



INSTYTUT BADAŃ
EDUKACYJNYCH
Państwowy Instytut Badawczy



Rzeczpospolita
Polska



Ministerstwo
Edukacji Narodowej

Kinga Wysieńska-Di Carlo

HISTORIA PEWNEJ REFORMY, czyli opowieść o dobrym i złym wykorzystywaniu dowodów naukowych w procesach decyzyjnych w oświacie



Spis treści

Historia pewnej reformy, czyli opowieść o dobrym i złym wykorzystywaniu dowodów naukowych w procesach decyzyjnych w oświacie	4
---	----------

Problem 1

Jaki jest stan obecny? Co dokładnie chcemy zmienić? Jakie są możliwe rozwiązania?	9
---	---

Problem 2

Jakie są dowody, że dana polityka lub interwencja przyniosła pozytywne efekty w przeszłości?	15
--	----

Problem 3

Jakie mechanizmy przyczynowe zostały zidentyfikowane i czy pozwalają one na stwierdzenie, że ten sam proces zadziała tu i teraz?	30
--	----

Problem 4

Jakie dodatkowe warunki muszą być spełnione, aby dany mechanizm zadziałał tu i teraz?	32
---	----

Podsumowanie

Korzyści z podejścia opartego na dowodach	35
---	----

Aneks 1

Jak nie dać się nabrać na nieprawdę?	42
--	----

Aneks 2

Poziom decyzji a wykorzystanie badań	45
--	----

Aneks 3

Gdzie szukać wiarygodnych dowodów?	48
--	----

Słowniczek	50
-------------------------	-----------

Bibliografia.....	52
--------------------------	-----------

Strony internetowe badań.....	55
--------------------------------------	-----------

Historia pewnej reformy, czyli opowieść o dobrym i złym wykorzystywaniu dowodów naukowych w procesach decyzyjnych w oświacie

Wyobraźmy sobie, że przy poparciu wszystkich sił politycznych i powszechnej akceptacji społecznej wprowadzamy dalekosiężną reformę oświatową. Wiąże się ona z wysokimi kosztami, ale mamy badanie naukowe, które sugeruje, że proponowane działania mogą istotnie poprawić wyniki szkolne dzieci i młodzieży. Kilkanaście lat i miliony złotych później oczekiwanej poprawy jednak nie widać. Z reformy po cichu się wycofujemy.

Czy taki scenariusz jest możliwy?

Opisana historia wydarzyła się naprawdę.



W 1996 roku Kalifornia wprowadziła program Class-Size Reduction (CSR). Miał on poprawić wyniki uczniów i uczennic dzięki zmniejszeniu liczebności klas przedszkolnych i wczesnoszkolnych do maksymalnie 20 osób. Program kosztował około 1,5 miliarda USD rocznie, a łączne wydatki sięgnęły 25 miliardów USD. Podobne reformy wprowadziły 24 stany USA, a w 2001 roku środki na zmniejszenie liczebności klas zostały wpisane do federalnego prawa oświatowego No Child Left Behind (Chingos, 2011). W 2013 roku program w Kalifornii wygaszono, choć projekty ewaluacyjne już dużo wcześniej wskazywały, że nie przynosi on spodziewanej poprawy rezultatów.



Jak to możliwe, że kosztowna i powszechnie popierana reforma, wspierana przez nauczycielki i nauczycieli, a także rodziców i osoby stanowiące prawo, nie przyczyniła się do znaczącej poprawy wyników szkolnych? Jakie błędy popełniono? Odpowiedź kryje się w jakości i interpretacji wykorzystanych dowodów naukowych oraz sposobie wdrożenia reformy. O tym jest ten materiał.

Podejmowanie decyzji dotyczących polityk i programów edukacyjnych wiąże się z ogromną odpowiedzialnością. Dlatego osoby je kształtujące wiedzą, że wykorzystywanie danych i badań naukowych jest kluczowe.

Badania i dowody naukowe stanowią najlepsze źródło odpowiedzi na pytania „czy” i „dlaczego” jedne programy lub interwencje są bardziej efektywne niż inne. Ich wykorzystanie zwiększa szanse na wdrożenie takich praktyk, które przynoszą korzyści wszystkim stronom procesu edukacyjnego – dzieciom, młodym dorosłym, rodzicom, kadrze pedagogicznej – a w dłuższej perspektywie także całemu społeczeństwu.

Podejmowanie decyzji na podstawie dowodów naukowych pozwala również na bardziej racjonalne wydawanie środków publicznych. **Decyzje oparte na dowodach** (*evidence-based decisions*) uwzględniają wyniki weryfikacji różnych praktyk, co umożliwia wybór tych, które przynoszą największe korzyści przy określonym poziomie inwestycji. Z kolei **decyzje oparte na opiniach** (*opinion-based decisions*) często okazują się nietrafne, ponieważ opinie nie są poddawane rygorystycznym testom i są podatne na błędy poznawcze (por. [Aneks 1](#)).

Wybór wiarygodnych dowodów naukowych spośród dostępnych danych, statystyk i opracowań stanowi wyzwanie. Materiał ten powstał po to, aby wesprzeć osoby kształtujące lub wdrażające polityki i programy oświatowe, a także uczestniczące w debacie na temat edukacji.



W debatach edukacyjnych często pojawiają się następujące terminy: **dane, badania, analizy, dowody**. Mają one różne znaczenie i nie powinny być mylone.



Dane to nic innego jak zbiór obserwacji lub informacji wyrażonych w formie tekstowej, liczbowej lub wizualnej. Przykładowe dane to: daty urodzenia osób w klasie, oceny na świadectwie, odpowiedzi na pytania kwestionariusza. Znaczenie i jakość danych można ocenić dopiero w kontekście konkretnego badania lub analizy.



Badanie to pewna struktura, która określa, jakie dane, dotyczące kogo, w jaki sposób i po co będą rejestrowane lub analizowane. Badanie zaczyna się od pytania (lub inaczej problemu badawczego). To ono wyznacza cel i metodę zbierania lub analizy informacji. Zatem tylko w kontekście badania możemy sensownie oceniać wartość i znaczenie danych (więcej o typach badań piszemy w kolejnych częściach materiału).



Analiza to sposób przetwarzania i interpretacji danych, który pozwala nadać im znaczenie. Może mieć charakter jakościowy lub ilościowy. Typ analizy wynika z celów i metodologii badania. W kontekście polityk oświatowych najważniejsze są analizy ilościowe, ponieważ pozwalają na opis całych populacji oraz obiektywne porównania efektywności różnych rozwiązań. Mogą one mieć charakter:

- **opisowy** – ich celem jest podsumowanie zebranych informacji i określenie głównych cech badanej grupy, np. za pomocą rozkładów częstości czy miar takich jak średnia, mediana, kwartyle lub odchylenie standardowe;
- **korelacyjny** – służą identyfikacji i opisowi związków między zmiennymi (np. za pomocą analiz regresji), wskazując, czy relacja jest pozytywna, negatywna lub czy jej brak;
- **przyczynowy** – pozwalają określić zależności przyczynowo-skutkowe i oszacować ich siłę.

Różnice między nimi można przedstawić na przykładzie hipotetycznego badania dotyczącego osiągnięć szkolnych dzieci w szkołach podstawowych. Analiza opisowa zebranych danych pokazała, że tylko 75% dzieci ma w domu komputer do użytku własnego. Analiza korelacyjna wykazała z kolei, że uczniowie i uczennice, którzy mają w domu komputer do użytku własnego, uzyskują systematycznie wyższe wyniki w matematyce (związek/korelacja pozytywna między posiadaniem komputera a wynikami). Korelacja między posiadaniem komputera a wynikami w nauce jest jednak pozorna, obie zmienne mają bowiem to samo źródło – sytuację społeczno-ekonomiczną gospodarstwa domowego. Bardziej zamożne rodziny stać na to, by kupić

swoim dzieciom komputer, a przy okazji zapewnić im warunki potrzebne do uzyskania lepszych wyników w nauce (relacja przyczynowa).



Dowód naukowy to logicznie spójna konstrukcja składająca się z:

1. argumentacji opisującej i wyjaśniającej związek przyczynowy między interwencją (np. wdrożonym programem) a jej skutkiem; argumentacja ta powinna opierać się na zweryfikowanych przesłankach i jasno określonych warunkach działania,
2. zestawu danych i analiz, które weryfikują poprawność zaproponowanego rozumowania.

Nienaukowe interpretacje danych, potoczne lub pojedyncze obserwacje nie stanowią dowodów! Innymi słowy opinie eksperckie, anegdoty, pojedyncze badania (zwłaszcza niskiej jakości) nie stanowią wystarczającej podstawy do podejmowania decyzji dotyczących polityk społecznych.

Rozróżnienie wiarygodnych badań i dowodów od tych niskiej jakości ma kluczowe znaczenie dla jakości podejmowanych decyzji.



Dobrze zaprojektowane badania, wykorzystujące odpowiednie analizy, dostarczają:

- wysokiej jakości danych przydatnych na różnych poziomach decyzyjnych (por. [Aneks 2](#)) i etapach projektowania interwencji w oświacie (patrz niżej),
- wysokiej jakości interpretacji nowych lub wcześniej zebranych danych.



Złe badania prowadzą do błędnych wniosków i decyzji. **Złe badania** to takie, które:

- wykorzystują **niesprawdzone narzędzia pomiarowe** (np. doraźnie przygotowane testy umiejętności),
- stosują **analizy niepasujące** do pytania badawczego lub rodzaju danych (np. dane dotyczą szkół, a analiza dotyczy indywidualnych uczniów i uczennic),

- mają **nieprzejrzyście opisaną metodologię**, której nie da się odtworzyć (np. brakuje szczegółowego opisu danej interwencji),
- **nie ujawniają ograniczeń** metodologicznych lub analitycznych,
- opierają się na **źle dobranej próbie** (np. próba nie jest reprezentatywna lub porównywane grupy nie są równoważne) (por. kryteria [WIEDZA](#)).

Jak ich unikać? Więcej o tym w dalszej części materiału.

Planując reformy lub programy edukacyjne, musimy opierać się na wysokiej jakości dowodach. Są one niezbędne, by określić, czy dana interwencja zadziałała „kiedyś” i „gdzieś”, a także czy ma szansę zadziałać „tu” i „teraz”, tj. uwzględniając aktualne uwarunkowania organizacyjne, społeczne i ekonomiczne.

Aby to ocenić, warto odpowiedzieć sobie na kilka pytań ([Cartwright i Hardie, 2012](#)):

- **Jaki jest stan obecny? Co dokładnie chcemy zmienić? Jakie są możliwe rozwiązania?**
- **Jakie są dowody, że dana polityka lub interwencja przyniosła pozytywne efekty w przeszłości?**
- **Jakie mechanizmy przyczynowe zostały zidentyfikowane i czy pozwalają one na stwierdzenie, że ten sam proces zadziała tu i teraz?**
- **Jakie warunki muszą być spełnione, aby ten sam mechanizm zadziałał tu i teraz?**





Problem 1

Jaki jest stan obecny? Co dokładnie chcemy zmienić? Jakie są możliwe rozwiązania?



Kalifornia, podobnie jak wiele innych stanów USA, monitoruje osiągnięcia edukacyjne dzieci w wieku szkolnym za pomocą ogólnokrajowych testów umiejętności NAEP (National Assessment of Education Progress). W ramach tego badania reprezentatywna próba losowa czwarto- i ósmoklasistów rozwiązuje co dwa lata zadania sprawdzające ich umiejętności matematyczne i w zakresie czytania. Testy te są obowiązkowo przeprowadzane we wszystkich stanach, które otrzymują fundusze federalne.

W 1994 roku średnie wyniki kalifornijskich dzieci z czwartych klas były znacząco niższe od średniej krajowej, zarówno w czytaniu, jak i matematyce. Odsetek dzieci, które osiągnęły przyjęty standard biegłości (*proficiency rates*), również należał do jednego z najniższych w porównaniu z innymi stanami.

