Polskie Forum Psychologiczne, 2018, tom 23, numer 2, s. 252-276
DOI: 10.14656/PFP20180203

METAANALIZA: O ILOŚCIOWYM SYNTETYZOWANIU USTALEŃ EMPIRYCZNYCH

Maciej Karwowski

Instytut Psychologii, Uniwersytet Wrocławski
Institute of Psychology, University of Wroclaw

META-ANALYSIS: ON QUANTITATIVE SYNTHESIZING OF EMPIRICAL RESULTS

Summary. This paper characterizes meta-analysis: the quantitative synthesis of previous studies, as a separate research method, of growing popularity in the social sciences. We focus on its main functions and goals and describe in details the process of conducting meta-analyses. Different statistical models within meta-analysis (fixed effects, random effects, multilevel meta-analysis) as well as traditional and more modern methods allowing for control for publication bias are discussed. The paper finishes with an example of a brief meta-analysis, allowing the readers for a step-by-step overview of analyses conducted and conducting their own meta-analyses.
Key words: meta-analysis, effect size, publication bias

Wprowadzenie

Metaanaliza to wystandaryzowana procedura ilościowej syntezy wcześniejszych wyników badań. Choć jej początków upatruje się w tekście Karla Pearsona (1904), powstanie metaanalizy w formie bliskiej dziś stosowanej, zawdzięczamy sporowi o skuteczność psychoterapii między sceptycznym (a opierającym swój sceptycyzm na przeglądzie literatury) Hansem Eysenckiem (1952) a będącym beneficjentem psychoterapii Genem Glassem (Smith i Glass, 1977). Po serii artykułów Eysencka kwestionujących skuteczność oddziaływań psychoterapeutycznych, Smith i Glass (1977) zagregowali niemal czterysta badań (w odróżnieniu od jedenastu uwzględnionych przez Eysencka), wykazując, że ważony wielkością próby efekt skuteczności psychoterapii, wyrażony miarą *d* Cohena (Cohen, 1988), to $d = 0{,}68$. Oznacza to, że przeciętny uczestnik psychoterapii po jej zakończeniu uzyskiwał rezultaty o $\frac{2}{3}$ odchylenia standardowego wyższe niż przeciętny reprezentant grupy kontrolnej bądź

Adres do korespondencji: Maciej Karwowski, e-mail: maciek.karwowski@gmail.com

też – ujmując ten wynik w inny sposób – że przeciętny uczestnik psychoterapii po jej zakończeniu funkcjonował lepiej niż 75% przedstawicieli grupy kontrolnej. I choć późniejsze metaanalizy pokazały, że efekt ten jest słabszy (Lipsey i Wilson, 1993, $d = 0{,}47$), spór Glass-Eysenck dał asumpt do narodzin metaanalizy.

Celem tego artykułu jest charakterystyka metaanalizy, jej kolejnych kroków oraz pułapek czekających na realizujących ją badaczy. Nie jest to jednak oczywiście pełen wykład na jej temat – nie przypadkiem w końcu poświęca się metaanalizie całe książki (Hedges i Olkin, 1985; Rosenthal, 1991; Lipsey i Wilson, 2001; Hunter i Schmidt, 2004) lub serie artykułów (DerSimonian i Laird, 1986; Brockwell i Gordon, 2001). Metaanaliza doczekała się też poświęconego sobie pisma: *Research Synthesis Methods*, a opisujące ją prace (np.: Simon, 2010a, 2010b, 2010c; Walecka i Zakrzewska-Bielawska, 2016) czy przykłady praktycznego zastosowania (np.: Wiśniewska i Karwowski, 2007) daje się odnaleźć również w rodzimej literaturze. Stąd też treści zawarte poniżej traktować trzeba jako z konieczności selektywne i zorientowane na kluczowe charakterystyki metaanalizy.

Dlaczego metaanaliza ma znaczenie?

Z jakich powodów warto zajmować się metaanalizą? Czy i w jaki sposób przynosi ona rezultaty bardziej wiarygodne i zasługujące na uznanie niż wyniki badań pierwotnych? To zapewne podstawowe pytania, jakie pojawiają się już na wstępie myślenia o jakimkolwiek własnym projekcie tego typu. Co zatem czyni studia metaanalityczne wartymi zainteresowania? Cztery argumenty mają tu szczególne znaczenie.

Po pierwsze, rola metaanaliz we współczesnych naukach społecznych rośnie na fali dyskusji o trudnościach z replikowalnością niektórych klasycznych efektów w psychologii (Pashler i Wagenmakers, 2012) i ogólniejszego zjawiska „wątpliwych praktyk badawczych" (*questionable research practices*, QRP; John, Loewenstein i Prelec, 2012)[1]. Metaanalityczne podsumowania nie tylko są w stanie wychwycić niespójno-

[1] Praktyki te obejmują szerokie spektrum działań, które łączy to, że prowadzą one do takiego analizowania wyników, aby rezultat był zgodny z oczekiwaniami badaczy. Przykładem QRP może być nieuzasadnione wykluczanie pewnych przypadków z analizy czy też pomijanie w doniesieniu informacji o wszystkich mierzonych zmiennych zależnych. Wykluczenie niektórych zmiennych bądź badanych może prowadzić do zmiany kluczowych rezultatów. Za QRP uznawane jest też błędne (lecz świadome) zaokrąglanie wartości p – na przykład raportowanie wartości $p = 0{,}051$ jako $p < 0{,}05$ oraz tzw. HARKing (*hypothesizing after the results are known*), tj. sugerowanie, że jakiś rezultat był oczekiwany *a priori*, podczas gdy jest on wynikiem analizy eksploracyjnej. Wreszcie QRP jest raportowanie wyłącznie tych eksperymentów, które przynoszą wyniki statystycznie istotne, bez wspomnienia o tych, których rezultaty dały wyniki zerowe. QRP są więc zróżnicowane i nie należy ich utożsamiać ze zwykłym fałszerstwem i manipulacją danymi.

ści pojedynczych badań, ale także wskazać na budzące wątpliwości ustalenia płynące z prac zespołów badawczych czy pojedynczych uczonych (Fanelli, 2009). W tym sensie metaanaliza jest więc audytem badawczym o szczególnym znaczeniu dla nomotetycznie zorientowanej nauki – pozwala ocenić generalizowalność ustaleń w danym obszarze; pokazuje co jest replikowalne, co zaś może być artefaktem, błędem I rodzaju lub też efektem selektywnego publikowania (skrzywienia publikacyjnego: *publication bias*) – tendencji redaktorów i recenzentów (ale i samych autorów) do publikowania rezultatów, które potwierdzają postawione hipotezy i przynoszą wyniki statystycznie istotne, pomijanie zaś wyników zerowych.

Po drugie, w dobie publikowanych tysięcy doniesień naukowych, badacze nie są w stanie szczegółowo analizować każdego pojawiającego się badania. Metaanaliza, syntetyzując te ustalenia, cieszy się wciąż rosnącą popularnością, bo pozwala na zredukowanie szumu informacyjnego. To niebagatelna zaleta.

Po trzecie, w niektórych obszarach takich jak edukacja (Hattie, 2008) czy psychoterapia (Weisz i in., 1995) znaczenie metaanaliz jest szczególne tak z powodów poznawczych, jak i praktycznych. Czy warto inwestować niemałe środki we wdrożenie nowej metody nauczania? Czy nowa terapia jest dostatecznie skuteczna, aby ją refundować? To kardynalne kwestie dla polityki oświatowej lub zdrowotnej, do odpowiedzi na które metaanaliza jest szczególnie przydatna. Nie oznacza to, że znaczenia nie mają tu oryginalne, dedykowane tym problemom studia, ale ich synteza ma oczywiste walory – przede wszystkim możliwość testowania poziomu korroboracji wcześniejszych ustaleń.

Po czwarte wreszcie, metaanaliza nie tylko dostarcza podsumowania stanu badań na jakiś temat, ale sama może – a wręcz powinna – być źródłem nowych hipotez i kolejnych badań. Metaanalityczna synteza niewiele nam powie o mechanizmach uzyskiwanego efektu, w szczególności możliwych efektach mediacyjnych, może jednak inspirować do podjęcia nowych studiów, wskazując na luki w obecnym stanie wiedzy i sprzyjając generowaniu nowych hipotez. Dobra metaanaliza podsumowuje więc stan badań, ale i wskazuje ich nowe kierunki.

Kluczowe elementy i rodzaje metaanaliz

Przed bardziej szczegółową prezentacją etapów realizacji metaanalizy scharakteryzujmy pokrótce jej kluczowe składowe i najczęstsze sposoby realizacji. Mowa tu o czterech podstawowych elementach: (1) wielkości efektu, (2) rodzajach wag i sposobach ważenia efektów w metaanalizie, (3) rodzaju zastosowanych analiz statystycznych oraz (4) typowych sposobach szacowania tzw. skrzywienia (obciążenia) publikacyjnego (*publication bias*).

Wielkość efektu (*effect size*: ES), to w dzisiejszej empirycznej psychologii kategoria dobrze znana: miara siły jakiegoś zjawiska. Może być nią związek między dwiema lub większą liczbą zmiennych: a więc dowolny współczynnik korelacji; może mieć ona jednak charakter odmienny, zależny od konkretnego schematu badawczego.

