

Co naprawdę oznaczają dane? Wyniki badawcze, metaanaliza i wiedza kumulatywna w psychologii *

Frank L. Schmidt **
College of Business, University of Iowa

WHAT DO DATA REALLY MEAN? RESEARCH FINDINGS, META-ANALYSIS, AND CUMULATIVE KNOWLEDGE ***

How should data be interpreted to optimize the possibilities for cumulative scientific knowledge? Many believe that traditional data interpretation procedures based on statistical significance tests reduce the impact of sampling error on scientific inference. Meta-analysis shows that the significance test actually obscures underlying regularities and processes in individual studies and in research literatures, leading to systematically erroneous conclusions. Meta-analysis methods can solve these problems – and have done so in some areas. However, meta-analysis represents more than merely a change in methods of data analysis. It requires major changes in the way psychologists view the general research process. Views of the scientific value of the individual empirical study, the current reward structure in research, and even the fundamental nature of scientific discovery may change.

Wiele osób jest rozczarowanych postępem, jaki dokonał się w psychologii w ciągu bieżącego stulecia. Przedstawiono różne przyczyny spowolnienia tego postępu. Wymieniano tu błędy filozofii założeń naukowych (np. przesadny nacisk na pozytywizm logiczny; Glymour, 1980; Schlagel, 1979; Suppe, 1977; Toulmin, 1979). Dyskutowano też nad negatywnym wpływem behawioryzmu (np. Koch, 1984; Mackenzie, 1977; McKeachie,

1976; Schmidt, Hunter i Pearlman, 1981). Niniejszy artykuł skupia się na rzadziej podawanej przyczynie — metodach, których psychologowie (oraz inni przedstawiciele nauk społecznych) zazwyczaj używają do analizy oraz interpretacji danych, zarówno w poszczególnych badaniach, jak i programach badawczych. W artykule tym przedstawione są trzy argumenty: (a) tradycyjna analiza danych oraz procedury interpretacji oparte na testach istotności statystycznej nie sprzyjają odkryciu podstawowych regularności oraz związków, które są podstawą postępu naukowego; (b) metody metaanalizy mogą rozwiązać ten problem, stało się już tak w pewnych dziedzinach; (c) metaanaliza nie jest jedynie no-

This article originally appeared in English as (include full bibliographic citation). Copyright © 1992 by the American Psychological Association. Translated by permission of the publisher and the author. The American Psychological Association is not responsible for the accuracy of this translation. Neither the original nor this translation can be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the American Psychological Association.

* Redaktorem niniejszego artykułu był Frederick A. King. Poprzednia, skrócona wersja została przedstawiona na Drugim Dorocznym Zjeździe Amerykańskiego Towarzystwa Psychologicznego 8 czerwca 1990 roku w Dallas w Teksasie.

** Korespondencję dotyczącą artykułu można kierować pod adresem: Frank L. Schmidt, Department of Management and Organization College of Business Administration, 108 Pappajohn Bus. Adm. Bldg. Iowa 52242-1000.

*** Tłumaczenie z j. angielskiego: Grzegorz Mizera
e-mail: szuwarek@psych2.psych.uw.edu.pl

wym sposobem dokonywania przeglądów literatury. Jest to nowy sposób myślenia o znaczeniu danych, wymagający od nas zmiany poglądów na pojedyncze badanie naukowe, a być może nawet i na podstawową naturę badań naukowych.

METODY TRADYCYJNE VERSUS METAANALIZA

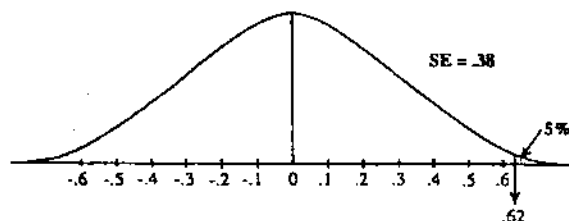
Tradycyjnie, psychologia i nauki społeczne opierały się przede wszystkim na teście istotności statystycznej w interpretowaniu znaczenia danych — zarówno w pojedynczych badaniach, jak i w programach badawczych. Podążając za pomysłem Fishera (1932), sprawdzanie istotności przy zastosowaniu hipotezy zerowej było dominującą procedurą analizy danych. Powszechnie przyjęta reguła decyzyjna, jak wykazał to empirycznie Oakes (1986), jest następująca: jeśli statystyka (t , F , itd.) jest istotna, zachodzi wtedy wpływ (lub związek): jeśli nie jest istotna, to takiego wpływu (lub związku) nie ma. (Patrz także Cohen, 1990; 1992.) Te powszechnie przyjęte procedury interpretacyjne skupiają się na kontrolowaniu błędów pierwszego rodzaju, przy niewielkiej uwadze skierowanej na kontrolowanie błędów drugiego rodzaju. Błąd pierwszego rodzaju (błąd alfa) polega na wnioskowaniu, że zachodzi związek (lub wpływ), podczas gdy w rzeczywistości nie ma on miejsca. Błąd drugiego rodzaju (błąd beta) polega na czymś zgoła przeciwnym — wnioskowaniu, że związek (lub wpływ) nie zachodzi, podczas gdy ma on miejsce. Błąd alfa kontrolowany jest na poziomie 0.05 lub 0.01, podczas gdy błędy beta mogą osiągać wysoki poziom, często rzędu 50%–80% (Cohen, 1962, 1988, 1990; Schmidt, Hunter i Urry, 1976). Aby to zobrazować, rozpatrzmy przykład z hipotetycznego obszaru psychologii eksperymentalnej.