Badanie NAEP to przykład ogólnokrajowego badania przekrojowego, które dostarcza wysokiej jakości danych o wiedzy i umiejętnościach uczennic i uczniów w określonym wieku i czasie. **Takie dane są bardzo przydatne do diagnozy i opisu stanu obecnego**, ale nie pozwalają na ocenę skuteczności konkretnych działań. Innymi słowy pokazują, *co się dzieje*, choć nie wskazują, *dlaczego*. Wyniki takiego badania można wykorzystywać do:

- **opisu trendów** w wynikach uczniów na przestrzeni lat;
- **analizy różnic między grupami** (np. ze względu na płeć, pochodzenie społeczne, język, miejsce zamieszkania) w celu identyfikacji obszarów rosnących lub malejących nierówności;
- **poszukiwania możliwych korelatów** osiągnięć edukacyjnych, czyli czynników współwystępujących z wynikami, które mogą stanowić punkt wyjścia do dalszych badań.



W Polsce są dostępne wyniki różnych badań przekrojowych dotyczących wiedzy i umiejętności uczniów szkół podstawowych i średnich. Badania te opierają się na testach wysokiej lub niskiej stawki. Oba typy testów dostarczają cennych informacji, choć służą innym celom.

Testy wysokiej stawki (np. egzamin ósmoklasisty, egzamin maturalny, egzamin SAT w USA) mają istotne konsekwencje dla rozwiązujących je osób, np. decydują o dalszej ścieżce edukacyjnej. Pełnią one funkcję selekcyjną i certyfikacyjną, tj. potwierdzają określone umiejętności lub kwalifikacje i są wykorzystywane w procesach rekrutacyjnych. Dane z takich testów są używane w polityce oświatowej przede wszystkim do oceny efektów nauczania i monitorowania jakości kształcenia.

Testy niskiej stawki (np. wspomniany wyżej NAEP, a także badania międzynarodowe PISA, TIMSS, PIRLS, ICCS, ICILS, SSES) nie wiążą się z indywidualnymi skutkami dla rozwiązujących je osób. Mierzą one nie tylko efekty nauczania, ale także motywację i umiejętności pozapoznawcze. Badania międzynarodowe pozwalają ponadto porównywać wyniki między krajami i badać czynniki sprzyjające uczeniu się.

Badania wykorzystujące wyniki testów wysokiej i niskiej stawki, podobnie jak badania krajowe i międzynarodowe, wzajemnie się uzupełniają.

Polska bierze udział w następujących międzynarodowych badaniach umiejętności dzieci i młodzieży:

PISA – Program Międzynarodowej Oceny Umiejętności Uczniów (Programme for International Student Assessment). Badanie mierzy wiedzę i umiejętności z zakresu rozumienia czytanego tekstu, matematyki i nauk przyrodniczych uczniów, którzy ukończyli 15 lat w roku poprzedzającym badanie.

TIMSS – Międzynarodowe Badanie Wyników Nauczania Matematyki i Nauk Przyrodniczych (Trends in International Mathematics and Science Study). Badanie mierzy wiedzę i umiejętności z zakresu matematyki i nauk przyrodniczych uczniów w czwartym i ósmym roku edukacji szkolnej.

PIRLS – Międzynarodowe Badanie Postępów Biegłości w Czytaniu (Progress in International Reading Literacy Study). Badanie mierzy umiejętności z zakresu rozumienia czytanego tekstu uczniów w czwartym roku edukacji szkolnej.

ICCS – Międzynarodowe Badanie Kompetencji Obywatelskich (International Civic and Citizenship Education Study). Badanie mierzy wiedzę i rozumienie zagadnień z zakresu edukacji obywatelskiej oraz postawy i zaangażowanie obywatelskie uczniów w ósmym roku edukacji szkolnej.

ICILS – Międzynarodowe Badanie Kompetencji Komputerowych i Informacyjnych (International Computer and Information Literacy Study). Badanie mierzy wiedzę i umiejętności z zakresu obsługi komputera oraz wyszukiwania informacji uczniów w ósmym roku edukacji szkolnej.

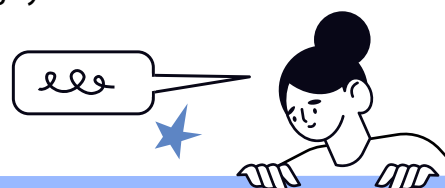
SSES – Międzynarodowe Badanie Kompetencji Społeczno-Emocjonalnych (Survey on Social and Emotional Skills). Badanie mierzy kompetencje społeczne i emocjonalne uczniów w dwóch grupach wiekowych (10 i 15 lat).

Wszystkie wymienione badania międzynarodowe oprócz testów wiedzy i umiejętności z określonego zakresu zawierają też pytania m.in. o opinie i postawy badanych dzieci, ich doświadczenia edukacyjne i środowisko domowe. Uwzględniają także informacje dotyczące procesów nauczania, zasobów i organizacji pracy szkół.

Dlaczego wyniki badań przekrojowych nie pozwalają ocenić skuteczności konkretnych działań lub polityk? Dlaczego **proste porównanie zagregowanych rezultatów w dwóch momentach w czasie**, np. przed i po wprowadzeniu jakiegoś programu, **nie stanowi dowodu, że to właśnie ten program wywołał obserwowane zmiany?**

Wynika to z faktu, że w każdej edycji badania mierzone są umiejętności innych roczników uczniów i uczennic, a ich osiągnięcia nie zależą tylko od szkoły (lub szerzej od czynników edukacyjnych), ale także od czynników pozaszkolnych – społecznych, ekonomicznych czy demograficznych. Zmiany w strukturze populacji, np. wzrost odsetka dzieci z doświadczeniem migracji lub pogorszenie sytuacji materialnej rodzin, mogą znacząco wpływać na średnie wyniki danego rocznika. Określenie, czy i jaki wpływ na wyniki miała każda z tych zmian, jest niezwykle trudne i wymaga dodatkowych danych lub zaawansowanych analiz ([Wyckoff, 2025](#); [Dee i Jacob, 2011](#)).

Podobne zastrzeżenie dotyczy porównań między krajami, regionami czy okręgami – różnice w wynikach często odzwierciedlają różnice w strukturze demograficzno-społecznej, a nie faktyczną skuteczność systemów edukacyjnych.



Przykład różnych interpretacji wyników badań przekrojowych (por. [Wyckoff, 2025](#); [Quintero, 2025](#))

Zagregowane wyniki testów NAEP pokazały, że czwartoklasiści w 2002 i 2003 roku osiągnęli lepsze rezultaty niż wcześniejsze roczniki. W kolejnych latach wyniki były stabilne, a po 2012 roku zaczęto obserwować ich spadek.

Jak to interpretować? Czy za poprawę wyników na początku XXI wieku odpowiadały reformy edukacyjne – takie jak federalny program No Child Left Behind (NCLB) i tzw. „reformy rynkowe”? A może większe znaczenie miały czynniki ekonomiczne, np. okres wzrostu gospodarczego w latach 90. XX wieku? A czy pogorszenie wyników po 2012 roku to efekt kryzysu finansowego, zmian w systemie odpowiedzialności szkół, a może powszechnego dostępu do smartfonów i mediów społecznościowych wśród dzieci i młodzieży po 2008 roku?

Zwolennicy NCLB i „rynkowych” reform edukacyjnych wprowadzanych od końca lat 90. interpretowali obserwowane zmiany jako „dowód” skuteczności reform.

Osoby sceptyczne wobec tych reform zwracają jednak uwagę, że wzrost wyników w pierwszej dekadzie XXI wieku mógł też być związany z faktem, że dzieci urodzone w latach 90. dorastały w lepszej sytuacji ekonomicznej w porównaniu z dziećmi urodzonymi w latach 80. Z kolei na spadek wyników po 2012 roku mógł wpłynąć kryzys ekonomiczny w drugiej połowie lat 2000.

Każda z tych interpretacji trendów w zagregowanych wynikach NAEP (lub podobnych badań przekrojowych) to jedynie hipoteza, którą można zweryfikować dopiero przy użyciu narzędzi do analiz przyczynowych.

Wyniki NAEP stanowiły główny impuls reformy w Kalifornii. Nie wzięto jednak pod uwagę, że na ich poziom mogły mieć wpływ różne czynniki (jak w przykładzie opisanym powyżej). Na przykład w latach 90. miały miejsce istotne zmiany w strukturze społeczno-demograficznej tego stanu. Zwłaszcza w dużych miastach wzrósł odsetek dzieci z rodzin o niskich dochodach oraz uczniów i uczennic z doświadczeniem migracji ([Jepsen i Rivkin, 2002](#)). Program objął jednak wszystkie publiczne szkoły podstawowe. Ten brak dogłębnej diagnozy stanu wyjściowego w praktyce oznaczał, że środki finansowe i organizacyjne nie trafiły tam, gdzie mogły przynieść największe efekty.

Nie był to jednak jedyny błąd. Zabrakło też analizy alternatywnych rozwiązań i ich kosztów, czyli tzw. **analizy horyzontalnej** ([Cartwright i Hardie, 2012](#)). Taka analiza polega na porównaniu różnych sposobów osiągnięcia tego samego celu i określeniu, jakie warunki muszą być spełnione, by każdy z nich zadziałał. Nawet tak ogólny cel, jak „poprawa umiejętności matematycznych i czytania dzieci na wczesnym etapie edukacji”, może zostać osiągnięty różnymi metodami, które nierzadko na siebie zachodzą.



Przykład analizy horyzontalnej

(na podstawie [Cartwright i Hardie, 2012](#); [Chingos, 2011](#); [Odden, 1990](#); [Kraft, 2020](#))

Zmniejszenie liczby dzieci w klasie np. z 30 do 20 wymaga:

1. zatrudnienia 50% więcej (wykwalifikowanego) personelu pedagogicznego,
2. wygospodarowania dodatkowej przestrzeni dla nowych klas, czyli albo rozbudowy szkoły, albo wynajęcia i wyposażenia dodatkowych sal lekcyjnych.

Koszty takiej zmiany są zatem ogromne. Dlatego przed jej wdrożeniem należy rozważyć inne sposoby osiągnięcia tego samego celu, czyli poprawy wyników dzieci w nauce czytania i matematyki.

Jakie są inne możliwe rozwiązania? Oto kilka przykładowych*:

1. zwiększenie liczby godzin nauki matematyki i czytania,
2. zajęcia dodatkowe dla dzieci potrzebujących większego wsparcia,
3. inwestycja w lepsze przygotowanie i poprawę motywacji nauczycieli i nauczycielek edukacji wczesnoszkolnej.

Każde z tych rozwiązań wiąże się oczywiście z kosztami, ograniczeniami i wyzwaniem (zagrożeniami). Jednak każde powinno być zostać poddane pod dyskusję. Zwłaszcza że inwestycja w redukcję liczebności klas oznacza zmniejszenie środków na pozostałe działania.

Analiza horyzontalna pomaga więc unikać decyzji kosztownych, lecz mało efektywnych – i kierować środki tam, gdzie przynoszą one największe korzyści edukacyjne.

* Więcej przykładów skutecznych interwencji w edukacji można znaleźć w wiarygodnych zestawieniach dowodów naukowych opisanych w [Aneksie 3](#).

Środki przeznaczane na edukację są ograniczone. Analizując istniejące dowody dotyczące skuteczności rozmaitych interwencji, musimy zatem zwracać uwagę na to, czy koszty ich wdrożenia są warte spodziewanych korzyści.



Problem 2

Jakie są dowody, że dana polityka lub interwencja przyniosła pozytywne efekty w przeszłości?



Program redukcji liczebności klas (CSR) w Kalifornii powoływał się na wyniki eksperymentu STAR (Student-Teacher Achievement Ratio). Jest to jeden z najbardziej znanych i przeprowadzonych na dużą skalę eksperymentów w edukacji ([Krueger, 1999](#)). Jego celem było oszacowanie wpływu liczebności klasy na wyniki edukacyjne dzieci. Eksperyment rozpoczął się w roku szkolnym 1985/1986 w stanie Tennessee i trwał 4 lata (miał więc charakter badania podłużnego, czyli takiego, w którym obserwuje się te same osoby przez dłuższy czas). Objął on 11600 dzieci z 80 szkół. Dzieci rozpoczynające edukację w zerówce, a także ich nauczyciele i nauczycielki, były losowo przydzielane do jednej z trzech grup:

- grupy 1 – **małej klasy**, liczącej 13–17 osób przypadających na jednego nauczyciela,
- grupy 2 – **regularnej klasy**, liczącej 22–25 osób przypadających na jednego nauczyciela,
- grupy 3 – **regularnej klasy**, liczącej 22–25 osób z **dodatkowym asystentem** nauczyciela.