W badaniach porównawczych i eksperymentalnych będzie to więc zazwyczaj standaryzowana różnica między średnimi: grupą eksperymentalną i kontrolną w poście bądź też grupą kryterialną i porównawczą w schemacie porównawczym (np.: między dziewczętami a chłopcami). Wielkości efektu oparte na współczynniku korelacji Pearsona, czasem odpowiednio przetransponowanym oraz na standaryzowanej różnicy średnich – zazwyczaj *d* Cohena, *g* Hedgesa bądź Δ Glassa[2] są najpopularniejsze w metaanalizach realizowanych w psychologii. Medycy stosunkowo często posługują się także ilorazem szans (OR: *odds ratio*) dla opisania prawdopodobieństwa wystąpienia jakiegoś zjawiska – na przykład wyleczenia pacjenta bądź jego zgonu. Różne rodzaje (rodziny) efektów daje się między sobą łatwo przeliczać – każdy podręcznik metaanalizy (np.: Lipsey i Wilson, 2001) oraz wiele kalkulatorów dostępnych online (np.: http://www.lyonsmorris.com/ma1/) pozwala na łatwe przejście z *r* na *d* i odwrotnie, jak również uzyskanie tych efektów z podstawowych statystyk raportowanych w badaniach (zob. np.: tabela 1).

Waga ma szczególne znaczenie w metaanalizie, nie jest bowiem tak, że każde włączone do niej badanie jest traktowane jednakowo. Najczęściej stosowane są dwa różne rodzaje wag, choć istota obu sprowadza się do uznania, że efekty uzyskane w badaniach zrealizowanych na większych próbach, w związku z mniejszym poziomem błędu standardowego, zasługują na większą (u)wagę. Hunter i Schmidt (2004) proponują ważenie uzyskiwanych efektów przez wielkość próby. W praktyce oznacza to zatem przemnożenie każdego efektu przez wielkość próby w badaniu, w którym efekt ten został uzyskany, zsumowanie tak uzyskanych wartości we włączonych do metaanalizy badaniach, a następnie podzielenie całości przez łączną liczbę osób badanych we wszystkich studiach. Alternatywą – rekomendowaną przez Larry'ego Hedgesa i jego współpracowników (Hedges i Olkin, 1985; Hedges i Vevea, 1998) – jest ważenie efektów przez odwrotność wariancji poszczególnych badań (bądź odwrotność ich błędu standardowego). Ponieważ błąd standardowy jest negatywnie powiązany z wielkością próby, oba te rodzaje wag skutkują podobnymi oszacowaniami.

Kluczową decyzją podczas realizacji każdej metaanalizy jest wybór modelu statystycznego zastosowanego następnie do analiz. Zdecydowana większość podręczników metaanalizy (zob. np.: Hunter i Schmidt, 2004) wspomina przy tej okazji o dwóch podstawowych modelach: modelu efektów stałych (*fixed effects*) i modelu efektów losowych (*random effects*). Przegląd metaanaliz (zob. Ioannidis i Trikalinos, 2007), wskazuje na wciąż większą popularność modelu efektów stałych w realizowanych metaanalizach. Jednak co najmniej z trzech powodów sytuacja ulega zmianie i publikowane w ostatnich latach metaanalizy sięgają raczej po modele efektów losowych.

Po pierwsze, model efektów stałych mało realnie zakłada, że każdy uwzględniony w metaanalizie efekt jest odwzorowaniem efektu populacyjnego, a różnice

[2] W dalszej części artykułu użycie współczynnika *d* oznacza *d* Cohena, podobnie jak samo *g* odnosi się do *g* Hedgesa, Δ do Δ Glassa, *r* do *r* Pearsona, zaś *z* do *z* Fischera.

między efektami są wyłącznie skutkiem błędu próbkowania. W sytuacji, gdy bardziej uzasadnione jest przekonanie o istnieniu różnych „populacji efektów", na przykład uzyskiwanych w różnych podgrupach, zastosowanie mieć powinien model efektów losowych. Po drugie, model efektów stałych bywa adekwatny wówczas, kiedy badacz nie ma ambicji generalizowania wyników uzyskanych w metaanalizie poza badania faktycznie w niej uwzględnione. Jeśli więc metaanalityk ma pewność, że uwzględnił wszystkie badania na jakiś temat bądź syntetyzuje jedynie wyniki własnych eksperymentów, model efektów stałych może mieć zastosowanie. Jeśli jednak ambicją jest generalizowanie rezultatów na „superpopulację" efektów – także pochodzących z badań, które jeszcze nie zostały przeprowadzone – wyborem powinien być model efektów losowych. Wreszcie po trzecie, pragmatyczno-statystycznym kryterium decyzji, jaki model zastosować, są parametry rozproszenia uzyskiwanych efektów, w postaci miar heterogeniczności (zazwyczaj współczynnik *Q* Cochrana; Patil, 1975). Jeśli heterogeniczność efektów jest niewielka i statystycznie nieistotna, wybór model efektów stałych daje się uzasadnić. W praktyce jednak parametry heterogeniczności są zwykle statystycznie istotne (w dużej mierze wynika to z niskiej odporności testów heterogeniczności na wielkość próby), to zaś prowadzi badaczy do sięgnięcia po model efektów losowych.

Model efektów losowych opiera się na założeniu, że różne efekty syntetyzowane w metaanalizie pochodzą z różnych populacji, w rezultacie należy więc kontrolować dwa źródła wariancji; wariancję płynącą z samych badań, a będącą pochodną błędów oszacowań pojedynczych studiów oraz wariancję między różnymi efektami (badaniami)[3]. W konsekwencji, nawet jeśli wynik oszacowany metodą efektów stałych i losowych bywa zbliżony, metaanaliza realizowana modelem efektów losowych daje zazwyczaj szersze przedziały ufności wokół punktowych oszacowań.

Zarówno model efektów stałych, jak i model efektów losowych zakłada niezależność efektów. W praktyce, każdy efekt włączony do analizy powinien więc pochodzić z niezależnego badania. Co jednak, gdy jakieś studium przynosi kilka efektów, a wszystkie są potencjalnie interesujące? – Na przykład we włączanym do metaanalizy badaniu na temat relacji pomiędzy wynikami w nauce a osobowością, odnajdujemy osobne korelacje pomiędzy otwartością na doświadczenie a rezultatami w testach osiągnięć z języka polskiego, angielskiego i matematyki? Aby skorzystać z modelu efektów stałych lub losowych należy albo uśrednić poszczególne efekty, albo wybrać jeden z nich (zob. dyskusja w Cheung, 2014). Taka procedura niesie jednak kłopotliwe konsekwencje. Po pierwsze, nie jesteśmy w stanie spraw-

[3] Jak słusznie zauważył anonimowy recenzent w opinii na temat tego tekstu: „[metoda efektów losowych] zakłada przede wszystkim losowość doboru [efektów] do metaanalizy z superpopulacji efektów, co jest praktycznie niemożliwe do spełnienia". Ta uwaga pokazuje, że także założenia modelu efektów losowych mogą być mało realne, a decyzja o wyborze modelu analizy nie powinna być automatyczna. Choć bowiem model efektów losowych jest modelem bardziej konserwatywnym, co bywa argumentem na rzecz jego użycia, jego wybór nie musi być wcale oczywisty.

dzić czy relacje między interesującymi nas charakterystykami (tu: osobowością) a efektami (osiągnięciami szkolnymi) nie są moderowane przez przedmiot kształcenia – jeśli uśrednimy efekty różnych przedmiotów, nie będziemy w stanie zestawić ich ze sobą. Jest też problematyczna konsekwencja statystyczna – redukcja liczby efektów z kilku bądź kilkunastu do jednego, osłabia moc analizy, a samo uśrednianie, zwłaszcza w przypadku dużej wariancji między efektami, może zaciemnić faktyczny obraz zamiast czynić go bardziej klarownym. Rozwiązaniem jest metaanaliza wielopoziomowa, która operuje na poziomie współzależnych (tj. poklastrowanych w obrębie badań) efektów, szacując zarówno wariancję między badaniami (poziom 3), jak też wewnątrz badań, tj. między efektami (poziom 2) oraz wewnątrz poszczególnych efektów (na podstawie wielkości próby). Metaanaliza wielopoziomowa jest uogólnionym trzypoziomowym modelem regresyjnym, gdzie ogólny efekt estymuje się jako stałą, zaś rolę ewentualnych czynników modyfikujących (moderatorów), testuje się wprowadzając je jako predyktory do modelu.

Wreszcie istotnym, a wręcz koniecznym elementem współcześnie realizowanych metaanaliz jest szacowanie wiarygodności uzyskiwanych wyników i ich odporności na zjawisko selektywnego publikowania. Dyskusja na temat powodów i szerszej charakterystyki selektywnego publikowania i zjawisk pokrewnych wykracza poza ramy tego artykułu. Wiemy jednak, że niemal cała literatura empiryczna w naukach społecznych pokazuje zawyżone efekty (Ioannidis, 2005, 2008), a zarówno proces recenzyjny (np.: uznawanie przez recenzentów wyników nieistotnych statystycznie za niekonkluzywne), jak też rozmaite decyzje badaczy (np.: niechęć do wysyłania nieistotnych rezultatów do druku) zwiększają ryzyko błędu I rodzaju. Konieczne staje się więc szacowanie odporności uzyskanych efektów na ryzyko skrzywienia publikacyjnego.

Etapy metaanalizy

Proces realizacji metaanalizy sprowadzić można do kilkunastu następujących po sobie kroków. Dotyczą one zarówno kwestii ogólnych i podstawowych, tj. określenia pytań badawczych oraz zakresu uwzględnionej literatury, jak też decyzji bardziej szczegółowych – kryteriów włączania i wyłączania badań, sposobu kodowania podstawowych moderatorów, decyzji co do zastosowanej miary efektu i stosowania (bądź nie) rozmaitych poprawek, wreszcie metod szacowania odporności uzyskanych efektów na problem selektywnego publikowania.

