kompromisów mogą opierać się na ocenie ekonomicznej i innych typach ewaluacji polityki, o których była mowa powyżej. Decyzje te są jednak czymś więcej niż jedynie działaniami technicznymi lub technokratycznymi, w zakresie, w jakim ujawniają, jak społeczeństwo ocenia różne grupy społeczne i różne rodzaje zachowania. W skrócie, obejmują rozważania dotyczące wartości społecznych i etyki społecznej.

Niedawne prace w zakresie etyki podejmowania decyzji w dziedzinie opieki zdrowotnej rozwinęły się w szerszą dyscyplinę, która próbuje ewaluować szerszy zakres kwestii politycznych (Hope). Ten typ ewalu-

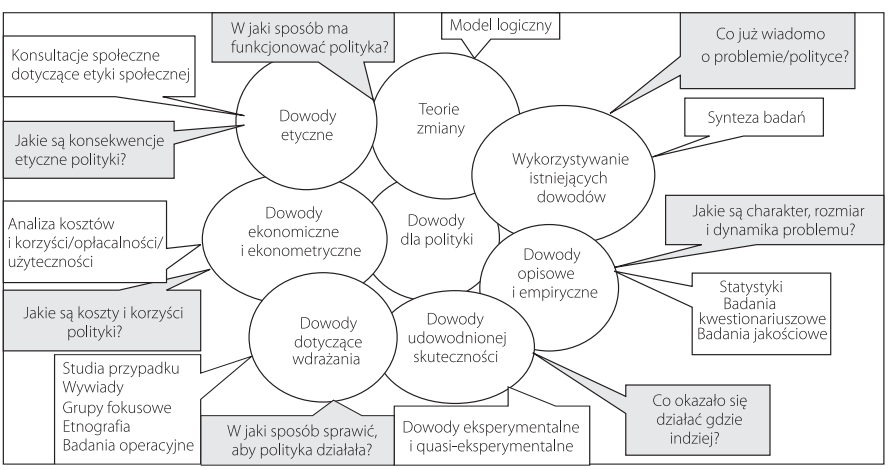
acji wiąże się z metodami partycypacyjnymi, takimi, jak analiza etyki społecznej, konsultacje z interesariuszami, posiedzenia w ratuszu, ławy przysięgłych, w skład których wchodzą obywatele i z innymi metodami konsultacyjnymi, takimi, jak technika delficka i grupy nominalnej, analiza przypadków krytycznych i konsultacje prowadzone drogą elektroniczną. Chociaż to podejście do ewaluacji polityki jest jeszcze w stadium początkowym i przez wielu ewaluatorów jest wykorzystywane w niewielkim stopniu, oferuje cenny wkład w szersze kwestie, które należy uwzględnić w procesie projektowania polityki.

Davies (2004) twierdzi, że istnieje wiele czynników innych niż dowody, które wpływają na proces kształtowania polityki, w tym wartości, przekonania i ideologie, które kierują projektowaniem polityki poprzez procesy polityczne. Osądy polityków, urzędników służby cywilnej, menedżerów polityk i pracowników świadczących usługi na pierwszej linii wpływają na projektowanie polityki i zawierają szereg wartości i etycznych osądów. Stosowanie analizy etyki społecznej i związanych z nią metod wspomnianych powyżej zapewnia zorganizowany i niezależny sposób ewaluacji tych kwestii etycznych. Pozwala również rozważać *a priori* kwestie etyczne, leżące u podstaw teorii zmiany, ustalając tym samym, w jaki sposób polityka ma funkcjonować na solidnej podstawie etycznej i teoretycznej.

## Podsumowanie – proces ewaluacji

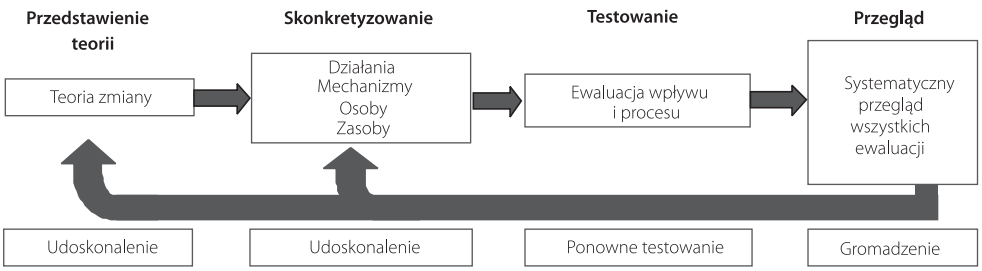
Na Rysunku 5 przedstawiono przegląd siedmiu pytań i związanych z nimi metod ewaluacji, omówionych powyżej. Pokazuje on wzajemne powiązanie wspomnianych siedmiu pytań (istnieją prawdopodobnie inne pytania powstałe w procesie kształtowania polityki) i różnych metod stosowanych w ewaluacji polityki. To z kolei uzasadnia stosowanie w ewaluacji podejścia opartego na metodach mieszanych i unika traktowania w sposób uprzywilejowany konkretnej metody jako najlepszej lub będącej złotym standardem. Adekwatność metody ewaluacji zależy od postawionego pytania lub pytań, których będzie wiele w trakcie przygotowywania, wdrażania i przeglądu polityki. Na Rysunku 5 i w analizie przedstawionej powyżej pokazano również, że zazwyczaj istnieje szereg metod w ramach każdego z typów ewaluacji. Najodpowiedniejszą metodą określenia z jak największą dokładnością i z jak najmniejszym obciążeniem statystycznym tego, który wariant polityki jest najskuteczniejszy pod względem zapewnienia konkretnego rezultatu jest zastosowanie RCT (randomizowanej próby kontrolnej). Jeżeli jednak polityka polega na przydzieleniu korzyści lub usługi na podstawie kwalifikującego kryterium administracyjnego (np. wiek, poziom dochodu, umiejętność czytania itd.), randomizacja nie jest możliwa i odpowiedniejszą metodą byłaby analiza nieciągłości w równaniu regresji lub, być może, analiza różnicy w różnicach.

**Rys. 5.** Wzajemne powiązania różnych metod oceny polityki

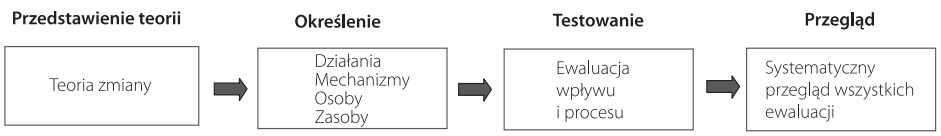


Na Rysunku 6a spróbowano podsumować różne etapy ewaluacji oparte na przedstawianiu teorii, doprecyzowywaniu, testowaniu i prowadzeniu przeglądów. Teoria zmiany wskazuje, co stara się osiągnąć polityka i czego wymaga osiągnięcie zamierzonych rezultatów. Działania, mechanizmy, ludzie i zasoby, które są wymagane w celu osiągnięcia zamierzonych rezultatów potrzebują bardziej szczegółowego określenia w świetle dowodów z systematycznych przeglądów i innych metod analizy teorii zmiany. Te bardziej szczegółowe specyfikacje z kolei należy następnie poddać ocenie wpływu, z wykorzystaniem metod eksperymentalnych lub quasi-eksperymentalnych, korzystając z danych pochodzących z badań kwestionariuszowych, spisów, danych administracyjnych i jakościowych. Może być również wymagana ewaluacja procesu dla różnych sposobów wdrażania z wykorzystaniem danych z monitoringu, jak również studiów przypadku, wywiadów pogłębionych, grup fokusowych, analiz etnograficznych i analiz operacyjnych. Warianty wdrażania mogą również wymagać ewaluacji empirycznej z wykorzystaniem metod eksperymentalnych i quasi-eksperymentalnych. Analizę kosztów, kosztów i korzyści, opłacalności oraz kosztów i użyteczności należy stosować we właściwy sposób, jako część etapu ewaluacji polegającego na „testowaniu". Należy również uwzględnić kwestie etyczne związane z proponowanymi wyborami i kompromisami. Ustalenia każdej z metod ewaluacji można następnie wykorzystać w celu ulepszenia zarówno ogólnej teorii zmiany, jak i konkretnych wkładów, mechanizmów, osób i zasobów, będących elementami teorii zmiany. Może to wymagać dalszego testowania (lub ponownego testowania) aż do uzyskania wystarczająco solidnych dostępnych dowodów o wewnętrznej i zewnętrznej spójności.

**Rys. 6a.** Proces oceny




**Rys. 6b.** Proces oceny – warunkowe przekazywanie środków pieniężnych



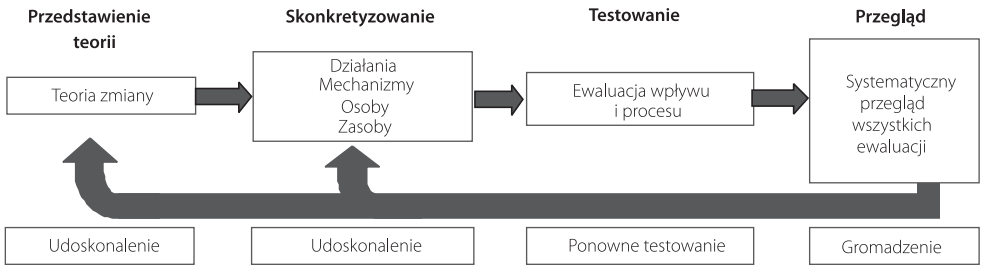

- Warunkowe przekazywanie środków zachęca rodziców do wysyłania dzieci do szkoły
- Zwiększenie liczby zapisów i frekwencji
- Poprawa efektów uczenia się

- Egzekwowanie wypełniania obowiązku szkolnego
- Zmiana zachowania rodziców i dzieci
- Zmiana względnych kosztów i korzyści nauki szkolnej kontra inne wykorzystanie czasu dzieci

- >100 ewaluacji wpływu
- Wiele ewaluacji procesu
- Testowanie warunkowego przekazywania środków w Ameryce Łacińskiej, Azji, Afryce

- Petrosino i in. (2012) systematyczny przegląd 23 badań
- Kształcenie wysokiej jakości dla wszystkich dzieci
- Warunkowe przekazywanie środków zwiększa liczbę zapisów i frekwencję w szkole
- Brak ogólnego wpływu na efekty uczenia się

**Rysunek 6c.** Proces oceny – ze zmianami




- Kwestie efektu deadweight
- Poziomy płatności
- Odbiorcy płatności
- Nagradzanie wyników

- Mierzenie efektów uczenia się
- Potrzeba interwencji po stronie podaży
- Jakość nauczycieli czynnikiem pośredniczącym
- Etapowe przekazywanie środków

- Potrzeba ponownego testowania udoskonaleń/mechanizmów
- Skoncentrowanie na specyfice kontekstowej

- Tworzenie bazy dowodowej na przestrzeni czasu
- Ustalenie możliwości uogólnienia/specyfiki kontekstu
- Unikanie błędów typu I/II

Systematyczne przeglądy odgrywają bardzo ważną rolę w ewaluacji polityki, identyfikując jak najwszechstronniej wszystkie badania dotyczące danej kwestii polityki i poddając je ocenie w zakresie związku z tematyką oraz jakości. Oddzielając ewaluacje wyższej jakości od tych o niższej jakości, systematyczne przeglądy są w stanie zidentyfikować najlepsze dostępne dowody i dokonać ich syntezy, „aby odkryć ich zgodność i wyjaśnić różnice pomiędzy badaniami sprawiającymi wrażenie podobnych" (Cooper i Hedges 1994, s. 4). W ten sposób można gromadzić na przestrzeni czasu dowody najwyższej jakości i stworzyć solidną bazę dowodową na potrzeby projektowania polityki i skutecznego świadczenia usług publicznych.

Na Rysunku 6b przedstawiono schemat takiego procesu ewaluacji na przykładzie warunkowego przekazywania środków pieniężnych jako działania mającego na celu poprawienie wyników kształcenia. Warunkowe przekazywanie środków „oferuje regularne płatności gotówkowe osobom lub rodzinom pod warunkiem pewnego zachowania, takiego jak zapisanie dziecka do szkoły, regularne uczęszczanie przez nie do szkoły i niekiedy wymóg dotyczący wyników osiąganych w szkole" (White, Krishneratne i Hombredos 2012, s. 6). Ogólna teoria zmiany leżąca u podstaw warunkowego przekazywania środków w zakresie edukacji ma celu zachęcenie rodziców, aby wysyłali dzieci do szkoły zamiast pozwolić im spędzać ten czas na innych czynnościach (np. pomaganie w domu, przedsiębiorstwie rodzinnym lub gospodarstwie rolnym itd.). Wysłanie dzieci do szkoły i zatrzymywanie ich tam zwiększa liczbę zapisów i frekwencję, co z kolei poprawia efekty uczenia się.

Ta ogólna teoria wymaga większej specyfikacji wkładów, mechanizmów i wyników, które są wymagane dla poprawy osiągnięć w zakresie liczby zapisów, frekwencji i efektów uczenia się. Głównym wkładem są pieniądze (środki pieniężne), które przekazuje się warunkowo rodzicom (mechanizm), tym samym zachęcając rodziców do wysyłania dzieci do szkoły (mechanizm) i zmieniając relatywne koszty i korzyści nauki szkolnej w stosunku do innych sposobów wykorzystania czasu dzieci (mechanizm), co z kolei zmieni zachowanie rodziców i dzieci (wynik). Ciągle na płaszczyźnie teoretycznej, zwiększenie liczby zapisów do szkoły i frekwencji doprowadzi z kolei do postępów dzieci w szkole (wynik) i poprawy wyników kształcenia (rezultat).

Teoria zmiany leżąca u podstaw warunkowego przekazywania środków w edukacji została zbadana empirycznie w wielu eksperymentalnych i quasi-eksperymentalnych ewaluacjach wpływu oraz ewaluacjach procesu w Ameryce Łacińskiej i Centralnej, Azji i Afryce (Fitzbein i Schady 2010). Wyniki poje-

dynczych ewaluacji wpływu są na ogół korzystne w zakresie zmiany zachowania rodziców i dzieci oraz zwiększenia liczby zapisów i frekwencji w szkole. W ramach systematycznego przeglądu przeprowadzonego przez Petrosino i in. (2012) zidentyfikowano 23 eksperymentalne i quasi-eksperymentalne ewaluacje wpływu warunkowego przekazywania środków w kształceniu podstawowym, które spełniały zakres przeglądu i kryteria oceny jakości. Podczas dalszej analizy tych dwudziestu trzech ocen przeprowadzonej przez White'a, Krishneratne'a i Hombradosa (2012) potwierdzono, że warunkowe przekazywanie środków zwiększyło liczbę zapisów do szkoły i poprawiło frekwencję, nie stwierdzono jednak żadnego ogólnego wpływu na efekty uczenia się.

Na rysunku 6c wskazano niektóre z cech teorii zmiany leżącej u podstaw warunkowego przekazywania środków i wyników kształcenia, które być może trzeba będzie udoskonalić i dalej zbadać, w celu określenia, jakie warunki powinny zaistnieć, aby uzyskać poprawę wyników kształcenia. Po pierwsze, wydaje się, że niewiele sensu ma przekazywanie środków pieniężnych rodzinom, w których poziom frekwencji szkolnej dzieci jest i tak wysoki. W takich okolicznościach warunkowe przekazywanie środków wiąże się ze znacznym efektem deadweight. White, Krishnaratne i Hombrados zauważają, że w Kolumbii „w związku z faktem, że praktycznie wszystkie dzieci i tak uczęszczają do szkoły podstawowej, płatności na rzecz dzieci w wieku szkoły podstawowej odpowiada raczej bezwarunkowemu przekazywaniu środków, co rodzi pytanie, czy nie skoncentrować tych środków na wyższych poziomach kształcenia" (White, Krishnaratne i Hombrados 2012, s. 7). Po drugie, poziomy płatności dla gospodarstw domowych mogą być zbyt niskie, aby wywrzeć znaczący wpływ na koszty zapisów dzieci do szkoły, tym samym czyniąc je nieskutecznymi. Po trzecie, odbiorca przekazywanych środków może być istotnym czynnikiem określenia, czy osiągnięto wpływ w zakresie zapisów do szkoły i frekwencji. Na ogół warunkowe przekazywanie środków odbywa się na ręce matek dzieci, co może mieć ograniczony wpływ w rodzinach, w których ważne decyzje dotyczące wydatków gospodarstwa podejmowane są przez innych członków rodziny. W przypadku starszych dzieci skuteczniejsze może być wypłacanie środków dziecku a nie rodzicowi, aby przeznaczyło zyski na potencjalne korzyści z kształcenia średniego. Po czwarte, bardziej skuteczne może być nagradzanie wyników ucznia zamiast nagradzania samego zapisu do szkoły lub frekwencji.

Można zasugerować dalsze doprecyzowanie teorii zmiany, opierając się na dowodach z istniejących systematycznych przeglądów. Jeżeli celem warunkowego przekazywania środków jest poprawa wyników uczniów, konieczne jest oczywiście, aby dane dotyczące efektów uczenia się zbierano wraz z pomiarami liczby zapisów do szkoły i frekwencji. White, Krishnaratne i Hombrados (2012) zasugerowali również, że interwencje po stronie podaży mogą być potrzebne jako mechanizmy pośrednie, aby warunkowe przekazywanie środków było skuteczne w poprawianiu wyników uczniów. Mogą one obejmować środki mające na celu poprawę i zapewnienie jakości nauczania, jak również dostęp do odpowiednich materiałów dydaktycznych i zasobów wykorzystywanych w procesie uczenia się. Potrzeba ustalenia skutków warunkowego przekazywania środków dla efektów uczenia się, a nie tylko zapisów do szkoły i frekwencji, sugeruje ponadto, że etapowe przekazywanie pieniędzy może być konieczne, aby zapewnić nagradzanie za początkowe i końcowe wyniki.

---

**Dr Philip Davies** jest szefem londyńskiego biura 3ie (Międzynarodowej Inicjatywy na Rzecz Oceny Oddziaływania). Jest odpowiedzialny za Program Systematic Review oraz reprezentuje 3ie w Europie, na Bliskim Wschodzie i w Afryce. Zanim zaczął pracować dla 3ie, dr Davies pełnił funkcję Dyrektora Wykonawczego w Oxford Evidentia, firmie konsultingowej, która specjalizuje się w analizie polityk publicznych, monitoringu i ewaluacji oraz transferze wiedzy. Od 2000 do 2007 dr Davies pracował jako starszy urzędnik w Kancelarii Rady Ministrów oraz w Ministerstwie Skarbu, gdzie był odpowiedzialny za analizy i ewaluacje polityk. Wcześniej zatrudniony był jako wykładowca w zakresie nauk społecznych i politycznych na Uniwersytecie

w Oxfordzie, wykładał również na Uniwersytecie w Aberdeen oraz na University of California w San Diego. Dr Davies ma bogate doświadczenie w dziedzinach: zdrowia i opieki zdrowotnej, edukacji, opieki społecznej, przestępczości i wymiaru sprawiedliwości oraz rozwoju międzynarodowego.