Zarówno dzieci, jak i nauczyciele pozostawali w danej grupie przez cały okres trwania eksperymentu (do końca 3 klasy). Na koniec każdego roku szkolnego dzieci rozwiązywały serię standaryzowanych testów mierzących ich umiejętności.

Po pierwszym roku różnice między grupami były bardzo obiecujące – dzieci z małych klas uzyskały średnio wyniki wyższe o 0,2 odchylenia standardowego (SD) – co można porównać do efektu około 16% roku nauki więcej (ok. 29 dni). Korzyści z małych klas były szczególnie widoczne wśród dzieci z rodzin o niskich dochodach, czarnym kolorze skóry oraz wśród chłopców. Pozytywny efekt utrzymywał się w kolejnych latach, choć na koniec 3 klasy różnice pomiędzy dziećmi z dużych i małych klas były mniejsze i wynosiły ok. 0,14 SD ([Chingos, 2011](#)).

Eksperyment STAR miał swoje ograniczenia, szczegółowo opisane w raporcie z badania:

- Eksperyment objął tylko szkoły, które dysponowały fizyczną przestrzenią na utworzenie przynajmniej trzech klas dla każdego rocznika objętego badaniem.
- Rodzice, a także nauczyciele wiedzieli, że uczestniczą w eksperymencie. Z tego też powodu, aby uspokoić rodziców dzieci przypisanych w zerówce do klas regularnych rozmiarów, przed rozpoczęciem 1 klasy były one ponownie losowo przypisywane do grupy z dodatkowym wsparciem asystenckim lub bez niego. Klasy „bez wsparcia” miały ponadto nieregularną pomoc asystencką.

Szereg innych czynników zakłócił ponadto pełną losowość eksperymentu. Na przykład, ponieważ w tym czasie zerówka nie była w Tennessee obowiązkowa, część dzieci rozpoczynała eksperyment w 1 klasie i była losowo przypisywana do określonych klas tylko w miarę możliwości (podobnie jak te dzieci, które nie otrzymały promocji do kolejnej klasy lub przeniosły się z innej szkoły). Więcej o zakłóceniach dotyczących losowego przydziału do grup w STAR można przeczytać w pracy [Kruegera \(1999\)](#).

Eksperymenty stanowią najlepszą metodę oceny wpływu i efektywności wprowadzanych interwencji w danym czasie i kontekście. Interpretując wyniki badań eksperymentalnych, należy jednak pamiętać, że interwencje testowane w eksperymentach są wdrażane w określonych warunkach, według precyzyjnych wytycznych i pod stałym nadzorem zespołu badawczego. W praktyce oznacza to, że wdrożenie programu na szeroką skalę – w zróżnicowanych szkołach i warunkach – rzadko przebiega w tak jednolity i uporządkowany sposób. To z kolei wpływa na efekty, jakich można realistycznie oczekiwać po upowszechnieniu danej interwencji.

Dlatego pojedyncze badanie, nawet bardzo dobre metodologicznie, nie powinno być jedyną podstawą podejmowanych decyzji. W momencie wprowadzania reformy w Kalifornii dysponowano jednak tylko eksperymentem STAR. Wskazywał on na pozytywny, choć umiarkowany w dłuższej perspektywie, wpływ małych klas na wyniki uczniów i uczennic. Zgodnie z zasadami dobrych badań zespół badawczy wskazał jego ograniczenia, ale zostały one zignorowane w kalifornijskiej debacie.



Kiedy możemy mówić o wpływie interwencji?

(na podstawie Schneider i in. 2007, s. 17–18)

Aby wiarygodnie ocenić, czy dana interwencja przynosi określony skutek, muszą być spełnione cztery kryteria. Najprościej spełnić je w eksperymentach, ale dotyczą one wszystkich badań.

- 1. Skutek danej interwencji musi być zawsze oceniany w porównaniu z innym rozwiązaniem.** Gdy szukamy skutecznych interwencji, interesuje nas nie tyle „czy dany program działa”, ale „czy działa lepiej niż inne rozwiązania”. Na przykład, gdy chcemy sprawdzić, czy redukcja liczebności klas jest efektywnym rozwiązaniem, musimy porównać jej skutki ze skutkami innej interwencji lub z wynikami nauczania w klasach o niezmienionej liczebności.
- 2. Musi istnieć możliwość losowego przypisania osób do interwencji.** Każda osoba uczestnicząca w badaniu powinna mieć taką samą szansę znalezienia się w każdej z porównywanych grup. Ponieważ możemy losowo przydzielać dzieci do określonych rozwiązań (np. sposobu nauczania czy małej klasy), możemy traktować te rozwiązania jako potencjalne przyczyny zaobserwowanych zmian. Natomiast cechy takie jak płeć, pochodzenie czy miejsce urodzenia nie są przyczynami w sensie badawczym, ponieważ nie można ich modyfikować ani losowo przypisać.
- 3. Ekspozycja na interwencję musi być wyraźnie określona w czasie.** Aby ustalić, czy dana interwencja była przyczyną zmian w wynikach, potrzebujemy zmierzyć je zarówno przed, jak i po jej wprowadzeniu. Tylko wtedy możemy ocenić, czy zaobserwowana zmiana może być skutkiem programu – bo wystąpiła po jego rozpoczęciu. Na przykład, aby stwierdzić, czy nowy program nauczania matematyki daje lepsze rezultaty niż dotychczasowy, należy porównać wyniki z testów z matematyki przed rozpoczęciem programu i po jego realizacji, w co najmniej w dwóch grupach: w grupie objętej nową metodą oraz w grupie uczącej się według dotychczasowego programu.

4. Metodologia badania musi umożliwiać wykluczenie alternatywnych wyjaśnień.

Należy zapewnić, że obserwowane różnice w wynikach mogą wynikać wyłącznie z wprowadzonej interwencji. W eksperymentach osiąga się to poprzez losowy przydział uczestników do porównywanych grup. Dzięki temu nawet jeśli dzieci lub szkoły różniły się między sobą na starcie, są one równomiernie reprezentowane w badanych grupach. W konsekwencji zaobserwowane różnice w wynikach można przypisać interwencji, a nie np. różnicom w zdolnościach dzieci, wcześniejszym doświadczeniom edukacyjnym czy warunkom domowym.

Eksperymenty a inne badania

Dobrze zaprojektowane badania eksperymentalne mają wysoką trafność wewnętrzną, bo pozwalają w prosty sposób spełnić wymienione wyżej kryteria i precyzyjnie oszacować wpływ interwencji na wyniki w kontrolowanych („idealnych”) warunkach. Badania quasiekperymentalne i analizy danych obserwacyjnych wymagają bardziej zaawansowanych narzędzi statystycznych, aby te kryteria spełnić, ale często mają wyższą trafność zewnętrzną (tj. pozwalają sprawdzić, jak dane rozwiązanie działa w różnych szkołach i kontekstach). **Jest tak dlatego, że badania i analizy quasiekperymentalne wykorzystują zazwyczaj duże zbiory danych zebrane na reprezentatywnych próbach, często na przestrzeni lat.**

Choć badania quasiekperymentalne nie opierają się na losowym przypisaniu do interwencji, istnieje szereg metod statystyczno-analitycznych pozwalających częściowo odtworzyć warunki zbliżone do eksperymentu. Wybór odpowiedniej metody (lub kombinacji metod) zależy m.in. od typu badania (np. czy jest ono przekrojowe czy podłużne (zob. [Słowniczek](#))) i dostępnych danych. Należy jednak pamiętać, że ze względu na brak randomizacji kluczowe jest upewnienie się, że porównywane grupy są równoważne – zarówno pod względem cech uwzględnionych w badaniu, jak i tych nieobserwowanych¹.

¹ Do najczęściej stosowanych metod ograniczania wpływu istniejących różnic między grupami na wyniki analiz przyczynowych (*selection bias/omitted variable bias*) w badaniach obserwacyjnych należą: modele z efektami stałymi (*fixed-effects*), ze zmienną instrumentalną (*instrumental variable*), z dopasowaniem według wyniku skłonności (*propensity score matching*) oraz modele regresji nieciągłej (*regression discontinuity*). Nieco inną metodą, polegającą na szacowaniu różnicy w wielkości zmian w jakimś przedziale czasu w dwóch grupach, jest metoda różnicy-w-różnicach (*difference-in-differences*). Metoda ta zakłada, że grupa, w której wprowadzono jakąś interwencję, bez tej interwencji miałaby tę samą dynamikę zmian, co grupa bez niej. Zatem różnica w dynamice zmian może być przypisana interwencji, ponieważ nieobserwowalne różnice między grupami są stałe w czasie. Każda z tych metod ma swoje ograniczenia, które należy brać pod uwagę przy interpretacji wyników (por. [Winship i Morgan, 1999](#); [Shadish i in., 2002](#); [Kim i Steiner, 2016](#); [Gopalan i in., 2020](#); [WWC, 2022](#)).

Oba typy badań wykorzystujących analizy przyczynowe są ważne. Eksperymenty pokazują, czy dana interwencja działa i jak silny jest jej efekt. Natomiast analizy danych obserwacyjnych pozwalają lepiej zrozumieć, w jakich warunkach program działa (lub nie działa), np. jak wpływają na niego skład klasy, organizacja pracy szkoły czy lokalne uwarunkowania (Schneider i in. 2007, s. 112). Temat rozwijamy w tabeli prezentującej wartość dowodową różnych analiz w kontekście oceny efektywności interwencji (por. tab. [Typologia analiz...](#)).

Innym źródłem dowodów, na które powoływano się w Kalifornii, była metaanaliza autorstwa [Glassa i Smith \(1979\)](#). Już w trakcie debaty na temat reformy wiadomo jednak było, że ma ona poważne ograniczenia:

- uwzględniała badania bardzo różnej jakości (z 77 uwzględnionych analiz tylko 14 było metodologicznie poprawnych),
- obejmowała wyniki dotyczące odmiennych typów „osiągnięć”, np. zarówno nauki czytania, jak i gry w tenisa),
- stosowała bardzo szeroką definicję „małej klasy”, obejmującą m.in. zajęcia indywidualne ([Odden, 1990](#)).

Dostępne w 1996 roku dowody były zatem niewystarczające, aby uzasadnić kosztowną reformę w całym stanie. Co zatem stanowiłoby solidne dowody na to, że istnieje relacja przyczynowa pomiędzy wielkością klasy a osiągnięciami uczniów? Jakich analiz warto szukać? Odpowiadając na to pytanie, warto przede wszystkim pamiętać o poniższych zasadach.

1. **Powinno się oceniać całość dostępnych dowodów i zachować ostrożność wobec pojedynczych badań, nawet tych najlepszych.** Pomijając kwestie metodologiczne, trzeba pamiętać, że badacze i badaczki nie są odporni na wpływ rozmaitych błędów i zniekształceń poznawczych. Błędy te mogą występować na każdym etapie procesu badawczego (por. [Aneks 1](#)).
2. **Przeglądy systematyczne i metaanalizy badań eksperymentalnych i quasiekperymentalnych dostarczają zazwyczaj najwyższej jakości dowodów efektywności różnych interwencji.** Muszą jednak opierać się na rygorystycznych metodach i surowych kryteriach dotyczących doboru badań

do analizy. Metaanalizy wykonane na słabych lub nieporównywalnych danych mogą wprowadzać w błąd. Tak było w przypadku analizy Glassa i Smith (1979), a także innych metaanaliz opisanych m.in. w przeglądzie [Bakera i Knighta \(2025\)](#).

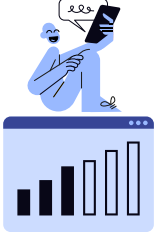
3. **Korzystanie z publikacji, które zostały poddane procesowi recenzji, zapewnia, że metodologia zbierania danych i ich interpretacja ma pewien poziom jakości.** Takie publikacje nadal wymagają jednak ostrożnej interpretacji, zwłaszcza jeśli dotyczą pojedynczych badań.

4. **Zgromadzenie solidnego zbioru dowodów dotyczących wpływu danego typu polityki, a nawet efektywności pojedynczej interwencji w konkretnym miejscu, wymaga czasu.** Jest tak z kilku powodów:

- Większość polityk edukacyjnych nie przynosi natychmiastowych efektów. W niektórych przypadkach może to zająć kilka lat, a w innych – nawet dekadę lub dłużej. Co więcej, nawet jeśli efekty pojawiają się szybko, musimy określić, czy utrzymują się w czasie.
- Zgromadzenie wysokiej jakości dowodów wymaga przeprowadzenia badań w różnych kontekstach i lokalizacjach, a polityki często rozprzestrzeniają się powoli – nierzadko w odpowiedzi na napływające stopniowo dowody.