Pierwszym i kluczowym krokiem każdego, nie tylko metaanalitycznego, procesu badawczego jest precyzyjne określenie pytań badawczych. Od ich charakteru zależeć będzie bowiem nie tylko ostateczna decyzja, co do wyboru określonego typu wielkości efektu, ale także bardziej szczegółowe kryteria włączania i wyłączania badań. W niektórych przypadkach charakter pytań badawczych w sposób naturalny wyznacza charakter włączanych badań – na przykład, gdy badacza interesują relacje pomiędzy różnymi charakterystykami (np.: związek między inteligencją a pozycją

społeczną), różnice między grupami (np.: zdolności przestrzenne mężczyzn i kobiet) lub efektywność różnych oddziaływań (np.: czy treningi kompetencji społecznych są skuteczne?). Czasami jednak pytania badawcze nie mają bezpośredniego przełożenia na charakter efektu – na przykład w jednej z metaanaliz (Szumski, Smogorzewska i Karwowski, 2017) pytano o efekty obecności uczniów z niepełnosprawnościami dla wyników w nauce ich sprawnych rówieśników. Kwestię tę można analizować zarówno sięgając po badania porównawcze (zestawienie wyników uzyskiwanych przez uczniów sprawnych w klasach bez uczniów z niepełnosprawnościami i w klasach do których uczęszczają uczniowie z niepełnosprawnościami), ale i korelacyjne – gdy analizuje się związek pomiędzy średnimi rezultatami w nauce uczniów sprawnych w zależności od liczby ich niepełnosprawnych kolegów.

Etap drugi, definiuje zakres danych niezbędnych do odpowiedzi na pytanie badawcze. Czy analiza obejmuje określony horyzont czasowy – na przykład od ostatniej opublikowanej metaanalizy bądź też z ostatnich 20 lat, czy też ambicją badacza jest prześledzenie całej literatury na dany temat? Czy w pierwszym etapie przesiewane są faktycznie wszystkie dostępne badania czy też może już w tym momencie bardziej uzasadnione byłoby wylosowanie próby badań – w niektórych przypadkach liczba dostępnych studiów może bowiem sięgać tysięcy. Jakie są kryteria lokalizowania i włączania oraz wyłączania badań? Standardem jest korzystanie z rozmaitych baz danych zbierających zasoby literatury, ale poważną decyzją jest włączenie albo wykluczenie niepublikowanych dysertacji lub raportów oraz doniesień konferencyjnych. Podobnie żywo dyskutowaną kwestią jest „problem Wieży Babel" (Gregoire, Derderian i LeLorier, 1995), a więc pytanie, czy i w jakiej mierze w metaanalizie powinno się polegać wyłącznie na – dominującej dziś – literaturze anglojęzycznej, w jakiej zaś włączane powinny być zidentyfikowane badania opublikowane w innych językach – często ważnych dla określonego problemu badawczego. Te kwestie urastają do szczególnej rangi, jeśli wziąć pod uwagę, że transparentność procesu doboru jest warunkiem replikowalności metaanaliz. Trudno tu też o stanowcze rekomendacje co do szczegółowych rozstrzygnięć, różne bywają bowiem praktyki w poszczególnych dyscyplinach i subdyscyplinach. Jedni badacze – idąc za głośną krytyką Eysencka (1978) – postulują włączanie do metaanaliz nie tylko wyłącznie prac opublikowanych, ale dodatkowo też takich, które spełniają zdefiniowane *a priori*, wyśrubowane kryteria jakości – na przykład odpowiednią rzetelność narzędzi czy randomizację grup. W medycznych metaanalizach na temat efektywności interwencji uwzględnia się niemal wyłącznie badania zrealizowane w schemacie RCT (*randomized controlled trial*), wykluczając studia poprzeczne i korelacyjne. Inni badacze, postulują włączanie danych niepublikowanych, jako pozwalających na określenie ryzyka *publication bias*, postulując dodatkowo kodowanie jakości badań i włączanie jej jako potencjalnego moderatora wyjaśniającego zróżnicowanie efektów lub też elementu wagi, która „gorszym" badaniom przypisuje mniejsze znaczenie (Greenland i O'Rourke, 2001).

Zmienia się również proces dostępu do publikacji. O ile dobór badań do wczesnych metaanaliz polegał na śledzeniu artykułów naukowych publikowanych w klu-

czowych periodykach, a następnie analizie pozycji zawartych w ich bibliografiach oraz w bibliografiach ich bibliografii, dziś przeszukiwanie jest znacznie bardziej zautomatyzowane. Dostęp do kluczowych baz danych: Scopus, Web of Science, PsychInfo, Academic Search Complete, etc., jest więc jedynie punktem wyjścia, który uzupełniany bywa pytaniami wysyłanymi do aktywnych badaczy w danej dziedzinie i umieszczaniem informacji o poszukiwanych, zwłaszcza niepublikowanych, wynikach na listach dyskusyjnych towarzystw naukowych (rozmaite: *listserv*). Dla replikowalności kluczowe jest zarówno podanie informacji o wykorzystanych źródłach, jak i szczegółowym procesie selekcji badań.

Zwykle wstępne przeszukiwanie na podstawie kilku najbardziej charakterystycznych haseł w tytułach, abstraktach i zestawieniach podawanych przez autorów słów kluczowych skutkuje identyfikacją bardzo wielu, czasem tysięcy, potencjalnie użytecznych tekstów. Ich szybki przegląd na podstawie treści abstraktów, pozwala na pierwszą selekcję: eliminowane są badania jakościowe, badania, w których próżno szukać koniecznych statystyk dla obliczenia efektów, badania, których autorzy sięgają po problematyczny (np.: mało trafny i rzetelny) pomiar. Istotne jest nie tylko zdanie sprawy z kryteriów eliminacji i włączania badań, ale i zdefiniowanie ich *a priori*, tak aby zmniejszyć ryzyko *selection bias*. Dobór badań do metaanalizy powinien mieć charakter przeglądu systematycznego, wraz ze wszystkimi konsekwencjami i wymogami obowiązującymi takie przeglądy (zob. np.: Boland, Cherry i Dickson, 2014; Matera i Czapska, 2014).

Badania zakwalifikowane do finalnej metaanalizy powinny zostać zakodowane – chodzi tu zarówno o wielkość efektu uzyskaną w każdym z nich, ale również o wiele moderatorów, tj. charakterystyk samych badań oraz wykorzystanych w nich miar. Decyzja co do liczby i charakteru moderatorów bywa trudna, zarówno ze względu na pracochłonność samego procesu kodowania, jak i problemy z przewidzeniem zawczasu jakie wymiary zasługują na uwzględnienie. Oczywistym jest kodowanie elementarnych danych o każdym z badań: roku realizacji, kraju, wykorzystanych narzędzi, odsetka kobiet (mężczyzn), średniego wieku uczestników czy rzetelności narzędzi. Faktycznie jednak liczba możliwych, a zarazem merytorycznie relewantnych, moderatorów może być znacznie większa. Sam proces kodowania wymaga sięgnięcia po co najmniej dwóch koderów, którzy niezależnie kodują moderatory we wszystkich badaniach, następnie zaś szacowana i raportowana jest ich spójność (np.: kappa: Cohen, 1968). W przypadku masywnych metaanaliz, zdarza się, że kodowana niezależnie jest jedynie część wszystkich badań (20%-30%), następnie zaś po upewnieniu się, że spójność jest wysoka, każdy z koderów samodzielnie koduje swoją część wszystkich badań.

Kluczowym krokiem jest obliczenie dla każdego z badań wielkości efektu wraz z właściwą wagą – wielkością próby bądź odwrotnością wariancji. W niektórych przypadkach, na przykład metaanaliz posługujących się wyłącznie badaniami korelacyjnymi, sytuacja jest stosunkowo prosta, efekty te są bowiem zazwyczaj zawarte w tabelach ze statystykami opisowymi i interkorelacjami w oryginalnych tekstach.

W badaniach eksperymentalnych lub podłużnych bywa to bardziej kłopotliwe, bowiem niezależnie od wymogów raportowania wielkości efektów, wciąż zdarza się, że próżno ich szukać w pracy. Dostępne kalkulatory pozwalają na uzyskanie najpopularniejszych miar efektów: *r* i *d*, nie tylko ze statystyk opisowych (średnich i odchyleń standardowych), ale także z testów statystycznych: testu *t* dla prób niezależnych, zależnych i jednej próby, analizy wariancji i kowariancji, w tym również wartości η^2 czy też proporcji (zob. tabela 1). Paradoksem tego etapu jest fakt, że problemem dla metaanalityków bywa rozwój statystyki i sięganie po badaczy po coraz bardziej złożone metody w badaniach pierwotnych. I tak, choć naturalną miarą wielkości efektu jest współczynnik korelacji, kłopotliwe jest uzyskanie go z doniesienia, gdzie badacze nie pokazują prostych korelacji parami, a posługują się modelem regresyjnym – zwłaszcza złożonym. Wprawdzie istnieją udane próby szacowania wartości *r* z raportowanych standaryzowanych współczynników regresji (Peterson i Brown, 2005), ale bywa to kłopotliwe, zwłaszcza, gdy modele regresji cechują się znaczną kompleksowością, a więc różnica między β a *r* może być znaczna z powodu kontroli kowariancji między predyktorami w modelu regresji. Podobnie ma się sytuacja w przypadku modeli równań strukturalnych czy regresji wielopoziomowych. Zwykle najefektywniejszym rozwiązaniem jest wówczas kontakt z autorami oryginalnej pracy i prośba o bardziej elementarne statystyki.