Przypuśćmy, że problemem badawczym jest pytanie o wpływ różnych narkotyków na proces uczenia się. Przypuśćmy też, że faktyczny wpływ poszczególnych dawek wynosi 0.50 odchylenia standardowego wzrostu ilości wyuczonej. Wielkość wpływu 0.50, rozważanego jako średni przez Cohena (1988), odpowiada różnicy między 50. a 69. centylem w rozkładzie normalnym. Przy tej wielkości wpływu 69% z grupy eksperymentalnej przekroczy średnią dla grupy kontrolnej, o ile rozkłady w obu grupach są normalne. W licz-

nych przeglądach literatury na ten temat zakłada się istnienie związku dla tej ogólnej wielkości (Hunter i Schmidt, 1990b). Przypuśćmy teraz, że przy takim dawkowaniu przeprowadzono większą liczbę badań, każde dla 15 szeszurów w grupie eksperymentalnej i 15 w grupie kontrolnej.

$$N_x = N_c = 15$$

$$\text{W sumie } N = 30$$



Wymagane dla istotności: $d_c = .62$; $d_c = [1.645(.38)]$
(Test jednostronny, $\alpha = .05$)

Rys. 1. Rozkład wynikający z hipotezy zerowej wartości d w szeregu eksperymentów

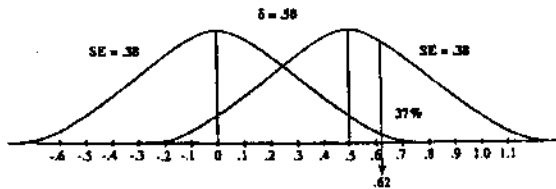
Na rysunku 1 przedstawiony jest oczekiwany rozkład wielkości wpływu (wartości d) pochodzący od hipotezy zerowej. Cała wariancja wokół średniej wartości zero wynika z błędu próby. Aby osiągnąć istotność na poziomie 0.05 (dla testu jednostronnego), wielkość wpływu musi wynosić 0.62 lub więcej. Jeśli hipoteza zerowa jest prawdziwa, tylko 5% będzie równe tej wielkości lub większe. Typowe jest, że w analizie własnych danych badacze skupiają się tylko na informacjach takich, jak na rysunku 1. Większość wierzy, że ich testy istotności ograniczają prawdopodobieństwo błędu do 5%.

W tym przykładzie prawdopodobieństwo błędu pierwszego rodzaju wynosi faktycznie zero — a nie 5%. Ponieważ rzeczywista wielkość wpływu wynosi zawsze 0.50, hipoteza zerowa jest zawsze fałszywa, i dlatego też nie ma możliwości wystąpienia błędu pierwszego rodzaju. Nie można przecież fałszywie wnioskować, że ma miejsce wpływ, kiedy, w istocie rzeczy, wpływ ten istnieje. Jeśli hipoteza zerowa jest fałszywa, jedynym rodzajem błędu, jaki może wystąpić, jest błąd drugiego rodzaju — niemożność wykrycia występującego wpływu. Jedynym rodzajem błędu, jaki więc może

się pojawić, jest błąd nie kontrolowany. Recenzent niniejszego artykułu poprosił mnie o rozwinięcie tego zagadnienia. W dowolnym badaniu, w populacji będącej jego przedmiotem, hipoteza zerowa albo jest fałszywa, albo nie jest fałszywa. Jeśli hipoteza zerowa jest fałszywa, jak w powyższym przykładzie, wówczas istnieje niezerowy wpływ w danej populacji. Błąd pierwszego rodzaju polega na wnioskowaniu, że występuje wpływ, podczas gdy wpływu nie ma. Ponieważ faktycznie wpływ tutaj istnieje, nie ma możliwości popełnienia błędu pierwszego rodzaju. (Hunter i Schmidt, 1990b, ss. 23–31).

$$N_E = N_C = 15$$

$$\text{W sumie } N = 30$$



Wymagane dla istotności: $d_c = .62$
 (Test jednostronny, $\alpha = .05$)
 Moc statystyczna = .37
 Wielkość błędu pierwszego rodzaju = 63%
 Wielkość błędu drugiego rodzaju = 0%

Rys. 2. Moc statystyczna w szeregu eksperymentów

Rysunek 2 pokazuje nie tylko nieistotny rozkład pochodzący od hipotezy zerowej, ale również faktyczny rozkład obserwowanego wpływu wśród badań. Średnią tego rozkładu jest prawdziwa wartość 0.50, ale błąd próby powoduje istotną zmienność w obserwowanych wielkościach wpływu. I znów, aby wpływ był istotny, jego wielkość musi wynieść 0.62 lub więcej. Tylko w 37% przeprowadzonych badań okaże się, że wpływ jest istotny, tak więc moc statystyczna dla tych badań jest równa tylko 0.37. Znaczy to, że prawdziwa wielkość wpływu tego narkotyku wynosi zawsze 0.50 — nigdy nie jest równa zeru. Dotąd jest wykrywana tylko w 37% badań. Wielkość błędu w tych przeglądach literatury wynosi 63%, nie 5% jak wiele osób mogłoby błędnie przypuszczać.

Większość badaczy w psychologii eksperymentalnej, aby dokonać analizy tych danych, zastosowałoby tradycyjnie analizę wariancji (ANOVA).

Oznacza to, że test istotności będzie raczej dwustronny aniżeli jednostronny, tak jak w naszym przykładzie. Przy dwustronnym teście (tj. jednoczynnikowej analizie wariancji) moc statystyczna jest nawet niższa — wynosi 0.26 zamiast 0.37. Wielkość błędu drugiego rodzaju (i stąd ogólna wielkość błędu) wynosiłaby 74%. A więc przykład ten zakłada użycie testu z , chociaż badacze nie stosujący analizy wariancji prawdopodobnie użyliby testu t . Dla jednostronnego testu t z $\alpha = 0.05$ i 28 stopniami swobody, wielkość wpływu (wartość d) musi wynieść 0.65, aby wpływ określić jako istotny (wartość t musi wynosić co najmniej 1.70, zamiast 1.645 wymaganego dla testu z). Przy teście t moc statystyczna byłaby również niższa — 0.35 zamiast 0.37. Tak więc oba alternatywne testy istotności, zastosowane łącznie, przyniosłyby nawet niższą moc statystyczną.