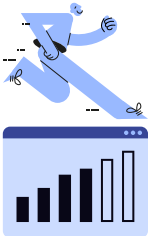
Najprościej mówiąc, prowadzenie rzetelnych badań to proces czasochłonny. Inwestycje w badania i dowody wysokiej jakości są jednak podstawą skutecznych interwencji. Ich wykorzystanie ogranicza ryzyko wdrożenia rozwiązań o niskiej lub nieustalonej skuteczności lub takich, które mogą przynieść negatywne efekty.

Typologia analiz wg wartości dowodowej w kontekście oceny efektywności programów i polityk edukacyjnych

Wartość dowodowa	Typ analiz	Charakterystyki
 Najniższa	Analizy jakościowe indywidualnych sytuacji i przypadków	Zalety Pozwalają szczegółowo opisać, jak przebiegała jakaś interwencja, jak reagowały osoby, które brały w niej udział, i jakie były jej skutki w danym miejscu i czasie. Ograniczenia Mogą stanowić inspirację do formułowania hipotez, ale nie mogą być uogólniane na całą populację i nie stanowią podstaw do wnioskowania o przyczynach. Typ badań, w których są wykorzystywane Stosowane najczęściej w badaniach jakościowych, np. studium przypadku. Przykład Stasz i Stecher (2001) ilustrują wyzwania związane z wdrażaniem programu małych klas w Kalifornii, opierając się na obserwacjach dwóch nauczycielek pracujących w różnych szkołach. Szczegółowy opis ich praktyk przed i po wprowadzeniu małych klas pokazał, że nie dostosowały one sposobu nauczania do zmniejszonej liczebności klasy.

Wartość dowodowa	Typ analiz	Charakterystyki
 Bardzo niska	Analizy opisowe dotyczące trendów w osiągnięciach uczniów i uczennic	Zalety Jeśli wdrożenie danej polityki zbiega się w czasie ze znacznymi i trwałymi zmianami wyników, można to potraktować jako przesłankę, że dana polityka wymaga dalszego zbadania – zwłaszcza jeśli podobne zmiany obserwuje się w wielu miejscach. Ograniczenia Nie mogą stanowić podstaw do wnioskowania o przyczynach ze względu na możliwość pominięcia różnych czynników mogących wpływać na wyniki. Typ badań, w których są wykorzystywane Najczęściej stosowane w ilościowych badaniach podłużnych i przekrojowych (zob. Słowniczek). Przykład Wyckoff (2025) wykorzystuje opisową analizę wyników badania NAEP z ostatnich 30 lat, aby pokazać trend spadkowy w osiągnięciach edukacyjnych dzieci, który rozpoczął się jeszcze przed pandemią.

Wartość dowodowa	Typ analiz	Charakterystyki
 Niska	Analizy korelacyjne wskazujące np. na poprawę wyników szkolnych tam, gdzie polityki zostały wdrożone, przy uwzględnieniu innych czynników mogących wpływać na wyniki	Zalety Mogą one sugerować skuteczność danej polityki (lub jej braku) i stanowić podstawę do pogłębionych analiz. Ograniczenia Nie stanowią podstaw do interpretacji przyczynowych, ze względu na możliwość pominięcia konkurencyjnych wyjaśnień i inne problemy. Typ badań, w których są wykorzystywane Najczęściej stosowane w ilościowych badaniach podłużnych i przekrojowych (zob. Słowniczek). Przykład Karpiński i in. (2023) dogłębnie analizują związek między umiejętnościami pozapoznawczymi a kompetencjami cyfrowymi mierzonymi w badaniu ICILS. Wskazują w nim na różnice w umiejętnościach pozapoznawczych ze względu na płeć, doświadczenie migracyjne i status społeczno-ekonomiczny. Podkreślają więc, że interpretacje nierówności w umiejętnościach cyfrowych muszą być ostrożne i uwzględniać dodatkowe czynniki kontekstowe.
 Wysoka	Nie-eksperymentalne lub quasyeksperymentalne strategie identyfikacji relacji przyczynowych	Zalety Gdy metody eksperymentalne nie są możliwe do zastosowania lub gdy chcemy ocenić wpływ interwencji w różnych kontekstach, podejścia quasyeksperymentalne są najlepszym sposobem ustalania kierunku i rozmiaru efektów przyczynowych. Ograniczenia Ze względu na brak losowego przypisania do porównywanych grup wymagają zaawansowanych technik statystycznych do zapewnienia, że są one równoważne (tj. niemal identyczne pod względem rozmaitych cech społeczno-demograficznych i innych).

Wartość dowodowa	Typ analiz	Charakterystyki
 Wysoka	Nie-eksperymentalne lub quasieksperymentalne strategie identyfikacji relacji przyczynowych	W związku z tym są podatne na różne błędy i zniekształcenia interpretacyjne (przede wszystkim stronniczość doboru (selection bias), które nie występują w eksperymentach. Ponadto każda strategia analityczno-statystyczna ma szczególne zastosowania i ograniczenia (por. Kim i Steiner, 2016). Metody analiz quasieksperymentalnych są jednak intensywnie rozwijane, m.in. dzięki krytycznej dyskusji dotyczącej ich rzetelności. Typ badań, w których są wykorzystywane Najczęściej stosowane w ilościowych badaniach podłużnych i przekrojowych wykorzystujących zróżnicowanie w czasie wprowadzania różnych programów (zob. Słowniczek). Przykład Caroline Hoxby (2000) przeprowadziła ciekawą analizę (z wykorzystaniem regresji nieciągłej) dotyczącą wpływu małych klas na wyniki szkolne. Wykorzystała w niej informacje o zmianach w wielkości klas, które pojawiają się wtedy, gdy naturalne wahania liczby dzieci rozpoczynających naukę powodują zmianę liczby oddziałów w danym roczniku. Według badania zmniejszenie liczebności klasy ma niewielki wpływ na poprawę osiągnięć szkolnych. Przykładem zastosowania nieco innej strategii identyfikacji relacji przyczynowych (sekwencji zdarzeń i różnic w różnicach) jest analiza Matthew Krafte i in. (2020) na temat wpływu reform dotyczących oceny kadry pedagogicznej na liczbę i kwalifikacje nowych nauczycieli. Łącząc informacje pochodzące z różnych badań prowadzonych różnym czasie badacze wykazali, że reforma miała negatywny wpływ na liczbę nowych nauczycieli. Miała jednak pozytywny wpływ na ich kwalifikacje.

Wartość dowodowa	Typ analiz	Charakterystyki
 Bardzo wysoka	Analizy przyczynowe na podstawie wyników badań eksperymentalnych	Zalety Randomizowane badania kontrolowane (RCT) stanowią jednoznaczne dowody, czy interwencja wywołuje dany skutek i jak silny jest ten związek. Ograniczenia Są bardzo kosztowne, co skutkuje tym, że zwykle mają ograniczony zasięg (np. jedno miejsce lub niewielka grupa miejsc). Zmniejsza to ich trafność zewnętrzną (możliwość generalizacji na całą populację), czasem w sposób nieoczywisty. Typ badań, w których są wykorzystywane Randomizowane badania kontrolowane (RCT) / badania eksperymentalne. Przykład Opisany wyżej eksperyment STAR stanowi przykład dobrze zaplanowanego i przeprowadzonego eksperymentu w edukacji. Przykład innego, wciąż trwającego eksperymentu opisują Hulme i in. (2025). Eksperyment ten dotyczy efektywności interwencji mających na celu poprawę umiejętności językowych wśród dzieci rozpoczynających naukę (Nuffield Early Language Intervention, NELI). Krótkoterminowe wyniki wpływu intensywnego 20-tygodniowego programu nauczania dzieci o niżej rozwiniętych zdolnościach językowych są obiecujące. Trwa natomiast ewaluacja efektów długoterminowych.

Wartość dowodowa	Typ analiz	Charakterystyki
 Najwyższa	Przeglądy systematyczne i metaanalizy opierające się na z góry przyjętych kryteriach selekcji badań i ocenie ich metodologii	Zalety Pozwalają na oszacowanie uśrednionego efektu skutków jakiejś interwencji, które zostały zmierzone w różnych kontekstach. Polegają na analizie wyników wielu różnych badań prowadzonych w różnych miejscach lub warunkach. Ograniczenia Wymagają wystarczającej liczby wysokiej jakości pojedynczych badań eksperymentalnych lub quasiekperymentalnych. Typ badań, w których są wykorzystywane Przeglądy systematyczne i metaanalizy mogą dotyczyć wszystkich typów badań ilościowych. Przykład Analiza autorstwa Filges i in. (2018) stanowi przykład wysokiej jakości przeglądu systematycznego zawierającego metaanalizę wyników badań dotyczących wpływu liczebności klas na osiągnięcia edukacyjne dzieci. Według badania małe klasy mają niewielki pozytywny wpływ na poprawę umiejętności czytania, ich wpływ na umiejętności matematyczne jest zaś nieistotny.

Interpretując wyniki badań, warto też przyjrzeć się wielkości podawanych w nich skutków danej interwencji lub polityki. Bardzo często w dyskusji o dowodach zwraca się uwagę tylko na to, czy zaobserwowane różnice między grupami są statystycznie istotne, tj. czy są one dziełem przypadku, czy nie. Jednak przy podejmowaniu decyzji edukacyjnych **równie ważne, o ile nie ważniejsze, jest określenie, jak duże były zmiany wywołane daną interwencją**. Czasem wyniki statystycznie nieistotne mogą wskazywać na obiecującą praktykę, zwłaszcza jeśli badana próba była mała, a efekt znaczący. Z drugiej strony, różnice statystycznie istotne mogą wiązać się z interwencją przynoszącą stosunkowo mały efekt.



Jak uniknąć typowych pułapek w interpretacji wielkości efektów?

(na podstawie [Kraft, 2020](#) oraz [Cheung i Slavin, 2016](#))

Wielkość efektu pokazuje, jak dużą zmianę wywołały dana interwencja lub działanie (po uwzględnieniu innych czynników, które mogły mieć na nią wpływ). Dla przykładu, w eksperymencie STAR różnica między średnimi wynikami dzieci z małych i dużych klas wyniosła 0,14 odchylenia standardowego (SD) na koniec 3 klasy, czyli efekt małej klasy oszacowano na 0,14 SD.

Dlaczego wielkość efektu najczęściej wyraża się w jednostkach odchylenia standardowego? Wyobraźmy sobie, że różnica między porównywanymi w eksperymencie grupami wyniosłaby 20 punktów na teście z matematyki na koniec 1 klasy i 25 punktów na teście na koniec 3 klasy. Czy 20 i 25 punktów to dużo czy mało? Bez informacji o tym, ile punktów w danym teście można było zdobyć (np. 100 czy 500) i jaki był rozrzut wyników wśród badanych dzieci (czy większość dzieci uzyskała rezultat zbliżony do średniej, czy też wyniki były bardzo zróżnicowane), interpretacja różnic wyrażonych w liczbach bezwzględnych jest trudna. Ponadto testy na koniec 1 i 3 klasy różnią się. Gdy chcemy porównywać skutki interwencji mierzonych różnymi narzędziami lub osiągnięcia edukacyjne różnych roczników na przestrzeni lat czy w różnych miejscach (np. wyniki dzieci w Kalifornii z wynikami dzieci w Tennessee) potrzebujemy ujednoliconej skali, która nam to ułatwi. Dzięki standaryzacji, tj. wyrażeniu wielkości efektu w jednostkach odchylenia standardowego (tzw. standaryzowana wielkość efektu), można zestawiać wyniki badań opartych na różnych skalach pomiaru.

Jakiej wielkości efektów możemy oczekiwać, wprowadzając dany program lub interwencję? Poniżej kilka wskazówek, jak interpretować wielkość efektów w kontekście polityk edukacyjnych.

- 5. Należy zwracać uwagę na typ badania.** Efekty oszacowane na podstawie analiz korelacyjnych lub opisowych nie stanowią wiarygodnych podstaw do wnioskowania o efektywności jakiejś polityki. Tylko efekty oszacowane na podstawie badań i analiz przyczynowych mogą stanowić podstawę do formułowania polityk edukacyjnych, tym bardziej że efekty pochodzące z analiz korelacyjnych lub opisowych mogą sugerować wyższy niż faktyczny wpływ jakiejś polityki.