Tabela 1. Przykładowe statystyki i sposoby ich przeliczania na *r* Pearsona oraz *d* Cohena

Statystyka	*r* Pearsona	*d* Cohena
Test *t* dla prób niezależnych	$r = \sqrt{\frac{t^2}{t^2 + df}}$	$d = \frac{2t}{\sqrt{df}}$
Statystyka *F*	$r = \sqrt{\frac{F}{F + df(\text{error})}}$	$d = \sqrt{F(\frac{n1 + n2}{n1\ x\ n2}) * (\frac{n1 + n2}{n1 + n2 - 2})}$
d Cohena / *r* Pearsona	$r = \frac{d}{\sqrt{d^2 + 4}}$	$d = \frac{2r}{\sqrt{1 - r^2}}$

Na etapie szacowania efektów konieczne jest ujednolicenie różnych miar pomiędzy badaniami – jest bowiem wielce prawdopodobne, że niektóre będą raportowały współczynniki korelacji, inne różnice w średnich czy miary regresji. Warto też zadbać o precyzję w przypadku zapisu wielkości efektu dla standaryzowanej różnicy w średnich. Najczęściej w metaanalizach spotyka się efekty wyrażone w postaci *d* Cohena, jednak miara ta bywa też traktowana jako generyczna nazwa każdego efektu opartego na standaryzowanej różnicy między średnimi. Różnice

między *d*, *g* a *Δ* sprowadzają się do nieco innej postaci mianownika – a więc obecności bądź braku rozmaitych poprawek na wielkość próby – w przypadku *g*, jak ilustrują to wzory (1.1-1.2) mamy do czynienia z nieco bardziej konserwatywnym charakterem – ponieważ w mianowniku pojawia się dodatkowa formuła -2, sprawia to, że wartości *g* będą zawsze nieznacznie niższe niż *d*, a różnica będzie większa w przypadku badań realizowanych na mniejszych próbach.

(1.1) $g = \frac{x1 - x2}{s*}$, gdzie s* = (1.2) $s^* = \sqrt{\frac{(n1-1)s1^2 + (n2-1)s2^2}{n1+n2-2}}$, gdzie: *g* = współczynnik *g* Hedgesa, *x1* = średni wynik w grupie pierwszej, *x2* = średni wynik w grupie drugiej, *s** = łączne odchylenie standardowe, *n1* = liczebność próby w grupie pierwszej, *n2* = liczebność próby w grupie 2, *s1* = odchylenie standardowe w grupie 1, *s2* = odchylenie standardowe w grupie 2.

Ważnym elementem przygotowywania pojedynczych efektów do metaanalizy, jeszcze przed przystąpieniem do faktycznego analizowania wyników, jest wprowadzenie poprawek, które rzutują na ostatecznie uzyskany rezultat. Procedurę rekomendowaną przez Hedgesa (np.: Hedges i Vevea, 1998), sprowadzić można do trzech możliwych korekt oryginalnych efektów tak, aby dawały one mniej obciążone rezultaty.

Po pierwsze, Hedges rekomenduje korektę efektów (standaryzowanej różnicy średnich – *g*) uzyskanych w małych badaniach ($n < 20$), poprzez zastosowanie formuły (2-3).

(2) $g^* = g \times (1 - \frac{3}{4N-9})$, a więc połączenie formuły 1.1 i 2, daje bardziej zgeneralizowany wzór na *g* (3) $g^* = \frac{x1 - x2}{s*} \times (1 - \frac{3}{4N-9})$, gdzie: *N* = łączna wielkość próby (pozostałe oznaczenia jak przy wzorze 1.1.)

Jak widać w tabeli 2, zastosowanie poprawki na wielkość próby powoduje bardzo nieznaczne zmiany: im mniejsze badania, tym wyraźniejsza korekta w dół. Ogółem jednak efekt pierwszej z rekomendowanej przez Hedgesa poprawki jest subtelny i jego stosowanie bądź nie, ma niewielki wpływ na rezultat metaanaliz.

Tabela 2. Przykład zastosowania poprawki Hedgesa dla trzech zasymulowanych wielkości próby: $n_1 = 10$, $n_2 = 15$ i $n_3 = 20$ i wielkości efektów wyrażonych standaryzowaną różnicą średnich (*g*), zawierających się w przedziale od $g = 0$ do $g = 2$

g uzyskane	g_1 ($n_1 = 10$)	g_2 ($n_2 = 15$)	g_3 ($n_3 = 20$)
0	0,00	0,00	0,00
0,1	0,09	0,09	0,10
0,2	0,18	0,19	0,19

cd. tabeli 2

0,3	0,27	0,28	0,29
0,4	0,36	0,38	0,38
0,5	0,45	0,47	0,48
0,6	0,54	0,56	0,57
0,7	0,63	0,66	0,67
0,8	0,72	0,75	0,77
0,9	0,81	0,85	0,86
1	0,90	0,94	0,96
1,1	0,99	1,04	1,05
1,2	1,08	1,13	1,15
1,3	1,17	1,22	1,25
1,4	1,26	1,32	1,34
1,5	1,35	1,41	1,44
1,6	1,45	1,51	1,53
1,7	1,54	1,60	1,63
1,8	1,63	1,69	1,72
1,9	1,72	1,79	1,82
2	1,81	1,88	1,92

Druga z rekomendowanych przez Hedgesa poprawek odnosi się do korekty uzyskiwanych efektów na ograniczoną rzetelność pomiaru. Stosowana jest w tym celu klasyczna formuła (4).

(4) $r' = \frac{r}{\sqrt{\alpha 1 \times \alpha 2}}$, gdzie: r' = poprawiony współczynnik korelacji, r = uzyskany współczynnik korelacji, $\alpha 1$ = rzetelność pierwszej zmiennej, $\alpha 2$ = rzetelność drugiej zmiennej.

Zgodnie z jedną z podstawowych zasad psychometrii, obserwowana korelacja pomiędzy dwiema zmiennymi nie może być wyższa niż rzetelność najmniej rzetelnej z nich – a więc w przypadku rzetelności przynajmniej jednej ze zmiennych na poziomie 0,6, uzyskana wielkość efektu (gdy mówimy o korelacji) nigdy nie będzie wyższa niż 0,6. Zastosowanie poprawki, zwłaszcza w odniesieniu do badań posługujących się mniej rzetelnymi miarami, może więc skutkować wyraźnym wzrostem efektów. Ilustrację symulacji takich zmian zawarto w tabeli 3.

Tabela 3. Zmiany wartości współczynnika korelacji *r* Pearsona w zależności od rzetelności pomiaru korelowanych zmiennych oraz relacja *r* Pearsona – *z* Fischera. Dla celów przykładu przyjęto, że rzetelność jednej zmiennej wynosi α = 0,75, drugiej α = 0,70. Z Fischera przeliczono z wartości surowych *r* Pearsona (bez poprawki na nierzetelność)

r Pearsona	*r'* (*r* Pearsona poprawione na nierzetelność)	*z* Fischera
0	0	0,00
0,1	0,14	0,10
0,2	0,28	0,20
0,3	0,41	0,31
0,4	0,55	0,42
0,5	0,69	0,55
0,6	0,83	0,69
0,7	0,97	0,87
0,8	*1,1*	1,10
0,9	*1,24*	1,47
0,99	*1,38*	2,65

Uwaga: kursywą podane wartości współczynników korelacji, które są matematycznie nieosiągalne, gdyż przekraczają 1.

Pokazane zmiany są już bardziej radykalne, co jednak wynika wyłącznie z przeciętnej rzetelności miar, którymi posłużono się w przykładzie. Przy rzetelności na poziomie 0,9, różnica między obserwowanym a poprawionym efektem będzie niewielka – na przykład obserwowana korelacja rzędu *r* = 0,3, po poprawieniu osiągnie wartość *r* = 0,33, a korelacja *r* = 0,5 – *r* = 0,56. Jako że badaczy interesują zwykle relacje między teoretycznymi wymiarami a nie wynikami uzyskanymi w testach czy kwestionariuszach, taka poprawka jest często uzasadniona. I choć w badaniach sprzed kilku dziesięcioleci niełatwo o odnalezienie wszystkich oszacowań rzetelności, w takich sytuacjach Hedges rekomenduje posłużenie się uśrednioną rzetelnością wyliczoną dla tych badań, gdzie wartości te są dostępne. Oczywiście tego typu poprawka może budzić wątpliwości jako przykład tzw. *p*-hackingu (zob. np.: Head i in., 2015) – bo wyższe punktowe oszacowanie wartości korelacji oznacza również wyższą dolną granicę jego 95% przedziału ufności, a więc większą szansę na wynik statystycznie istotny. Nie jest to jednak wielkim problemem z dwóch powodów. Po pierwsze, metaanaliza nie skupia się na istotności statystycznej – jej kluczowym wynikiem jest właśnie wielkość efektu i jej interpretacja w języku siły relacji pomiędzy zmiennymi. Po drugie, poprawka na nierzetelność sprawia, że wielkość efektu oszacowana korelacyjnie staje się niemal dokładnym ekwiwalentem relacji ścieżkowej w modelu strukturalnym[4]. Tak więc zastosowanie poprawki Hedgesa ma też walor praktyczny

– pozwala bowiem włączyć do metaanaliz wyniki studiów opartych na modelach strukturalnych.

Trzecia z poprawek proponowanych przez Hedgesa ma zastosowanie wyłącznie do wielkości efektu szacowanej jako współczynnik korelacji. Biorąc pod uwagę jego specyfikę, a więc niesymetryczne przedziały ufności wokół oszacowań (dolna wartość przedziału ufności nie może przekroczyć -1, a górna +1), dla większej stabilności estymacji Hedges zaleca przeliczanie uzyskanych wartości *r* na *z* Fischera, następnie procedurę metaanalizy na wartościach *z*, aby finalnie uzyskany efekt przeliczyć ponownie na *r*. Samo przeliczanie *r* na *z* odbywa się według wzoru (5). Różnice między *r* a *z* są widoczne jedynie w przypadku bardzo wysokich (lub bardzo niskich) wartości *r* (zob. tabela 3, ostatnia kolumna), stąd w zdecydowanej większości metaanaliz wspominana transformacja jest raczej opcjonalna.