Ponadto badania, które są znaczące, przynoszą zniekształcone oszacowania wielkości wpływu. Prawdziwa wielkość wpływu wynosi zawsze 0.50, wszelkie odstępstwa od 0.50 są spowodowane wyłącznie błędem próby. Ale najmniejsza wartość wymagana dla istotności wynosi 0.62. Uzyskana wartość d musi być o 0.12 większa niż wartość prawdziwa — 24% większa niż jej wartość rzeczywista — aby można było określić ją jako istotną. Średnią istotnych wartości d jest 0.89, czyli o 78% więcej niż wynosi wartość prawdziwa.

W każdym badaniu w tym przykładzie, które przypadkowo (szczęśliwie) przyniosło prawidłową wielkość 0.50, wniosek przy ogólnie obowiązującej regule decyzyjnej będzie taki, że związku nie ma. To znaczy, że tylko badania, które przypadkowo są całkiem nieprawidłowe prowadzą do prawidłowego wniosku, że istnieje związek.

Jak byłoby takie badanie interpretowane w tradycyjnym przeglądzie literatury? Istnieją dwie, najczęściej przyjmowane, interpretacje. Pierwsza oparta jest na tradycyjnej metodzie głosowania (Hedges i Olkin, 1980; Light i Smith, 1971). Stosując tę metodę można zauważyć, że 69% badań odkrywa „brak związku”. Ponieważ owe 69% stanowi większość wszystkich przeprowadzonych badań, można by wysnuć wniosek, że związek nie istnieje. Jednak łatwo zauważyć, że wniosek taki byłby błędny. Wiele dotychczasowych przeglądów zostało napisanych zgodnie z tym wzorcem. Druga interpretacja brzmi następująco: w 63% badań nie wykryto wpływu narkotyku. Jednakże w 37% badań wpływ taki wykryto. Potrzebne jest

więc badanie, aby zidentyfikować zmienne pośredniczące (interakcje), które powodują, iż obserwowalny jest wpływ narkotyku w niektórych badaniach, podczas gdy w pozostałych takiego wpływu zaobserwować nie można. Możliwe jest, przykładowo, że rasa szczura użytego w badaniach lub sposób wstrzykiwania narkotyku wpływa na wyniki badań. Jednak i ta interpretacja jest błędna.

I. Obliczenie rzeczywistej wariancji wielkości wpływu (S_d^2)

1. $S_d^2 = .1444$
(Obserwowana wariancja wartości d)
2. $S_e^2 = .1444$
(Wariancja przewidywana z błędu próby)
3. $S_d^2 = S_d^2 \cdot S_e^2$
4. $S_d^2 = .1444 - .1444 = 0$
(Prawdziwa wariancja wartości δ)

II. Obliczenie średniej wielkości wpływu (δ)

1. $\bar{d} = .50$
(Średnia obserwowanych wartości d)
2. $\bar{\delta} = .50$
3. $SD_\delta = 0$

III. Wniosek: Jest tylko jedna wielkość wpływu, jej wartość wynosi .50 standardowego odchylenia

Rys. 3. Metaanaliza badań z rysunku 1

Jak metaanaliza mogłaby interpretować te badania? Różne podejścia do metaanalizy używają nieco odmiennych procedur ilościowych (Bangert-Drowns, 1986; Glass, McGaw i Smith; 1981; Hedges i Olkin, 1985; Hunter, Schmidt i Jackson, 1982; Hunter i Schmidt; 1990b; Rosenthal, 1984, 1991). Zilustruję ten przykład stosując metody przedstawione przez Huntera i innych (1982) oraz Huntera i Schmidta (1990b). Rysunek 3 pokazuje, że użycie metaanalizy umożliwia wyciągnięcie

prawidłowego wniosku. Metaanaliza oblicza najpierw wariancję obserwowanych wartości d . Następnie stosuje standardowy wzór na wariancję błędu próby (dla wartości d) w celu określenia, ile wariancji, pochodzącej od samego błędu próby, można spodziewać się w obserwowanych wartościach d . Ilość rzeczywistej wariancji w populacji (wartości d) jest szacowana jako różnica między nimi. W naszym przykładzie różnica ta wynosi zero, wskazując prawidłowo, że istnieje tylko jedna wartość dla populacji. Ta pojedyncza wartość dla populacji jest szacowana jako średnia obserwowanej wartości, która wynosi w tym przypadku 0.50, co jest wartością prawidłową. Jeśli liczba badań jest wielka, średnia wartość d będzie bliska prawdziwej wartości, ponieważ błędy próby są przypadkowe, stąd ich uśrednienie do zera. Zauważmy, że te metody metaanalizy nie polegają na statystycznych testach istotności. Stosowane są tylko wielkości wpływu, a testy istotności nie znajdują zastosowania w analizie wielkości wpływu.

Tabela 1
21 badań trafności ($N = 68$ w każdym)

Badanie	Obserwowana trafność	Badanie	Obserwowana trafność
1	0.04	12	0.11
2	0.14	13	0.21
3	0.31*	14	0.37*
4	0.12	15	0.14
5	0.38*	16	0.29*
6	0.27*	17	0.26*
7	0.16	18	0.17
8	0.36	19	0.39*
9	0.20	20	0.22
10	0.02	21	0.21
11	0.23		

* $p < 0.05$, dwustronnie

Dane w tym przykładzie są hipotetyczne. Niemniej jednak, jeśli ktoś przyjmuje ważność podstawowych wzorów statystycznych dla błędów próby, nie będzie miał zastrzeżeń do tego przykładu. Jednak te same zasady stosują się także do prawdziwych danych, jak pokazano to dalej na przykładzie z doбором personelu. Tabela 1 pokazuje zaobserwowane współczynniki trafności z 21

badani pojedynczego testu pracowników umysłowych, oraz pojedynczej miary wykonywania zawodu. W każdym badaniu $N = 68$ (mediana N w literaturze psychologii pracy) i każde badanie jest przypadkowym ciągnięciem (bez przestawień) z pojedynczego większego badania z 1428 badanymi. Korelacja w badaniu ogólnym (nieprawidłowa dla błędu pomiaru, ograniczenia zakresu lub innych artefaktów) wynosi 0.22 (Schmidt, Ocasio, Hillery i Hunter, 1985).