6. **Należy zwracać uwagę na to, co i kiedy było poddane interwencji.** Wpływ jakiegoś działania na rezultaty bezpośrednio z nim związane i mierzone wkrótce po jego wprowadzeniu będzie większy niż efekt długoterminowy lub tylko pośrednio związany z interwencją. Przykładowo dodatkowa lekcja matematyki w 8 klasie będzie miała większy wpływ na wyniki egzaminu ósmoklasisty niż na wyniki matury z tego przedmiotu lub na rezultaty testów ogólnych umiejętności analitycznych.
7. **Należy zwracać uwagę na to, jaka była intensywność interwencji w porównywanych grupach.** Innymi słowy „dawka” ma znaczenie. Badania, w których jedna grupa korzysta z danego programu, a druga nie, zazwyczaj pokazują większe efekty niż badania porównujące dwie grupy, z których obie mają dostęp do jakiejś formy wsparcia. Na przykład w eksperymencie STAR dzieci z dużych klas otrzymywały pomoc asystentki lub asystenta na wybranych lekcjach.
8. **Należy sprawdzić, jakie grupy porównywano ze sobą.** Porównywanie bardziej zróżnicowanych grup lub szkół zwykle daje mniejsze efekty niż porównywanie grup bardziej podobnych. Programy obejmujące wszystkie szkoły pozwalają badać całą populację, ale zazwyczaj pokazują mniejsze efekty niż interwencje skierowane do konkretnej grupy, np. szkół wiejskich. Jest tak po pierwsze dlatego, że badania w wybranych, mniej zróżnicowanych subpopulacjach mierzą zmiany u tych, którzy najbardziej mogą skorzystać z programu; po drugie w mniej zróżnicowanych grupach wyniki są bardziej wyrównane.
9. **Należy ocenić koszty uzyskania efektu.** Niektóre efekty są droższe do osiągnięcia niż inne – i choć wydaje się to oczywiste, często się o tym zapomina przy podejmowaniu decyzji edukacyjnych. Liczą się nie tylko koszty programu w przeliczeniu na jedną osobę, ale też tzw. koszty alternatywne: wybór jednego rozwiązania zwykle oznacza rezygnację z innego. Dlatego czasem warto postawić na tańsze programy o mniejszych efektach, jeśli droższe rozwiązania wymagają bardzo dużych nakładów finansowych i organizacyjnych.
10. **Należy ocenić, jak łatwo program da się wdrożyć na dużą skalę.** Gdy interwencje lub programy są wprowadzane na dużą skalę (patrz pkt. 5), mają zazwyczaj dużo mniejszy wpływ na mierzone rezultaty niż te szacowane na podstawie ich pilotażu albo ograniczonej implementacji. Dzieje się tak z wielu powodów. Po pierwsze, co wiąże się z trafnością zewnętrzną, interwencja przeprowadzona w kontrolowany

sposób na ograniczonej lub szczególnie dobranej próbie (np. tylko w szkołach podstawowych podlegających danemu kuratorium), może nie mieć tego samego efektu w bardziej zróżnicowanej populacji (wszystkich szkołach podstawowych). Po drugie, nawet jeśli oszacowane na podstawie badań eksperymentalnych efekty cechowały się wysoką trafnością zewnętrzną, przy implementacji na dużą skalę mogą wystąpić różne nieprzewidziane okoliczności (np. opór części rodziców czy grona pedagogicznego).

Jakie efekty są zatem małe, a jakie duże? W kontekście interwencji edukacyjnych badacze i badaczki zgadzają się, że wielkości efektów na poziomie 0,20 lub więcej odchylenia standardowego są duże (por. [WWC, 2014, s. 21](#)). Takie wyniki są jednak rzadkie w odniesieniu do osiągnięć szkolnych dzieci i młodzieży. Metaanalizy eksperymentów pokazują, że przeciętne efekty dla osiągnięć szkolnych mieszczą się w zakresie 0,05 a 0,16 SD (czyli są umiarkowane) (por. [Cheung i Slavin, 2016](#); [Fryer, 2017](#)). Metaanalizy badań quasyeksperymentalnych dają zazwyczaj nieco wyższe oszacowania (por. stroniczość doboru próby, [Aneks 1](#)), podobnie jak analizy badań opublikowanych w recenzowanych czasopismach (por. błąd publikacji, [Aneks 1](#)).



Problem 3

Jakie mechanizmy przyczynowe zostały zidentyfikowane i czy pozwalają one na stwierdzenie, że ten sam proces zadziała tu i teraz?



Analizy dotyczące wpływu wielkości klas na osiągnięcia uczniów, które były dostępne przed wprowadzeniem reformy w Kalifornii, wskazywały na następujące możliwe wyjaśnienia tej zależności.

- W małych klasach nauczyciele spędzali mniej czasu na korygowaniu zachowania i dyscyplinowaniu dzieci. Co za tym idzie – mieli więcej czasu na realizację planu lekcji i dodatkowe ćwiczenia.
- W małych klasach nauczyciele mieli więcej możliwości poświęcenia uwagi każdemu dziecku z osobna i dzięki temu oferowali bardziej zindywidualizowane wsparcie i informację zwrotną.
- W małych klasach nauczyciele mieli wyższe poczucie satysfakcji, co przekładało się na większe zaangażowanie w pracę.
- Uczniowie i uczennice mieli więcej możliwości uczestniczenia w lekcji, czekali krócej na oceny swojej pracy i byli bardziej uważni (Odden, 1990; Filges i in., 2018).

Interwencje i polityki powinny przynieść oczekiwane efekty dzięki zadziałaniu jakiegoś określonego mechanizmu przyczynowego. Aby przewidzieć, czy dana interwencja zadziała „tu i teraz”, trzeba ten mechanizm najpierw opisać i zrozumieć. Przyczyny nie występują bowiem w próżni, ale są powiązane z szeregiem różnych czynników, które dany mechanizm wspierają lub blokują. Mechanizmy przyczynowe nie są zatem uniwersalne. To, że jakaś interwencja (np. zmniejszenie liczby dzieci w klasie do maksymalnie 20) przyniosła spodziewane efekty w stanie Tennessee, nie oznacza, że przyniesie takie same efekty w Kalifornii, Polsce czy gdziekolwiek indziej.

W przypadku kalifornijskiej reformy założono, że zmniejszenie liczebności klas na wczesnym etapie edukacji automatycznie uruchomi procesy prowadzące

do wyższych osiągnięć edukacyjnych dzieci – to jest, że nauczyciele będą spędzać mniej czasu na zarządzaniu zachowaniem, a więcej na dopasowaniu metod nauczania do indywidualnych potrzeb uczniów, co w konsekwencji doprowadzi też do większego zaangażowania i satysfakcji zawodowej. Problem polega na tym, że aby opisywany proces wystąpił, muszą być spełnione pewne dodatkowe warunki:

1. kadra pedagogiczna musi mieć odpowiednie kompetencje do tego, żeby dostosować sposób nauczania do indywidualnych potrzeb i predyspozycji dziecka,
2. nauczyciele muszą rzeczywiście dopasowywać swoje zachowania i style działania w zależności od wielkości klasy,
3. satysfakcja i zaangażowanie nauczycieli muszą rosnąć niezależnie od wynagrodzenia lub innych czynników niezwiązanych z liczebnością klasy.

Kalifornijska reforma zakładała również, że wszystkie dzieci skorzystają z niej tak samo, niezależnie od swojej sytuacji. Tymczasem analizy badań przekrojowych (omówionych wyżej) pokazują, że dzieci z rodzin o niższym statusie społeczno-ekonomicznym osiągają słabsze wyniki. Co więcej, eksperyment STAR wykazał, że największe korzyści z mniejszych klas czerpią właśnie dzieci z grup nieuprzywilejowanych i chłopcy.

Błędne okazało się zatem zarówno przekonanie, że samo zmniejszenie liczebności klas automatycznie uruchomi ten sam proces, który zadziałał w Tennessee, jak i że reforma przyniesie podobne efekty dla wszystkich dzieci. Błędy te były związane zarówno z faktem, że nie przeprowadzono szczegółowej analizy mechanizmu przyczynowego ani czynników, które go wspierają, jak i że niewłaściwie zinterpretowano wyniki eksperymentu STAR.



Problem 4

Jakie dodatkowe warunki muszą być spełnione, aby dany mechanizm zadziałał tu i teraz?



Program redukcji liczebności klas w Kalifornii wprowadzono niemal z dnia na dzień. Przyjęta regulacja zobowiązała wszystkie szkoły publiczne do redukcji o $\frac{1}{3}$ liczebności klas na etapie edukacji wczesnoszkolnej. Spowodowało to nagłą potrzebę zatrudnienia tysięcy nowych nauczycielek i nauczycieli oraz wygospodarowania odpowiedniej przestrzeni dla nowopowstałych oddziałów.

Przypomnijmy, że opisany wcześniej mechanizm przyczynowy (mniejsze klasy → więcej indywidualnej uwagi → lepsze uczenie się) ma szansę zadziałać tylko wtedy, gdy spełniony jest szereg warunków, określanych czasem jako **warunki zakresowe**. Jednym z nich jest dostępność odpowiednio wykwalifikowanej, **doświadczonej i wynagradzanej kadry nauczycielskiej**.

W momencie wprowadzenia reformy liczba osób mających odpowiednie uprawnienia do nauczania w klasach edukacji początkowej była w Kalifornii niewystarczająca. Co gorsza, gwałtowne zapotrzebowanie na nowy personel we wszystkich szkołach doprowadziło do odpływu doświadczonej kadry z dużych, miejskich okręgów szkolnych do szkół podmiejskich i pozamiejskich. W wielu szkołach brakowało nawet kadry do pracy w starszych klasach podstawówki, ponieważ część nauczycieli wybierała pracę w młodszych, nowo powstałych oddziałach. Aby uzupełnić braki kadrowe, zatrudniane były osoby bez doświadczenia pedagogicznego, mające niepełne kwalifikacje lub w trakcie uzyskiwania kwalifikacji do nauczania początkowego. Dane wskazują, że w 1997 roku (rok po wprowadzeniu reformy), 25% nauczycieli i nauczycielek w Kalifornii miało rok lub mniej doświadczenia zawodowego ([Jepsen i Rivkin, 2002](#)). W rezultacie miejskie szkoły o wyższym odsetku dzieci z rodzin o niższym statusie społeczno-ekonomicznym zostały pozbawione najważniejszego elementu odpowiedzialnego za sukces programu, tj. odpowiednio przygotowanego personelu nauczycielskiego.

Drugim warunkiem powodzenia reformy była dostępność odpowiedniej infrastruktury szkolnej. Program STAR w Tennessee objął tylko te szkoły, które miały przestrzeń do utworzenia dodatkowych oddziałów. W Kalifornii wygospodarowanie odpowiedniej przestrzeni było chaotyczne, często kosztem świetlic, zajęć dodatkowych lub innych miejsc wspierających naukę i rozwój. W wielu szkołach pogorszyło to warunki uczenia się – zwłaszcza tam, gdzie zasoby były ograniczone już przed reformą.

Ostatecznie reforma redukcji liczebności klas (CSR) nie przyniosła spodziewanych efektów mimo przeznaczonych na nią miliardów dolarów. Zwłaszcza te dzieci, które miały na programie skorzystać najbardziej, dotkliwie odczuły jej negatywne skutki. Dokładne oszacowanie efektów (pozytywnych i negatywnych) reformy nie jest jednak możliwe, ponieważ przed jej wprowadzeniem Kalifornia nie przeprowadzała ogólnostanowych pomiarów osiągnięć edukacyjnych dzieci po każdym roku nauki.



Czy można było uniknąć popełnionych błędów dzięki lepszemu wykorzystaniu dowodów naukowych, w tym z eksperymentu STAR?
Odpowiedź brzmi: tak.

W tym samym czasie, gdy była wprowadzana kalifornijska reforma, podobną inicjatywę wdrażano także w stanie Wisconsin w ramach programu SAGE (Student Achievement Guarantee in Education). Program miał ten sam cel – poprawę osiągnięć edukacyjnych dzieci dzięki mniejszym klasom. Jego wdrożenie wyglądało zupełnie inaczej. SAGE był wprowadzany stopniowo (zmiany rozłożono na lata 1996–1999) i był skierowany przede wszystkim do szkół z dużym odsetkiem dzieci z rodzin o niskich dochodach. Oprócz zmniejszenia klas obejmował także dodatkowe elementy wspierające: specjalny program nauczania, zajęcia pozaszkolne dla dzieci i społeczności lokalnej oraz system wsparcia zawodowego dla kadry nauczycielskiej, połączony z oceną postępów. W ramach programu dopuszczono też możliwość prowadzenia wszystkich lub części lekcji przez dwie osoby zamiast podziału oddziałów na mniejsze.