## Bibliografia


- *Bickman L., Using program theory in evaluation. [w:] New Directions for Program Evaluation, 1987, 33, s. 5-18.*
- *Campbell D.T., i Russo M.J., Social Experimentation, Thousand Oaks, Sage Publications, 1998.*
- *Chen H.T., Theory-Driven Evaluations. Newbury Park CA: Sage Publications, 1994.*
- *Chen H.T., Practical Program Evaluation: Assessing and Improving Planning, Implementation, and Effectiveness. Thousands Oaks: Sage Publications, 2004.*
- *Chen H.T., Theory-Driven Evaluation; [w:] Mathison, S. (wyd.) Encyclopedia of evaluation, Thousand Oaks, Calif.; Londyn: SAGE Publications, 2005, s. 415-419.*
- *Chen H.-T. i Rossi, P.H., Evaluating with sense: The theory-driven approach. „Evaluation Review", 7, 1983, s. 283-302.*
- *Connell J.P i Kubisch A.C., Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects, and Problems, 1998.*
- *HMT, The Green Book: Appraisal and Evaluation in Central Government, Londyn, TSO, 2003.*
- *HMT, The Magenta Book: Guidance for Evaluation, Londyn, Her Majesty's Treasury, 2007.*
- *McLennan D., Working with Microdata as a Source of Evidence, Oxford Institute of Social Policy, Department of Social Policy and Intervention, University of Oxford, 2012.*
- *Noble M**.,** Cheung S Y., Smith G., i Smith T., Using Census Data to predict Income Support Dependency, „Policy and Politics" vol. 23, nr 4,1995.*
- *Noble M., McLennan D., Wilkinson K., Whitworth A., Barnes H. i Dibben C., The English Indices of Deprivation 2007. Londyn: Department for Communities and Local Government, 2008.*
- *Noble M., Barnes H., Wright G., McLennan D., Avenell D., Whitworth A., Roberts B., The South African Index of Multiple Deprivation 2001 at Datazone Level. Pretoria: Department of Social Development, 2009.*
- *O'Connor T., How The Prime Minister Monitors Performance and Assesses Delivery, Presentation to GORS Induction, Tony O'Connor CBE, Chief Operational Research Analyst. Prime Minister's Delivery Unit, 8 maja 2008.*
- *Olejniczak K., Theory Driven Evaluation: tracing links between assumptions and effects, prezentacja na VI Europejskiej Konferencji Ewaluacyjnej Polityki Spójności, Warszawa, 30 listopada 2009.*
- *Park A., Clery E., Phillips, British Social Attitudes 29th Report, Londyn, Sage Publications, 2012.*
- *Patton M.Q., Utilization-focused evaluation. 4th edition. Los Angeles, Londyn: Sage Publications, 2008.*
- *Pawson R., Evidence-based Policy: In Search of a Method. „Evaluation", 8(2), 2002, s. 157-181.*
- *Petrosino A., Rogers P., Huebner T. i Hacsi T., Program Theory in Evaluation: Challenges and Opportunities, „New Directions for Evaluation", 2000.*
- *Ravallion M., The Mystery of Vanishing Benefits: Ms Speedy Analyst's Introduction to Evaluation, Waszyngton, Bank Światowy, 1999.*
- *Ravallion, M., Evaluating Anti-Poverty Programs, „Policy Research Working Papers Series" Nr 3625, Waszyngton, Bank Światowy, 2005.*
- *Robinson R, Cost Utility Analysis, „British Medical Journal", 307, 1993, s. 859-862.*
- *Shadish W.R., Cook T.D., i Campbell D.T., Experimental and Quasi-Experimental Designs for Generalised Causal Inference, Belmont, California, Wadsworth Cengage Learning, 2002.*

*Daniel Fujiwara*

# Wykorzystywanie ewaluacji wpływu do podejmowania decyzji o polityce: Od ewaluacji do wyceny

## 1. Wprowadzenie

Ewaluacja wpływu służy *ostatecznie* doskonaleniu procesu podejmowania decyzji dotyczących polityk. W zasadzie ewaluacja wpływu dostarcza szacunków wpływu polityki na rezultaty będące przedmiotem zainteresowania. Wyniki mogą na przykład wskazywać, że program szkoleń pracowniczych doprowadził do 20% wzrostu płac lub że dzięki nowej kampanii informacyjnej występowanie danej choroby spadło o 10%. Załóżmy na moment, że są to solidne szacunki (powiedzmy, że interwencje poddano badaniom randomizowanym). Pojawia się wówczas pytanie: jak wykorzystać te szacunki do podejmowania decyzji dotyczących polityk? Czasami ocena dotyczy pojedynczej interwencji czy szeregu interwencji zmierzających do określonego rezultatu bądź kilku różnych rezultatów. Może się to odbywać z punktu widzenia jednego wydziału czy ministerstwa lub z perspektywy powiedzmy całego rządu, który musi zdecydować o tym, jak wydać środki publiczne na poszczególne interwencje, z których każda przynosi inne rezultaty.

W takich okolicznościach ewaluacja wpływu jest niezbędnym elementem oceny **wartości** danej polityki bądź polityk. Wniosek, jaki można wyciągnąć z samej ewaluacji wpływu dotyczy **skuteczności** polityki, ale do oceny jej wartości potrzebujemy czegoś więcej, co pozwoli ostatecznie zdecydować, czy daną politykę warto wprowadzać w życie, czy nie. Innymi słowy, odwołując się do powyższego przykładu, 20% wzrost płac wskazuje na to, że szkolenie pracownicze było skuteczne, ale czy taki wpływ można określić jako ‚dobry'? Czy jest na tyle wartościowy, aby zmotywować decydentów do prowadzenia programów szkoleń pracowników na szerszą skalę?

Aby poznać wartość polityki, musimy mieć informacje na temat skuteczności, *ale także* koszty i korzyści z nią związane. Przy takich założeniach polityka, dla której korzyści przewyższają koszty, byłaby warta realizacji, podczas gdy polityka, dla której korzyści netto są większe niż w przypadku innych polityk, powinna być traktowana przez decydentów priorytetowo.

W większości krajów OECD zaleca się, aby przy ocenie wartości polityki korzystać z **analizy kosztów i korzyści** (ang. Cost Benefit Anlysis, CBA), która mierzy koszty i korzyści polityki pod kątem zmian w zakresie dobrobytu społeczeństwa. Zmiany w zakresie dobrobytu materialnego społeczeństwa są wyrażane w ujęciu pieniężnym, aby koszty i korzyści były w CBA wyrażane w tych samych porównywalnych jednostkach. Ewaluacja wpływu (tj. ocena efektów przyczynowo-skutkowych polityki) stanowi istotny element CBA, ponieważ musimy znać rezultaty polityki, jednak w procesie oceny pojawia się dodatkowy aspekt polegający na przypisaniu rezultatom polityki wartości pieniężnych. Oczywiście CBA nie jest jedyną metodą oceny polityki. To właśnie ze względu na założenia konsekwencjalistyczne największe znaczenie przypisywane jest dobrobytowi materialnemu. Istnieją jednak mocne argumenty za tym, aby w procesie tworzenia polityk brać pod uwagę także (jeżeli nie wyłącznie) prawa i obowiązki – innymi słowy za deontologicznym podejściem do tworzenia polityk. Chociaż takie prawa, jak wolność polityczna, prawo do nauki, itp. można by włączyć do konsewkencjalistycznych założeń CBA, rzadko tak się dzieje. Istnieją też inne zarzuty skierowane przeciwko CBA, zarówno filozoficzne (normatywne), jak i techniczne. Wiele z nich dotyczy także

innych technik oceny polityki, takich, jak **analiza efektywności kosztowej** (ang. Cost-Effectiveness Analysis, CEA) i **analiza kosztów i użyteczności** (ang. Cost-Utility Analysis, CUA).

Szczegółowe omówienie tych kwestii nie jest jednak celem niniejszej pracy. Ze względu na popularność i przydatność CBA przy tworzeniu polityk, traktuję ją jak ustalone narzędzie oceny polityki i omawiam jedno z głównych wyzwań wobec metodologii wyceny wyników polityki. Chodzi tu o sposób pomiaru zmian w poziomie dobrobytu materialnego zachodzących dzięki interwencjom w ramach polityki. Ponieważ organizacje sektora publicznego coraz częściej korzystają z ewaluacji wpływu przy podejmowaniu decyzji o polityce, będą musiały także przyjrzeć się temu, na czym im ostatecznie zależy, czy mówiąc językiem ekonomicznym, co chciałyby „zmaksymalizować". Ekonomiści tradycyjnie wykorzystywali zaspokojenie preferencji ludzi do oceny poziomu dobrobytu materialnego i wartości przypisywanych wynikom polityki na potrzeby CBA, dlatego też możemy powiedzieć, że zadaniem polityki tradycyjnie było maksymalizowanie zaspokojenia preferencji ludzi – co ekonomiści określają terminem „użyteczność". Coraz obszerniejsza literatura z zakresu ekonomi behawioralnej, psychologii i neuronauki (a także jej pochodnej neuroekonomii) kwestionuje rolę, jaką w przypadku polityk publicznych odgrywają preferencje ludzi. Rządy krajów OECD w coraz większym stopniu polegają obecnie na pomiarach subiektywnego poczucia dobrobytu (raporty sporządzane przez ludzi na temat ich dobrobytu) przy podejmowaniu decyzji dotyczących polityki, a rzadziej (lub dodatkowo) na zaspokajaniu preferencji.

W niniejszym tekście dokonuję przeglądu koncepcji analizy kosztów i korzyści, dobrobytu mierzonego za pomocą zaspokajania preferencji i subiektywnego poczucia dobrobytu, omawiam główne wady i zalety każdej z miar i pokazuję, w jaki sposób mogą być wykorzystywane przy CBA na potrzeby podejmowania decyzji dotyczących polityki. Są to kwestie i pytania, z którymi będą musiały się zmierzyć organizacje sektora publicznego z uwagi na to, że w coraz większym stopniu wykorzystują ewaluację wpływu przy podejmowaniu decyzji dotyczących polityki.

## 2. Ocena wartości polityki

W centrum techniki oceny polityki, jaką jest CBA leży pojęcie **dobrobytu społecznego**, miara jakości życia jednostek tworzących społeczeństwo. W zasadzie, ramy CBA pozwalają nam ocenić, które polityki przyczynią się do maksymalizacji dobrobytu społecznego, czy subiektywnego poczucia dobrobytu społecznego[1]. Ramy ocen polityk, np. CBA, są **welfarystyczne**. Teorie welfarystyczne głoszą, że istotny jest *wyłącznie* dobrobyt materialny jednostki. Warto zauważyć, że są to podstawowe *normatywne* założenia dotyczące roli rządu i polityki publicznej. Politykę ocenia się przede wszystkim biorąc pod uwagę dobrobyt ponieważ zakłada się, że dobrobyt jest ostatecznym rzeczywistym dobrem w takim sensie, że nasza dbałość i starania dotyczą tylko tego, co może mieć ostateczny wpływ na nasz dobrobyt materialny.Wyjaśniając tę kwestię dalej, nie chcę przez to powiedzieć, że takie wyniki jak poprawa zdrowia, lepsza oświata, większa wydajność gospodarcza, itp. nie są ważne, chodzi raczej o to, że wyniki te są ważne, z tego tylko powodu, iż mają wypływ na dobrobyt jednostek tworzących społeczeństwo. Z kolei, w tych ramach logicznie rzecz ujmując, na polityce publicznej spoczywa *moralny obowiązek* dążenia do maksymalizacji ogólnie pojętego dobrobytu społeczeństwa (Bentham 1983) i od tego zależy sposób pomiaru skuteczności interwencji w ramach polityki, powinniśmy starać się ocenić, w jaki sposób polityka wpływa na ogólnie pojęte subiektywne poczucie dobrobytu w społeczeństwie. Taki właśnie pogląd stanowi podstawy teoretyczne ekonomii dobrobytu i standardowe podejście do oceny polityk w Wielkiej Brytanii, opisane w Zielonej



[1] W niniejszym tekście terminy dobrobyt materialny i subiektywne poczucie dobrobytu stosowane są zamiennie.

Księdze stanowiącej podręcznik ewaluacji polityki (HM Treasury 2011, s. 57) i wielu krajach OECD, takich jak Australia (Departament Finansów i Administracji)[2], Kanada (Departament Skarbu)[3], Nowa Zelandia (Departament Skarbu)[4].

Fakt, że polityka wywiera wpływ na różne jednostki i grupy tworzące społeczeństwo jest operacjonalizowany za pomocą **funkcji dobrobytu społecznego** (ang. *Social Welfare Function* SWF), która po prostu sumuje dobrobyt. Typowa SWF przedstawia się następująco:

$$(1) \qquad SW = a_1 w_1 SWF = w_1 + a_2 w_2 w_2 + a_3 w_3 w_3 + ... + a_N w_N w_N$$

Gdzie SW to ogólny dobrobyt, *w* to dobrobyt jednostki a *1....N* w indeksie dolnym oznaczają poszczególnych członków społeczeństwa. $a_i$ jest wagą przypisaną dobrobytowi poszczególnych jednostek. Dzięki SWF wpływ różnych polityk można odnieść do ogólnego dobrobytu. Ważnym założeniem przyjętym w standardowej analizie kosztów i korzyści jest to, że dobrobyt poszczególnych jednostek jest porównywalny i dla każdej jednostki ma takie samo znaczenie. Określa to termin **utalitarystyczna funkcja dobrobytu społecznego**. Możliwe są także nie-utalitarystyczne SWF, w których subiektywne poczucie dobrobytu odbierane przez niektóre jednostki, czy grupy, jest z założenia ważniejsze, jednak rzadko wykorzystuje się je na potrzeby CBA, chociaż mogą się okazać słuszne i przydatne na etapie podejmowania decyzji na szczeblu politycznym. Innymi słowy, politycy mogą odrzucić wyniki i rekomendacje CBA i skłaniać się ku konkretnej polityce, ponieważ przynosi ona korzyści określonej grupie społeczno-ekonomicznej, co pozwala domniemywać, że z założenia dobro tej grupy jest ważniejsze. Analityk nie ma wpływu na te kwestie, które w dużej mierze zależą od aktualnego kontekstu politycznego i społecznego.

W przypadku każdej typowej polityki, SWF będzie składała się z wpływów pozytywnych i negatywnych, tj. danej polityki może być $\Delta w_i > 0$ jak również $\Delta w_j \Delta w_{-i} < 0$. W tym przypadku *i* to grupa(y), które na polityce zyskały a *j* to grupa, która odczuła jej koszty. Ta druga to zwykle grupa finansująca interwencję. Jeżeli $\Delta SW > 0$, oznacza to, że polityka jest zarówno skuteczna jak i opłacalna. Jeżeli $\Delta SW_{(polityka\ A)} \Delta SWF_{(polityka\ A)} > \Delta SW_{(polityka\ B)} \Delta SWF_{(polityka\ B)} > 0$, oznacza to, że zarówno polityka A jak i B są opłacalne, ale priorytetowo należy potraktować politykę A.

Kluczowym aspektem CBA jest definicja dobrobytu materialnego (*w*) w równaniu (1). CBA ocenia zmiany w zakresie indywidualnego i ostatecznego społecznego dobrobytu definiując miarę dobrobytu a następnie przekształcając wszystkie elementy mające wpływ na dobrobyt na porównywalną miarę pieniężną. Dzięki temu zarówno finansowe koszty ponoszone przez grupę *j*, jak i niefinansowe (ale także finansowe) korzyści osiągane przez grupę *i* można porównywać. Przy założeniu interpersonalnych porównań dobrobytu można następnie zebrać wszystkie wyniki, aby oszacować *ΔSWF*, którą wyraża się w ujęciu pieniężnym. O wartości polityki świadczy więc stosunek korzyści do kosztów, który w przypadku społecznych korzyści netto z polityki wynosi > 1.

Oczywiście pojawiają się też głosy krytyki skierowane przeciwko założeniom utylitarnej SWF i koncepcji interpersonalnych porównań dobrobytu. Z tego powodu istnieje silna grupa przeciwników CBA, jednak prawdopodobnie CBA jest wciąż najchętniej wykorzystywana przez rządy do oceny polityk. Zagadnienie, na którym się tu skupimy to, jak definiowany jest dobrobyt i jakie są tego skutki.

[2] http://www.finance.gov.au/publications/finance-circulars/2006/docs/Handbook_of_CB_analysis.pdf
[3] http://www.tbs-sct.gc.ca/rtrap-parfa/analys/analys-eng.pdf
[4] http://www.treasury.govt.nz/publications/guidance/planning/costbenefitanalysis/primer/cba-primer-v12.pdf

## 3. Analiza kosztów i korzyści

Najważniejsze dla CBA jest pojęcie dobrobytu, a każda analiza składa się z dwóch ważnych etapów metodologicznych:

*i. Ewaluacji wpływów przyczynowo-skutkowych polityki.*

*ii. Wyceny i zsumowania tych wpływów polityki.*

Ewaluacja wpływu jest najważniejszym elementem CBA ponieważ stanowi (1) etap procesu metodologicznego. Jeżeli chodzi o ścisłość, przedmiotem zainteresowania powinien być jedynie pomiar wpływów na wyniki, które mają znaczenie dla dobrobytu społecznego na etapie (1). „Złoty standard" przy ewaluacji wpływu to badania randomizowane (lub naturalne eksperymenty terenowe, gdzie na przykład efekt Hawthorna, efekt Johna Henry'ego, czy efekt reaktywny mogą wykrzywiać wyniki). Jednak w celu zbadania wpływu polityki w CBA wykorzystuje się także techniki quasi-eksperymentalne. W niniejszym tekście wychodzimy z założenia, że solidne efekty przyczynowo-skutkowe interwencji zostały już zmierzone; innymi słowy, że z sukcesem zakończono etap (1). Powstaje pytanie o to, jak można wykorzystać informacje zdobyte dzięki ewaluacji wpływu, aby kierować polityką z pomocą CBA? Aby tego dokonać trzeba zmierzyć dobrobyt społeczny i wyrazić zmiany w ujęciu pieniężnym.

### 3.1. Wycena w analizie kosztów i korzyści

Wycena w CBA polega na przypisaniu wartości pieniężnych zmianom w poziomie dobrobytu jednostki i dobrobytu społecznego, jakie nastąpiły dzięki polityce. Bazuje to na teorii wartości opracowanej przez Hicksa i Allena (1934) oraz powszechnie obecnie stosowanych terminów: zmiana kompensacyjna, zmiana ekwiwalentna i nadwyżka. Nadwyżka kompensacyjna (ang. Compensating Surplus, CS) to kwota, wypłacona lub otrzymana, dzięki której konsument zachowuje pierwotny poziom dobrobytu, po zmianie dotyczącej (poziomu) dobra. Nadwyżka ekwiwalentna (ang. Equivalent Surplus, ES) to kwota, która powinna zostać wypłacona lub otrzymana i która zmieni status konsumenta w zakresie dobrobytu w przypadku braku zmiany dotyczącej (poziomu) dobra[5]. W tym sensie monetyzacja polega po prostu na wyrażaniu wpływu na dobrobyt za pomocą ekwiwalentnej jednostki pieniężnej. Faktycznie nie istnieją jakiekolwiek z góry ustalone powody czy teoretyczne podstawy dla wykorzystywania w CBA jednostek pieniężnych, ma to jednak sens ponieważ koszty programów są zwykle z założenia prezentowane w ujęciu pieniężnym.

Teoria leżąca u podstaw wyrażania dobrobytu w jednostkach pieniężnych została szczegółowo opisana w ekonomii dobrobytu (za pomocą CS i ES) i jest standardowo wykorzystywana w CBA. Najnowsze osiągnięcia w dziedzinie CBA dotyczą raczej miary dobrobytu niż metodologii jego wyrażania w jednostkach pieniężnych. To, jakich miar dobrobytu należy używać jest kwestią normatywną. Warto wspomnieć, że w początkowych pionierskich pracach Hicksa i Allena na temat teorii wyceny nie było odniesień do miary dobrobytu. Jednak ich prace zbiegły się chronologicznie z pracami Paula Samuelsona na temat ujawnionych preferencji, za które Samuelson otrzymał nagrodę Nobla. Samuelson wykazał, że przy niewielkiej liczbie (rozsądnych) założeń preferencje ludzi mogą dostarczyć danych na temat ich dobrobytu lub użyteczności przy różnych stanach świata. Ekonomiści dobrze wiedzą, że pod warunkiem, że preferencje ludzi są spójne i trwałe, można je opisać za pomocą stabilnej funkcji użyteczności. Wykazano, że zaspokojenie preferencji zwiększa użyteczność (lub dobrobyt materialny) co można wykorzystać na potrzeby teorii ekonomicznej (w świetle tego Hicks i Allen (1934) dokonali później przemodelowania teorii wartości zgodnie z ujęciem dobrobytu w oparciu o zaspokojenie preferencji). Zaspokajanie preferencji stalo się dominującą



[5] Definicje za Bockstaelem i McConnellem (1980).

miarą dobrobytu w ekonomii, a co za tym idzie w CBA – preferencje traktowano jako miarę dobrobytu, względem której można mierzyć wartości pieniężne zgodnie z CS i ES korzystając z takich technik jak rynki hedoniczne i wycena warunkowa (więcej na temat tych technik w dalszej części tekstu).