Trafność jest istotna dla ośmiu (czyli 38%) z przeprowadzonych badań, dla wielkości błędu 62%. Tradycyjny wniosek byłby taki, że test ten jest trafny dla 38% organizacji i nietrafny dla pozostałych, oraz że jego średnia obserwowanej trafności w organizacjach, gdzie uznano go za trafny, wynosi 0.33 (co stanowi o 50% więcej niż jego wartość rzeczywista). Metaanaliza tych trafności wskazuje, iż średnia wynosi 0.22 i cała wariancja współczynników jest spowodowana jedynie przez błąd próby. Wnioski płynące z metaanalizy są prawidłowe, zaś wnioski wyciągane na podstawie tradycyjnego podejścia są fałszywe.

Poleganie na testowaniu statystycznej istotności w psychologii i naukach społecznych prowadziło często, przez dłuższy okres, do poważnych błędów w interpretowaniu znaczenia danych (Hunter i Schmidt, 1990b, ss. 29–42, 483–484), które systematycznie opóźniały przyrost wiedzy kumulatywnej. Jest jednak niezmiernie trudno oduczyć badaczy nauk społecznych zachwyty, którym darzą testowanie istotności (Oakes, 1986). Można jedynie żywić nadzieję, że lekcja, jakiej udzieliła metaanaliza, przyspieszy w końcu zmiany. Jednak nawet w ostatnich latach zdarzało się mnie i moim kolegom-badaczom spotkać badaczy, u których zrozumienie zwodniczości testowania istotności zanika, w chwilach słabości, prowadząc do nieuleczalnego niemal nawyku polegania na testach statystycznej istotności. Sam przypadkowo także tego doświadczyłem. Psychologia uzależnienia od testów istotności byłaby fascynującym obszarem badań. Uzależnienie to musi zostać przewyciężone, jednak dla większości badaczy nie będzie to łatwe.

W powyższych przykładach jedynym rodzajem kontrolowanego błędu jest błąd pierwszego rodzaju — błąd, który nie ma prawa wystąpić. Jest prawdopodobne, że w większości obszarów, których badania dotyczą, w miarę upływu czasu oraz w miarę coraz lepszego zrozumienia przez ba-

daczy procesów, których badania te dotyczą coraz rzadziej pojawiać się będzie przypadek, że hipoteza zerowa będzie prawdziwa, natomiast coraz częściej — że będzie fałszywa. Tak więc znaczenie błędu pierwszego rodzaju maleje, zaś znaczenie błędu drugiego rodzaju rośnie. Oznacza to, iż badacz, w miarę upływu czasu, winni zwracać większą uwagę na błąd drugiego rodzaju oraz na moc statystyczną. Jednakże niedawny przegląd opublikowany w *Psychological Bulletin* (Sedlmeier i Gigerenzer, 1989) przyniósł konkluzję, iż średnia moc statystyczna badań w jednym z czasopism American Psychological Association spadła z 46% do 37% w ciągu 22 lat, mimo że już w 1962 Cohen apelował do badaczy, aby zwracali uwagę na moc statystyczną. Tylko w dwóch z 64 eksperymentów wymienionych w przeglądzie wspomniano o mocy statystycznej, ale w żadnym nie dokonano oszacowań tejże mocy. W konkluzji przeglądu stwierdzono, że obniżenie się mocy było spowodowane wzrostem użycia procedur korygujących błąd alfa (*alpha-adjusted procedures*) (takich, jak procedury Newmana-Keulsa, Duncana czy Scheffego). To znaczy, zamiast starać się zmniejszyć wielkość błędu drugiego rodzaju, badacze narzucali coraz ściślejszą kontrolę nad błędami pierwszego rodzaju — które przypuszczalnie nie mogły wystąpić w większości badań. Wynikiem tego jest dalszy wzrost wielkości błędów; średnio o 17%.

W niniejszych przykładach zastosowań metaanalizy badany jest wpływ jedynie błędu próby. Istnieją jednak i inne statystyczne oraz pomiarowe artefakty powodujące artefaktową zmienność w wielkościach wpływu i korelacji między badaniami — na przykład różnice między badaniami co do błędu pomiaru, ograniczenia zakresu oraz dychotomizacji miar. Również w metaanalizie wartości średniej d oraz korelacje średnich muszą być korygowane w celu złagodzenia wpływu takich artefaktów, jak błąd pomiaru i dychotomizacja miar. Artefakty te wykraczają poza zakres niniejszego artykułu, są niemniej szczegółowo przedstawione w innych publikacjach (Hunter & Schmidt, 1990a, 1990b). Naszym celem jest tylko pokazanie, że tradycyjne analizy danych oraz metody interpretacji w logiczny sposób prowadzą do błędnych wniosków, natomiast metaanaliza jest w stanie rozwiązać te problemy.