Analiza wyników tego kontrolowanego – choć nierandomizowanego – programu z Wisconsin wykazała, że dzieci należące do mniejszości i grup nieuprzywilejowanych istotnie skorzystały z wprowadzonych rozwiązań ([Filges i in., 2018](#)).

Dlaczego inicjatywa Sage przyniosła oczekiwane rezultaty, a kalifornijska reforma nie? Przyczyn jest oczywiście wiele, z czego niektóre są związane z różnicami w strukturze demograficznej Wisconsin i Kalifornii. Porównując podejście obu stanów, możemy jednak zauważyć, że w Wisconsin poświęcono więcej czasu i uwagi na analizę dostępnych dowodów, w tym dotyczących czynników wspierających mechanizm interwencji. Tymczasem Kalifornia skupiła się na szybkim wdrożeniu rozwiązania bez analizy warunków jego skuteczności.

Nieudana kalifornijska reforma ilustruje dwie ważne prawidłowości:

1. proces decyzyjny, który w luźnym stopniu opiera się na dowodach naukowych, nie tylko nie przynosi spodziewanych efektów, ale może przynieść więcej szkody niż pożytku;
2. wysokiej jakości dowody wspierają proces decyzyjny na każdym jego etapie.

Podsumowanie

Korzyści z podejścia opartego na dowodach

Budowanie polityk społecznych, w tym edukacyjnych, w oparciu na dowodach nie gwarantuje, że zostaną osiągnięte spodziewane efekty. Ale pominięcie dowodów naukowych zwiększa szanse niepowodzenia. Dlatego poniżej przedstawiamy prosty sposób na wstępną selekcję dostępnych badań tak, aby wyłuskać te, które są wyższej jakości i mogą zostać poddane dalszej ocenie z punktu widzenia wartości dowodowej.



Ocena wiarygodności badań przy użyciu kryteriów „WIEDZA”

Kryterium	Opis
 Wnioski	Zielone światło: • Główne wnioski są spójne z zakresem badania i typem przeprowadzonych analiz. • Wnioski opisują wszystkie oszacowane zależności, nie tylko te statystycznie istotne lub zgodne z przewidywaniami (hipotezami). • Wnioski zawierają informacje o ograniczeniach badania lub analiz.
	Czerwone światło: • Główne wnioski nie są spójne z zakresem badania lub typem analiz. Na przykład: • wnioski są przedstawione jako wskazujące na relację przyczynową, mimo że analizy miały charakter obserwacyjny lub korelacyjny; • badanie dotyczyło opinii kadry pedagogicznej na temat jakiegoś rozwiązania, a wnioski dotyczą wpływu tego rozwiązania na proces uczenia. • Wnioski koncentrują się tylko na statystycznie istotnych zależnościach, zgodnych z postawionymi hipotezami. Na przykład: • w badaniu sformułowano trzy hipotezy o spodziewanych relacjach między wynikami testów a płcią dzieci je rozwiązujących, ale przy przedstawieniu wniosków skoncentrowano się tylko na opisaniu dwóch, które pokazują istotne różnice między chłopcami i dziewczynkami. • Wnioski nie zawierają informacji o ograniczeniach badania, ze szczególnym uwzględnieniem ograniczeń metodologicznych lub analitycznych.

Kryterium	Opis
 Implementacja	Zielone światło: • Badanie wyraźnie identyfikuje możliwe sytuacje lub populacje, do których opisane mechanizmy bądź interwencje mogą być aplikowane. • Badanie wskazuje na możliwe zagrożenia związane z implementacją jakiegoś rozwiązania na dużą skalę. Czerwone światło: • Badanie, mimo że przeprowadzone na małej próbie lub szczególnej populacji, wskazuje szeroki wachlarz możliwości implementacji. Na przykład: • choć rezultaty opisanej w badaniu interwencji dotyczyły efektu ustrukturalizowanych prac domowych ćwiczących umiejętności pisania wśród dzieci w wieku przedszkolnym, autorzy i autorki badania przekonują, że wdrożenie prac domowych w klasach 1–3 będzie równie efektywne. • Badanie nie rozważa potencjalnych problemów związanych z implementacją polityki na dużą skalę. Na przykład: • badanie dotyczące efektywności prac domowych nie wskazało, że wdrożenie wymogu prac domowych może się wiązać z oporem części rodziców lub grona pedagogicznego.
 Efektywność	Zielone światło: • Badanie określa warunki wspierające lub ograniczające efektywność jakiejś interwencji lub wystąpienie jakiejś zależności. • Badanie wyraźnie wskazuje na koszty uzyskania zaobserwowanego efektu. Czerwone światło: • Badania nie identyfikują kosztów i warunków wspierających lub ograniczających obserwowane zależności czy efekty. Na przykład: • badanie dotyczące efektów ustrukturalizowanych prac domowych w przedszkolu nie wskazuje, że im bardziej liczna grupa przedszkolna, tym mniejszy pozytywny efekt prac domowych, oraz że taka interwencja wymaga przeszkolenia nauczycieli w zakresie budowania określonego typu instrukcji do prac domowych i ich oceny.

Kryterium	Opis
	Zielone światło: - Narzędzia wykorzystane do zebrania określonych informacji (np. do pomiaru wiedzy lub umiejętności uczniów) charakteryzują się wysokim poziomem trafności i rzetelności. - Sposób aplikacji narzędzi badawczych (zbierania danych) jest przejrzysto opisany (np. kiedy, jak i kto aplikował dane narzędzia). - Surowe dane są zdeponowane w uznanym repozytorium danych wraz z materiałem pozwalającym na odtworzenie krok po kroku wszelkich przekształceń dokonanych na zbiorze.
	Czerwone światło: - Narzędzia wykorzystane do zebrania określonych informacji zostały stworzone <i>ad hoc</i>. Sposób budowy narzędzia nie jest precyzyjnie opisany lub nie opiera się na sprawdzonych teoriach pomiaru. Trafność i rzetelność są ograniczone tylko do danego badania lub ich poziom jest nieokreślony. Na przykład: - w badaniu wykorzystano niestandardowe testy umiejętności, których trafność i rzetelność nie została sprawdzona na szerszą skalę; - do badania użyto „nowatorskiego” narzędzia, które jednak nie spełnia kryteriów przyjętych w naukach społecznych. - Sposób aplikacji narzędzi badawczych (zbierania danych) nie jest przejrzysto opisany (np. kiedy, jak i kto aplikował dane narzędzia). - Surowe dane nie zostały nigdzie zdeponowane i dostęp do nich jest uzależniony od dobrej woli auterek lub autorów badania.

Kryterium	Opis
-----------	------

Zielone światło:

- W badaniu jasno zdefiniowano cel, kluczowe pojęcia i zmienne oraz inne ważne parametry (np. miejsce i czas badania).
- Populacja docelowa została precyzyjnie opisana.
- Sposób doboru próby gwarantuje, że jest ona reprezentatywna dla populacji docelowej.
- W przypadku badań lub analiz przyczynowych sposób zapewnienia równoważności porównywanych grup jest szczegółowo opisany i zrealizowany.

Czerwone światło:

- Badanie nie przedstawia wprost, jaki jest jego główny cel, nie zawiera definicji kluczowych pojęć lub zmiennych i nie precyzuje parametrów geograficznych i czasowych, w jakich zostało przeprowadzone. Na przykład:
 - opis badania jest zbyt ogólny i nie pozwala na zrozumienie, na jakie dokładnie pytanie odpowiada (np. „celem niniejszego badania było zrozumienie wpływu nowoczesnych technologii na wyniki szkolne”);
 - kluczowe pojęcia lub zmienne nie zostały zdefiniowane lub zostały zdefiniowane nieprecyzyjnie (wykorzystując powyższy przykład, „nowoczesne technologie”, „wpływ”, jak i „wyniki szkolne” powinny być zdefiniowane w sposób umożliwiający powiązanie ich z odpowiednimi zmiennymi i wskaźnikami wykorzystanymi w badaniu);
 - w opisie badania powinna się znaleźć czytelna informacja dotycząca tego, w jakim okresie mierzono ew. „wpływ nowych technologii na wyniki” (tj. miesiąc, rok, kilka lat?) i gdzie.



Zakres

Kryterium	Opis
 Zakres	• Populacja docelowa nie została precyzyjnie opisana. Na przykład: • jeśli interesuje nas, jaki związek ma korzystanie z internetu z wynikami szkolnymi dzieci kończących podstawówkę, musimy tę populację bardzo dokładnie zdefiniować (czy chodzi o osoby kończące ósmą klasę, czy może dzieci w określonym wieku, czy chodzi o dzieci uczęszczające do wszystkich typów szkół podstawowych, czy tylko te kończące szkoły publiczne itd.) Bez dokładnej definicji nie do końca wiemy, czy poddana badaniu próba rzeczywiście jest czy nie jest reprezentatywna dla populacji docelowej. • Opis doboru próby jest niedokładny lub metoda doboru nie gwarantuje, że jest ona reprezentatywna dla populacji docelowej. • W sekcji metodologicznej badania brakuje szczegółowego opisu metody doboru próby albo opisana metoda jest nieadekwatna. Na przykład populacja docelowa została zdefiniowana jako osoby w wieku 15 lat, uczestników dobrano wyłącznie spośród ósmoklasistów i ósmoklasistek. • W przypadku badań lub analiz przyczynowych sposób osiągnięcia równoważności porównywanych grup nie jest zagwarantowany lub wystarczająco precyzyjnie opisany. • Mechanizmy losowego przypisania do grup eksperymentalnych lub zasady równoważności nie gwarantują, że grupy są zbilansowane (np. grupa poddana interwencji była dużo mniej liczna niż ta, która jej nie doświadczyła, lub więcej osób z grupy niepoddanej interwencji wycofało się z badania).

Kryterium	Opis
 Argumentacja	Zielone światło: • Przedstawione w badaniu wnioskowanie jest logiczne i zawiera opis przesłanek, na podstawie których sformułowano określone hipotezy. • Badanie opiera się na uprzednio opisanym i zweryfikowanym mechanizmie (modelu) przyczynowym lub teorii naukowej.
	Czerwone światło: • Przedstawione wnioskowanie nie jest logicznie spójne, hipotezy nie zostały wywiedzione z przesłanek, tylko zostały sformułowane ad hoc. Na przykład: • w badaniu została sformułowana hipoteza, że dziewczynki osiągają wyższe wyniki w matematyce, gdy osoba ucząca jest kobietą niż gdy jest mężczyzną; dlaczego jednak płeć nauczyciela(-ki) ma znaczenie i na jakiej podstawie wysnuto tę hipotezę, nie zostało wyjaśnione. • Badanie nie buduje na istniejącej wiedzy (czyli jej nie rozwija), tylko oferuje wyjaśnienia post hoc. Na przykład: • wyższe wyniki uzyskiwane przez dziewczynki uczone przez kobietę w porównaniu z tymi uczonymi przez mężczyznę wyjaśnione jest innym stylem komunikacji mężczyzn i kobiet. Tymczasem istnieje szereg teorii psychologicznych i socjologicznych, które mogłyby lepiej i precyzyjniej wyjaśnić zaobserwowaną korelację.

Aneks 1

Jak nie dać się nabrać na nieprawdę?

Nasze wartości, cele i poglądy mogą wpływać na sposób, w jaki selekcjonujemy dane i jak interpretujemy dowody. Każda osoba podejmująca decyzje, niezależnie od tego, czy jest badaczką, decycentką czy rodzicem, jest podatna na tzw. rozumowanie tendencyjne (*motivated reasoning*) i skłonna do wykorzystywania prostych, ale błędnych schematów myślenia.

Rozumowanie tendencyjne to termin naukowy. Dotyczy on mentalnej skłonności do poszukiwania, zapamiętywania, opisywania i przekazywania informacji w sposób, który jest spójny z dotychczasowymi przekonaniami.