Zaspokajanie preferencji jest jedną z trzech szerokich miar dobrobytu zdefiniowanych i używanych przez filozofów (Parfit 1984):

1. *Ujęcie zaspokojenia preferencji,*
2. *Ujęcie stanu umysłu,*
3. *Listy obiektywne.*

Ujęcia zaspokojenia pragnień (lub preferencji) oparte są na założeniu, że wnioski na temat dobrobytu ludzi można wyciągać na podstawie ich wyborów, ponieważ – „najlepsze dla jednostki jest to, co najlepiej spełniłoby wszystkie jej pragnienia" (Parfit, 1984, s.494). Ujęcie zaspokojenia preferencji stworzyło podstawy nowoczesnej ekonomii oraz CBA, jednak polityki publiczne w coraz większym stopniu polegają na ujęciu stanu umysłu i listach obiektywnych. Ujęcia stanów umysłu odnoszą się do subiektywnych doświadczeń ludzi dotyczących ich własnego dobrobytu, które są zwykle mierzone ich własnymi odpowiedziami w kwestionariuszach. Te miary są często nazywane subiektywnym poczuciem dobrobytu (ang. *Subjective Wellbeing* – SWB) i istnieje szeroki zakres pytań dotyczących SWB, w tym pytań o szczęście, emocje, zadowolenie z życia, cel w życiu, smutek, zmartwienia oraz poziom osiągnięcia celów życiowych. Każde z nich mieści się w innej koncepcji teoretycznej dobrobytu. Ujęcie dobrobytu w formie list obiektywnych opiera się na założeniach dotyczących podstawowych potrzeb i praw ludzi (Dolan et al., 2011) a listy takie są często wykorzystywane do mierzenia dobrobytu w kontekście rozwoju (np. wskaźnik rozwoju społecznego ONZ).

Jednym z istotnych nowych osiągnięć w zakresie teorii wyceny i wobec tego bardziej ogólnie w zakresie metodologii CBA było opracowanie podejścia obejmującego **wycenę subiektywnego poczucia dobrobytu.** Ze względu na to, że, obok standardowych wskaźników ekonomicznych, dostępnych jest coraz więcej krajowych zbiorów danych dotyczących SWB, coraz więcej ekonomistów w badaniach stosowanych mierzy dobrobyt korzystając raczej z SWB niż preferencji. Innymi słowy, przeszli oni od ujęcia zaspokojenia preferencji do ujęcia dobrobytu na gruncie stanu umysłu. Takie wnioski można wyciągnąć na podstawie coraz większej liczby artykułów poświęconych badaniom subiektywnego poczucia dobrobytu ukazujących się w czołowych czasopismach ekonomicznych. Jednym z rezultatów jest pojawienie się, na razie skromnego pod względem ilości, piśmiennictwa z zakresu ekonomii, gdzie na potrzeby CBA z zastosowaniem podejścia wyceny subiektywnego poczucia dobrobytu, miary SWB wykorzystywane są do określania wartości dóbr pozarynkowych. Potwierdzają to także zmiany w Wielkiej Brytanii dotyczące oficjalnego wykorzystywania tego podejścia na potrzeby CBA i w procesie tworzenia polityk – Fujiwara i Campbell (2011) są autorami pierwszego dokumentu zawierającego wytyczne na temat wyceny subiektywnego poczucia dobrobytu, także Zielona Księga Skarbu Wielkiej Brytanii została uzupełniona o metodę wyceny subiektywnego poczucia dobrobytu.

Oznacza to tyle, że obecnie teoretycznie można prowadzić CBA z wykorzystaniem zaspokojenia preferencji *lub* SWB jako wybranego sposobu pomiaru dobrobytu, na podstawie którego wpływy polityki można wyrazić w ujęciu pieniężnym[6]. W efekcie po otrzymaniu solidnego szacunku efektu interwencji (dzięki ewaluacji wpływu) decyzję o ewentualnym wprowadzeniu polityki w życie można podejmować na podstawie tego, czy rezultaty polityki zaspokajają preferencje społeczeństwa lub czy wpływają na wzrost SWB jednostek w społeczeństwie. Ważne jest to, że ponieważ preferencje i SWB czerpią z odmiennych pojęć dobrobytu, w wyniku CBA przeprowadzonej z wykorzystaniem ujęcia preferencji można uzyskać inne

[6] W ekonomii podejmowane są także próby wykorzystywania ujęcia dobrobytu w formie list obiektywnych przy ocenie polityk (szczególnie w sektorze zdrowia) (np. zobacz Arnaud, itd.).

wyniki i rekomendacje dotyczące polityki niż w przypadku CBA z wykorzystaniem ujęcia stanu umysłu. Można to uzasadnić tym, że wiele rzeczy w życiu, ważnych z punktu widzenia preferencji ludzi, może nie mieć żadnego wpływu na ich stan umysłu i odwrotnie. Kwestia, na którym ujęciu dobrobytu się opierać stanowi pytanie normatywne, na które nie da się udzielić jednoznacznej odpowiedzi (chociaż wiele przemawia za każdym z możliwych pomiarów). Organizacje sektora publicznego muszą określić, która jednostka dobrobytu powinna stanowić podstawowy komponent CBA i decyzji dotyczących polityk. Szereg krajów OECD zaczęło na szczeblu krajowym zbierać dane o SWB. Najbardziej aktywny w tym zakresie jest rząd Wielkiej Brytanii, gdzie Narodowy Urząd Statystyczny (Office for National Statistics – ONS) – angielski odpowiednik polskiego GUS-u - przeprowadził szerokie konsultacje na temat pomiaru subiektywnego poczucia dobrobytu i zawarł szereg kwestii dotyczących SWB w większości swoich głównych zbiorów danych.

## 3.2. Wycena przez preferencje

W ujęciu zaspokojenia preferencji CS i ES mogą być szacowane na podstawie pośredniej funkcji użyteczności w następujący sposób:

(2) $v(\rho^0,Q^0,M^0)=v(\rho^1,Q^1,M^1-CS)$

(3) $v(\rho^0,Q^0,M^0+ES)=v(\rho^1,Q^1,M^1)$

gdzie $v(.)$ to *pośrednia funkcja użyteczności, M* to pieniądze/dochód a $\rho$ to ceny. Indeksy dolne 0 i 1 odnoszą się odpowiednio do stanu przed i po konsumpcji dobra bądź doświadczeniu dobra *Q*, co zgodnie z przyjętym założeniem ma pozytywny wpływ na użyteczność. Terminy te można wyrazić za pomocą następującej intuicyjnej struktury preferencji wykorzystującej Gotowość do płacenia (ang. *Willingness To Pay* – WTP) oraz Gotowość do przyjęcia rekompensaty (ang. *Willingness to Accept* – WTA).

**Tabela 1.** Nadwyżka kompensacyjna i ekwiwalentna a preferencje

| | **Nadwyżka kompensacyjna (CS)** | **Nadwyżka ekwiwalentna (ES)** |
|---|---|---|
| **Wzrost dobrobytu** | *Gotowość do płacenia (WTP) za zmianę na lepsze* | *Gotowość do przyjęcia rekompensaty (WTA) za doświadczenie zmiany na lepsze* |
| **Spadek dobrobytu** | *WTA za zmianę na gorsze* | *WTP za uniknięcie zmiany na gorsze* |

### 3.2.1 Metody z wykorzystaniem preferencji deklarowanych

W metodach z wykorzystaniem preferencji deklarowanych stosuje się specjalne kwestionariusze pozwalające uzyskać szacunki WTP i WTA dla określonego rezultatu. W metodach wyceny warunkowej respondenci odpowiadający na pytania kwestionariuszowe stykają się z hipotetycznym rynkiem. Zwykle przedstawia się szczegółowy opis dobra, kanał dystrybucji oraz sposób i częstotliwość wnoszenia opłat. Następnie, stawia się pytania pozwalające na wyciągnięcie wniosków na temat WTP i WTA respondenta. Pytania dotyczące wyceny mogą być formułowane na wiele różnych sposobów, m.in. jako pytania otwarte, gry przetargowe, karty płatnicze, procedury uzyskiwania odpowiedzi w oparciu o dychotomiczny wybór. Najważniejszym rezultatem analizy odpowiedzi jest oszacowanie przeciętnej WTP lub WTA dla próby z osób poddanych badaniu ankietowemu.

Natomiast w metodach modelowania wyboru dobra nierynkowe są opisywane według atrybutów, a w celu ujawnienia szacunków, co do ich wartości, respondentom przedstawia się w kwestionariuszach szereg alternatywnych opisów dobra. Opisy alternatywne powstają poprzez różnicowanie poziomów atrybutów

dobra. Zależnie od przyjętej metody modelowania wyboru, respondentów prosi się następnie o sklasyfikowanie (warunkowe klasyfikowanie), dokonanie wyboru (wybory eksperymentalne), dokonanie oceny (warunkowa ocena) lub dokonanie najpierw wyboru a potem oceny (porównywanie parami) zaprezentowanych opisów (Fujiwara i Campbell 2011). W przypadku tych metod, o ile wśród atrybutów znajdzie się także koszt i cena, techniki statystyczne mogą posłużyć do określenia szacunkowego WTP dla atrybutów danego dobra.

### 3.2.2. Metody z wykorzystaniem preferencji ujawnionych

Metody z wykorzystaniem preferencji ujawnionych pozwalają poznać szacunkowe wartości dóbr nierynkowych na podstawie dowodów na zachowania ludzi w obliczu rzeczywistych wyborów. Podstawowym założeniem metody ceny hedonicznej jest na przykład to, że dobra nierynkowe wpływają na ceny dóbr rynkowych na innych prawidłowo funkcjonujących rynkach. Różnice cen na tych rynkach pozwalają oszacować wartości WTP i WTA (Fujiwara i Campbell 2011).

Na potrzeby metod z wykorzystaniem preferencji ujawnionych powszechnie wykorzystuje się rynek mieszkaniowy i rynek pracy, aby dokonać odpowiednio wyceny udogodnień środowiskowych/lokalnych, ryzyka i innych czynników odnoszących się do pracy, oraz rynki turystyczne (metoda kosztów podróży), aby dokonać wyceny miejsc rekreacji i odpoczynku.

Bardziej szczegółowe omówienie metod wyceny w oparciu o preferencje znaleźć można w Champ i in. (2003) oraz Fujiwara i Campbell (2011).

## 3.3. Wycena z wykorzystaniem subiektywnego poczucia dobrobytu

I odwrotnie , zmiany w poziomie dobrobytu określone za pomocą CS i ES można oszacować korzystając z danych SWB. Aby tego dokonać wystarczy przenieść zainteresowanie z dobrobytu mierzonego w oparciu o ujęcie preferencji na pomiar w oparciu o ujęcie stanu umysłu. Można wtedy bezpośrednio obserwować funkcję użyteczności i jej poziomy (krzywe obojętności) oraz oszacować krańcowe stopy substytucji (ang. *Marginal Rates of Substitution, MRS*) pomiędzy dochodem a dobrem nierynkowym, aby uzyskać szacunkową wartość ES lub CS. Na przykład, z faktu, że 20% spadek przestępczości na szczeblu lokalnym powoduje wzrost SWB jednostki o 1 punkt indeksowy a wzrost dochodu gospodarstwa domowego o 5 000 GBP rocznie także powoduje przyrost SWB o 1 punkt procentowy, można wyciągnąć wniosek, że 20% spadek przestępczości ma dla mieszkańców wartość 5 000 GBP rocznie. W praktyce dane dotyczące SWB są zwykle analizowane przy wykorzystaniu modeli ekonometrycznych takich jak:

$$(4) \quad SWB_i = \alpha + \beta_1 Q_i + \beta_2 ln(M_i) + \beta_3 X_i + \varepsilon_i$$

$$(5) \quad SWB\,(Q^0, X^0, M^0) = SWB(Q^1, X^1, M^1 - CS)$$

gdzie $M$ oznacza dochód, $Q$ – oceniany rezultat polityki (dobro nierynkowe) oraz $X$ – wektor innych determinant SWB. Dochód przedstawiony jest w formacie logarytmicznym, aby uchwycić malejącą krańcową użyteczność dochodu.

CS dla próby może być przybliżana ze współczynnika regresji z funkcji SWB takiej jak (4) w następujący sposób (skupiam się na CS, jako że jest to standardowa miara wartości wykorzystywanej w funkcji CBA)[7]:

[7] Por. Fujiwara i Campbell (2011) oraz Fujiwara (2013), aby zapoznać się z pełnym sposobem wyprowadzenia CS w metodzie wyceny dobrobytu.

$$SWB_i = \alpha + \beta_1 Q_i + \beta_2 ln(M_i) + \beta_3 X_i + \varepsilon_i \quad (6)$$

aby otrzymać:

$$CS = M^0 - e\left[ln(M^0) - \frac{(\beta_1 (Q^1 - Q^0)}{\beta_2}\right] \quad (7)$$

gdzie $M^0$ to przeciętny dochód dla próby.

Miarą SWB najpowszechniej wykorzystywaną w literaturze na potrzeby wyceny subiektywnego poczucia dobrobytu jest zadowolenie z życia, gdzie respondentów zwykle prosi się o wskazanie ogólnego poziomu zadowolenia z życia w skali od 1 do 7 lub od 0 do 10, itd. Należy jednak zwrócić uwagę na to, że każdy pomiar dobrobytu z wykorzystaniem ujęcia stanu umysłu, na przykład efekt pozytywny i negatywny, można by wykorzystać zamiast zadowolenia z życia na potrzeby wyceny subiektywnego poczucia dobrobytu (zob. Powdthavee i van der Berg 2011 dla odniesienia do przykładów i porównań wycen subiektywnego poczucia dobrobytu z wykorzystaniem różnych miar SWB). Tendencja do tego, aby na potrzeby wyceny subiektywnego poczucia dobrobytu wybierać raczej zadowolenie z życia jako miarę SWB prawdopodobnie wynika z faktu, że zadawane w tym przypadku pytania dotyczą życia w ogóle, co pozostaje w zgodności z koncepcją użyteczności przyjętą przez ekonomistów (Frey i in. 2009, Frey i Stutzer 2002, MacKerron 2011) oraz że to właśnie pytanie o poczucie dobrobytu najczęściej pada w dużych badaniach ankietowych o zasięgu krajowym i w związku z tym dostępność danych na ten temat jest wysoka.

Należy zauważyć, że jeżeli ludzie nie zaspokajają preferencji jedynie po to, aby zwiększyć swoje zadowolenie z życia (bądź inną miarę przyjętą do wyceny subiektywnego poczucia dobrobytu), wówczas wartości wyceny poczucia dobrobytu i wartości preferencji mogą nie być zgodne. Podsumowując, wartości otrzymanych w wyniku wyceny subiektywnego poczucia dobrobytu nie należy traktować jako jednoznacznych wartości WTP czy WTA, jako że są one wyprowadzane z preferencji. Są one natomiast miarami CS i ES i w tym sensie, są one równie słusznymi miarami wartości w ujęciu pieniężnym i mogą być wykorzystywane na potrzeby CBA (równania (5) – (7) można przeformułować ujmując ES zamiast CS).

Tabela 2 w aneksie pokazuje niektóre wartości otrzymane dzięki zastosowaniu matody wyceny dobrobytu przy wykorzystaniu zadowolenia z życia.

## 4. Różnice w wycenie przy zastosowaniu metody bazującej na subiektywnym poczuciu dobrobytu i metody bazującej na preferencjach

Powyżej dowodziłem, że wartości otrzymane w wyniku wyceny subiektywnego poczucia dobrobytu nie mogą być zgodne z wartościami otrzymanymi w wyniku podejścia opartego na preferencjach. Wynika to z faktu, iż SWB i preferencje to dwa odmienne ujęcia dobrobytu, które opierają się na innych składowych jakości życia ludzi. Zarówno ujęcia z wykorzystaniem zaspokojenia preferencji jaki i stanu umysłu to subiektywne ujęcia poczucia dobrobytu, oznaczające suwerenność na poziomie jednostki; to jednostka decyduje o swoim życiu i o tym, co dla niej ważne. Stoi to w opozycji do ujęcia z wykorzystaniem list obiektywnych, które mierzą dobrobyt jednostki w oparciu o wcześniej ustalony zestaw wskaźników subiektywnego poczucia dobrobytu. W subiektywistycznym ujęciu dobrobytu wartości otrzymane w oparciu o preferencje oraz wartości wynikające z subiektywnego poczucia dobrobytu mogą się różnić, ponieważ ludzie mogą zaspokajać preferencje z wielu różnych powodów, które mogą nie mieć związku z ich stanem umysłu czy SWB.

Zadowolenie z życia, domyślną miarę wykorzystywaną przy wycenie subiektywnego poczucia dobrobytu, można postrzegać jak połączenie wpływu (pozytywnych i negatywnych emocji i uczuć) oraz oceny poznawczej stopnia realizacji aspiracji i celów jednostki (Diener 1984; Kahneman i in. 2006). Odpowiedzi dotyczące zadowolenia z życia będą w pewnym stopniu zawierały sądy retrospektywne na temat życia jednostki oraz jej odczucia dotyczące sytuacji bieżącej (Kahneman i Krueger, 2006). Tak więc, ujęcia wykorzystujące preferencje i stan umysłu, np. zadowolenie z życia, wyraźnie opierają się na różnych aspektach ogólnie pojętego dobrobytu jednostki, co oznacza, że wartości dla określonego dobra uzyskane z zastosowaniem dwóch różnych metodologii wyceny mogą się znacznie różnić. Na przykład, jednym z typowych wniosków jest to, że dzięki przystosowaniu, pogorszenie stanu zdrowia i choroby mogą być postrzegane jako bardziej lub mniej istotne dla określania preferencji i subiektywnego poczucia dobrobytu. Okazało się, że w przypadku kilku chorób, poziom przystosowania jest wysoki i ludzie twierdzą, że są umiarkowanie zadowoleni z życia, natomiast na podstawie preferencji można wnioskować, że oddaliby wiele lat życia by odzyskać dobre zdrowie (Dolan i Kahneman 2008). Dolan (2011) stwierdza, że w przypadku preferencji duże znaczenie ma stan zdrowia fizycznego, podczas gdy dla SWB znacznie ważniejsze od zdrowia fizycznego jest zdrowie psychiczne (zob. również Powdthavee i van der Berg 2011). W innej pracy, Dolan i Metcalfe (2010) stwierdzają, że ludzie zdecydowanie wolą, aby elektrownie wiatrowe nie powstawały w pobliżu ich miejsc zamieszkania, a jednocześnie szybko okazuje się, że z czasem mieszkanie w pobliżu farm wiatrowych ma niewielki wpływ na to, jak oceniają swoje życie, czy na deklarowane subiektywne poczucie dobrobytu. Z takich różnic konceptualnych wynika szereg względnych zalet i wad dwóch metodologii wyceny.

### 4.1. Ocena metod wyceny z wykorzystaniem preferencji

Głównym elementem metod wyceny z wykorzystaniem preferencji jest podejmowanie decyzji; decyzje podejmowane przez jednostki w badaniach ankietowych lub na rynkach imitujących rynki rzeczywiste są obserwowane na potrzeby wyciągania wniosków na temat wartości. O ile preferencje ludzi są racjonalne (tj. spójne) oraz dysponują oni pełnymi/wystarczającymi informacjami na temat danego dobra, wówczas wartości otrzymane na bazie preferencji dostarczą ważnych danych o zmianach w poziomie dobrobytu związanych z danym dobrem.