ZASTOSOWANIE METAANALIZY W PSYCHOLOGII ORGANIZACJI I PRACY

Metody metaanalityczne zostały zastosowane do różnych przeglądów badań w psychologii organizacji pracy. Oto kilka przykładów: (a) korelaty konfliktu i dwuznaczności ról (Fisher i Gittelson, 1983; Jackson i Schuler, 1985); (b) związek między zadowoleniem z pracy a absencją (Hackett i Guion, 1985; Terborg i Lee, 1982); (c) relacja między poziomem wykonania pracy a fluktuacją na stanowiskach pracowniczych (McEvoy i Cascio, 1987); (d) związek między zadowoleniem z pracy a poziomem wykonania pracy (Iaffaldano i Muchinsky, 1985; Petty, McGee i Cavender, 1984); (e) wpływ nieselektywnych interwencji w funkcjonowanie organizacji na produktywność oraz efekty pracy osiągnięte przez pracownika (Guzzo, Jette i Katzell, 1985); (f) wpływ realistycznych oczekiwań pracownika co do pracy na fluktuację na stanowiskach pracowniczych (McEvoy i Cascio, 1985; Premack i Wanous, 1985); (g) ocena teorii przywództwa Fiedlera (Peters, Harthe i Pohlman, 1985) oraz (h) dokładność samooceny zdolności i umiejętności (Mabe i West, 1982). Powyższe relacje zastosowano zarówno do korelacyjnych, jak i eksperymentalnych przeglądów. W psychologii organizacji i pracy opisano wystarczającą liczbę przypadków zastosowań metaanalizy, aby można było opublikować teraz przegląd badań metaanalitycznych w tym obszarze. Ten długi przegląd (Hunter i Hirsh, 1987) odzwierciedla fakt, iż obecnie jest sporo literatury na ten temat. Warto zauważyć, że przegląd poświęca znaczną część na rozwinięcie i prezentację propozycji teoretycznych; jest to możliwe dzięki temu, że wyjaśnienie przeglądów badań dokonane przez metaanalizę dostarcza podstaw do rozwoju teorii, która dotąd nie istniała.

Mimo że do tej pory opublikowano około 50 zastosowań metaanalizy w psychologii organizacji i pracy, najczęstszym po dzień dzisiejszy zastosowaniem jest badanie trafności testów zatrudnienia i innych metod stosowanych przy selekcji personelu. Metaanaliza była używana do testowania hipotezy trafności sytuacyjnie specyficznej (*situation specific validity*). W selekcji personelu dość długo panowało przekonanie, że trafność jest ściśle określona dla danych sytuacji, to znaczy, sądzono, że trafność tego samego testu dla tak samo wyglądającej pracy zmienia się od pra-

cownika do pracownika, od regionu do regionu, w ciągu określonego czasu itd. (Schmidt i in., 1976). W rzeczywistości sądzono, że ten sam test może mieć wysoką trafność (tj. wysoką korelację z poziomem wykonania pracy w jednym miejscu czy też organizacji) oraz być zupełnie nietrafny (tj. mieć zerową trafność) w innych. Przekonanie to oparte było na obserwacji, że uzyskiwane współczynniki trafności oraz poziomy istotności statystycznej dla podobnych bądź identycznych testów oraz zawodów zmieniały się istotnie na przestrzeni różnych badań prowadzonych w różnych środowiskach; to znaczy, było ono oparte na wynikach podobnych do tych z tabeli 1. Zróżnicowanie to było tłumaczone tym, że zawody, które wydawały się takie same, faktycznie różniły się w istotny sposób cechami oraz zdolnościami wymaganymi do ich wykonywania. Przekonanie to doprowadziło do konieczności przeprowadzania badań trafności lokalnej (lub sytuacyjnej). Utrzymywano, że trafność powinna być oszacowana osobno dla każdej sytuacji poprzez badanie przeprowadzone w danym środowisku, to jest wyniki trafności nie mogą być uogólniane na inne środowiska, sytuacje, pracowników itp. (Schmidt i Hunter, 1981). W późnych latach siedemdziesiątych oraz w latach osiemdziesiątych zostały przeprowadzone metaanalizy współczynników trafności, zwane badaniami uogólniania trafności (*validity generalization studies*), w celu sprawdzenia, czy faktycznie nie da się uogólniać trafności (Callender i Osburn, 1981; Hirsh, Northrop i Schmidt, 1986; Pearlman, Schmidt i Hunter, 1980); Schmidt, Gast-Rosenberg i Hunter, 1980; Schmidt i Hunter, 1977; Schmidt, Hunter, Pearlman i Shane, 1979). Gdyby całość lub większa część zmienności z badania na badanie spowodowana była artefaktami, tradycyjne przekonanie o sytuacyjnej specyfice trafności postrzegane byłoby jako błędne, co w efekcie pozwoliłoby na generalizację wyników trafności.

Do tej pory metaanaliza została zastosowana w ponad 500 publikacjach dotyczących selekcji pracowników, przy czym każdy reprezentuje związek predyktor-poziom wykonania pracy. Znalaziono i oceniono kilka nieznacznie różniących się procedur obliczeniowych dla oszacowania wpływu artefaktów. Dają one bardzo podobne wyniki i konkluzje (Callender i Osburn, 1980; Raju i Burke, 1983; Schmidt i in., 1980). Dodatkowo poza testami zdolności i odpowiedniości, badania pre-

dyktorów włączają procedury nietestowe, takie jak: oceny wykształcenia i doświadczenia (McDaniel, Schmidt i Hunter, 1988), wywiady (McDaniel, Whetzel, Schmidt i Mauer, 1991) i skale danych biologicznych (*biodata scales*) (Rothstein, Schmidt, Erwin, Owens i Sparks, 1990). W wielu przypadkach artefakty odpowiadały za całą wariancję w badaniach trafności; średni udział artefaktów w wariancji wyniósł 80%–90% (Schmidt, Hunter, Pearlman i Hirsh, 1985; Schmidt i in., w druku). Jako przykład rozważmy związek między zdolnościami ilościowymi oraz ogólnym poziomem wykonania pracy przez pracowników umysłowych (Pearlman i in., 1980). Badanie to oparte było na 453 korelacjach obliczonych na próbie 39 584 osób. 77% wariancji obserwowanych trafności pochodziło od artefaktów, pozostawiając mało istotną wariancję w wielkości 0.019. Średnia wartość wyniosła 0.47. Tak więc, podsumowanie tej ogromnej liczby danych prowadzi do ogólnej (i dającej się uogólnić) zasady, że związek między zdolnościami ilościowymi i poziomem wykonania pracy umysłowej jest równy w przybliżeniu 0.47, z bardzo niewielką (o ile w ogóle istniejącą) wariancją prawdziwą wokół tej wartości. Tak jak i inne wyniki, rezultat ten pokazuje, że stare przekonanie, iż trafności zależą od specyficznej sytuacji, jest błędne (Schmidt i Hunter, 1981).