Najczęstsze błędy poznawcze występujące w procesie badawczym

Typ błędu	Opis	Przykład
Błąd potwierdzenia (<i>confirmation bias</i>)	Skłonność do zauważania, interpretowania lub zapamiętywania informacji, które potwierdzają własną hipotezę.	Badacz lub badaczka przywiązuje większą wagę do danych potwierdzających hipotezę, pomijając dane, które im przeczą.
Stronniczość doboru (<i>selection bias</i>)	Systematyczne zniekształcenie wynikające ze sposobu wyboru badanych jednostek lub danych.	Wybieranie do programu pilotażowego szkół, które już osiągały dobre wyniki.
Stronniczość pomiaru (<i>measurement bias</i>)	Metody zbierania danych sprzyjają określonym wynikom.	Zadania na teście są sformułowane w sposób, który jest bardziej zrozumiały dla osób z rodzin o wyższym statusie społeczno-ekonomicznym, np. „wyobraź sobie, że Twoja rodzina planuje wakacje na Malediwach...”.
Stronniczość interpretacji (<i>interpretation bias</i>)	Stronnicze rozumowanie podczas analizy lub wyjaśniania wyników.	Opisywanie wyników badania jako jednoznacznie wspierających jakąś hipotezę, mimo że analizy nie pozwalają na tak kategoryczne wnioski.
Błąd publikacji (<i>publication bias</i>)	Czasopisma częściej publikują badania z pozytywnymi lub istotnymi wynikami.	Wyniki badań, w których nie zaobserwowano żadnych istotnych efektów, nie są publikowane, co zniekształca bazę dowodową.

Najczęstsze błędy poznawcze występujące w procesie decyzyjnym

Typ błędu	Opis	Przykład
Selektywne dobieranie danych (<i>cherry picking</i>)	Wybieranie tylko tych danych, które pasują do określonej tezy lub wcześniejszych założeń.	Osoby popierające reformę zmniejszania liczebności klas przedstawiają wyłącznie wyniki badań pokazujących jej pozytywny efekt, a ignorują badania wskazujące na brak efektu lub efekt negatywny. I na odwrót – osoby przeciwnie reformie ignorują dane wskazujące na pozytywny efekt w określonych kontekstach.
Odwołanie do autorytetu (<i>appeal to authority</i>)	Używanie opinii lub pozycji autorytetu (osoby lub instytucji) zamiast rzeczywistego argumentu.	W debacie na temat skutecznych metod nauczania matematyki, zamiast wyników badań przywoływane są nazwiska i tytuły naukowe autorów lub autorek konkurencyjnych podejść.
Fałszywa alternatywa (<i>false alternative</i>)	Przedstawienie dwóch możliwości jako jedynych, mimo że istnieją inne opcje.	W dyskusji na temat reformy na rzecz poprawy wyników nauczania początkowego, wybór między inwestycją w zmniejszenie liczebności klas a inwestycją w podniesienie kompetencji nauczycielskich przedstawiony jest jako jedyny możliwy. Tymczasem istnieje wachlarz możliwych rozwiązań.
Dowód anegdotyczny (<i>anecdotal evidence</i>)	Wykorzystywanie osobistego doświadczenia lub pojedynczego przypadku zamiast solidnych danych, często w celu odrzucenia statystyk.	Osoby uczestniczące w debacie na temat prac domowych odwołują się do opinii swoich dzieci zamiast do wyników badań naukowych.
Błąd ekologiczny/ błąd indywidualizmu (<i>ecological/ individualistic fallacy</i>)	Zakładanie, że to, co prawdziwe dla jednej części, musi być prawdziwe dla całości (lub odwrotnie).	Błąd ekologiczny: interpretowanie niższych średnich wyników testów po pandemii COVID-19 jako wskazujących na obniżenie umiejętności wśród wszystkich dzieci, podczas gdy ich obniżenie odnotowano głównie wśród dzieci z rodzin o niższym statusie społeczno-ekonomicznym. Błąd indywidualizmu: interpretowanie niższych od średniej krajowej wyników testów dzieci w danym okręgu jako wskazujących na niską jakość nauczania w nim.

Typ błędu	Opis	Przykład
Błąd fałszywej przyczyny (<i>false cause</i>)	Przypisywanie relacji przyczynowej między zdarzeniami, które jedynie występują razem.	Interpretowanie spadku osiągnięć szkolnych dzieci w trakcie i po pandemii jako jej wyniku.
Odwołanie do emocji (<i>appeal to emotions</i>)	Wykorzystywanie emocji zamiast logicznych argumentów.	Argumenty typu: „Chodzi przecież o dobro dzieci! Nie możemy czekać na wyniki badań, bo dzieci nie mogą czekać!”.
Stronniczość grupowa (<i>in-group bias</i>)	Faworyzowanie osób należących do „naszej” grupy i deprecjonowanie osób z zewnątrz.	Nauczyciele preferują interwencje oparte na praktyce innych nauczycieli, a odnoszą się z niechęcią do badań akademickich.
Efekt zakotwiczenia (<i>anchoring</i>)	Poleganie zbyt mocno na pierwszej informacji (kotwicy) przy podejmowaniu decyzji.	Oparcie decyzji o wprowadzeniu małych klas na wyniku pierwszego eksperymentu dotyczącego efektów tej interwencji.
Błąd utopionych kosztów (<i>sunk cost fallacy</i>)	Kontynuowanie działań, ponieważ włożono w nie już dużo wysiłku lub środków, mimo że nie ma to sensu.	Kontynuowanie inwestycji w małe klasy w Kalifornii do 2013 roku, mimo napływu dowodów wskazujących na efekt niewspółmierny do kosztów.
Efekt sformułowania (<i>framing effect</i>)	Sposób przedstawienia informacji wpływa na jej interpretację.	Przedstawianie decyzji o wprowadzeniu edukacji zdrowotnej jako „zamachu na prawa rodziców” lub decyzji o ograniczeniu prac domowych jako „obniżania standardów edukacyjnych.”
Błąd nadmiernego optymizmu (<i>optimism bias</i>)	Skłonność do przeceniania prawdopodobieństwa pozytywnych wyników.	Przecenianie prawdopodobieństwa, że interwencja, która przyniosła pozytywne efekty w pewnej grupie szkół, przyniesie takie same pozytywne efekty we wszystkich szkołach.
Pesymizm retrospektywny (<i>declinism</i>)	Przekonanie, że świat (lub dana dziedzina) się pogarsza w porównaniu ze stanem z przeszłości.	Pogląd, że umiejętności poznawcze i społeczne wśród dzieci są dramatycznie niższe niż kiedy (i oparte na tym przeświadczeniu zakazy).

Aneks 2

Poziom decyzji a wykorzystanie badań

Osoby podejmujące decyzje na różnych szczeblach systemu oświaty mogą potrzebować odmiennych badań i analiz. Poniżej wyróżniono trzy ogólne „poziomy” decyzyjne i pięć wymiarów, które w przybliżeniu można traktować jako wskazówkę przy poszukiwaniu odpowiednich publikacji i analiz naukowych.

Typ badań a poziom decyzyjny

Poziom decyzyjny	Wymiar	Opis
MAKRO Ustawodawca i osoby decyzyjne na szczeblu krajowym lub wojewódzkim	Zakres	Nacisk na trafność zewnętrzną i wdrażanie polityk na szeroką skalę. Potrzeba analiz i metaanaliz opartych na badaniach prowadzonych w różnych kontekstach.
	Mierzone rezultaty	Potrzeba analiz krótko- i długoterminowych dotyczących nie tylko osiągnięć edukacyjnych, ale także pozaedukacyjnych rezultatów, takich jak ogólny dobrostan psychospołeczny dzieci i młodzieży, przygotowanie do rynku pracy, kapitał ludzki, nierówności, retencja i jakość kadry pedagogicznej itp.
	Jakość analiz	Zróznicowane potrzeby co do jakości dowodowej: badania korelacyjne lub opisowe mogą być przydatne do diagnozowania obszarów wymagających wsparcia lub identyfikowania grup osiągających wysokie i niskie wyniki w celu dalszych analiz.
	Potencjał zlecania badań	Wysoki poziom możliwości w zakresie prowadzenia i zlecania badań na dużych próbach (w tym badań eksperymentalnych), a także gromadzenia i interpretowania danych.
	Koszty	Możliwość wpływania na poziom i strukturę finansowania wymaga dowodów dotyczących efektywności kosztowej interwencji.

Poziom decyzyjny	Wymiar	Opis
MEZO Przedstawiciele i przedstawicielki agencji edukacyjnych lub lokalne decydentki i decydenci odpowiedzialni za szkoły w danym regionie/ samorządzie	Zakres	Większa przydatność badań o ograniczonym zakresie (np. dotyczących danego miasta lub gminy). Potrzeba analiz przyczynowych uwzględniających specyficzne uwarunkowania.
	Mierzone rezultaty	Główny nacisk na oszacowania wpływu polityk na krótkoterminowe wyniki edukacyjne, takie jak wyniki testów, ukończenie szkoły czy dobrostan psychospołeczny lub zachowania negatywne (np. przemoc rówieśnicza).
	Jakość analiz	Wysokie wymagania wobec jakości badań ze względu na ograniczone możliwości „testowania” alternatywnych rozwiązań. Duży nacisk na szczegółowy opis mechanizmów przyczynowych.
	Potencjał zlecania badań	Z wyjątkiem nielicznych przypadków (np. dużych aglomeracji), zazwyczaj ograniczone możliwości gromadzenia, interpretowania i zlecania badań. Kluczowe znaczenie ma zatem kontrola jakości metodologii i analiz w ew. zlecanych badaniach.
	Koszty	Decyzje polityczne podejmowane są w warunkach ograniczonych środków finansowych. Efektywność kosztowa szeroko zakrojonych reform jest zazwyczaj określona na wyższym poziomie decyzyjnym.

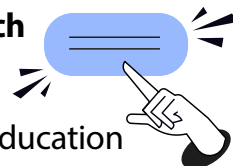
Poziom decyzyjny	Wymiar	Opis
MIKRO Personel szkolny	Zakres	Decyzje skoncentrowane na lokalnym kontekście klasy lub szkoły. Potrzeba wyników analiz w podobnych kontekstach.
	Mierzone rezultaty	Krótkoterminowe wyniki uczniów, w tym również pozaszkolne, ze szczególnym uwzględnieniem dowodów dotyczących wdrażania określonych interwencji.
	Jakość analiz	Możliwe wykorzystywanie analiz jakościowych dotyczących przebiegu jakiejś interwencji i opisowych dotyczących wyników uczniów na poziomie szkoły w dłuższym okresie. Potrzeba wysokiej jakości dowodów skuteczności interwencji związanych z procesem nauczania i uczenia się.
	Potencjał zlecania badań	Bardzo niski. Badania są raczej zlecane przez związki zawodowe lub agencje edukacyjne, choć rozpowszechniane z wykorzystaniem sieci nauczycielskich. Wysoki poziom uzależnienia od instytucji nadrzędnych i stron trzecich przy interpretacji wyników.
	Koszty	Przeważnie osobiste lub analizowane na poziomie szkoły.

Gdzie szukać wiarygodnych dowodów?

Źródła wiarygodnych dowodów

1. Instytut Badań Edukacyjnych: Przewodnik po strategiach edukacyjnych

Przewodnik po strategiach edukacyjnych (ang. *Teaching and learning toolkit*)



jest adaptacją narzędzia opracowanego przez brytyjską fundację non-profit Education

Endowment Foundation (EEF). Narzędzie to zostało zaadaptowane do polskiego kontekstu edukacyjnego przez Instytut Badań Edukacyjnych – Państwowy Instytut Badawczy.

Celem Przewodnika jest dostarczanie syntetycznych, opartych na dowodach naukowych informacji na temat skuteczności wybranych strategii edukacyjnych, rozumianych jako sposoby pracy dydaktycznej, organizacji nauczania i wspierania uczenia się uczniów.

Przewodnik porządkuje dostępne wyniki badań empirycznych (w tym metaanaliz i przeglądów systematycznych), wskazując względną siłę dowodów, potencjalny wpływ strategii na osiągnięcia uczniów oraz warunki ich skutecznego stosowania.

Narzędzie może być wykorzystywane zarówno przez nauczycieli oraz dyrekcję szkół, jak i przez osoby odpowiedzialne za projektowanie i ocenę działań oraz polityk edukacyjnych opartych na dowodach.