Poczynając jednak od teorii ograniczonej racjonalności autorstwa Simona (1955) ekonomiści i psychologowie coraz bardziej krytykowali ten racjonalny pogląd na świat. Ekonomiści behawioralni podkreślali rolę percepcji, poznania i uczenia się w procesie podejmowania decyzji. W wyniku tego uznano, że preferencje powstają często w momencie, gdy pojawia się prośba o ich ujawnienie i w związku z tym mogą być uzależnione od kontekstu (Slovic i Lichtenstein 2006). W procesie podejmowania decyzji, ludzie mogą stosować skróty poznawcze (heurystyka), szczególnie gdy problemy, z którymi się stykają są nieznane bądź złożone. Skróty te przyspieszają i ułatwiają jednostkom proces podejmowania decyzji, mogą jednak prowadzić do powstania nieracjonalnych bądź niespójnych wyborów. Kwestie te są dobrze znane i nie zostały tu szczegółowo opisane, ale wystarczy wspomnieć, że mogą doprowadzić do takich problemów jak odwrócenie preferencji czy nieoptymalne wybory (zob. teksty w Slovic i Lichtenstein 2006). Z powodu tych problemów preferencje mogą być słabo, bądź wcale nie powiązane z dobrobytem jednostki, co czyni ich użycie na potrzeby tworzenia polityk problematycznym. Błędy kontekstowe mogą prowadzić do:

- *Efektu zakotwiczenia*. Gdy uzyskane wartości są zakotwiczone w pierwszej wartości zasugerowanej w wyliczeniu badania ankietowego lub wartości, która w danym momencie się wyróżnia (Ariely i in. 2003). Wyliczenia w ankiecie mogą wpływać na WTP i WTA określone przez respondentów.

- *Efekt osadzenia*. Gdy w ankietach dotyczących preferencji deklarowanych ludzie nie zwracając uwagi na zakres (nie są gotowi zapłacić więcej za większą ilość dobra) lub są niewrażliwi na efekty sekwencjonowania (WTP dla dobra zależy od tego, kiedy jest zaprezentowane w ankiecie) (Desvonges i in. 1992, Fujiwara i Campbell 2011).
- *Efekt obciążenia informacyjnego*. Ludzie nie mają wystarczającej ilości informacji na temat danego dobra (Frey et al. 2004a; Frey i Stutzer 2005) i mogą być podatni na błędne informacje (Fujiwara i Campbell 2011) i w związku z tym nie są zdolni do wyrażenia swoich prawdziwych preferencji.

W przypadku technik preferencji deklarowanych możliwe są także błędy związane z badaniami ankietowymi. Należą do nich: brak odpowiedzi, wartości protestacyjne (gdy ludzie deklarują zerowe WTP chociaż cenią sobie dane dobro, ponieważ nie chcą przyporządkować dobru wartości pieniężnej) i błędy strategiczne (gdy zawyżają lub zaniżają wartość jakiegoś dobra aby wpłynąć an politykę). Natomiast w przypadku preferencji ujawnionych konieczne jest efektywne działanie rynków, na podstawie czego zebrać można przydatne informacje o preferencjach. Jednak nie zawsze musi tak być. W wyborach ludzi dotyczących rynku mieszkaniowego nie ujawniają się ich preferencje co do niższych poziomów przestępczości oraz czystych ulic, niezależnie od tego, jak ważne są dla nich te wartości, jeśli rząd wprowadzi program wpływający na te wybory. Implikacje dotyczące zarówno technik preferencji deklarowanych jak i ujawnionych wskazują na to, że w pewnych okolicznościach, preferencje mogą nie być dobrymi wskaźnikami dobrobytu, tzn. wartości określone na ich podstawie mogą być błędne lub mylące. Musimy jednak również zauważyć niektóre zalety metod związanych z preferencjami. Metody preferencji ujawnionych wykorzystują dane dotyczące rzeczywistych wyborów i zachowania. Metody preferencji deklarowanych są w dużym stopniu elastyczne – teoretycznie można stworzyć ankietę w celu oceny każdego typu rezultatu lub dobra i istnieje bogata literatura na temat tej techniki, co pomogło położyć fundamenty pod dobre praktyki. Na przykład dwa numery *Journal of Environmental and Resources Economics* (2005 i 2010) poświęcono metodom zwalczania nieprawidłowości w zakresie preferencji w badaniach preferencji deklarowanych. Jednym z najważniejszych mechanizmów zmniejszenia skali nieprawidłowości w tych badaniach jest uczenie się przez powtórzenia i doświadczenie, co jak pokazuje praktyka, pozwoliło wyeliminować wiele problemów dotyczących metod wykorzystujących preferencje deklarowane.

### 4.2. Ocena wyceny subiektywnego poczucia dobrobytu

Można stwierdzić, że wycena subiektywnego poczucia dobrobytu powstała dzięki krytyce skierowanej przeciwko metodom wyceny z wykorzystaniem preferencji. Jak widzieliśmy, w wycenie subiektywnego poczucia dobrobytu możliwe jest określanie wartości dóbr nierynkowych bez konieczności polegania na preferencjach ludzi. Oznacza to, że można uniknąć takich problemów jak nieprawidłowości w zakresie preferencji, błędów kontekstowych. Nie polegamy na założeniach dotyczących racjonalności (w wycenie subiektywnego poczucia dobrobytu ludzie po prostu wskazują swoje obecne poziomy SWB, a my analizujemy, co wpłynęło na ich SWB), a problemy takie, jak efekt zakotwiczenia czy efekt osadzenia są likwidowane. Ponadto błędy związane z badaniami ankietowymi właściwe dla metod preferencji deklarowanych, jak obciążenie strategiczne, są również likwidowane. Wycena subiektywnego poczucia dobrobytu nie wymaga rynków zastępczych do generowania wartości, jak metody preferencji ujawnionych, dlatego też wycena subiektywnego poczucia dobrobytu ma znacznie szersze zastosowanie niż preferencje ujawnione. Są to niektóre z zalet podejścia opartego o wycenę subiektywnego poczucia dobrobytu, istnieje jednak kilka problemów, które należy wziąć pod uwagę. Ważność wyceny subiektywnego poczucia dobrobytu zależy od ważności miary SWB (zwykle zadowolenie z życia) i w związku z tym musimy zastosować solidne metody statystyczne do oszacowania równań (6) i (7).

Dwa najważniejsze wyzwania dotyczące miar SWB to *ważność* i *spójność* miar. Czyli po prostu: „*czy miary SWB odzwierciedlają faktyczne subiektywne poczucie dobrobytu oraz czy obraz ten jest spójny?*" (Fedderson i in. 2012).

Na ważność miary SWB może wpłynąć szereg czynników. Po pierwsze, ludzie mogą niezbyt dobrze pamiętać to, czego doświadczyli w przeszłości. W eksperymentach udowodniono, że subiektywne poczucie dobrobytu jakie zapamiętują ludzie może być obciążone błędem ze względu na tendencję do przyjmowania zasady „szczytu i końca"; dokonując oceny retrospektywnej ludzie większą wagę przywiązują do szczytu (okresu najintensywniejszego nasilenia) doświadczenia i do okresu końcowego. Mniejszą wagę przywiązują do okresu środkowego. Tak więc, mogą pojawić się rozbieżności pomiędzy rzeczywistymi doświadczeniami ludzi a retrospektywną ewaluacją tych doświadczeń w odpowiedziach udzielanych w badaniach ankietowych (Kahneman et al. 1993; Schwarz 2010).

Po drugie, podobnie jak w przypadku preferencji, dla SWB znaczenie mają także czynniki kontekstowe. Osoby pytane o subiektywne poczucie dobrobytu mogą opierać się na informacjach najłatwiej dostępnych w danym momencie. Na przykład, znaczenie ma kolejność pytań, ponieważ pojawia się prawdopodobieństwo, że respondent w momencie udzielania odpowiedzi na pytanie o zadowolenie z życia przypomni sobie informacje przytoczone przy okazji udzielania odpowiedzi na pytanie poprzednie (Bertrand i Mullainathan, 2001). Udowodniono także, że wpływ na raporty dotyczące zadowolenia z życia ma pogoda, znalezienie dziesięciogroszówki na kserokopiarce, przebywanie w przyjemnym (a nie nieprzyjemnym) pomieszczeniu czy obejrzenie zwycięskiego meczu piłkarskiego ulubionej drużyny (Schwarz i Strack, 1999). O ile takie czynniki prawdopodobnie wpływają na bieżący nastrój, nie powinny mieć widocznego wpływu na rzeczywisty ogólny poziom zadowolenia z życia.

Po trzecie, badane osoby mogą dostosowywać poziom zadowolenia z życia deklarowany w raportach, aby udzielone odpowiedzi były bardziej akceptowalne społecznie. Na przykład, gdy wywiad prowadzi osoba niepełnosprawna, respondenci wykazują tendencję do zaniżania zadowolenia z życia w odpowiedziach, jakich udzielają. Prawdopodobne jest, że pewien wpływ na poziom zadowolenia z życia deklarowany w odpowiedziach ma porównywanie jakości życia w różnych okresach i porównywanie się z innymi w danym okresie (Schwartz i Strach 1999, Dolan i White 2006, Diener i Suh 1997).

Z drugiej strony, istnieje też wiele dowodów na to, że zadowolenie z życia jest ogólnie dobrą miarą subiektywnego poczucia dobrobytu. Pavot i Diener (1993), Eid i Diener (2004), Fujita i Diener (2005) jak również Schimmack i Oishi (2005) dowodzą, że nastrój i kontekst mają na nią ograniczony wpływ. W kilku badaniach wskazuje się na silną korelację pomiędzy odpowiedziami na temat SWB i bardziej obiektywnymi miarami subiektywnego poczucia dobrobytu (trafność zbieżna) a rezultatami, które intuicyjnie *powinny* odnosić się do subiektywnego poczucia dobrobytu (trafność teoretyczna). Na przykład, Sandvik i in. (1993) wskazują na silną korelację pozytywną pomiędzy klasyfikacją subiektywnego poczucia dobrobytu, a takimi emocjami jak śmiech czy marszczenie brwi. Badania pokazują, że uśmiech Duchenne'a (tj. rodzaj uśmiechu, który powoduje napięcie mięśni wokół oczu, co stanowi dowód prawdziwego a nie udawanego rozbawienia) jest skorelowany z SWB (Ekman et al. 1990). Urry et al. (2004) wskazują na korelację pomiędzy deklarowanym w raportach zadowoleniem z życia a aktywnością lewego obszaru kory przedczołowej, który odpowiada za odczuwanie pozytywnych emocji i przyjemności. Ponadto, subiektywne poczucie dobrobytu to dobry prognostyk zdrowia (zob. przykłady w Fujiwara i Campbell 2011). Cohen i in. (2003) jak również Kiecolt-Glaser i in. (2002) wskazują, że ludzie deklarujący wyższy poziom zadowolenia z życia rzadziej zapadają na choroby.

Jeśli chodzi o spójność w zakresie poziomu zadowolenia z życia, Krueger i Schkade (2008) oceniają wiarygodność odpowiedzi na temat zadowolenia z życia udzielonych w pierwszym i kolejnych badań ankietowych. Stwierdzają oni, że korelacja pomiędzy odpowiedziami dotyczącymi zadowolenia z życia była

na poziomie *r = 0,59* dla tej samy próby na przestrzeni czasu. Krueger i Schkade wnioskują, że poziomy wiarygodności uzyskane w ponownych badaniach *„są prawdopodobnie wystarczająco wysokie, aby dostarczyć przydatnych danych szacunkowych na potrzeby … badań naukowych"*.

Odchodząc na chwilę od kwestii pomiarów, oczywiste jest to , że należy także uzyskać solidne dane szacunkowe dotyczące efektu przyczynowo-skutkowego dochodu i dóbr nierynkowych dla SWB. Innymi słowy, potrzebne są obiektywne dane szacunkowe dla $\beta_1$ i $\beta_2$ w równaniu (7), co prowadzi nas do zagadnień pojawiających się we wszystkich ewaluacjach wpływu, a mianowicie samozadowolenia, odwrotnej zależności przyczynowo-skutkowej, błędu pomiaru, itd. Są to kwestie dobrze znane we wnioskowaniu na temat zależności przyczynowo-skutkowych i nie zostały omówione w dalszej części niniejszego tekstu. Kwestie statystyczne wyraźnie związane z podejściem opartym na wycenie subiektywnego poczucia dobrobytu zostały szczegółowo omówione w Fujiwara i Campbell (2011) oraz Fujiwara (2013). W tym miejscu warto zauważyć, że przyczynowo-skutkowy wpływ dochodu ($\beta_2$) nie był wystarczająco dobrze rozumiany w dotychczasowej literaturze na temat subiektywnego poczucia dobrobytu. Często stwierdza się, że współczynnik dochodu jest w zbyt dużym stopniu zmanipulowany, co prowadzi do zawyżania wartości (lub CS) przy zastosowaniu metody bazującej na wycenie subiektywnego poczucia dobrobytu (zob. równanie (7)). Był on niewiarygodnie wysoki dla kilku dóbr nierynkowych, w przypadku których korzystano z podejścia opartego na wycenie subiektywnego poczucia dobrobytu – na przykład dla zatrudnienia (Clark i Oswald), zdrowia (Powdthavee i van der Berg 2011) i relacji społecznych (Powdthavee 2008). Zarówno Levinson (2012) jak i Luechinger (2009) stwierdzili, że w przypadku dóbr środowiskowych wartości otrzymane w wyniku pomiarów subiektywnego poczucia dobrobytu są o kilka rzędów wielkości większe niż w te otrzymane w wyniku badania ujawnionych i deklarowanych preferencji. Jednak obecnie opracowywane są rozwiązania tych technicznych problemów (zob. Fujiwara 2013).

### 4.3. Dyskusja

Oczywiście żadna z metod wyceny nie jest doskonała, co więcej, liczne błędy mogą się pojawić zarówno w metodach wyceny bazujących na preferencjach, jak i na subiektywnym poczuciu dobrobytu. Odnoszą się one do różnych obszarów technicznych, nie ma reguł decydujących o tym, które z obciążeń są poważniejsze, stąd nie można stwierdzić, że jedno podejście jest lepsze od drugiego. Badanie wartości związanych z odnową miast przeprowadzone przez Dolana i Metcalfa było pierwszym badaniem, w którym bezpośrednio porównano wartości wygenerowane z wykorzystaniem metod wyceny bazujących na preferencjach i subiektywnym poczuciu dobrobytu dla tych samych dóbr nierynkowych. Stwierdzili oni, że „regeneracja" miast (remonty ulic, okolicznych terenów i domów) nie miała wpływu na ceny domów w Walii, choć gotowość do płacenia za program regeneracji była na poziomie 250 GBP. Ponadto, jak stwierdzili, regeneracja miast miała pozytywny wpływ na zadowolenie z życia mieszkańców, co odpowiadało wartości ok. 7 000 GBP zgodnie z zastosowanym podejściem bazującym na wycenie subiektywnego poczucia dobrobytu (Dolan i Metcalfe 2008). Podobnie jak w niniejszej pracy, także w innych badaniach stwierdzono, że wartości otrzymane na bazie wyceny subiektywnego poczucia dobrobytu są wyższe niż te wynikające z preferencji (Fujiwara i Campbell). Prawdopodobnie spowodowane jest to problemami wynikającymi z obciążeń współczynnika dochodu, jak zostało to omówione powyżej.

Ostatnio Dolan i Fujiwara (2012) zajmują się porównywaniem wartości dla edukacji dorosłych otrzymanych w oparciu o subiektywne poczucie dobrobytu i preferencje deklarowane. Pytali respondentów o WTP dla różnych kursów, które prowadziły do szeregu rezultatów, takich jak zdobycie kwalifikacji, doskonalenie umiejętności zawodowych, doskonalenie umiejętności rodzicielskich, itp., jak również otrzymali wartość wyceny subiektywnego poczucia dobrobytu dla edukacji dorosłych bazując na Brytyjskim Badaniu Panelowym Gospodarstw Domowych (ang. British Household Panel Survey, BHPS). Istniały pewne roz-

bieżności pomiędzy wartościami otrzymanymi w oparciu o preferencje, a tymi otrzymanymi dzięki pomiarom subiektywnego poczucia dobrobytu, jednak badacze ci stwierdzili, że jeżeli w badaniach preferencji deklarowanych respondentom zadaje się pytanie o WTP dla kursu, który prowadzi do podniesienia poziomu zadowolenia z życia, otrzymane wartości były w dużej mierze podobne do tych wygenerowanych z zastosowaniem metody wyceny subiektywnego poczucia dobrobytu z wykorzystaniem zadowolenia z życia. Powyższe badanie oraz wnioski Dolana i Metcalfa (2008) potwierdzają do pewnego stopnia, że wartości otrzymane na bazie preferencji i subiektywnego poczucia dobrobytu będą zwykle różne, zgodnie z teorią przedstawioną powyżej (ponieważ opierają się one na różnych ujęciach dobrobytu), jednak w pewnych okolicznościach – gdy respondentów wyraźnie prosi się o dokonanie wyceny pod kątem zadowolenia z życia – można dostrzec określone podobieństwa.

Podsumowując, okazuje się, że metoda wyceny bazująca na subiektywnym poczuciu dobrobytu powinna stanowić raczej *alternatywne* niż *uzupełniające* podejście techniczne do wyceny na potrzeby CBA, ponadto nie ma zgody co do tego, która z metod wyceny jest tą‚właściwą'. Będzie to w dużej mierze zależało od normatywnego osądu osób tworzących polityki i tego, które miary dobrobytu wybiorą. Tak więc, ważne jest, aby miary dobrobytu były w sposób wyraźny traktowane jako element ogólnego procesu oceny polityk oraz podejmowania decyzji. Korzystając z CBA przy podejmowaniu decyzji dotyczących polityki konieczne jest wzięcie pod uwagę odpowiednich zalet i wad każdej miary dobrobytu i związanych z nimi technik wyceny.

## 5. Podsumowanie

Ewaluacje wpływu są kluczowym elementem analizy polityk. CBA wymaga aby wpływ polityk był mierzony w kontekście tego, jak wpływają one na dobrobyt ludzi. Tradycyjnie, dobrobyt był mierzony w kontekście zaspokojenia preferencji z wykorzystaniem metod takich jak preferencje deklarowane oraz ujawnione, ale ostatnio analizy polityk oraz analiza kosztów i korzyści w coraz większym stopniu opierają się na ujęciach stanu umysłu, przede wszystkim na zadowoleniu z życia.

Oceny wyprowadzone z danych dotyczących preferencji oraz subiektywnego poczucia dobrobytu będą miały tendencję do odchyleń, czasami znacznych, co będzie miało wpływ na wyniki oraz rekomendacji dla polityk otrzymane z analizy kosztów i korzyści. Ponad 2000 lat debaty filozoficznej nie przyniosło miary dobrobytu do wykorzystania na potrzeby polityk, która stanowiłaby konsensus i z tego powodu organizacje sektora publicznego nie powinny czuć potrzeby wyróżniania jednej miary spośród pozostałych. Jednak, dane dotyczące SWB wprowadzają dodatkowy wymiar do procesów oceny polityk i stąd powinny być częścią danych zbieranych przez narodowe biura statystyczne na temat obywateli. Rozrastająca się dziedzina badań nad subiektywnym poczuciem dobrobytu z pewnością może przynieść wartościowe informacje na potrzeby procesu tworzenia polityk.