(5) $z = 0{,}5 \ln \frac{1+r}{1-r}$, gdzie: *z* = wartość statystyki *z* Fischera, *ln* = logarytm naturalny, *r* = współczynnik korelacji *r* Pearsona.

Dysponując bazą danych z odpowiednio przygotowanymi efektami – poprawionymi bądź nie – oraz zakodowanymi właściwościami badań, mogącymi pełnić funkcję potencjalnych moderatorów, jesteśmy gotowi do właściwej metaanalizy. Jak wspomniano wyżej, efekty można ważyć zarówno wielkością próby, jak również odwrotnością wariancji. Tę ostatnią wyliczyć można korzystając ze wzoru (6) w przypadku dysponowania *d* lub *g* oraz wzoru (7) dla *r*.

(6) $var(d) = \frac{(n1+n2)}{(n1 \times n2)} + \frac{d^2}{2(n1+n2)}$

(7) $var(r) = \frac{(1-r^2)^2}{n-1}$, gdzie: *d* = *d* Cohena, pozostałe oznaczenia jak we wzorach powyżej.

Metaanaliza w praktyce – krótki przykład

Ta część stanowi rozbudowany przykład, pokazujący zastosowanie omawianych wcześniej zagadnień. Jego prześledzenie powinno pozwolić czytelnikom nie tylko na odtworzenie wszystkich opisanych operacji, ale również realizację własnych metaanaliz.

[4] Oczywiście z wyłączeniem jej regresyjnego charakteru, tj. omawiana poprawka owszem „znosi" nierzetelność, ale nie jest sobie w stanie poradzić z wcześniej wspominanym regresyjnym charakterem modelu strukturalnego, tj. obserwowaną niemal zawsze (poza rzadkimi zjawiskami supresji, Paulhus i in., 2004) inflacją wartości β w stosunku do *r*, gdy w modelu znajdują się predyktory o choćby niewielkiej kowariancji. Stąd też w sensie ścisłym przywołany przykład traktować należy jako odnoszący się do modelu strukturalnego z dwiema zmiennymi latentnymi – jednym predyktorem i jedną zmienną wyjaśnianą.

Prosta metaanalityczna baza[5] została przedstawiona w tabeli 4. Dla celów ilustracji przyjmijmy, że nasza metaanaliza obejmuje 10 badań[6]. Dla każdego z nich dysponujemy wartością współczynnika korelacji (kolumna [2]), którego w tym przypadku nie przeliczamy na wartość *z*, ani nie poprawiamy na nierzetelność oraz wielkością próby (kolumna [3]). Stosując więc wzór (7) dla każdego z badań jesteśmy w stanie obliczyć jego wariancję (kolumna [4]). Naszą wagą (W) jest odwrotność wariancji, która znajduje się w kolumnie [5]. Mając tak przygotowaną bazę, jesteśmy w stanie przystąpić do właściwych analiz.

Tabela 4. Przykładowe dane metaanalityczne dla 10 badań (kolumny [2-3] oraz [10-11] oraz kolejne przekształcenia omówione w tekście (kolumny [4-9], pozwalające na realizację metaanalizy krok po kroku)

Badanie	*r*	*N*	*var*	*W* (1/*var*)	*r*W*	r^2*W	W^2	*W*2	Moderator 1	Moderator 2
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
1	0,32	161	0,0056	198,59	63,55	20,34	39437,43	33,64	1	0
2	-0,03	21	0,04991	20,04	-0,60	0,02	401,44	13,41	1	1
3	0,3	119	0,007018	142,50	42,75	12,83	20304,79	31,54	1	0
4	0,51	277	0,001984	504,15	257,12	131,13	254171	37,49	1	1
5	0,32	203	0,003989	250,72	80,23	25,67	62859,56	34,87	1	0
6	0,56	204	0,002321	430,87	241,29	135,12	185644,9	37,02	2	1
7	0,22	790	0,001148	871,31	191,69	42,17	759165,2	38,70	2	0
8	0,34	75	0,01057	94,61	32,17	10,94	8950,94	28,36	2	1
9	0,02	92	0,01098	91,07	1,82	0,04	8294,26	28,03	2	0
10	0,16	417	0,002282	438,15	70,10	11,22	191971,9	37,07	2	1

Jaka zatem będzie metaanalitycznie oszacowana wielkość naszego efektu? Gdybyśmy obliczyli ją jako prostą średnią arytmetyczną, uzyskalibyśmy wartość $r = 0,27$.

[5] Należy zauważyć, że prezentowana baza ma już postać przetworzoną – prezentuję w niej już gotowe wielkości efektu: w tym przypadku współczynniki korelacji *r* Pearsona wraz z koniecznymi przeliczeniami, o których mowa w tekście. Faktycznie w pracy nad własną metaanalizą, zwłaszcza gdy uwzględnia ona badania eksperymentalne, porównawcze bądź podłużne, przeliczenie różnie raportowanych danych do postaci wielkości efektu: *d* Cohena, *g* Hedgesa czy *r* Pearsona, wymaga osobnej i uważnej pracy.

[6] Nie jest to przykład całkowicie fikcyjny – dla celów ilustracji wartości współczynników korelacji oraz wielkości prób zostały zapożyczone z metaanalizy Gajdy i współpracowników (2017, tabela 1). Przy ewentualnym bezpośrednim porównaniu trzeba jednak pamiętać, że w tej metaanalizie uwzględniono łącznie dane ze 120 badań, analizowane wielopoziomowo. Przykłady moderatorów są fikcyjne i dodane wyłącznie dla celów propedeutycznych.

Wiemy jednak, że byłoby to oszacowanie niepoprawne, bo ignorujące różną wielkość próby w uwzględnionych badaniach. Jak zatem uzyskać efekt zważony odwrotnością wariancji? Jeśli każdą wartość *r* przemnożymy przez *W* (odwrotność wariancji) – tak jak pokazano to w kolumnie [6] – a następnie sumę wszystkich wartości *r*W* podzielimy przez sumę wartości *W* (a więc w naszym przypadku sumę wartości zawartych w kolumnie [5]) uzyskamy zważoną wartość efektu. Będzie to więc $\frac{980{,}107\ (suma\ wartości\ kolumny\ [6])}{3041{,}986\ (suma\ wartości\ kolumny\ [5])} = 0{,}32$. Poszukiwany efekt – w naszym przykładzie metaanalitycznie oszacowana korelacja pomiędzy zdolnościami twórczymi a osiągnięciami szkolnymi, to *r* = 0,32, wartość, którą w świetle zwyczajowych standardów (np.: Cohen, 1988; Ellis, 2010) można określić jako umiarkowaną. Czy jednak jest ona statystycznie istotna? Jakie są jej przedziały ufności? Czy jest homogeniczna?

Aby wyznaczyć przedziały ufności dla *r*, potrzebujemy jego błędu standardowego. Błąd standardowy jest pierwiastkiem kwadratowym z wariancji, a suma kolumny [5] to suma odwrotności wariancji. Zatem pierwiastek kwadratowy z tej wartości będzie wielkością błędu standardowego oszacowania. Tak więc $SE(r) = \sqrt{\frac{1}{3041{,}986}} = 0{,}018$. Czy zatem *r* = 0,32 jest statystycznie istotne? Owszem – wiedząc, że górny i dolny 95% przedział ufności wyznaczają wartości $r \pm 1{,}96 \times SE$, łatwo obliczyć, że 95% procentowe przedziały ufności dla uzyskanego oszacowania zawierają się w przedziale od 0,29 do 0,36 – ponieważ ten przedział ufności nie przecina zera, możemy być pewni, że jest od niego istotnie różny, a więc, że jest to wartość istotna statystycznie na poziomie przynajmniej $p < 0{,}05$. Alternatywnym sposobem upewnienia się, że jest tak w istocie, jest obliczenie wartości statystyki *z*. Wiedząc, że $z = \frac{r}{SE(r)}$, a więc w naszym przypadku $z = \frac{0{,}32}{0{,}018} = 17{,}77$ oraz że krytyczne wartości *z* dla $p = 0{,}05$ to $z = 1{,}96$ a dla $p = 0{,}01$ $z = 2{,}58$, mamy pewność, że oszacowana wartość jest różna od 0.

Widzimy więc, że w kilku prostych krokach, w niewielkiej metaanalizie o $k = 10$ (liczba badań), na łącznej próbie $N = 2359$, udało się nam wykazać istnienie pozytywnej relacji: $r = 0{,}32$ 95% PU: (0,29-0,36), pomiędzy zdolnościami twórczymi a osiągnięciami szkolnymi. Trzeba jednak wspomnieć, że cała ta procedura to w istocie metaanaliza metodą efektów stałych. Czy zaś zastosowanie tego modelu było uzasadnione? Aby się o tym przekonać, należy oszacować heterogeniczność uzyskanego efektu. Jej wartość obliczana jest według wzoru (8)

(8) $Q = \sum(W \times r^2) - \frac{[\sum(W \times r]^2}{\sum W}$, gdzie: *Q* = współczynnik *Q* Cochrana, *W* = współczynnik będący iloczynem wagi i wielkości współczynnika korelacji *r* Pearsona. Pozostałe oznaczenia jak we wzorach wyżej.

Potrzebujemy więc wartości iloczynu wagi i kwadratu *r* – wyliczamy go w kolumnie [7]. Gdy zsumujemy cząstkowe wartości w ramach tej kolumny, uzyskamy wartość 389,46. Zatem nasze $Q = 389{,}46 - \frac{980{,}107^2}{3041{,}986} = 73{,}68$. Statystyka *Q* ma rozkład testu χ^2 o liczbie stopni swobody (*df*) o 1 mniejszej niż liczba badań, a więc w naszym

przypadku *df* = 10-1 = 9. Przy *Q* = 73,68 i *df* = 9, *p* < 0,001, zatem mamy podstawy do odrzucenia hipotezy zerowej, zakładającej, że nasze oszacowanie jest homogeniczne. Pozostaje więc uznanie, że uzyskany efekt jest heterogeniczny, a w związku z tym zastosowany model efektów stałych może dawać nieprawidłowe oszacowania.