W ostatnich czasach liczne organizacje — w tym rząd federalny oraz Urząd do Spraw Zatrudnienia Stanów Zjednoczonych (*U.S. Employment Service*), a także niektóre wielkie korporacje — stosują rezultaty badań generalizacji trafności jako podstawę ich własnych programów selekcyjno-testowych. Generalizacja trafności została włączona do standardowych tekstów (tj. Anastasi, 1982, 1988) w Standardach dla Testów Pedagogicznych i Psychologicznych (*Standards for Educational and Psychological Testing*) (American Educational Research Association, American Psychological Association i National Council on Measurement in Education, 1985). Wysłunięto propozycje, aby włączyć generalizację do rządowego programu zatytułowanego Ujednolicone Wskazówki Dotyczące Procedury Doboru Pracowników (*Principles for the Validation and Use of Personnel Selection Procedures*) (APA, 1987), gdy w niedalekiej przyszłości ten dokument będzie poprawiany w celu odzwierciedlenia zmian w ostatnio uchwalonym Akcie Praw Cywilnych (*Civil Rights Act*) w 1991 roku. Ostatnie spr-

wozдание dokonane przez Narodową Akademię Nauk (*National Academy of Sciences*) (Hartigan i Wigdor, 1989) poświęca cały rozdział (rozdz. 6) generalizacji trafności oraz potwierdza jej metody i założenia.

ROLA METAANALIZY W ROZWIJANIU TEORII

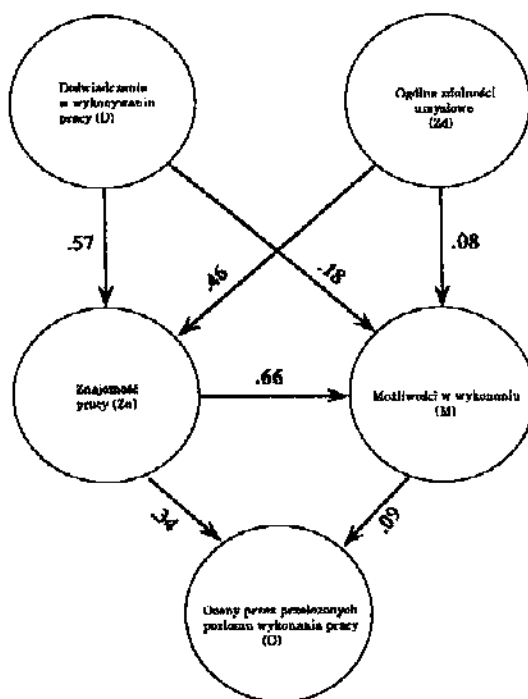
Głównym celem każdej dyscypliny naukowej jest rozwijanie teorii. Dobra teoria jest po prostu dobrym wyjaśnieniem procesów, które faktycznie zachodzą w zjawisku. Przykładowo biorąc, co zdarza się, kiedy pracownicy osiągają wysoki poziom zaangażowania w funkcjonowanie organizacji? Czy zadowolenie z pracy rozwija się najpierw, a potem powoduje rozwój zaangażowania? Jeśli tak, to co powoduje, że zadowolenie z pracy rozwija się i jaki ma to wpływ na zaangażowanie? Jako kolejny przykład — jak wyższe poziomy zdolności umysłowych pociągają za sobą wyższe poziomy wykonania pracy — czy tylko przez poszerzanie wiedzy zawodowej, czy także poprzez bezpośrednie doskonalenie rozwiązywania problemów, na którym praca polega? Naukowiec społeczny jest w istocie detektywem: jego zajęciem jest odkrywanie, jak się dzieje to, co się dzieje i dlaczego czyni to właśnie w taki, a nie inny sposób. Jednak aby budować teorie, należy najpierw poznać pewne podstawowe związki, jak np. empiryczne związki pomiędzy zmiennymi. Związki te są cegiełkami, z których wznosi się gmach teorii. Na przykład, jeśli istnieje wysoka i zgodna korelacja między zadowoleniem z pracy a zaangażowaniem w funkcjonowanie organizacji, sprawi to, że teoria rozwinię się w konkretnych kierunkach. Jeśli korelacja między tymi zmiennymi jest niska i zgodna, rozwój teorii nastąpi w różnych kierunkach. Jeśli wiadomo, że związek jest zmienny wśród organizacji i środowisk, badacze będą zachęcani do rozwijania teorii interaktywnych i opartych na moderatorach (*moderator-based theory*). Metaanaliza dostarcza tych empirycznych cegiełek dla teorii. Wyniki metaanalizy ujawniają, czym jest to coś, co ma być wyjaśnione przez teorię.

Teorie są przyczynowymi wyjaśnieniami. Celem każdej dyscypliny naukowej jest wyjaśnienie, a wyjaśnienia zawsze są przyczynowe. W naukach behawioralnych i społecznych metody a-

nalizy ścieżkowej (tj., patrz Hunter i Gerbing, 1982; Kenny, 1979; Loehlin, 1987) mogą być użyte w celu sprawdzenia teorii przyczynowych, gdy dane wychodzą naprzeciw założeniom metody. Związki ujawnione przez metaanalizę — empiryczne cegielki dla teorii — mogą być użyte w analizie ścieżkowej, aby sprawdzić teorie przyczynowe, nawet jeśli wszystkie naszkicowane związki mają charakter raczej obserwacyjny niż eksperymentalny. Eksperymentalnie zdeterminowane związki mogą być również wprowadzone do analizy ścieżkowej razem ze związkami na podstawie obserwacji. Konieczne jest jedynie przekształcenie wartości d na korelacje (Hunter i Schmidt, 1990b, rozdz. 7). Tak więc analiza ścieżkowa może być „mieszana”. Analiza ścieżkowa może stanowić użyteczne narzędzie do redukcji liczby teorii, które mogą być zgodne z danymi — czasami do bardzo małej ich liczby, a czasami tylko do jednej (Hunter, 1988). Każde takie zredukowanie liczby możliwych teorii stanowi postępowanie w rozumieniu.