## Aneks

**Tabela 2.** Wcześniejsze prace poświęcone wycenie subiektywnego poczucia dobrobytu i ich wyniki *(Uwaga: poprzednio w literaturze stosowano termin „kompensacyjna zmiana dochodu", który używany był jako ogólne pojęcie obejmujące ES i CS. Bibliografia dostępna jest w pracy Fujiwary i Campbella, 2011).*

| Autor(rzy) | Kraj | Wyceniane dobro | Kompensacyjna zmiana dochodu |
|---|---|---|---|
| ***Blanchflower i Oswald (2004)*** | USA i Wielka Brytania | Różne | Na przykład: Koszty bezrobocia: -60 000 USD rocznie (dodatkowo, oprócz utraty pensji). |
| ***Carroll et al. (2009)*** | Australia | Susze i inne zdarzenia życiowe | Susza (w okresie wiosennym): -18 000 AUD; Małżeństwo: 67 000 AUD rocznie; Zatrudnienie: 7200 AUD rocznie (dodatkowo oprócz podwyżki płac). |
| ***CASE (2010)*** | Wielka Brytania | Kultura i sport | Zaangażowanie w sport (11 000 GBP rocznie, chodzenie na koncerty (9000 GBP rocznie), chodzenie do kina (9000 GBP rocznie). Wszystkie wartości przy zaangażowaniu: „przynajmniej raz w tygodniu". |
| ***Clark i Oswald (2002)*** | Wielka Brytania | Różne | Zatrudnienie w stosunku do bezrobocia: – 15 000 GBP miesięcznie (Kwestionariusz Ogólnego Stanu Zdrowia, GHQ) i -23 000 GBP miesięcznie (SWB) (dodatkowo oprócz utraty pensji); Świetny stan zdrowia w stosunku do dobrego stanu zdrowia: -10 000 GBP miesięcznie (GHQ), -12 000 GBP miesięcznie (SWB); Świetny stan zdrowia w stosunku do dostatecznego stanu zdrowia: -32 000 GBP miesięcznie (GHQ), -41 000 GBP miesięcznie (SWB). |
| ***Cohen (2008)*** | USA | Przestępczość i zdrowie | Przestępczość: -49 USD rocznie w przyp. 10% wzrostu przestępczości. Zdrowie: Dobry stan zdrowia w stosunku do dostatecznego stanu zdrowia: -161 060 USD rocznie; Dobry stan zdrowia w stosunku do złego stanu zdrowia: -276 624 USD rocznie. |
| ***DCLG (2010)*** | Wielka Brytania | Regeneracja miast | 59 600 GBP rocznie za przejście od „niezadowolony" do „zadowolony" z najbliższego otoczenia. Podane są także wartości dla innych wyników regeneracji. |
| ***Deaton et al. (2008)*** | Afryka | Wartość życia | Nieliczne dane szacunkowe na temat wartości życia wśród Afrykanów. |
| ***Tella et al. (2003)*** | USA i Europa | Różne | Wartości szacowane dla stopy bezrobocia i inflacji na poziomie makro. |
| ***Dolan i Metcalfe (2008)*** | Wielka Brytania | Regeneracja miast | Regeneracja otoczenia: 19 000 GBP. 6400 GBP przy zinstrumentalizowanym dochodzie. |
| ***Feeer-i-Carbonell i van Praag (2002)*** | Niemcy | Choroby przewlekłe | Na przykład, koszty cukrzycy: 59% dochodu; koszty artretyzmu: 43% dochodu; koszty problemów ze słuchem: 18% dochodu. |
| ***Feeriera i Moro (2009)*** | Irlandia | Jakość powietrza i klimat | Mniejsze zanieczyszczenie powietrza: 645 EUR za mikrogram PM10 na metr sześcienny (5% poprawa w stosunku do przeciętnego poziomu). Klimat: 15 585 EUR za wzrost temperatury o jeden stopień Celsjusza w styczniu i 5759 EUR za wzrost temperatury o jeden stopień Celsjusza w lipcu (wzrost temperatury został wyceniony pozytywnie). |
| ***Frey et al. (2004b)*** | Paryż, Londyn, Irlandia Północna | Terroryzm | Wartość zmniejszenia nasilenia działań terrorystycznych do niskiego poziomu (charakteryzującego inne części świata): 14% – 41% dochodu per capita. |
| ***Groot et al. (2004)*** | Holandia | Choroby sercowo-naczyniowe | 12 000 EUR – 25 000 EUR rocznie w przyp. osób 25-letnich. Wartość wyceny spada z wiekiem. W oparciu o podejście z wykorzystaniem miar zadowolenia z dochodu a niezadowolenia z życia. |

| | | | |
|---|---|---|---|
| ***Groot van den Brink (2006)*** | Wielka Brytania | Choroby sercowo-naczyniowe | Koszty chorób serca: -49 564 GBP (mężczyźni) i -17 503 GBP (kobiety). 93 532 GBP w przyp. 25-letniego mężczyzny i 1808 GBP w przyp. 75-letniego mężczyzny. |
| ***Helliwell i Huang (2005)*** | USA | Niefinansowa charakterystyka pracy | Jednopunktowy spadek zadowolenia z pracy (przy 10-punktowej skali) to koszt rzędu od 30 000 USD do 55 000 USD rocznie. |
| ***Levinson (2009)*** | USA | Jakość powietrza | Koszt 464 USD rocznie za mikrogram PM10 na metr sześcienny (wskazuje, że to więcej niż wartości wynikające z preferencji ujawnionych). |
| ***Leuchinger (2009)*** | Niemcy | Jakość powietrza | Wartość od 183 GBP do 313 GBP za spadek poziomu $SO_2$ o 1 mikrogram na metr sześcienny (w porównaniu do 6 GBP – 34 GBP z wykorzystaniem metody bazującej na preferencjach ujawnionych). |
| ***Leuchinger i Raschky (2009)*** | Europa | Powodzie | Wartość zapobiegania powodziom: 6500 USD; Wartość spadku prawdopodobieństwa powodzi w danym roku o średnią: 190 USD (należy zauważyć, że to tyle samo co kompensacja na rynkach hedonicznych). |
| ***Mackerron i Mourato (2009)*** | Wielka Brytania | Jakość powietrza w Londynie | Koszty wzrostu poziomu $NO_2$ o 1%: 5,3% dochodu (należy zauważyć, że to wyjątkowo dużo w porównaniu do badań bazujących na preferencjach deklarowanych i ujawnionych). |
| ***Oswald i Powdthavee (2008)*** | Wielka Brytania | Śmierć członka rodziny | Śmierć matki: -20 000 GBP rocznie (-10 000 GBP przy zinstrumentalizowanym dochodzie); Śmierć dziecka: -41 000 GBP rocznie (-34 000 GBP przy zinstrumentalizowanym dochodzie); Śmierć partnera: -64 000 GBP rocznie (-36 000 GBP przy zinstrumentalizowanym dochodzie). |
| ***Powdthavee (2008)*** | Wielka Brytania | Relacje społeczne | Koszty przeprowadzki, zmiana od możliwości widywania przyjaciół i krewnych rzadziej niż raz w miesiącu do nigdy: -63 000 GBP rocznie; Małżeństwo: 68 000 GBP rocznie; Wartość poprawy zdrowia z bardzo złego stanu do świetnego stanu zdrowia 300 000 GBP; Koszty bezrobocia: -74 000 GBP rocznie (dodatkowo oprócz utraty pensji). |
| ***Powdthavee i van den Berg (2011)*** | Wielka Brytania | Stan zdrowia | Koszty problemów dotyczących rąk, nóg, dłoni, stóp, pleców, itp. (7000 GBP rocznie), Cukrzyca (6000 GBP rocznie), Problemy z sercem, ciśnieniem krwi lub krążeniem krwi (8000 GBP rocznie). Raporty zawierają wiele innych danych szacunkowych. Wykorzystano kilka miar subiektywnego poczucia dobrobytu – zaprezentowane tu wyniki dotyczą wyłącznie pomiarów w oparciu o zadowolenie z życia. |
| ***Rehdanz i Maddison (2005)*** | Panel kilku krajów | Klimat | Szereg wartości oszacowanych dla 67 krajów. |
| ***Stutzer i Frey (2004)*** | Niemcy | Dojazdy do pracy | Koszty dojazdów do pracy zabierających 23 minuty dziennie (wartość średnia próbki): -242 EUR miesięcznie (19% średniej miesięcznej pensji). |
| ***van den Berg i Ferrer i Caronell (2007)*** | Holandia | Opieka nieformalna | Koszty opieki: 0d 8 EUR do 9 EUR za godzinę w przypadku członka rodziny. Od 7 EUR do 9 EUR za godzinę gdy nie chodzi o członka rodziny. |
| ***Van Praag i Baarsma (2005)*** | Holandia | Hałas lotniczy | Koszty hałasu na przelot: 253 EUR. |
| ***Welsch (2002)*** | W różnych krajach | Zanieczyszczenie powietrza | Koszt 70 USD rocznie za jedną kilotonę dwutlenku azotu per capita. |

| | | | |
|---|---|---|---|
| ***Welsch (2006)*** | 10 państw europejskich | Zanieczyszczenie powietrza | Zmniejszenie poziomu pyłu całkowitego (TSP) wyceniono na od 13 USD do 211 USD rocznie za mikrogram (na metr sześcienny) (wskazuje, że wartości te są porównywalne do wartości otrzymanych na podstawie amerykańskich modeli hedonicznych). |
| ***Welsch (2007)*** | Międzynarodowe – 54 państwa | Zanieczyszczenie powietrza | Koszty rzędu „kilkuset dolarów amerykańskich" za tonę dwutlenku azotu w przypadku efektu bezpośredniego. Efekt pośredni zanieczyszczenia powietrza na SWB jest pozytywny, ponieważ to środek produkcji, jednak mniejszy niż efekt bezpośredni w ujęciu bezwzględnym. |
| ***Welsch (2008a)*** | Międzynarodowe – 21 państw z historią konfliktu | Konflikty cywilne | Koszty jednej ofiary śmiertelnej z powodu konfliktu: -108 000 USD. |
| ***Welsch (2008b)*** | Międzynarodowe | Korupcja | Wzrost korupcji o 1 punkt indeksowy na 10-punktowej skali Transparency International (stosunkowo duża zmiana) to koszt rzędu -900 USD per capita rocznie (włącznie z efektami pośrednimi). |

---

**Daniel Fujiwara** jest starszym ekonomistą w Kancelarii Rady Ministrów Wielkiej Brytanii oraz badaczem w London School of Economics and Political Science (LSE). Pełni funkcję głównego doradcy Rządu Brytyjskiego w sprawach technik wyceny dóbr nierynkowych oraz analizy kosztów i korzyści, a także kieruje analizami ekonometrycznymi danych dotyczących poziomu życia w Wielkiej Brytanii wykonywanymi dla Narodowego Biura Statystycznego. Jest autorem poradników dla Rządu Brytyjskiego dotyczących ewaluacji, w tym współautorem najnowszej wersji Zielonej Księgi Ministerstwa Skarbu (2011). W przeszłości Daniel Fujiwara był odpowiedzialny za analizy kosztów i korzyści w Departamencie Pracy i Emerytur. W latach 2007-2009 był Starszym Ekonomistą w ministerstwie Finansów Tanzanii, gdzie prowadził dla Banku Światowego i ONZ ewaluacje projektów w ramach Milenijnych Celów Rozwoju. Obecnie kończy rozprawę doktorską z dziedziny nauk behawioralnych w LSE, a jego głównymi polami zainteresowania są teorie mikroekonomiczne, ekonometria, psychologia poznawcza i neurobiologia oraz aplikacja tych nauk do ewaluacji polityk. Dodatkowo jest recenzentem kilku czasopism akademickich.

## Bibliografia


- Ariely D., Loewenstein G., Prelec D., *"Coherent Arbitrariness": Stable Demand Curves without Stable Preferences,* "The Quarterly Journal of Economics", 118, 2003, s. 73-105.
- Bertrand M., Mullainathan S., *Do People Mean What They Say? Implications for Subjective Survey Data*, "The American Economic Review", 91, 2001, s. 67-72.
- Bockstael N., Mcconnell K., *Calculating Equivalent and Compensating Variation for Natural Resource Facilities*, "Land Economics", 56, 1980, s. 56-63.
- Champ P., Boyle K., Brown T., *A Primer on Nonmarket Valuation,* Boston, Kluwer Academic Press 2003.
- Clark A. E., Oswald A. J., *A simple statistical method for measuring how life events affect happiness*, "International Journal of Epidemiology", 31, 2002, s. 1139-1144.
- Cohen S., Doyle W., Turner R., Alper C., Skoner D., *Emotional Style and Susceptibility to the Common Cold*, "Psychosomatic Medicine", 65, 2003, s. 652–57.
- Deontology, Together with a Table of the Springs of Action and the Article on Utilitarianism, Bentham J. (red.), Oxford: Clarendon Press, 1983.
- Desvousges W. H., Johnson F., Dunford R., Boyle K., Hudson S., Wilson, N., *Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy*, RTI Press, 1992.
- Diener E., *Subjective well-being*, "Psychological Bulletin", 95, 1984, s. 542-575.
- Diener E., Suh E., *Measuring Quality of Life: Economic, Social, and Subjective Indicators. Social Indicators Research*, 40, 1997, s. 189-216.

- Dolan P., *Using Happiness to Value Health*. [w:] *Office of Health Economics Monograph*, 2011.
- Dolan P., Fujiwara D., *Valuing Adult Learning: Comparing Wellbeing Valuation to Contingent Valuation*. BIS Research Paper, 85, 2012.
- Dolan P., Kahneman D., *Interpretations Of Utility And Their Implications For The Valuation Of Health*, "Economic Journal", 118, 2008, s. 215-234.
- Dolan P., Layard R., Metcalfe R., *Measuring Subjective Wellbeing for Public Policy: Recommendations on Measures*, Centre for Economic Performance, London School of Economics and Political Science, Wydanie specjalne, 2011.
- Dolan P., Metcalfe R., *Comparing willingness to pay and subjective wellbeing in the context of non-market goods.,* Centre for Economic Performance (London School of Economics) Discussion paper 890, 2008.
- Dolan P., Metcalfe R., *Valuing wind farms: does experience matter?,* Niepublikowany dokument roboczy, 2010.
- Dolan P., White M. P., *How Can Measures of Subjective Well-Being Be Used to Inform Public Policy?*, "Perspectives on Psychological Science", 2, 2007, s. 71-85.
- Eid M., Diener E., *Global Judgments of Subjective Well-Being: Situational Variability and Long-Term Stability*, Social Indicators Research, 65, 2004, s. 245-277.
- Ekman P., Davidson R., Friesen W., *The Duchenne Smile: Emotional Expression and Brain Physiology II.*, "Journal of Persotutlity and Social Psychology", 58, 1990, s. 342-53.
- Fedderson J., Metrcalfe R., Wooden M., *Subjective Well-Being: Weather Matters; Climate Doesn't.,* University of Oxford. Department of Economics Working Paper Series, 2012.
- Frey B. S., Luechinger S., Stutzer A., *Valuing Public Goods: The Life Satisfaction Approach. Public Choice*, 138, 2009, s. 317–345.
- Frey B. S., Stutzer A., *What Can Economists Learn from Happiness Research?*, "Journal of Economic Literature", 40, 2002, s. 402-435.
- Frey B. S., Stutzer A., *Happiness research: State and prospoects.*, University of Zurich Dokument roboczy, 2005.
- Fujita F., Diener E., *Life Satisfaction Set Point: Stability and Change*. "Journal of Personality and Social Psychology", 88, 2005, s. 158-164.
- Fujiwara D., *A General Method for Valuing Non-Market Goods using Wellbeing Data: Three-Stage Wellbeing Valuation*, Centre for Economic Performance Discussion Paper 1233, 2013.
- Fujiwara D., Campbell R., *Valuation Techniques for Social Cost-Benefit Analysis: Stated Preference, Revealed Preference and Subjective Well-Being Approaches*. [w:] Pensions, H. T. A. D. F. W. A. (red.). Londyn, 2011.
- Hicks J. R., Allen R. G. D., *A Reconsideration of the Theory of Value*. Część I. "Economica", 1, 1934, s. 52-76.
- Kahneman D., Fredrickson B. L., Schreiber C. A., Redelmeier D. A., *When More Pain Is Preferred to Less: Adding a Better End*, "Psychological Science", 4, 1993, s. 401-405.
- Kahneman D., Krueger A., Schkade D., Schwarz N., Stone A., *Would You Be Happier If You Were Richer? A Focusing Illusion*, Dokument roboczy CEPS, 125, 2006.
- Kiecolt-Glaser J., Mcguire L., Robles T., *Glaser R. Psychoneuroimmunology: Psychological Influences on Immune Function and Health*. "Journal of Consulting and Clinical Psychology", 70, 2002, s. 537–47.
- Krueger A. B., Schkade D. A., *The reliability of subjective well-being measures*. "Journal of Public Economics", 92, 2008, s. 1833-1845.
- Levinson A., *Valuing public goods using happiness data: The case of air quality*. "Journal of Public Economics", 96, 2012, s. 869-880.
- Luechinger S., *Valuing Air Quality Using the Life Satisfaction Approach*. "Economic Journal", 119, 2009, s. 482-515.
- Mackerron G., *Happiness Economics From 35 000 Feet*. "Journal of Economic Surveys", vol.26, nr.4, 2011.
- Parfit D., *Reasons and Persons*, Oxford Scholarship Online,1984.
- Pavot W., Diener E., *Review of the Satisfaction With Life Scale*. "Psychological Assessment", 5, 1993, s. 164-172.
- Powdthavee N., *Putting a price tag on friends, relatives, and neighbours: Using surveys of life satisfaction to value social relationships*, "Journal of Socio-Economics", 37, 2008, s.1459-1480.
- Powdthavee N., Van Den Berg B., *Putting different price tags on the same health condition: Re-evaluating the well-being valuation approach*, "Journal of Health Economics", 30, 2011, s. 1032-1043.
- *Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications*, Schwarz N., Strack F. (red.), New York: Russell Sage Foundation, 1999.
- Sandvik E., Diener E., Seidlitz L., *Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures*, "Journal of Personality", 61, 1993, s. 317-342.
- Schimmack U., Oishi S., *The influence of chronically and temporarily accessible information on life satisfaction judgments*, "Journal of Personality and Social Psychology", 89, 2005, s. 395-406.
- Simon H. A., *A Behavioral Model of Rational Choice*, "The Quarterly Journal of Economics", 69, 1955, s. 99-118.
- Slovic P., Lichtenstein S., *The Construction of Preference*, New York, Cambridge University Press, 2006.
- *Treasury*, Green Book, 2003.
- Urry H. L., Nitschke J. B., Dolski I., Jackson D. C., Dalton K. M., Mueller C. J., Rosenkranz M. A., Ryff C. D., Burton H. S., Davidson R. J., *Making a Life Worth Living: Neural Correlates of Well-Being*, "Psychological Science", 15, 2004, s. 367-372.
- *Why Researchers Should Think "Real-Time": A Cognitive Rationale*, Schwarz N. (red.), Nowy Jork: Guilford, 2010.