Jaka zatem jest wielkość efektu obliczona w modelu efektów losowych? Przypomnijmy, że model ten zakłada istnienie dwóch źródeł błędu – płynącego z wewnątrz badań, a w naszym przypadku kontrolowanych przez ważenie przez odwrotność wariancji, ale również wariancji między badaniami. Aby oszacować interesującą nas relację modelem efektów losowych, musimy zmodyfikować wagę. Nie będzie nią odwrotność wariancji, ale suma odwrotności wariancji oraz jej efektu losowego. Tak więc nowa waga będzie efektem zastosowania formuły W2 = $\frac{1}{var + v_0}$, gdzie *W2* = nowa waga, *var* = wariancja, i v_0 = współczynnik losowy. Wartość v_0 można uzyskać korzystając z formuły (9)

(9) $v_0 = \frac{Q - k - 1}{\sum W - (\frac{\sum W^2}{\sum W})}$, gdzie: v_0 = współczynnik losowy, *Q* = *Q* Cochrana, *k* = liczba badań, *W* = współczynnik będący iloczynem wagi i *r* Pearsona.

W kolumnie [8] znalazły się wartości W^2 dla każdego z badań. Ich suma to 1 531 201,358. Zatem $v_0 = \frac{73,68 - 10 - 1}{3041,986 - \frac{1531201,358}{3041,986}} = \frac{62,68}{2538,63} = 0,02469$. Zatem nowa waga (kolumna [9]): W2 = $\frac{1}{W + 0,02469}$.

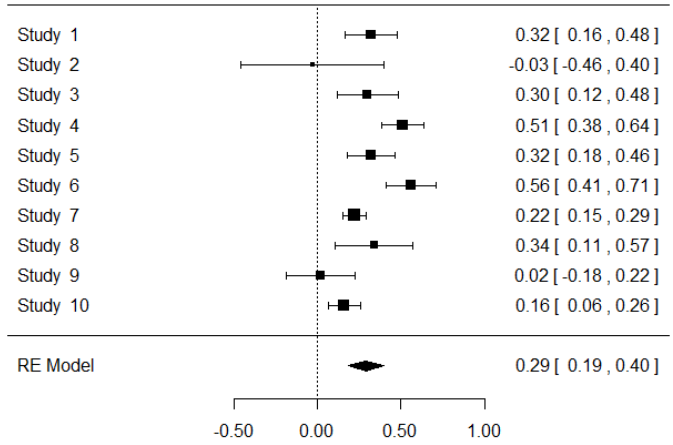
Gdybyśmy teraz powtórzyli całą opisaną wcześniej procedurę – szczegóły pomijam z braku miejsca – okazałoby się, że uzyskujemy *r* = 0,29, *SE* = 0,05, a więc 95% PU zawierałby się w przedziale od 0,19 do 0,40. Zgodnie z przewidywaniami uzyskaliśmy więc zbliżoną wartość *r*, jednak większy błąd standardowy, a więc i szersze przedziały ufności wokół punktowego oszacowania. Zarówno stosunkowo prostą, jak i bardziej złożoną metaanalizę według kroków opisanych wyżej można zrealizować korzystając z kalkulatora i kartki papieru bądź arkusza kalkulacyjnego.

Przykładowym podsumowaniem dyskutowanych wyników może być tzw. wykres drzewa (*forest plot*), tu stworzony dla naszego przykładu w środowisku R.

Bywa, że badaczom wystarcza oszacowanie ogólnego efektu i informacja o jego sile lub istotności. Zwykle jednak naturalnym kolejnym pytaniem jest sprawdzenie czy i jakie czynniki mogą stać za zróżnicowaniem uzyskiwanych efektów – czy wielkość efektu zmienia się w czasie, a więc jest różna w badaniach nowszych i starszych? Czy inne oszacowania przynoszą badania korelacyjne i eksperymentalne? Czy pomiar określoną skalą daje inne wyniki niż zastosowanie innych narzędzi?

Procedura analizy moderatorów najczęściej przybiera jedną z dwóch postaci. Najpopularniejszą jest porównanie wielkości efektów dla różnych poziomów moderatorów nominalnych bądź porządkowych – takich jak kraj badania czy rodzaj pomiaru. W praktyce realizuje się wówczas odpowiednik analizy wariancji dla metaanalizy. W naszym przykładzie łatwo wyobrazić sobie na przykład, że pierwszych 5 efektów zostało uzyskanych w USA (zakodowane 1), zaś kolejne 5 w innych kra-

jach (zakodowane 2). Gdybyśmy więc powtórzyli procedurę osobno dla pierwszych 5 efektów, następnie zaś dla kolejnych 5, dodatkowo zaś policzylibyśmy wartość *Q* w każdej z grup, wówczas suma wartości *Q* w obu grupach dałaby nam oszacowanie tzw. *Q* wewnątrzgrupowego. Chcąc sprawdzić, czy nasz moderator istotnie różnicuje wielkość uzyskiwanych efektów potrzebujemy *Q* międzygrupowego – jest to różnica pomiędzy oszacowanym wcześniej *Q* całkowitym, a *Q* wewnątrzgrupowym. Wiedząc, że liczba stopni swobody *Q* międzygrupowego to *j*-1, gdzie *j* = liczba grup, a więc w naszym przypadku *df* = 1 oraz pamiętając o wspominanym już fakcie, że *Q* ma rozkład χ^2, jesteśmy w stanie sprawdzić czy uzyskana wartość jest statystycznie istotna na poziomie $p < 0{,}05$. Jeśli tak, mamy powody, aby odrzucić hipotezę zerową zakładającą, że różne wartości moderatora dają te same wielkości efektu. Innymi słowy, możemy uznać, że dany wymiar istotnie moderuje wielkość efektu uzyskiwanego w metaanalizie. Oczywiście w praktyce nie ma konieczności realizowania takich porównań krok po kroku – dostępne programy dedykowane metaanalizie (np.: Comprehensive meta-analysis, zob. Biostat, 2008) lub zestawy poleceń dla najpopularniejszych pakietów statystycznych (np.: SPSS, zob. np.: zestaw makr Wilsona – http://mason.gmu.edu/~dwilsonb/ma.html), pozwalają na automatyzację tego procesu.



Rysunek 1. Przykład prezentacji wyników w metaanalizy w postaci tzw. wykresu drzewa (*forest plot*)

W przypadku moderatorów o charakterze ciągłym (np.: rok badania, odsetek mężczyzn/kobiet) lub zmiennych binarnych, bardziej właściwe będzie sięgnięcie po metaregresję. Zmienną zależną jest w niej wielkość efektu, zaś poszczególne moderatory, to predyktory. Należy pamiętać, że interpretacja ich wkładu zależy od skali na jakiej są mierzone i posługiwać się niestandaryzowanymi współczynnikami regresji oraz wartością stałej. Tu również należy korzystać z dedykowanych rozwiązań

– prosta analiza regresji w dostępnych pakietach – na przykład SPSS, nawet z włączoną wagą – nie da właściwych oszacowań.

Szacowanie *publication bias*

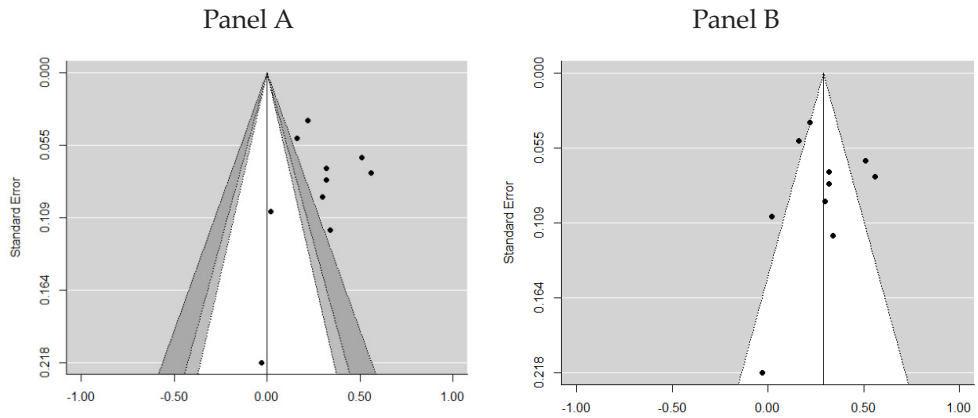
Właściwe wyestymowanie zagregowanej wielkości efektu oraz przetestowanie i wykazanie ewentualnej różnicującej roli moderatorów, to kluczowe etapy metaanalizy. Jednak dla stwierdzenia czy uzyskane podsumowanie jest wiarygodne, niezbędne jest również wykazanie, że analizowane dane nie są obarczone ryzykiem rozmaitych zaburzeń, które mogą tworzyć zafałszowany obraz relacji. Kluczowym jest wspominana już kilkukrotnie tendencja do publikowania wyników pokazujących wyniki statystycznie istotne. Wiele dobrze zrealizowanych studiów, w których nie udaje się potwierdzić zakładanych hipotez ląduje w szufladach autorów bądź jest odrzucanych w fazie recenzji jako niekonkluzywne. I choć sytuacja zmienia się za sprawą rozwoju rejestrowanych raportów (Nosek i Lakens, 2014) oraz rosnącej popularności statystyk Bayesowskich (Lee, 2012) pozwalających na skwantyfikowanie prawdopodobieństwa prawdziwości hipotezy zerowej (a nie jedynie stwierdzenia o braku podstaw do jej odrzucenia), problem skrzywienia publikacyjnego wciąż zaburza obraz literatury przedmiotu. Jest to szczególnie problematyczne w przypadku publikowanych badań, które są wyraźnie „niedomocowane", a więc ich statystyczna moc, wynikająca z małych prób, pozwala na wykrycie jedynie bardzo silnych efektów. Fakt, że często takie właśnie efekty w tych badaniach się uzyskuje, a dla ich siły nie ma przekonującego uzasadnienia teoretycznego, prowadzi do podejrzeń, że albo uzyskany wynik jest efektem błędu I rodzaju, albo wynikiem selektywnego publikowania – a więc autor zapewne ma wiele „zerowych" rezultatów w swojej szufladzie, bądź też – co z tej listy najgorsze – że wynik jest efektem problematycznych praktyk badawczych. Jak zatem z problemem selektywnego publikowania i skrzywienia publikacyjnego radzą sobie metaanalitycy? Omówmy pokrótce kilka metod, zarówno klasycznych, jak również rozwijanych współcześnie.