Niedawne badanie (Schmidt, Hunter i Outerbridge, 1986) jest tego przykładem. Zastosowaliśmy metaanalizę do badań korelacji pomiędzy następującymi zmiennymi: ogólne zdolności umysłowe, znajomość pracy (*job knowledge*), możliwości w wykonywaniu pracy (*job performance capability*), oceny przez przełożonych poziomu wykonywanej pracy (*supervisory ratings of job performance*) oraz doświadczenie w wykonywaniu pracy. Ogólne zdolności umysłowe mierzone były przy użyciu wystandaryzowanej grupy testów inteligencji.; znajomość pracy była szacowana przez trafne treściowo pisemne miary faktów, zasad oraz metod, których znajomość była potrzebna do wykonywania danej pracy; możliwości w wykonywaniu pracy były mierzone przy użyciu próbek pracy, które symulowały lub reprodukowały istotne zadania występujące w pracy, które ujawniła analiza danej pracy; zaś doświadczenie w pracy było mierzone kodyfikacją miesięcy w pracy będącej przedmiotem rozważań. Korelacje te zostały skorygowane ze względu na błąd pomiaru, ich wynikiem była macierz metaanalitycznie estymowanych korelacji przedstawiona w tabeli 2. Użyliśmy tych korelacji w celu sprawdzenia teorii o wspólnym oddziaływaniu zdolności umysłowych i doświadczenia na znajomość oraz możliwości, na podstawie metod opisanych przez Huntera i Gerbinga (1982). Wyniki analizy ścieżkowej są po-

kazane na rysunku 4. Wyniki te wskazują, że główne oddziaływanie zdolności umysłowych na możliwości (mierzone na próbce pracy) jest pośrednie: wyższe zdolności umysłowe prowadzą do zwiększenia znajomości, która z kolei ma silny wpływ na możliwości $[(0.46)(0.66)=.30]$. Ten pośredni wpływ zdolności jest prawie czterokrotnie większy niż bezpośredni wpływ zdolności na możliwości. Układ dla doświadczenia jest podobny: pośredni wpływ poprzez zwiększanie znajomości jest znacznie większy niż bezpośredni wpływ na możliwości. Analiza również ujawnia, że oceny wykonania pracy dokonywane przez przełożonych są silniej zdeterminowane przez znajomość niż przez możliwości mierzone miarami próbek pracy. Nie chcę rozwodzić się dłużej nad wynikami tego badania. Istotne jest, że badanie to jest przykładem dwustopniowego procesu, który stosuje metaanalizę w rozwoju teorii.



Rys. 4. Model ścieżkowy oraz współczynniki ścieżkowe

Krok 1. Zastosowanie metaanalizy w celu uzyskania dokładnych oszacowań związków między zmiennymi. Metaanaliza dokonuje uśrednienia odchylenia błędu próby od wartości poprawnych i koryguje średnie ze względu na zniekształcenia

spowodowane przez błąd pomiaru i inne artefakty. Na wyniki metaanalizy nie mają wpływu problemy, które w logiczny sposób mogą zniekształcać wyniki i interpretacje testów istotności.

Krok 2. Zastosowanie pochodzących z metaanalizy oszacowań związków do analizy ścieżkowej w celu sprawdzenia teorii.

SZERSZE ODDZIAŁYWANIE METAANALIZY

Niektórzy twierdzą, że metaanaliza nie jest niczym więcej niż nową, ilościową metodą prowadzenia przeglądów literatury (Guzzo, Jackson i Katzell, 1986). Gdyby tak faktycznie było, jej oddziaływanie byłoby w pełni oceniane jedynie poprzez sprawdzanie różnic we wnioskach płynących z metaanalizy vs. tradycyjnych przeglądów literatury. Różnice te są istotne, wskazują bowiem, że wnioski z przeglądów opisowych, opartych na tradycyjnych procedurach testów statystycznej istotności, są często błędne. Niemniej jednak, metaanaliza to coś znacznie więcej niż nowa metoda dokonywania przeglądów. Rzeczywistość dotycząca danych i wyników naukowych, ujawniona przez zasady metaanalizy, wymaga wielkich zmian w naszych poglądach na poszczególne badanie empiryczne, w naturze kumulatywnej wiedzy badawczej oraz w strukturze wynagrodzeń w przedsięwzięciach badawczych.

Metaanaliza wyjaśnia krytyczną rolę błędów próby, błędów pomiaru oraz innych artefaktów w determinowaniu obserwowanych rezultatów oraz mocy statystycznej poszczególnych badań. W ten sposób metaanaliza ujawnia, jak niewiele informacji zawartych jest w pojedynczym badaniu, co dowodzi, inaczej niż głosi to powszechnie przyjęty pogląd, że żadne jednostkowe badanie nie jest w stanie rozwiązać danej kwestii czy znaleźć odpowiedzi na dane pytanie. Rozważenie zasad metaanalizy sugeruje, że w naukach behawioralnych i społecznych panuje przesadny kult empiryzmu, to znaczy, nadmierna wiara w dane, jako bezpośrednie źródło prawd naukowych, oraz niedocenywanie, jak mylna jest większość danych w naukach społecznych, gdy przyjmowane są jako wartość nominalna i interpretowane naiwnie. Powszechne przekonanie, że postęp badawczy nastąpi tylko wtedy, gdy „oddamy głos danym”, jest niestety błędne. Ze względu na wpływy artefak-

tów takich, jak błąd próby i błąd pomiaru, trafniejsze byłoby stwierdzenie, że dane przychodzą do nas zaszyfrowane i, aby zrozumieć ich znaczenie, najpierw musimy ten szyfr złamać. Metaanaliza tego właśnie wymaga. Dlatego każde pojedyncze badanie musi być rozważane jako pojedynczy element danych stanowiący wkład do przyszłej metaanalizy. Tak więc status naukowy oraz wartość pojedynczego badania są z konieczności zredukowane.