*Jochen Kluve*

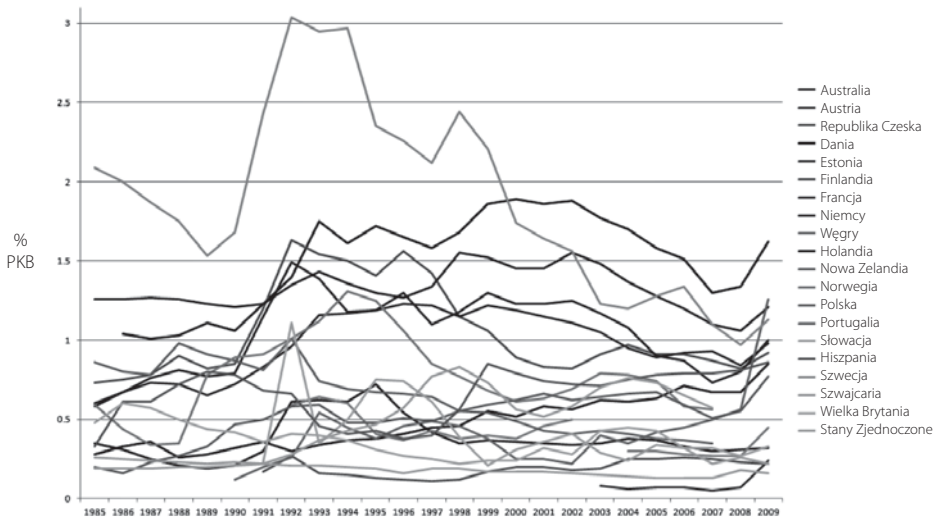
# Skuteczność Aktywnych Polityk Rynku Pracy: wyniki metaanaliz

## Wprowadzenie

W celu zmniejszenia bezrobocia, a bardziej ogólnie – zwiększenia szans na rynku pracy osób bezrobotnych lub pracowników o niskich kwalifikacjach, kraje OECD od kilku dekad stosują Aktywne Polityki Rynku Pracy (ang. *Active Labour Market Policies* – ALMP). Aktywne Polityki Rynku Pracy można zaklasyfikować do 4 głównych kategorii: szkolenie z zakresu rynku pracy, zatrudnienie w sektorze prywatnym, zatrudnienie w sektorze publicznym oraz pomoc w szukaniu zatrudnienia. W szczególności w ciągu ostatnich dziesięciu lat działania w ramach Aktywnych Polityk Rynku Pracy w krajach OECD w coraz większym stopniu były powiązane z systemem biernego wsparcia w ramach „strategii aktywizacji" poszczególnych krajów (OECD 2007). Występowanie tego zjawiska sugeruje, że mogą być zastosowane elementy ograniczenia świadczeń (w przypadku, gdy nie są przestrzegane zalecenia dotyczące poszukiwania pracy) lub obowiązkowe objęcie Aktywną Polityką Rynku Pracy (po pewnym okresie pozostawania bezrobotnym).

Skuteczność Aktywnych Polityk Rynku Pracy wzbudza kontrowersje od czasów ich pierwszego zastosowania w latach 40. ubiegłego wieku w Stanach Zjednoczonych. Rysunek 1 pokazuje, że co najmniej od lat 80. XX w. wiele krajów OECD stosowało Aktywne Polityki Rynku Pracy na odpowiednią skalę. Badanie skuteczności polityk było nieco opóźnione w stosunku do momentu ich wdrażania, ale w miarę poprawy jakości i ilości danych (administracyjnych) i metod ekonometrycznych (Heckman i in. 1999), w szczególności dla badań nieeksperymentalnych, od 20 lat pula dostępnych dowodów na skuteczność stale wzrasta.

**Rys. 1.** Wydatki na aktywną politykę rynku pracy w krajach OECD, 1985-2009




Źródło danych: stats.oecd.org

Oprócz analiz wpływu dla poszczególnych programów przeprowadzono zatem dodatkowo szereg badań, które podsumowują wyniki ewaluacji poszczególnych programów i których celem jest znalezienie systematycznych tendencji dotyczących skuteczności Aktywnych Polityk Rynku Pracy dla poszczególnych krajów i typów programów. OECD ze szczególnym zaangażowaniem śledzi doświadczenia krajów członkowskich (np. Martin i Grubb 2001) i regularnie aktualizuje swoją wiedzę na temat stosowania i skuteczności Aktywnych Polityk Rynku Pracy w swojej serii *Employment Outlook*. Heckman i in. (1999) prezentują kompleksowy przegląd metodologii oraz ewaluacji programów w ramach ALMP, dla badań przeprowadzonych do końca lat 90. ubiegłego wieku.

Ostatnio, metaanalizy skuteczności Aktywnych Polityk Rynku Pracy zapewniły systematyczną ocenę wpływu programów. Greenberg i in. (2003) analizują 31 ewaluacji programów rządowych dla osób w niekorzystnej sytuacji życiowej w Stanach Zjednoczonych. Kluve (2010) analizuje programy w Europie zestawiając w bazie danych 137 szacunków wpływu programów. Card i in. (2010) wykorzystuje nową, obszerną próbę 97 badań i 199 szacunków wpływu, z których większość pochodzi z krajów OECD[1].

Celem tych metaanaliz jest znalezienie ilościowych, systematycznych tendencji dla skuteczności programów poprzez zestawienie miary wpływu programu z szeregiem zmiennych objaśniających, takich jak np. typ programu, charakterystyka grupy docelowej, metodologia ewaluacji itd. Idealną miarą wpływu programu byłby szacunek wielkości efektu (wykorzystywany w metaanalizach prowadzonych w innych dziedzinach, takich jak np. epidemiologia). Jednak ze względu na niejednorodność danych i metod stosowanych w ewaluacji Aktywnych Polityk Rynku Pracy w poszczególnych krajach, można tego dokonać jedynie dla niewielkiej części dostępnych badań (Card i in. 2010). Metaanalizy Aktywnych Polityk Rynku Pracy skupiły się więc na zestawianiu trójmianowej miary skuteczności programu (pokazującej, czy szacunek wpływu jest wyraźnie pozytywny, wyraźnie negatywny lub nieznacznie różny od zera) z charakterystyką programu i jego ewaluacji.

## Rodzaje Aktywnych Polityk Rynku Pracy

Zgodnie z zaleceniami zawartymi w literaturze ewaluacyjnej oraz praktyką stosowaną przez np. OECD i Eurostat wyróżnia się zazwyczaj cztery typy Aktywnych Polityk Rynku Pracy.

Pierwszy typ, **szkolenie (z zakresu rynku pracy),** obejmuje programy takie, jak szkolenie w szkole, szkolenie na stanowisku pracy i doświadczenie zawodowe. Interwencje mogą zapewnić bardziej ogólne wykształcenie (obejmujące wszystkie rodzaje kursów podstawowych) lub specjalistyczne umiejętności zawodowe (kursy z zakresu np. umiejętności technicznych i produkcyjnych). Głównym celem programów jest zwiększenie wydajności i szans na zatrudnienie uczestników oraz rozwój kapitału ludzkiego poprzez podniesienie kwalifikacji. Programy szkoleniowe można więc uznać za „klasyczną" Aktywną Politykę Rynku Pracy. Są to najczęściej stosowane na świecie typy programów (Betcherman i in. 2004; Card i in. 2010).

Drugi typ, **programy zachęt w sektorze prywatnym**, obejmuje wszystkie interwencje mające na celu stworzenie zachęt, które zmieniają podejście pracodawcy lub pracownika do zatrudnienia w sektorze prywatnym. Najważniejszym działaniem w tej kategorii – w szczególności w krajach OECD – jest subsydiowanie wynagrodzeń. Celem subsydiów jest zachęcenie pracodawców do zatrudniania nowych pracowników lub do utrzymania miejsc pracy, które w innym wypadku zostałyby zlikwidowane. Subsydia te mogą być bezpośrednimi dopłatami do wynagrodzeń dla pracodawców lub finansowymi zachętami dla pracowników ograniczonymi czasowo. Są one często kierowane do długotrwale bezrobotnych i osób w niekorzystnej sytuacji życiowej. Innym rodzajem subsydiowanego zatrudnienia w sektorze prywatnym jest wsparcie samozatrudnienia. Osoby bezrobotne, które zakładają własną firmę, otrzymują dotacje lub pożyczki, a czasem nawet

[1] Analizy Kluve (2010) i Card i in. (2010) nakładają się w przypadku 37 badań. Wcześniejsza analiza zawiera badania do lat ok. 2002/2003 a późniejsza – do 2007 r.

wsparcie w postaci doradztwa, przez określony czas. Takie programy na rzecz przedsiębiorczości, które łączą wsparcie finansowe i szkolenia są coraz częściej stosowane w gospodarkach wschodzących i krajach rozwijających się, często z większym naciskiem na element szkoleniowy w stosunku do elementu dotacji/pożyczki. Szkolenie techniczne w zakresie samozatrudnienia może obejmować umiejętności biznesowe (np. mentoring lub prowadzenie księgowości), umiejętność czytania i pisania, „umiejętności życiowe".

Trzeci typ Aktywnych Polityk Rynku Pracy – **programy bezpośredniego zatrudnienia w sektorze publicznym** – koncentruje się, w przeciwieństwie do subsydiów dla sektora prywatnego, na bezpośrednim kreowaniu prac publicznych lub innych aktywności, które generują publiczne towary lub usługi. Działania te są zazwyczaj skierowane do osób w najbardziej niekorzystnej sytuacji życiowej i mają na celu utrzymanie przez te osoby kontaktu z rynkiem pracy i zapobieganie utracie kapitału ludzkiego w okresie bezrobocia. Tworzone miejsca pracy są jednak często dodatkowo generowanymi miejscami pracy i nie mają przełożenia na rzeczywisty rynek pracy.

Czwarty typ polityk, **usługi i sankcje**, obejmuje wszelkie działania mające na celu zwiększenie efektywności procesu szukania pracy. Definicja ta jest stosowana np. u Kluve (2010) i Card i in. (2010) i nieznacznie zmienia standardową kategorię „Pomoc w szukaniu pracy", głównie przez dodanie sankcji. Interwencje zwyczajowo wpisujące się w tę kategorię – kursy z zakresu szukania pracy, kluby pracy, poradnictwo zawodowe, doradztwo i monitoring oraz sankcje w przypadku niepodporządkowania się zaleceniom dotyczącym szukania pracy – mają wspólny cel, ponieważ dążą do zwiększenia efektywności procesu dopasowywania miejsc pracy. O ile usługi te mogą zasadniczo być świadczone przez podmioty publiczne lub prywatne, w poszczególnych krajach dominują usługi publiczne. Wśród Aktywnych Polityk Rynku Pracy programy wsparcia w szukaniu pracy są zazwyczaj najmniej kosztowne. Sankcje w zakresie świadczeń (np. zmniejszenie zasiłku dla bezrobotnych) są obecnie stosowane w większości krajów OECD, jeżeli z obserwacji wynika, że bezrobotny nie szuka pracy wystarczająco intensywnie lub jeżeli odmawia on przyjęcia oferty pracy (np. OECD 2007).

Rysunek 1 pokazuje tendencje w wydatkach na Aktywne Polityki Rynku Pracy (mierzone jako procent PKB) w krajach OECD oraz sumarycznie wydatki dla wszystkich powyższych typów. Można zaobserwować ogólny trend spadkowy ze znacznym wzrostem w poszczególnych krajach w 2009 r., na początku kryzysu.

## Sposoby generowania wiedzy

### Ewaluacje pojedynczych programów

Aby dowiedzieć się, którą Aktywną Politykę Rynku Pracy wykorzystać w danym kontekście dla danej grupy docelowej, kluczowa jest ocena skuteczności poszczególnych pojedynczych interwencji. Taka ewaluacja programu (analiza skuteczności, ewaluacja wpływu) nie tylko informuje wdrażającego program, czy polityka osiągnęła swój cel (cele), ale także wpływa na ewentualną kontynuację, zmianę lub zakończenie programu. Ponadto ewaluacje poszczególnych programów zazwyczaj generują wiedzę, którą można zastosować w przypadku podobnych programów w innych kontekstach.

W ciągu ostatniego dwudziestolecia miały miejsce dwa ważne zjawiska dotyczące ewaluacji Aktywnych Polityk Rynku Pracy, jedno w środowisku akademickim a drugie w środowisku politycznym. Zjawiska te nastąpiły w pewnym stopniu równolegle, lecz są ze sobą blisko powiązane.

Po pierwsze, wśród osób odpowiedzialnych za tworzenie polityk w Europie wzrasta zainteresowanie ewaluacją polityk publicznych, ponieważ chcą one uzyskać wiedzę na temat efektów wdrażanych polityk. Ma to swoje korzenie w Stanach Zjednoczonych, gdzie już w latach 60. i 70. ubiegłego wieku – kiedy wprowadzano pierwsze Aktywne Polityki Rynku Pracy – zauważono, że ewaluacja empiryczna tych programów

jest kluczowa ze względów informacyjnych[2]. Obserwacja ta stanowi początek ogólnego dążenia do wdrażania tak zwanej „polityki opartej na dowodach".

W Europie takie podejście w zasadzie przyjęło się dopiero w latach 90. XX w., a rozwój w kierunku europejskiej „kultury ewaluacji" został zainicjowany (przynajmniej częściowo) przez Komisję Europejską zalecającą ewaluacje Aktywnych Polityk Rynku Pracy jako część Europejskiej Strategii Zatrudnienia (dalsze omówienie tej tendencji zob. Kluve i in. 2007). Oczywiście, między poszczególnymi krajami występują w tym względzie różnice dotyczące stopnia postrzegania ewaluacji polityk publicznych jako koniecznej oraz stopnia promowania wykorzystania wyników ewaluacji. Ogólna tendencja jest jednak optymistyczna. Konkretne kamienie milowe to na przykład eksperymentalna ewaluacja programu „Restart" w Wielkiej Brytanii w latach 90. ubiegłego wieku (Dolton i O'Neill 1996) i formalne włączenie wymogu ewaluacji do przepisów Hartza w Niemczech na początku pierwszej dekady XXI wieku (Jacobi i Kluve 2007).

Drugi, równoległy, trend, to stworzenie przez ekonomistów zajmujących się rynkiem pracy szeregu narzędzi statystycznych na potrzeby ewaluacji Aktywnych Polityk Rynku Pracy. Ta metodologiczna debata znacznie przyczyniła się do postępu w dziedzinie ewaluacji programów (por. Heckman i in. 1999). Ponadto metodologiczny postęp w poszczególnych krajach wzmocniła większa dostępność dużych ilości danych administracyjnych dla badaczy. Wiele ewaluacji poszczególnych programów także generuje dane, np. z kwestionariuszy specjalnie dopasowanych do określonej ewaluacji.

Łącznie te dwie zmiany – czyli większe zainteresowanie polityków wynikami ewaluacji i polityką opartą na dowodach oraz większe możliwości dostarczania takich dowodów przez badaczy – przyczyniły się do powstania dużej ilości informacji na temat skuteczności Aktywnych Polityk Rynku Pracy w Europie i krajach OECD. Kolejne części tekstu prezentują, w jaki sposób można podsumować tę wiedzę i które z rezultatów są wciąż aktualne dziś.

## Podsumowanie dowodów z ewaluacji poszczególnych programów z wykorzystaniem metaanalizy

Dużą liczbę pojedynczych ewaluacji wpływu programów, które przeprowadzone zostały w różnych krajach, można zasadniczo zbadać i podsumować na dwa różne sposoby. Pierwszym sposobem jest tradycyjna analiza literatury, tak zwany „przegląd piśmiennictwa". W kontekście skuteczności Aktywnych Polityk Rynku Pracy, wielokrotnie robiła to OECD; zob. np. Martin i Grubb (2001) i OECD (2007).

Drugim sposobem na podsumowanie dowodów jest „przegląd ilościowy" z wykorzystaniem metaanalizy. Wiedza na temat skuteczności Aktywnych Polityk Rynku Pracy została na przykład ostatnio podsumowana w metaanalizach autorstwa Kluve (2010) i Card i in. (2010). Metaanaliza jest badaniem, które zbiera szereg badań analizujących ten sam (lub podobny) problem badawczy i generuje jeden zestaw metadanych. Zbieranie wielu badań odbywa się zgodnie z tak zwanym „protokołem" określającym kryteria, na podstawie których badania są uwzględniane w zestawie metadanych lub nie. Badania Kluve (2010) i Card i in. (2010) obejmują na przykład tylko te ewaluacje Aktywnych Polityk Rynku Pracy, które oceniają skutki programu z wykorzystaniem pewnego wariantu doboru grupy kontrolnej. Po zebraniu metadanych można je poddać analizie z wykorzystaniem (zazwyczaj prostych) narzędzi statystycznych w celu zidentyfikowania w danych systematycznych tendencji. Tabela 1 prezentuje przegląd danych zebranych przez Kluve (2010), w ramach których zgromadzono 137 ewaluacji programów z 19 krajów europejskich.

[2] W Stanach Zjednoczonych dyskusja na temat ewaluacji polityk publicznych została natychmiast powiązana z dyskusją metodologiczną, tj. uznano fakt, że do odpowiedniej oceny skutków programu i tym samym dostarczania informacji na potrzeby polityki niezbędne są dowody eksperymentalne.

**Tabela 1.** Statystyki dotyczące metadanych dla skuteczności Aktywnych Polityk Rynku Pracy

| | Liczba badań | Średnia (odchylenie standardowe) |
|---|---|---|
| a) Rodzaje programów i grup docelowych | | |
| Szkolenie | 70 | |
| Program zatrudnienia bezpośredniego | 26 | |
| Program zachęt dla sektora prywatnego | 23 | |
| Usługi i sankcje | 21 | |
| Programy dla młodych pracowników | 35 | |
| b) Plan badania i przedział czasowy | | |
| Eksperyment | 9 | |
| Estymator dopasowania | 51 | |
| Model okresowy | 42 | |
| OLS, wybór, inne | 39 | |
| Program wdrażany w latach 70. | 4 | |
| Program wdrażany w latach 80. | 36 | |
| Program wdrażany w latach 90. | 81 | |
| Program wdrażany w latach 2000 | 16 | |
| c) Kontekst instytucjonalny na rynku pracy | | |
| Wskaźnik prawnej ochrony zatrudnienia | | 2,29 (0,75) |
| Wskaźnik umów zawartych na czas określony | | 2,16 (1,16) |
| Wskaźnik dla zjawiska pracy tymczasowej | | 2,34 (1,66) |
| Stopa zastąpienia brutto | | 35,65 |
| d) Otoczenie makroekonomiczne | | |
| Wskaźnik bezrobocia | | 7,82 (3,27) |
| Wydatki na aktywną politykę rynku pracy (% PKB) | | 1,23 (0,59) |
| Wzrost PKB | | 2,54 (1,35) |
| e) Najważniejsze kraje | | |
| Austria | 4 | |
| Dania | 15 | |
| Francja | 12 | |
| Finlandia | 8 | |
| Niemcy | 16 | |
| Irlandia | 5 | |
| Holandia | 11 | |
| Norwegia | 12 | |
| Szwecja | 23 | |
| Szwajcaria | 5 | |
| Zjednoczone Królestwo | 12 | |

Źródło: Kluve (2010).

Uwaga: wszystkie zmienne instytucjonalne c) i zmienne makro d) są uzależnione od czasu i zawsze mierzone w momencie realizowania danego programu. W danych wskaźnik ochrony zatrudnienia OECD waha się od 0,8 do 4,3, wskaźnik umów zawartych na czas określony OECD waha się od 0 do 5,3 a wskaźnik dla zjawiska pracy tymczasowej OECD waha się od 0,5 do 5,5.

Bez wchodzenia w szczegóły (tabela ma służyć głównie jako ilustracja), pięć kategorii od a) do e) pokazuje, że poszczególne ewaluacje programów można sklasyfikować pod względem kilku cech:

a) Typ programu;

b) Typ analizy empirycznej („projekt badania i przedział czasowy");

c) Kontekst rynku pracy, w którym program był realizowany;

d) Kontekst ekonomiczny, w którym program był realizowany;

e) Kraj.

Przykładem ewaluacji konkretnego programu uwzględnionej w badaniu Kluve (2010) jest:

a) Program szkoleniowy dla dorosłych …

b) … wdrożony w latach 2000-2001 w Hiszpanii e) i oceniony z wykorzystaniem metod analizy okresowej.

c) W tym czasie w Hiszpanii wartości wskaźników: prawnej ochrony zatrudnienia, umów zawartych na czas określony oraz zjawiska pracy tymczasowej plasowały się powyżej średniej (wartości odpowiednio: 2,6; 2,5; 4). Stopa zastąpienia była nieznacznie poniżej średniej i wyniosła 31 procent.

d) Sytuację makroekonomiczną w tym czasie charakteryzowała stopa bezrobocia wynosząca 12,3%, wydatki na Aktywne Polityki Rynku Pracy wynoszące 0,9 % PKB i stopa wzrostu PKB 3,6%.