Pierwszą, najbardziej naturalną metodą kontroli *publication bias* jest włączanie do metaanaliz „szarej literatury", a więc prac magisterskich i doktorskich, referatów na konferencjach czy niepublikowanych raportów. Dysponując w bazie danych zmienną opisującą poszczególne badania jako niepublikowane (kodowane np. jako 0) lub publikowane (kodowane jako 1), stosując ANOVA, można łatwo sprawdzić, czy wielkości efektów niepublikowanych i publikowanych różnią się od siebie w sposób statystycznie istotny. Jeżeli efekt uzyskany w badaniach niepublikowanych jest istotnie słabszy niż ten z badań publikowanych, moglibyśmy uznać to za dowód na istnienie skrzywienia publikacyjnego. Faktycznie jednak jest to jedynie sugestia. Niewykluczone bowiem, że badania niepublikowane nie znalazły swojego miejsca na łamach profesjonalnych periodyków nie dlatego, że uzyskiwano w nich mniej spektakularne rezultaty, ale zostały gorzej zrealizowane, albo kluczowe zmienne zostały zmierzone w sposób mniej rzetelny. Jako że nierzetelność obniża siłę efektów, niższe

korelacje czy *d*/*g* w tych badaniach nie muszą być więc wcale świadectwem *publication bias.* Stąd też pierwsza metoda, choć użyteczna, bywa zawodna.

Drugą, klasyczną metodą szacowania odporności oszacowań metaanalitycznych, dziś rzadko już rekomendowaną, jest opracowana przez Rosenthala (1979) technika *fail-safe N:* koniecznej liczby badań o wynikach zerowych, których dodanie do metaanalizy spowodowałoby zredukowanie jej efektu do wartości nieistotnej statystycznie. Im więcej takich badań należałoby dodać, tym odporniejsza pierwotna analiza. Na przykład oszacowana wartość *fail-safe N* dla naszego przykładu z 10 badaniami to wartość ponad 500: z jednej strony wskazująca na odporność zaprezentowanej tu mini-metaanalizy, z drugiej jednak mało wiarygodna, szczególnie, że nic nie wiadomo o wielkości próby w każdym z badań. Podobne wątpliwości pojawiały się pod adresem metody Rosenthala już wcześniej (Becker, 2005) i *fail-safe N* ma dziś historyczny charakter – na przykład autorzy *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins i Green, 2011) odradzają korzystanie z tego rozwiązania jako mało czułego i niezbyt konkluzywnego.

Trzecim sposobem szacowania ryzyka selektywnego publikowania jest cała rodzina metod opartych na jednoczesnej analizie wielkości efektów oraz ich błędów standardowych. Analiza ta może mieć charakter czysto jakościowy – polega wtedy na oglądzie wykresu lejka (*funnel plot*), gdzie na osi X znajdują się uzyskane w kolejnych badaniach efekty, na osi Y zaś ich błąd standardowy. Co do zasady te dwie wartości powinny być nieskorelowane. Jeśli daje się zauważyć systematycznie wyższe efekty płynące z badań obarczonych większym błędem – a więc studiów realizowanych na mniejszych próbach, może być to symptomem skrzywienia. Nie oznacza to rzecz jasna automatycznie uznania, że dane są nieprawdziwe – są jednak powody by sądzić, że małe badania, które przyniosły bardzo silne efekty mogą być obarczone błędem I rodzaju. Istnieje też graniczącą z pewnością szansa, że równie małe badania przynoszące wyniki zerowe nie zostałyby opublikowane, jako niekonkluzywne. Dlatego uzupełnieniem jakościowej analizy wykresu lejka jest metoda polegająca na „wycinaniu i wklejaniu" (*trim-and-fill*) (Duval i Tweedie, 2000), polegająca na tym, że w przypadku wykrycia niesymetryczności lejka, imputowane są dodatkowe efekty, które zapewniają jego pełną symetryczność, a całkowity efekt szacowany jest na nowo z efektami dodanymi. Metoda ta jest krytykowana jako zbyt mechaniczna i opierająca się na niewłaściwych założeniach (zob. Peters i in., 2007). Zwraca się na przykład uwagę, że skrzywienie publikacyjne oznacza, że łatwiej jest opublikować wyniki statystycznie istotne, podczas gdy metoda przycinania-wypełniania mechanicznie zastępuje bardzo silne efekty pozytywne równie silnymi efektami negatywnymi, co nie ma żadnego logicznego uzasadnienia, a wprowadzane korekty mogą radykalnie zmieniać uzyskiwane efekty. Tytułem przykładu przyjrzyjmy się jak wygląda wykres lejkowy w przypadku przedstawionej powyżej mini-metaanalizy. Jego ilustracja została zawarta na rysunku 2.



Rysunek 2. Wykres lejkowy dla danych zasymulowanych w metaanalizie prezentowanej powyżej. Panel A i panel B prezentują te same dane, różnią się linią referencyjną – w przypadku panelu A jest nią wartość $r = 0$, w przypadku panelu B – $r = 0{,}29$: średni oszacowany efekt. Białe tło panelu A oznacza efekty nieistotne statystycznie, efekty poza lejkiem są istotne na poziomie $p < 0{,}001$.

Symetryczność uzyskanego rozwiązania – dobrze zilustrowana zwłaszcza na panelu B – nie budzi wątpliwości. Istotnie, zastosowanie metody *trim-and-fill* nie wskazuje na konieczność dodawania dodatkowych efektów, aby wymusić bardziej równomierny rozkład. Można by to więc uznać za potwierdzenie, że nie mamy do czynienia ze zniekształceniem powodowanym przez małe badania.

Alternatywą dla jakościowego oglądu wykresu lejka jest obliczenie korelacji rangowej pomiędzy wielkością efektu w poszczególnych badaniach a ich błędem standardowym (Begg i Mazumdar, 1994). W naszym przypadku jest to $\tau = -0{,}07$, $p = 0{,}87$ – a więc rzeczywiście wielkość efektu i błąd standardowy są od siebie niezależne. Często zdarza się jednak, że jest inaczej – na przykład metaanaliza wpływu zagrożenia stereotypem na wyniki dziewcząt w stereotypowo chłopięcych sferach matematyki i zdolności przestrzennych (Flore i Wicherts, 2015) pokazała, że choć oszacowany efekt jest istotny i negatywny ($g = -0{,}22$), to rozkład efektów jest wyraźnie niesymetryczny ($\tau = -0{,}27$, $p = 0{,}01$), a dołączenie wyestymowanych w procedurze przycinania i wypełniania 11 efektów całkowicie zniosło ogólną zależność.

Stosunkowo nowym, wciąż intensywnie badanym (i coraz częściej krytykowanym, zob. np.: http://datacolada.org/30 bądź http://daniellakens.blogspot.com/2014/12/p-curves-are-better-at-effect-size.html) sposobem korekty wpływu małych badań jest metoda określana skrótem PET-PEESE, za którym stoi „Precision-Effect Testing – Precision-Effect-Estimate with Standard Error" (a więc „test dokładny – oszacowanie na podstawie błędów standardowych"). Jest to korekta oparta na modelu regresji, gdzie wielkość efektu jest przewidywana błędem standardowym (model

PET), bądź wariancją (PEESE) badań. Model ten nie tylko szacuje stopień symetryczności wykresu lejka, ale także koryguje uzyskany efekt o wpływ badań realizowanych na niewielkich próbach. W praktyce rekomenduje się (Stanley i Doucouliagos, 2014) traktowanie estymatora PET-PEESE jako warunkowego – jeśli stała (a więc przewidywana wielkość efektu) w modelu regresji z błędem standardowym jako predyktorem jest statystycznie istotna, wówczas nieobciążone oszacowanie przynosić będzie model PEESE – a więc z wariancją jako predyktorem. Jeśli natomiast stała w modelu PET jest statystycznie nieistotna, to właśnie wartość stałej z tego modelu powinna być traktowana jako skorygowana wielkość efektu. W naszym przypadku stała w modelu PET marginalnie nieistotna (p = 0,07), a więc to wartość stałej z modelu PEESE możemy uznać za najbliższe, nieobciążone oszacowanie faktycznego efektu. Oszacowanie to, to r = 0,29 (95% PU: 0,14-0,44), a więc niemal identyczne, jak uzyskane. Widać więc, że w przykładowej metaanalizie prezentowanej w tym artykule nie mieliśmy do czynienia z wynikami pochodzącymi z małych badań, które zaburzałyby ogólny efekt. Często jednak tak właśnie się dzieje, a procedura PET-PEESE, choć wymaga wciąż symulacji i badań, bo nie jest wolna od problemów (Stanley, 2017), może być w przyszłości użyteczną metodą korekty uzyskiwanych wyników.