<https://przewodnik.ibe.edu.pl/>

2. IES: What Works Clearinghouse

What Works Clearinghouse (WWC) jest źródłem dowodów naukowych

dotyczących efektywności różnych programów, produktów, praktyk i polityk

edukacyjnych. WWC jest inicjatywą stworzoną w ramach Institute of Education

Sciences (IES) Departamentu Edukacji USA. Institute of Education Sciences (IES) jest jednostką zajmującą się statystyką, badaniami i ewaluacją. Celem WWC jest przegląd istniejących badań, ocena, które z nich spełniają rygorystyczne standardy, i podsumowanie wyników.



Aby odpowiedzieć na pytanie: „Co działa w edukacji?”, WWC wychodzi z założenia, że nie wszystkie badania edukacyjne są sobie równe. Dlatego koncentruje się na opisywaniu standardów i wskazywaniu dobrze zaprojektowanych badań, wiarygodnych wyników oraz znaczących ustaleń, które mogą wspierać podejmowanie decyzji edukacyjnych.

<https://ies.ed.gov/ncee/WWC/>

3. Best Evidence Encyclopedia

Best Evidence Encyclopedia (BEE) została stworzona przez Center for Research and Reform in Education (CRRE) przy Johns Hopkins University School of Education. Jej celem jest dostarczanie nauczycielom oraz badaczom i badaczom rzetelnych i użytecznych informacji na temat siły dowodów naukowych dotyczących różnych programów edukacyjnych dostępnych dla uczniów od przedszkola do końca liceum.



BEE zawiera głównie systematyczne metaanalizy badań nad skutecznymi programami w zakresie czytania, matematyki, pisania, nauk przyrodniczych, edukacji wczesnoszkolnej i innych obszarów. Zawiera również artykuły dotyczące metod przeglądów badawczych oraz zagadnień takich jak polityka edukacji specjalnej i reforma oparta na dowodach.

<https://bestevidence.org/>

4. Education Endowment Foundation

Education Endowment Foundation (EEF) została założona w 2011 roku przez Sutton Trust dzięki funduszowi w wysokości 125 mln funtów przekazanemu przez Department for Education. EEF i Sutton Trust zgodnie z decyzją brytyjskiego rządu od 2013 roku tworzą What Works Centre for Education.

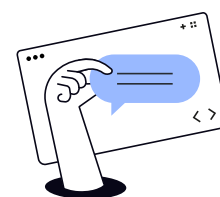


EEF zakłada, że najlepszym sposobem na poprawę nauczania i uczenia się jest zapewnienie sektorowi edukacji wiarygodnych i praktycznych informacji o tym, jak podnosić wyniki. Baza opartych na dowodach naukowych opracowań dotyczących efektywnych praktyk edukacyjnych obejmuje wachlarz interwencji dotyczących dzieci w wieku od 2 do 19 lat.

<https://educationendowmentfoundation.org.uk/>

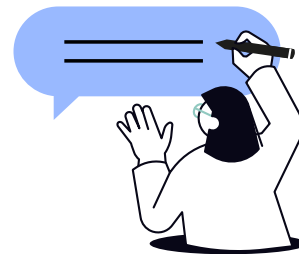
5. Campbell Collaboration

Campbell Collaboration jest ośrodkiem recenzującym i publikującym syntezy dowodów naukowych. Campbell Systematic Reviews stosują najwyższe standardy metodologiczne dla przeglądów systematycznych badań i map luk dowodowych. Przeglądy badań w ramach Campbell Collaboration podsumowują najnowszy stan wiedzy w formie informacji gotowych do wykorzystania w procesach decyzyjnych.



<https://www.campbellcollaboration.org/> i <https://onlinelibrary.wiley.com/index/18911803?=&startPage=&ConceptID=229187>

Słowniczek



- Badania podłużne – polegają na tym, że te same osoby są badane wielokrotnie w regularnych odstępach czasu. Przykład badania podłużnego dotyczącego struktury społecznej w Polsce: badanie [POLPAN](#).
- Badania przekrojowe – polegają na tym, że różne osoby są badane w jednym wybranym momencie. Przykłady: badania [NAEP](#), [PISA](#), [TIMSS](#), [PIRLS](#), [ICCS](#).
- Eksperyment – taka konstrukcja badania, która umożliwia kontrolowany sposób wprowadzenia interwencji (zmienną niezależną) i pomiar jej skutków (zmienną zależną) wśród co najmniej dwóch grup, przy czym osoby są losowo przypisane do tych grup.
- Korelacja – związek/współwystępowanie dwóch zmiennych. Miara statystyczna korelacji opisuje siłę i kierunek związku między dwiema zmiennymi – pokazuje, w jakim stopniu zmiany jednej zmiennej wiążą się ze zmianami drugiej, ale nie oznacza, że jedna powoduje drugą.
- Odchylenie standardowe – stanowi (podobnie jak wariancja) jeden ze sposobów mierzenia rozproszenia/rozrzutu wyników wokół średniej. Im mniej obserwacje są skupione wokół średniej, tym większe odchylenie standardowe.
- Trafność wewnętrzna – w kontekście badań eksperymentalnych to stopień, w jakim badanie pozwala na wiarygodne ustalenie związku przyczynowego między interwencją (zmienną niezależną) a obserwowanym wynikiem. Badanie ma wysoką trafność wewnętrzną, jeśli można z dużą pewnością stwierdzić, że to interwencja, a nie inne czynniki, spowodowała zaobserwowaną zmianę.
- Trafność zewnętrzna – w kontekście badań eksperymentalnych to stopień, w jakim wyniki badania można uogólniać na inne osoby, sytuacje, miejsca lub czas poza badanymi próbami lub warunkami eksperymentu. Im wyższa trafność zewnętrzna, tym bardziej można zakładać, że te same zależności co w badaniu wystąpią również w rzeczywistych codziennych warunkach lub w innych populacjach.

- Wariancja (zróźnicowanie) – miara pokazująca, jak bardzo wyniki różnią się od średniej. Duża wariancja oznacza duże rozproszenie danych, a mała – że wyniki są blisko średniej. Pierwiastek kwadratowy z wariancji to odchylenie standardowe.
- Wielkość efektu – w badaniach eksperymentalnych to miara tego, jak duży wpływ miała interwencja na wyniki, czyli jak bardzo różnią się wyniki grupy eksperymentalnej i kontrolnej.
- Zmienna – to cecha osób lub innych jednostek, która może przyjmować różne wartości.

Bibliografia



Baker, B.D., Knight, D. (2025). *Does money matter in education?* (3rd ed.). Albert Shanker Institute. <https://www.shankerinstitute.org/resource/does-money-matter-in-education>

California Department of Education. (2023). *Class size reduction K-3*.
<https://www.cde.ca.gov/re/pr/csrk3.asp>

Cartwright, N., Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.
<https://doi.org/10.1093/acprof:osobl/9780199841608.001.0001>

Cheung, A. C. K., Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283–292.
<https://doi.org/10.3102/0013189X16656615>

Chingos, M. (2011). *The false promise of class-size reduction*. Center for American Progress.
<https://files.eric.ed.gov/fulltext/ED536071.pdf>

Dee, T. S., Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30(3), 418–446.
<https://www.jstor.org/stable/23018959>

Filges, T., Sonne-Schmidt, C. S., Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, 14(1), 1–107. <https://doi.org/10.4073/csr.2018.10>

Glass, G. V., Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1(1), 2–16.
<https://doi.org/10.3102/01623737001001002>

Gopalan, M., Rosinger, K., Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, 44(1), 218–243. <https://doi.org/10.3102/0091732X20903302>

Hoxby, C. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, 115(4), 1239–1285.

<https://doi.org/10.1162/003355300555060>

Hulme, C., West, G., Rios Diaz, M., Hearne, S., Korell, C., Duta, M., Snowling, M. J. (2025). The Nuffield Early Language Intervention (NELI) programme is associated with lasting improvements in children's language and reading skills. *Journal of Child Psychology and Psychiatry*, 66, 1357–1365. <https://doi.org/10.1111/jcpp.14157>

Jepsen, C., Rivkin, S. (2002). *Class size reduction, teacher quality, and academic achievement in California public elementary schools*. Public Policy Institute of California.

Karpiński, Z., Di Pietro, G., Biagi, F. (2023). Non-cognitive skills and social gaps in digital skills: Evidence from ICILS 2018. *Learning and Individual Differences*, 102, 102254.

<https://doi.org/10.1016/j.lindif.2022.102254>

Kim, Y., Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychologist*, 51(3–4), 395–405. <https://doi.org/10.1080/00461520.2016.1207177>

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>

Kraft, M. A., Brunner, E. J., Dougherty, S. M., Schwegman, D. J. (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics*, 188, 104212. <https://doi.org/10.1016/j.jpubeco.2020.104212>

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2), 497–532. <http://www.jstor.org/stable/2587015>

Odden, A. (1990). Class size and student achievement: Research-based policy alternatives. *Educational Evaluation and Policy Analysis*, 12(2), 213–227.

<https://doi.org/10.3102/01623737012002213>

Quintero, E. (2025). *Literacy policy and NAEP*. Albert Shanker Institute.

<https://www.shankerinstitute.org/blog/literacy-policy-and-naep>

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs (Report from the Governing Board of the American Educational Research Association Grants Program)*. American Educational Research Association.

<https://www.aera.net/portals/38/docs/causal%20effects.pdf>

Shadish, W. R., Cook, T. D., Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

https://moodle2.units.it/pluginfile.php/132646/mod_resource/content/1/Estratto_ShadishCookCampbellExperimental2002.pdf

Stasz, C., Stecher, B. M. (2001). *Before and after class size reduction: A tale of two teachers*. RAND Corporation. <https://www.rand.org/pubs/drafts/DRU2454.html>

WWC. (2014). *Procedures and standards handbook (Version 3.0)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_draft_standards_handbook.pdf

WWC. (2022). *Procedures and standards handbook (Version 5.0)*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/WWC/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

Winship, C., Morgan, S. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659–706. <https://doi.org/10.1146/annurev.soc.25.1.659>

Wyckoff, J. H. (2025). *Puzzling over declining academic achievement* (EdWorkingPaper 25–1197). Annenberg Institute at Brown University. <https://doi.org/10.26300/kjaa-ec47>

Strony internetowe badań

Boston College. Lynch School of Education and Human Development. (2025).

IEA TIMMS & PIRLS. <https://timssandpirls.bc.edu>

International Association for the Evaluation of Educational Achievement. (2025).

ICCS (International Civic and Citizenship Education Study).

<https://www.iea.nl/studies/iea/iccs>

National Center for Education Statistics. (2025). NAEP (National Assessment of Educational Progress). <https://nces.ed.gov/nationsreportcard>

The OECD (Organisation for Economic Co-operation and Development). (2025).

PISA (Programme for International Student Assessment).

<https://www.oecd.org/en/about/programmes/pisa.html>



Autorka: Kinga Wysieńska-Di Carlo

Redakcja merytoryczna: Aleksandra Jasińska-Maciążek

Współpraca: Kinga Karteczka-Świątek, Marzena Binkiewicz, Sylwia Opozda-Suder, Artur Pokropek, Anna Ziółkowska

Redakcja językowa: Michał Pranke

Projekt okładki: Anna Nowak

Skład: Anna Nowak

Ilustracje: Shutterstock

Wydawca:



INSTYTUT BADAŃ
EDUKACYJNYCH
Państwowy Instytut Badawczy

Instytut Badań Edukacyjnych – Państwowy Instytut Badawczy

ul. Górczewska 8, 01-180 Warszawa

tel. (22) 241 71 00; www.ibe.edu.pl

ISBN: 978-83-68747-15-7

DOI: 10.24131/9788368747157

Opracowano w ramach zadania: „Ewaluacja programów i przedsięwzięć istotnych dla rozwoju systemu kształcenia i uczenia się przez całe życie”, finansowanego ze środków Ministerstwa Edukacji Narodowej na podstawie umowy nr MEN/2026/DIR/38 z dnia 22 stycznia 2026 roku.

Copyright © Instytut Badań Edukacyjnych – Państwowy Instytut Badawczy 2026

Wzór cytowania:

Wysieńska Di-Carlo, K. (2026). *Historia pewnej reformy, czyli opowieść o dobrym i złym wykorzystywaniu dowodów naukowych w procesach decyzyjnych w oświacie*. Instytut Badań Edukacyjnych – Państwowy Instytut Badawczy.

Warszawa 2026

Egzemplarz bezpłatny