W ramach ewaluacji oszacowano pozytywny wpływ udziału w szkoleniu na prawdopodobieństwo zatrudnienia uczestników szkolenia.

Informacja ta należy do zestawu 137 ewaluacji składających się na metadane. Możliwa jest zatem kombinacja informacji na temat tego, czy ewaluacja wykazała pozytywny czy negatywny skutek, czy też jego brak, z informacjami a) do e) opisanymi powyżej w celu ustalenia, czy systematyczne tendencje dotyczące skuteczności programu uzależnione są od wymienionych parametrów.

## Skuteczność Aktywnych Polityk Rynku Pracy

Przeprowadzenie opisanej powyżej analizy korelacji (w postaci regresji) prowadzi do uzyskania szczegółowych wyników liczbowych zebranych w Tabeli 2. Dla przejrzystości wywodu Tabelę 2 można jednak pominąć i przejść do Tabeli 3, która pokazuje zestawienie najważniejszych wyników. W Tabeli 3 więc „+" oznacza pozytywną korelację, „(+)" oznacza nieznacznie pozytywną korelację, „0" oznacza brak istotnego związku, „(-)" oznacza nieznacznie negatywną korelację a „-" oznacza negatywną korelację.

**Tabela 2.** Korelaty skuteczności Aktywnych Polityk Rynku Pracy

| | **(1)** | | **(2)** | |
|---|---|---|---|---|
| | **Szacunek negatywny** | | **Szacunek pozytywny** | |
| | **Skutek krańcowy *t*** | | **Skutek krańcowy *t*** | |
| a) Rodzaj programu i grupy docelowej (pominięto: szkolenie; dorosłych/niepełnosprawnych) | | | | |
| Program zatrudnienia bezpośredniego | 0,155 | 1,92 | -0,216 | -2,13 |
| Program zachęt dla sektora prywatnego | -0,144 | -3,52 | 0,280 | 2,91 |
| Usługi i sankcje | -0,205 | -3,87 | 0,436 | 4,63 |
| Programy dla młodych pracowników | 0,140 | 1,79 | -0,202 | -1,94 |
| b) Plan badania i przedział czasowy (pominięto: OLS/wybór/inne; badania z lat 70. i 80.) | | | | |
| Eksperyment | 0,314 | 1,32 | -0,356 | -1,87 |
| Dopasowanie | 0,061 | 0,88 | -0,095 | -0,94 |
| Model okresowy | 0,041 | 0,50 | -0,064 | -0,52 |
| Program wdrażany w latach 90. | 0,115 | 1,45 | -0,192 | -1,50 |
| Program wdrażany w latach 2000 | 0,190 | 1,30 | -0,248 | -1,61 |
| c) Kontekst instytucjonalny na rynku pracy | | | | |
| Wskaźnik prawnej ochrony zatrudnienia | 0,067 | 1,77 | -0,109 | -1,76 |
| Wskaźnik umów zawartych na czas określony | -0,023 | -0,80 | 0,037 | 0,80 |
| Wskaźnik dla zjawiska pracy tymczasowej | 0,001 | 0,03 | -0,001 | -0,03 |
| Stopa zastąpienia brutto | 0,004 | 1,40 | -0,006 | -1,41 |

| | | | | |
|---|---|---|---|---|
| d) Otoczenie makroekonomiczne | | | | |
| Wskaźnik bezrobocia | -0,022 | -2,13 | 0,035 | 1,95 |
| Wydatki na aktywną politykę rynku pracy (% PKB) | 0,060 | 1,12 | -0,097 | -1,13 |
| Wzrost PKB | 0,009 | 0,35 | -0,015 | -0,35 |

Źródło: Kluve (2010).

Uwaga: Zmienna zależna jest zmienną kategoryczną pokazującą, czy szacunek dla skutku programu jest negatywny (-1), neutralny (0) czy pozytywny (+1). Dane w kolumnach (1) – (4) pokazują skutki krańcowe (obliczone dla przykładowej średniej) z odpowiedniej uporządkowanej regresji probitowej odpowiednio dla negatywnych i pozytywnych skutków. Różnica w przewidywanym prawdopodobieństwie osiągnięcia negatywnego (pozytywnego) skutku działania, który wynika z krańcowej zmiany w stałym czynniku wyjaśniającym (takim jak wzrost PKB) lub który wynika ze zmiany wskaźnika spośród czynników wyjaśniających (takiego jak wskaźnik dla planu badania eksperymentalnego) mieści się w przedziale od 0 do 1. Wskaźniki T krańcowych skutków podane są w sąsiedniej kolumnie. Występujące standardowe błędy są poprawiane w poszczególnych badaniach.

**Tabela 3.** Korelaty skuteczności Aktywnej Polityki Rynku Pracy – podsumowanie

| | |
|---|---|
| a) Rodzaj programu | |
| Szkolenie | (+) |
| Program zatrudnienia bezpośredniego | – |
| Programy zachęt w sektorze prywatnym | + |
| Usługi i sankcje | + |
| Program dla osób młodych | – |
| b) Plan badania i przedział czasowy | |
| Eksperyment | – |
| Badanie od lat 90. do lat 2000 | (–) |
| c) Kontekst instytucjonalny na rynku pracy | |
| Przepisy dotyczące ochrony zatrudnienia | (–) |
| Przepisy dotyczące umów na czas określony | 0 |
| Przepisy regulujące pracę tymczasową | 0 |
| Stopa zastąpienia brutto | 0 |
| d) Otoczenie makroekonomiczne | |
| Wskaźnik bezrobocia | (+) |
| Wydatki na aktywną politykę rynku pracy | 0 |
| Wzrost PKB | 0 |

Wyniki w Tabeli 3 podsumowują najważniejsze ustalenia Kluve (2010). Metaanaliza w Card i in. (2010) wykorzystuje podobną metodę, korzystając jednak z nowego zestawu danych obejmującego 199 ewaluacji wpływu. W ostatnim fragmencie niniejszej części zebrano najważniejsze ustalenia i tendencje dotyczące skuteczności Aktywnych Polityk Rynku Pracy na podstawie dwóch wyżej wymienionych badań, wcześniejszego kompleksowego przeglądu w Heckman i in. (1999) oraz metaanalizy tylko dla Stanów Zjednoczonych przeprowadzonej przez Greenberg i in. (2003).

1. Metabadania pokazują raczej wyraźną tendencję dla skuteczności polityki w zależności od typu programu: a) programy szkoleniowe są skuteczne w niewielkim stopniu (lecz widać w ich przypadku potencjalny wpływ w dłuższym okresie, zob. poniżej); b) subsydiowanie wynagrodzeń zazwyczaj daje pozytywne skutki; c) tworzenie miejsc pracy w sektorze publicznym negatywnie oddziałuje na szanse uczestników na zatrudnienie; d) pomoc w poszukiwaniu pracy jest skuteczna, w większości przypadków także pozytywnie wypada w jej przypadku stosunek kosztów do korzyści.

2. Pochodzące z krajów OECD informacje dotyczące Aktywnych Polityk Rynku Pracy pokazują, że grupą docelową, której szczególnie trudno udzielić skutecznej pomocy są ludzie młodzi. W porównaniu do

programów skierowanych do osób dorosłych, w przypadku programów dla ludzi młodych istnieje znacznie mniejsze prawdopodobieństwo, że przyniosą one pozytywne skutki.

3. Ta utrwalona tendencja różni się znacznie od danych z innych regionów, szczególnie regionu Ameryki Łacińskiej i Karaibów, gdzie programy dla osób młodych zazwyczaj odnoszą większe sukcesy (zob. np. Ibarrarán i Rosas 2009).

4. W kwestii przyczyn nieskuteczności programów skierowanych do młodych ludzi w krajach OECD można tylko spekulować: oficjalne systemy szkolnictwa w tych krajach są zazwyczaj dobrze rozwinięte.

5. Grupa młodych ludzi dorosłych, którzy są (długotrwale) bezrobotni składa się z osób o niskich kwalifikacjach i umiejętnościach oraz osób, które porzuciły szkołę i nie mają średniego wykształcenia.

6. Wśród pracowników, którzy przeciętnie mają wysokie kwalifikacje i z których duża część ma wyższe wykształcenie, ludzie młodzi objęci Aktywną Polityką Rynku Pracy stanowią grupę w bardzo niekorzystnej sytuacji i objęcie ich wsparciem może być trudne. Na tle innych regionów kraje rozwinięte odnotowują największą negatywną liniową korelację między poziomem zdobytego wykształcenia a zagrożeniem bezrobociem.

7. Nieliczne programy na rzecz ludzi młodych, które zdają się działać, to te, które są szeroko zakrojone i intensywnie wdrażane. Dwa najważniejsze przykłady skutecznych programów skierowanych do ludzi młodych w krajach OECD to Job Corps w Stanach Zjednoczonych (Schochet i in. 2008) i New Deal for Young People w Wielkiej Brytanii (NDYP; e.g. van Reenen 2003, Dorsett 2006). Zostaną one bardziej szczegółowo omówione jako przykłady dobrych praktyk w części 4 niniejszego tekstu. O ile obydwa programy różnią się w wielu szczegółach, najważniejsze ich cechy, czyli kompleksowość i duża intensywność, są wspólne.

8. W każdym przypadku elementy programu obejmują wsparcie w szukaniu pracy, poradnictwo, szkolenie i usługi pośrednictwa pracy. Podobne pozytywne rezultaty dotyczące kompleksowych programów zaobserwowano także poza OECD, chodzi w szczególności o programy „Jóvenes" w kilku krajach Ameryki Łacińskiej (Ibarrarán i Rosas 2009; Urzúa i Puentes 2010).

9. W większości przypadków, w których Aktywna Polityka Rynku Pracy na rzecz ludzi młodych nie przynosi pozytywnych skutków, znaczenie mogą mieć inne czynniki: dwupoziomowe rynki pracy, na których ich uczestnicy są raczej dobrze chronieni, co powoduje utrudniony dostęp dla osób z zewnątrz, w szczególności osób młodych i o niskich kwalifikacjach (jako przykłady zazwyczaj wymienia się Francję i Hiszpanię). To strukturalne zjawisko może także odgrywać ważną rolę w wychodzeniu z kryzysu finansowego, ponieważ z uwagi na dużą liczbę bezrobotnych młodych ludzi, w wielu krajach grupa młodych ludzi potrzebujących pomocy obejmuje nie tylko osoby o niskich kwalifikacjach oraz młodzież NEET (ang. *Not in Education, Employment or Training* – osoba nieucząca się, niepracująca ani nieszkoląca się), ale także wiele osób o wysokich kwalifikacjach i większej motywacji.

10. Kluve (2010) pokazuje, że Aktywne Polityki Rynku Pracy zazwyczaj są mniej skuteczne na rynkach, gdzie przepisy dotyczące ochrony zatrudnienia są surowsze.

11. Programy wsparcia w szukaniu pracy, tj. usługi i sankcje, są często skuteczne. Ponieważ są to zazwyczaj względnie mało kosztowne interwencje, istnieje także większe prawdopodobieństwo, że będą opłacalne.

12. Programy subsydiowania wynagrodzeń także zdają się bardzo skuteczne, w przeciwieństwie do zatrudnienia w sektorze publicznym. Ten ostatni środek często powoduje nawet negatywne skutki, prawdopodobnie z powodu stygmatyzacji oraz/lub rodzajów wykonywanych w ramach programu prac publicznych, które nie są nawet w stanie utrzymać kapitału ludzkiego uczestników sprzed interwencji.

13. Pytania dotyczące subsydiowania wynagrodzeń, są następujące: a) czy występuje jakikolwiek pozytywny wpływ na zatrudnienie w długim okresie? i b) czy można wyeliminować efekty zniekształcające ogólną równowagę, takie jak zastępowanie (subsydiowany pracownik zastępuje niesubsydiowanego pracownika), przesunięcie (firmy z subsydiowanymi pracownikami mogą przejmować rynek kosztem firm niesubsydiowanych) i efekt bezwładności – deadweight (zatrudnienie pracownika nastąpiłoby także bez otrzymania subsydium). Jak dotąd kwestie te nie zostały dostatecznie przekonująco omówione w ewaluacji programów. Inną

kwestią dotyczącą subsydiowania wynagrodzeń jest wzrost prawdopodobieństwa zakłóceń na rynku pracy równoległy do wzrostu zakresu interwencji. Oznacza to, że subsydiowanie wynagrodzeń może być właściwe dla określonych grup celowych w odpowiednio zdefiniowanych kontekstach (sektory, regiony), ale nie jest dobrym rozwiązaniem dla polityk publicznych prowadzonych na szeroką skalę.

14. Zazwyczaj wpływ programu nie stawał się bardziej pozytywny wraz z upływem czasu. Jak pokazują dwie metaanalizy, jest tak zarówno w przypadku Stanów Zjednoczonych (Greenberg i in. 2003), jak i programów na świecie (Card i in. 2010, większość obserwacji na podstawie danych pochodzących z krajów OECD). Ponieważ badania dla Stanów Zjednoczonych opierają się na randomizowanych próbach kontrolnych, stwierdzenie to oznacza prawdopodobnie, że programy rzeczywiście nie przynosiły lepszych skutków wraz z upływem czasu. Z drugiej strony dla większej próby badań ewaluacyjnych z całego świata występuje tendencja, zgodnie z którą programy w istocie przynosiły do pewnego stopnia lepsze skutki w czasie, ale dla zagregowanych danych można uznać, iż tendencję tę neutralizuje fakt, że w przypadku wyników wczesnych ewaluacji programów opartych na ograniczonej ilości danych i ograniczonej metodologii istniało większe prawdopodobieństwo pokazywania nadmiernie pozytywnych rezultatów, podczas gdy wyniki późniejszych ewaluacji wykorzystujących duże ilości danych i rygorystyczne metody były bardziej zbliżone do „prawdziwych" skutków programu.

15. Z perspektywy uśrednionego wyniku dla dotąd przeprowadzonych ewaluacji programy szkoleniowe z zakresu rynku pracy są umiarkowanie skuteczne. Ponieważ szkolenia podnoszące kwalifikacje są najbardziej popularnym, najczęściej stosowanym i teoretycznie najbardziej obiecującym działaniem – ze względu na element generowania kapitału ludzkiego – warto spojrzeć na dwie prawidłowości zaobserwowane w ostatnim badaniu dotyczącym szkoleń.

16. Po pierwsze, wpływ szkolenia może uwidocznić się w długim okresie, czasem nawet po bardzo długim czasie (Lechner i in. 2011). Istnieje coraz więcej dowodów na to, że najskuteczniejszą sekwencją programów dla bezrobotnych (w krajach OECD) jest po pierwsze, intensywna pomoc w szukaniu pracy (wraz z doradztwem i monitoringiem), przynosząca efekty w krótkim czasie i w drugim etapie, szkolenie przynoszące skutki w średnim i długim okresie ze względu na akumulację kapitału ludzkiego (Hotz i in. 2006).

17. Po drugie, ostatnie badania pokazują, że programy szkoleniowe osiągają swoją maksymalną skuteczność w okresie 4-5 miesięcy i że dłuższe interwencje nie wywierają dodatkowego wpływu na zwiększenie szans na zatrudnienie uczestników po interwencji (Kluve i in. 2011). Jest to przypadek programów szkoleniowych, które nie kończą się dyplomem zawodowym. Programy szkoleniowe, które prowadzą do uzyskania takiego dyplomu są zazwyczaj dużo dłuższe (do dwóch lat szkolenia), a także wykazują pozytywne skutki (Lechner i in. 2011).

18. Jeden ogólny wniosek z badań Aktywnych Polityk Rynku Pracy mówi, że wczesne interwencje są lepsze niż późne. Wniosek ten ma uzasadnienie ekonomiczne (wcześniejsze kształcenie daje efekty w dłuższym okresie) oraz uzasadnienie w postaci znaczenia budowania potencjału, w tym umiejętności społecznych, przed osiągnięciem wieku dorosłego (Urzúa i Puentes 2010).

19. Dla skuteczności kompleksowych programów (programy Job Corps, New Deal for the Young People, Jóvenes) istotne jest także budowanie zintegrowanych struktur zdobywania umiejętności. Jednym z aspektów tej kwestii jest instytucjonalny związek między programami szkolenia zawodowego a oficjalnym systemem edukacji.

20. Niezależnie od prawidłowości wskazanych powyżej, literatura poświęcona ewaluacji Aktywnych Polityk Rynku Pracy pokazuje, że konieczne jest wzmacnianie wysiłków w kierunku prowadzenia dalszych, ciągłych ewaluacji. Pozyskane dotąd dowody przyczyniły się znacznie do zrozumienia, który typ Aktywnych Polityk Rynku Pracy wydaje się działać. Jednocześnie wiele pytań pozostaje otwartych. Na przykład większość ewaluacji pokazuje szacunki wpływu w krótkim i średnim okresie, a niewiele wiadomo na temat długoterminowego oddziaływania Aktywnych Polityk Rynku Pracy. Dodatkowo, przydatne byłyby dalsze informacje na temat dokładnego składu programów złożonych z wielu elementów. Dalszego zbadania wy-

maga także zależność między długością okresu wdrażania a skutecznością programu. Te przykłady otwartych pytań pokazują znaczenie dalszego prowadzenia ewaluacji wpływu Aktywnych Polityk Rynku Pracy.

21. Zatem, prowadzenie ewaluacji pojedynczych programów jest istotne, aby odpowiednio informować polityków oraz osoby odpowiedzialne za wdrażanie konkretnego programu, a jednocześnie przyczynia się do bardziej ogólnego procesu poszukiwania wiedzy.

## Dwa przykłady „najlepszych praktyk"

### Przykład 1: The New Deal for Young People (NDYP), Wielka Brytania

Mimo raczej zaskakujących rezultatów Aktywnej Polityki Rynku Pracy skierowanej do ludzi młodych, dwa programy można uznać za sukces: The New Deal for Young People w Zjednoczonym Królestwie i Job Corps w Stanach Zjednoczonych. Zostaną one omówione w niniejszej części.

W 1998 roku rząd brytyjski wprowadził program dla ludzi poniżej 25. roku życia New Deal for Young People (NDYP) jako kluczowy element swojej strategii „od zasiłku do zatrudnienia". Celem jest wsparcie młodych bezrobotnych w znalezieniu pracy i zwiększenie ich szans na zatrudnienie. Udział jest obowiązkowy dla wszystkich osób między 18. a 24. rokiem życia, które otrzymują zasiłek dla bezrobotnych (ang. *Jobseeker's allowance,* JSA) przez sześć miesięcy lub dłużej.

W ramach programu NDYP młodzi ludzie najpierw uzyskują wsparcie w szukaniu pracy, a potem otrzymują ofertę szkoleniową lub programy alternatywne. Na New Deal składają się trzy etapy: najpierw uczestnicy przechodzą przez etap wstępny, podczas którego są przydzielani do osobistego doradcy, który udziela im wsparcia w szukaniu pracy. Jeżeli młoda osoba bezrobotna nie jest w stanie znaleźć niesubsydiowanego zatrudnienia i pozostaje na zasiłku po zakończeniu etapu wstępnego (do 4 miesięcy), stosowane jest jedno z czterech działań New Deal: 1) kształcenie i szkolenie w pełnym wymiarze czasu, 2) subsydiowanie zatrudnienia („działanie dla pracodawcy"), 3) zatrudnienie w sektorze publicznym („środowiskowa grupa zadaniowa") lub 4) wolontariat. Wszystkie działania trwają do sześciu miesięcy, z wyjątkiem kształcenia i szkolenia w pełnym wymiarze czasu, które może trwać do 12 miesięcy. W przypadku wszystkich pozostałych działań pracodawcy są zobowiązani do oferowania kształcenia i szkolenia co najmniej przez jeden dzień w tygodniu, co powinno także prowadzić do uzyskania oficjalnego potwierdzenia zdobytego wykształcenia. Ostatni, trzeci etap to etap końcowy z kontynuacją poradnictwa i wsparcia dla osób pozostających na zasiłku, po tym, jak skierowane do nich działanie dobiegło końca.