Zamiast podsumowania

Przedstawiony w tym artykule proces realizacji metaanaliz może sugerować, że mamy do czynienia z metodą łatwą. Nic bardziej mylnego. Faktyczne projekty metaanalityczne potrafią trwać latami i niemal zawsze najeżone są pułapkami pominięcia istotnych badań na skutek selektywnego przeglądu bądź też błędami w wyznaczaniu wielkości efektów. Dlatego metaanaliza zajmuje wiele czasu i jest procedurą, w której ogromną rolę odgrywa dokładność i wielokrotne powracanie do już zrealizowanych obliczeń.

Szczęśliwie dla zainteresowanych metaanalizą badaczy jej popularność sprawia, że coraz liczniej pojawia się nie tyko poświęcona metaanalizie literatura, ale również oprogramowanie pozwalające na usprawnienie procesu jej realizacji. Mamy więc do czynienia z programami stworzonymi wyłącznie w celu realizacji metaanaliz – na przykład Comprehensive meta-analysis (Biostat – program płatny), MetAnalysis (płatny), WEasyMA (płatny) czy też Mix bądź Review Manager (programy darmowe) (zob. Bax i in., 2007). Istnieją również dedykowane pakiety popularnych programów statystycznych – na przykład kilka niezależnych pakietów dla środowiska R (meta, Schwartzer, 2007; metafor, Viechtbauer, 2010 czy metaSEM, Cheung, 2015) oraz zestawy makr pozwalające realizować metaanalizę w programie SPSS. Wart uwagi jest MetaModel – moduł do realizacji metaanaliz w bezpłatnym programie jamovi (www.jamovi.org). Swoje metaanalityczne rozwiązania mają też inne popularne pakiety statystyczne – Stata, HLM, MPlus i SAS. Jeśli dodać do tego liczne kal-

kulatory – zarówno dostępne online, jak i wersje lokalne możliwe do ściągnięcia – których dziesiątki pokazuje dowolna internetowa wyszukiwarka, sam proces analiz jest stosunkowo prosty. To, co rodzi problemy, to nieoczywiste schematy badawcze, gdzie uzyskanie wielkości efektu nie jest wcale prostą sprawą. Jednak najistotniejszą charakterystyką współczesnych metaanaliz, stanowiącą wyzwanie dla metaanalityków, jest fakt, że stopniowo acz nieubłaganie, metaanaliza przestaje być prostym przeglądem literatury. Redaktorzy i recenzenci słusznie oczekują, że metaanaliza nie tylko podsumuje istniejący efekt relacji między zmiennymi, ale także pozwoli przetestować bardziej wyrafinowane teoretycznie predykcje. W praktyce oznacza to coraz częściej konieczność realizacji kilku metaanaliz, a następnie – na przykład sięgając po modele ścieżkowe – testowania złożonych relacji między zmiennymi, na przykład w poszukiwaniu efektów mediacyjnych (zob. np.: von Stumm, Hell i Chamorro-Premuzic, 2011). Zapewne więc metaanalizy stawać się będą bardziej wyrafinowane statystycznie, aby móc efektywnie służyć teorii.

Literatura cytowana

Bax, L., Yu, L.M., Ikeda, N. i Moons, K.G.M. (2007). A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Medical Research Methodology*, 7, 40.

Becker, B.J. (2005). Failsafe N of File-Drawer number. W: H.R. Rothstein, A.J. Sutton i M. Bornstein (red.), *Publication bias in meta-analysis: Prevention, assessment and adjustment* (s. 111-127). New York: Wiley.

Begg, C.B. i Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50, 1088-1101.

Biostat. (2008). *Comprehensive meta-analysis software* (Version 2.2.064) [Computer software]. Englewood, NJ: Author.

Boland, A., Cherry, M.G. i Dickson, R. (2014). *Doing a systematic review: A student's guide*. London: Sage.

Brockwell, S.E. i Gordon, I.R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in medicine*, 20, 825-840.

Cheung, M.W.L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods*, 46, 29-40.

Cheung, M.W.L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, 5, 1521.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (wyd. 2). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

DerSimonian, R. i Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177-188.

Duval, S. i Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455-463.
Egger, M., Davey Smith, G., Schneider, M. i Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629-634.
Ellis, P.D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* New York: Cambridge University Press.
Eysenck, H.J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, 16, 319-324.
Eysenck, H.J. (1978). An exercise in mega-silliness. *American Psychologist*, 33, 517.
Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, 4 (5), e5738.
Flore, P.C. i Wicherts, J.M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, 53, 25-44.
Gajda, A., Karwowski, M. i Beghetto, R.A. (2017). Creativity and academic achievement: A meta-analysis. *Journal of Educational Psychology*, 109, 269-299.
Greenland, S. i O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*, 2, 463-471.
Gregoire, G.F., Derderian, L. i LeLorier, J. (1995). Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias? *Journal of Clinical Epidemiology*, 48, 159-163.
Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.
Head, M.L., Holman, L., Lanfear, R., Kahn, A.T. i Jennions, M.D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13 (3), e1002106.
Hedges, L.V. i Olkin, I. (1985). *Statistical methods for meta-analysis.* London: Academic Press.
Hedges, L.V. i Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486-504.
Higgins, J.P. i Green, S. (red.) (2011). *Cochrane handbook for systematic reviews of interventions*. T. 4. New York: John Wiley & Sons.
Hunter, J.E. i Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (wyd. 2). Thousand Oaks, CA: Sage.
Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, 2 (8), e124, 696-701. Website: http://medicine.plosjournals.org/ [accessed March 20, 2017].
Ioannidis, J.P.A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19, 640-648.
Ioannidis, J.P.A. i Trikalinos, T.A. (2007). The appropriateness of asymmetry test for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, 176, 1091-1096.
John, L.K., Loewenstein, G. i Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524-532.

Lee, P.M. (2012). *Bayesian statistics: An introduction*. New York: John Wiley & Sons.
Lipsey, M.W. i Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181-1209.
Lipsey, M.W. i Wilson, D.B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.
Matera, J. i Czapska, J. (2014). *Zarys metody przeglądu systematycznego w naukach społecznych*. Warszawa: Instytut Badań Edukacyjnych.
Nosek, B.A. i Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45, 137-141.
Pashler, H. i Wagenmakers, E.J. (2012). Editors' introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science*, 7, 528-530.
Patil, K.D. (1975). Cochran's Q test: Exact distribution. *Journal of the American Statistical Association*, 70, 186-189.
Paulhus, D.L., Robins, R.W., Trzesniewski, K.H. i Tracy, J.L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research*, 39, 303-328.
Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 5, 1243-1246.
Peters, J.L., Sutton, A.J., Jones, D.R., Abrams, K.R. i Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, 26, 4544-4562.
Peterson, R.A. i Brown, S.P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, 90, 175-181.
Rosenthal, R. (1979). The 'file drawer problem' and the tolerance for null results. *Psychological Bulletin*, 86, 638-641.
Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.
Schwarzer, G. (2007). Meta: An R package for meta-analysis. *R news*, 7, 40-45.
Simon, W. (2010a). Meta-analiza w badaniach nad skutecznością psychoterapii. Cz. I: Pytania badawcze, przegląd literatury, kodowanie danych. *Psychiatria i Psychoterapia*, 6, 3-12.
Simon, W. (2010b). Meta-analiza w badaniach nad skutecznością psychoterapii. Cz. II: Rodzaje wielkości efektu, binominalna wielkość efektu, testowanie homogeniczności, zmienne mediujące i moderujące. *Psychiatria i Psychoterapia*, 6, 13-24.
Simon, W. (2010c). Meta-analiza w badaniach nad skutecznością psychoterapii. Cz. III: Losowy i stały model efektu, ważenie wielkości efektu, przycinanie danych, replikacje, przykłady meta-analiz. *Psychiatria i Psychoterapia*, 6, 25-32.
Smith, M.L. i Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-760.
Stanley, T.D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science* [accessed April 18, 2017]. Article first published online: January 1, 2017, doi: 10.1177/1948550617693062

Stanley, T.D. i Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5, 6078.
Szumski, G., Smogorzewska, J. i Karwowski, M. (2017). Academic achievement of students without special educational needs in inclusive classrooms: A meta-analysis. *Educational Research Review*, 21, 33-54.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metaphor package. *Journal of Statistical Software*, 36, 1-48.
Von Stumm, S., Hell, B. i Chamorro-Premuzic, T. (2011). The hungry mind: Intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science*, 6, 574-588.
Walecka, A. i Zakrzewska-Bielawska, A. (2016). Metodyka metaanalizy – egzemplifikacja wykorzystania w naukach o zarządzaniu. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, 6, 63-80.
Weisz, J.R., Weiss, B., Han, S.S., Granger, D.A. i Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin*, 117, 450-468.
Wiśniewska, E. i Karwowski, M. (2007). Efektywność treningów twórczości – podejście metaanalityczne. *Ruch Pedagogiczny*, 3-4, 31-50.

Streszczenie. Artykuł charakteryzuje metaanalizę: ilościową syntezę wcześniejszych badań, jako odrębną metodę badawczą, coraz powszechniej stosowaną w naukach społecznych. Omówione zostały podstawowe funkcje i cele oraz proces realizowania metaanaliz. Dyskutowane są kwestie zastosowania różnych modeli statystycznych w ramach metaanalizy (metaanaliza metodą efektów stałych, efektów losowych i metaanaliza wielopoziomowa) oraz tradycyjnych i współczesnych metod kontroli tzw. selektywnego publikowania. Artykuł wieńczy przykład niewielkiej metaanalizy, pozwalający czytelnikom na prześledzenie analiz krok po kroku i zrealizowanie własnych metaanaliz.
Słowa kluczowe: metaanaliza, wielkość efektu, selektywne publikowanie

Data wpłynięcia: 20.10.2017
Data wpłynięcia po poprawkach: 5.03.2018
Data zatwierdzenia tekstu do druku: 31.03.2018