Wyniki ewaluacji pokazują, że nastąpił znaczny wzrost zatrudnienia spowodowany programem New Deal oraz że zyski dla społeczeństwa przewyższają koszty. Bezrobotni młodzi mężczyźni mają o 20% większe szanse na znalezienie pracy w wyniku programu. Efekt ten w dużej części może być spowodowany zastosowaniem subsydium do wynagrodzenia, ale co najmniej jedna piąta tego efektu spowodowana jest intensywniejszym poszukiwaniem pracy. Znaczenia bardziej intensywnego, dłużej trwającego kształcenia i szkolenia nie można w pełni ocenić, ponieważ dane z długoterminowej oceny nie zostały jeszcze poddane ewaluacji. Intensywne inwestycje w kapitał ludzki mogą jednak przynieść zyski szczególnie w długim okresie, a szkolenie w ramach New Deal jest bardziej intensywnym działaniem niż inne. Ponieważ wsparcie w szukaniu pracy w ramach New Deal jest działaniem bardziej opłacalnym niż inne działania w ramach Aktywnych Polityk Rynku Pracy, New Deal okazuje się najmniej kosztowną wszechstronną interwencją dla młodych ludzi w krajach OECD. Koszt w przeliczeniu na beneficjenta waha się od 734 USD do 1277 USD (wartości z roku 1999). Ponadto, koszt utworzonego miejsca pracy nie przekracza 6500 USD przy pośrednictwie dla średnio 17 250 osób rocznie.

Programowi New Deal towarzyszyły istotne reformy w zakresie świadczenia usług. Od 2001 r. nowe „Jobcentre Plus" (pol. „Centrum Pracy plus") świadczy usługi, za które wcześniej odpowiadały Employment

Service i Benefits Agency, i staje się „jednym okienkiem" dla spraw związanych z zatrudnieniem, doradztwem w zakresie zasiłków i wsparcia. Jego celem jest usprawnienie oceny i świadczeń, aby klienci otrzymywali odpowiednie, dopasowane do swoich potrzeb usługi. Z ewaluacji Jobcentre Plus wynika, że dzięki jego funkcjonowaniu udało się zwiększyć zatrudnienie w grupie docelowej (Corkett i in. 2005).

### Przykład 2: Job Corps

Job Corps jest programem ogólnokrajowym realizowanym przez Departament Pracy Stanów Zjednoczonych, który rozpoczęto w 1964 roku celem wyszkolenia kwalifikujących się młodych ludzi dorosłych w zakresie umiejętności sprzyjających zatrudnieniu i niezależności oraz umożliwienia im wartościowego zatrudnienia lub dalszego kształcenia. Do najważniejszych kryteriów kwalifikowalności należy wiek od 16 do 24 lat, legalny pobyt w Stanach Zjednoczonych, niekorzystna sytuacja ekonomiczna oraz potrzeba dodatkowego kształcenia, szkolenia lub kwalifikacji zawodowych. Program ma zapewniać bezpieczne środowisko edukacyjne wolne od narkotyków. Uczestnicy zapisują się na 30-tygodniowy kurs, aby nauczyć się zawodu, uzyskać dyplom ukończenia szkoły średniej lub dyplom potwierdzający wykształcenie ogólne oraz otrzymać wsparcie w znalezieniu pracy. Cykl programu, który przechodzą uczestnicy składa się z czterech części: wejście w cykl, przygotowanie do kariery, rozwój kariery i zmiana zawodu. Są to cztery elementy programu, wśród których drugi jest etapem profilowania, trzeci – głównym etapem szkoleniowym, a czwarty polega na pośrednictwie pracy. Uczestnicy programu przez cały okres szkolenia otrzymują miesięczne świadczenia i korzystają z doradztwa zawodowego i wsparcia dla zmiany, przez okres do 12 miesięcy od zakończenia edukacji.

Job Corps jest największym i najszerzej zakrojonym programem edukacyjno-szkoleniowym dla młodych w niekorzystnej sytuacji życiowej w Stanach Zjednoczonych, z którego korzysta ponad 60 000 nowych uczestników rocznie i który kosztuje 1,5 mld USD. Ze względu na znaczne koszty zaangażowane w program w celu zbadania jego skuteczności, Departament Pracy Stanów Zjednoczonych sfinansował badanie Job Corps, które przeprowadzono w latach 1993-2004. Wyniki zostały opublikowane w serii raportów oraz podsumowane w artykule Schochet i in. (2008). Ewaluacja wpływu opierała się na randomizowanej próbie kontrolnej obejmującej około 9400 młodych ludzi w grupie objętej programem i prawie 6000 młodych ludzi w grupie kontrolnej.

Ewaluacja pokazała między innymi, że dzięki programowi rozwinęły się usługi szkoleniowe i edukacyjne świadczone na rzecz ludzi młodych. Łącznie ich liczba wzrosła o około 1000 godzin, co odpowiada 10 miesiącom zwykłego roku szkolnego. Jednocześnie Job Corps znacznie poprawił umiejętność czytania i pisania. Jeżeli chodzi o najważniejsze skutki dla rynku pracy, zaobserwowano statystycznie istotne wzrosty zarobków w ciągu pierwszych dwóch lat po wyjściu z programu. Różnice w zarobkach między grupą objętą programem a grupą kontrolną nie utrwaliły się w jednak kolejnych latach. Jedyna podgrupa, w której różnice te były trwałe, to ludzie w wieku 20-24 lata. Grupa ta stanowi około jedną czwartą uczestników Job Corps i zazwyczaj pozostaje w programie dłużej, jest bardziej zmotywowana i zdyscyplinowana.

Analiza kosztów i korzyści programu jest szczególnie ciekawa, biorąc pod uwagę, że Job Corps przyczynia się także znacznie do zmniejszenia przestępczości we wszystkich podgrupach. Schochet i in. (2008) wnioskują, że ponieważ wzrost zarobków nie utrzymuje się, płynące z programu Job Corps korzyści dla społeczeństwa są mniejsze niż koszty programu. Autorzy szacują, że koszty Job Corps przekraczają jego korzyści dla społeczeństwa o około 10 300 USD na uczestnika. Korzyści płynące ze wzrostu zarobków (1119 USD), stosowania innych programów i usług w mniejszym zakresie (2186 USD) i mniejszej przestępczości (1240 USD) są rzeczywiście niewielkie w porównaniu z kosztami. Program wydaje się jednak opłacalny dla podgrupy młodych ludzi w wieku 20-24 lata, których większe zarobki utrzymują się nawet przez 3 do 8 lat po wyjściu z programu. Ponadto z perspektywy uczestników programu korzyści przewyższają koszty.

**Jochen Kluve** od 2011 roku pełni funkcję profesora Ekonomiki Pracy na Uniwersytecie Humboldta w Berlinie. Studiował ekonomię w Heidelbergu, Dublinie i Amsterdamie, w 2002 został doktorem Uniwersytetu w Heidelbergu, następnie został zatrudniony jako pracownik naukowy na Uniwersytecie w Berkeley. Od 2003 roku pracuje dla RWI (Rheinisch-Westfälisches Institut), instytutu zajmującego sie badaniami ekonomicznymi, z siedzibą w Essen, gdzie zajmował stanowiska szefa jednostki zajmującej się badaniami rynku pracy oraz od 2007 – szefa oddziału w Berlinie. Jego badania skupiają się na ewaluacji polityk rynku pracy, a główne zainteresowania badawcze to: metody oceny wpływu i ich zastosowanie w programach edukacyjnych i szkoleniowych, zarówno w krajach rozwiniętych, jak i rozwijających się/rynkach wschodzących. Brał udział w projektach badawczych np. dla kilku ministerstw niemieckich, Komisji Europejskiej, Banku Światowego czy Międzyamerykańskiego Banku Rozwoju. Jego prace były publikowane w wielu czasopismach naukowych, takich jak: The Economic Journal, Labour Economics, The Journal of Development Effectiveness oraz The Journal of the Royal Statistical Society (Series A).

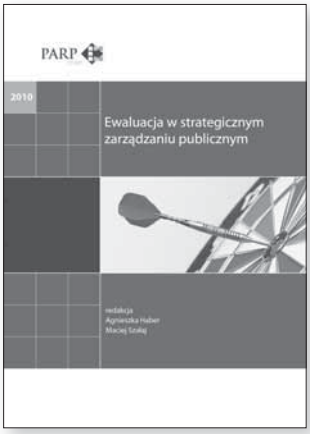
## Bibliografia


- *Betcherman G., Olivas K., Dar A., Impacts of Active Labor Market Programs: New Evidence from Evaluations with Particular Attention to Developing and Transition Countries, Social Protection Discussion Paper Series 0402, Waszyngton, Bank Światowy, 2004.*
- *Card D., Kluve J., Weber A., Active Labour Market Policy Evaluations: A Meta-analysis, "The Economic Journal", 120, 2010, s. F452-F477.*
- *Corkett J., Bennett S., Stafford J., Frogner M., Shrapnell K., Jobcentre Plus evaluation: summary of evidence, Department for Work and Pensions Research Report No 252, UK, 2005.*
- *Dolton P., O'Neill D., Unemployment duration and the Restart effect: some experimental evidence, "The Economic Journal" 106, 1996, s. 387-400.*
- *Dorsett R., The new deal for young people: effect on the labour market status of young men, "Labour Economics", 13, 2006, s. 405-422.*
- *Greenberg D.H., Michalopoulos C., Robins P.K., A Meta-Analysis of Government-Sponsored Training Programs, "Industrial and Labor Relations Review", 57 (1), 2003, s. 31-53.*
- *Heckman J.J., LaLonde R.J., Smith J.A., The economics and econometrics of active labour market programs, [w:] Ashenfelter, O., Card, D. (red.). "Handbook of Labor Economics", 3. Elsevier, Amsterdam, 1999.*
- *Hotz V.J., Imbens G.W., Klerman J.A., Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program, "Journal of Labor Economics", 24, 2006, s. 521-566.*
- *Ibarraran P., Rosas D. Evaluating the Impact of Job Training Programs in Latin America: Evidence from IDB funded operations, "Journal of Development Effectiveness", 1(2), 2009, s. 195-216.*
- *Jacobi L., Kluve J., Before and After the Hartz Reforms: The Performance of Active Labour Market Policy in Germany, "Journal for Labour Market Research", 40, 2007, s. 45-64.*
- *Kluve J., The effectiveness of European active labor market programs, "Labour Economics", 17, 2010, s. 904-918.*
- *Kluve J., Card D., Fertig M., Góra M., Jacobi L., Jensen P., Leetmaa R., Nima L., Patacchini E., Schaffner S., Schmidt C.M., van der Klaauw B., Weber A., Active Labor Market Policy in Europe: Performance and Perspectives, Berlin, Springer, 2007.*
- *Kluve J., Schneider H., Uhlendorff A., Zhao Z., Evaluating continuous training programs using the Generalized Propensity score, "Journal of the Royal Statistical Society Series A", 2012.*
- *Lechner M., Miquel R., Wunsch C., Long-Run Effects of Public Sector Sponsored Training in West Germany, "Journal of the European Economic Association", 2011.*
- *Martin J.P., Grubb D., What works and for whom: a review of OECD countries' experiences with active labour market policies, "Swedish Economic Policy Review", 8, 2001, s. 9-56.*
- *OECD, Activating the unemployed: what countries do, rozdział 5 [w:] "Employment Outlook OECD", OECD: Paryż, 2007.*
- *Scarpetta S., Sonnet A., Manfredi T., Rising Youth Unemployment During The Crisis – How to prevent negative long-term consequences on a generation?, "Social, Employment and Migration Working Papers OECD", nr 106, Publikacje OECD: Paryż, 2010.*
- *Schochet P.Z., Burghardt J., McConnell S., Does Job Corps Work? Impact Findings from the National Job Corps Study, "American Economic Review", 98, 2008, s. 1864-1886.*
- *Urzúa S., Puentes E., La evidencia del impacto de los programas de capacitación en el desempeño en el mercado laboral, Banco Interamericano de Desarrollo – Unidad de Mercados Laborales del Sector Social, Notas Técnicas, 268, BID: Washington, DC, 2010.*
- *Van Reenen, J., Active labor market policies and the British new deal for the young unemployed in context, Dokument roboczy NBER, 9576, 2003.*

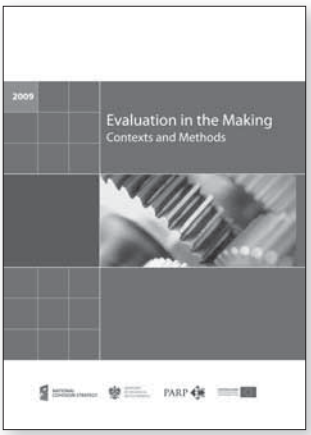
## W ramach serii *Ewaluacja* ukazały się następujące publikacje:


http://www.parp.gov.pl/index/more/25949


http://www.parp.gov.pl/index/more/22858




http://www.parp.gov.pl/index/more/19489




http://www.parp.gov.pl/index/more/14819


http://www.parp.gov.pl/index/more/12416


http://www.parp.gov.pl/index/more/13335

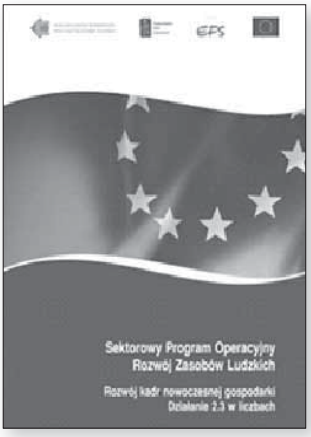

http://www.parp.gov.pl/index/more/9658


http://www.parp.gov.pl/index/more/2046

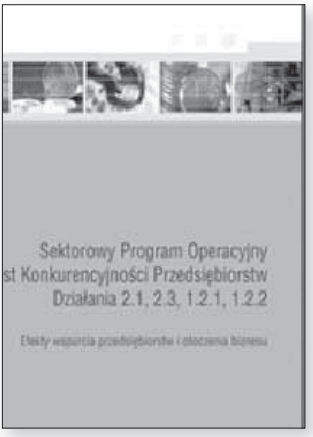

http://www.parp.gov.pl/index/more/24238


http://www.parp.gov.pl/index/more/19735


http://www.parp.gov.pl/index/more/9850




http://www.parp.gov.pl/index/more/5474




http://www.parp.gov.pl/index/more/5475

**Polska Agencja Rozwoju Przedsiębiorczości** (PARP) jest agencją rządową, która od 2000 roku wspiera przedsiębiorców. Celem działania PARP jest rozwój małych i średnich firm w Polsce – powstawanie nowych podmiotów, podnoszenie kwalifikacji i wzrost potencjału, wzmocnienie pozycji konkurencyjnej w oparciu o innowacyjność i nowoczesne technologie, kształtowanie przyjaznego otoczenia biznesowego, tworzenie warunków do prowadzenia działalności gospodarczej. Realizując działania wspierające przedsiębiorców (a także: instytucje otoczenia biznesu, jednostki samorządu terytorialnego, państwowe jednostki budżetowe, uczelnie), PARP korzysta ze środków budżetu państwa oraz funduszy europejskich. Zarówno w okresie przedakcesyjnym, jak i po wejściu przez Polskę do Unii Europejskiej, PARP oferowała przedsiębiorcom wsparcie finansowe i szkoleniowo-doradcze. W latach 2007–2015 Agencja jest odpowiedzialna za realizację działań w ramach trzech programów operacyjnych: **Innowacyjna Gospodarka**, **Kapitał Ludzki** oraz **Rozwój Polski Wschodniej** oraz aktywnie uczestniczy w opracowaniu założeń programów pomocowych w perspektywie finansowej 2014–2020.

PARP posiada unikalne doświadczenie nie tylko w przekazywaniu pomocy unijnej przedsiębiorcom. Od kilku lat w Agencji działa **Ośrodek Badań nad Przedsiębiorczością**, którego zadaniem jest prowadzenie badań z zakresu przedsiębiorczości, innowacyjności, zasobów ludzkich i usług wspierających prowadzenie działalności gospodarczej. W oparciu o ich wyniki powstają założenia dla kolejnych programów pomocowych, które odpowiadają na zidentyfikowane potrzeby przedsiębiorców.

Aby pomoc była skuteczna, przedsiębiorca musi mieć łatwy dostęp do informacji na jej temat. PARP zainicjowała utworzenie **Krajowego Systemu Usług dla MŚP (KSU)**. KSU oferuje doradztwo dla firm na każdym etapie prowadzenia działalności: od rejestracji działalności, poprzez sprawne prowadzenie i zarządzanie firmą, aż po zawieszenie lub zakończenie działalności. Wszystkie ośrodki KSU (około 170) działają na podstawie wypracowanych Standardów Usług, dzięki czemu przedsiębiorca może być pewien, że otrzyma usługę najwyższej jakości. Przedsiębiorca chcący skorzystać z usługi doradztwa biznesowego ma do wyboru: Punkty Konsultacyjne KSU, ośrodki Krajowej Sieci Innowacji KSU oraz ośrodki realizujące usługi w zakresie ochrony środowiska, szybkiej optymalizacji kosztów, a także ośrodki testujące nowe usługi pilotażowe. Dodatkowo może otrzymać pożyczkę lub poręczenie ze współpracującego funduszu. Wiele organizacji tworzących KSU współpracuje jednocześnie z innymi znanymi sieciami, takimi jak Enterprise Europe Network (konsorcja dawnych Centrów Euro Info, EIC i Ośrodków Przekazu Innowacji, IRC).

Działający przy PARP ośrodek sieci **Enterprise Europe Network** daje szansę przedsiębiorcom na skorzystanie z możliwości rynku ogólnoeuropejskiego. Ośrodek oferuje nieodpłatne, kompleksowe usługi obejmujące informacje, szkolenia i doradztwo, przede wszystkim z zakresu prawa i polityk Unii Europejskiej, prowadzenia działalności gospodarczej w Polsce i za granicą, dostępu do źródeł finansowania, internacjonalizacji przedsiębiorstw, transferu technologii oraz udziału w programach ramowych UE.

PARP stale dopasowuje ofertę informacyjno-doradczą do zmieniających się potrzeb przedsiębiorców oraz pojawiających się nowych kanałów komunikacji. Obecnie Agencja dysponuje kilkunastoma **specjalistycznymi portalami internetowymi i społecznościowymi**, oferującymi szkolenia e-learningowe, e-booki, transmisje ze spotkań szkoleniowych i konferencji, informacje na temat możliwości ubiegania się o wsparcie, bazy wiedzy, publikacje, wyniki badań. Z informacji i narzędzi zawartych we wszystkich portalach PARP dostępnych za pośrednictwem głównego portalu Agencji **www.parp.gov.pl** korzysta blisko milion internautów miesięcznie.

Osoby zainteresowane uzyskaniem dostępnych w PARP informacji na temat programów wsparcia dla przedsiębiorców oraz instytucji otoczenia biznesu, mogą skorzystać z infolinii prowadzonej w ramach **Informatorium** PARP. Konsultanci udzielają informacji telefonicznie i mailowo oraz biorą udział w spotkaniach z zainteresowanymi osobami.

Zapraszamy do skorzystania z naszych usług!

**PARP**
ul. Pańska 81/83, 00-834 Warszawa
tel.: + 48 22 432 80 80
faks: + 48 22 432 86 20
**biuro@parp.gov.pl**
**www.parp.gov.pl**

**Punkt informacyjny PARP**
tel.: + 48 22 432 89 91-93
0 801 332 202
**info@parp.gov.pl**

ISBN 978-83-7633-272-7