Wynikiem jest przesunięcie uwagi w odkryciu naukowym z pojedynczego badania wstępnego (*primary research*) na metaanalizę, tworząc wielką zmianę we względnym statusie przeglądów. Czasopisma, które dotąd publikowały tylko badania wstępne i odmawiały publikacji przeglądów, publikują teraz mnóstwo przeglądów metaanalizy. W przeszłości przeglądy badań były oparte na metodzie narratywno-subiektywnej (*narrative-subjective method*), miały ograniczony status oraz zyskiwały niewielkie poparcie w kręgach akademickich. Wynagrodzenia szły dla tych, którzy dokonywali badań wstępnych. Być może było to właściwe, ponieważ w retrospektywie widać, że takie przeglądy często przynosiły niewielki wkład w gromadzenie wiedzy (Glass i in., 1981; Hedges i Olkin, 1985). Na szczęście nastąpił olbrzymi rozwój. Dzisiaj wiele odkryć i postępów w kumulowaniu wiedzy jest dokonywanych nie przez tych, którzy przeprowadzają badania wstępne, ale przez tych, którzy stosują metaanalizę w celu odkrywania ukrytych znaczeń istniejących już przeglądów badań. Dzisiaj jest możliwe, żeby badacze z nauk behawioralnych oraz społecznych, z niezbędną praktyką i umiejętnościami, mogli dokonywać wielkich, oryginalnych odkryć oraz wносить wkład bez prowadzenia pierwotnych badań naukowych — po prostu przez wydobywanie informacji z zebranych przeglądów badań. Proces ten jest już dzisiaj zaawansowany. Psychologia organizacji i pracy oraz badanie zachowań związanych z funkcjonowaniem w organizacji — które znam najlepiej — są procesowi temu coraz częściej poddawane. Jest to widoczne nie tylko w liczbie publikowanych metaanaliz, ale, co ważniejsze — w przesunięciu układu cytowań w literaturze i książkach z badań wstępnych na metaanalizy. To samo ma miejsce w psychologii wychowawczej, psychologii społecznej, medycynie, finansach, marketingu i innych obszarach (Hunter i Schmidt, 1990b, rozdz. 1).

Tabela 2

Pierwotna macierz korelacji, macierz korelacji reprodukowana z modelu ścieżkowego oraz macierz różnic

Zmienna	Zd	Zn	M	O	D
Macierz pierwotna					
Zdolności (Zd)	—	0.46	0.38	0.16	0.00
Znajomość pracy (Zn)	0.46	—	0.80	0.42	0.57
Możliwości w wykonywaniu pracy (M)	0.38	0.80	—	0.37	0.56
Oceny przez przełożonych (O)	0.16	0.42	0.37	—	0.24
Doświadczenie w wykonywaniu pracy (D)	0.00	0.57	0.56	0.24	—
Macierz reprodukowana					
Zdolności (Zd)	—	0.46	0.38	0.19	0.00
Znajomość pracy (Zn)	0.46	—	0.80	0.41	0.57
Możliwości w wykonywaniu pracy (M)	0.38	0.80	—	0.36	0.56
Oceny przez przełożonych (O)	0.19	0.41	0.36	—	0.24
Doświadczenie w wykonywaniu pracy (D)	0.00	0.57	0.56	0.24	—
Macierz różnic (pierwotna minus reprodukowana)					
Zdolności (Zd)	—	0.00	0.00	-0.03	0.00
Znajomość pracy (Zn)	0.00	—	0.00	0.01	0.00
Możliwości w wykonywaniu pracy (M)	0.00	0.00	—	0.01	0.00
Oceny przez przełożonych (O)	-0.03	0.01	0.01	—	0.00
Doświadczenie w wykonywaniu pracy (D)	0.00	0.00	0.00	0.00	—

Proces metaanalizyczny, w porządkowaniu i nadawaniu sensu publikacjom badawczym, nie tylko ujawnia tkwiącą w tym wiedzę kumulatywną, ale również zapobiega przetworzeniu wartościowych zasobów badawczych w niepotrzebne badania. Zastosowania metaanalizy ujawniły, że istnieją pytania, dla których dodatkowe badania zmarnowałyby tylko naukowo i społecznie wartościowe środki. Przykładowo, w 1980 roku przeprowadzono 882 badań, porównujących wpływ szybkości spostrzeżeniowej na poziom wykonania pracy przez pracowników umysłowych, na całkowitej próbie 70 935 osób. Na podstawie tych

badania, nasze metaanalizyczne oszacowanie tej korelacji wynosi 0.47 ($S^2 = 0.05$; Pearlman i in., 1980). Dla innych zdolności przeprowadzono często 200-300 badań. Jest oczywiste, że dalsze badania nad tymi związkami nie są najlepszym wykorzystaniem dostępnych źródeł.

Tylko metaanalizyczna integracja wyników z różnych badań może kontrolować szanse oraz inne statystyczne i pomiarowe artefakty, jak również zapewnić wiarygodne podstawy wnioskowania. A jednak metaanaliza nie jest możliwa, jeśli nie są przeprowadzone niezbędne badania wstępne. Czy jest możliwe, że metaanaliza zabije

motywację do prowadzenia pierwotnych badań? W nowych obszarach badań ten potencjalny problem nie jest aż tak niepokojący. Pierwsze badanie prowadzone nad daną kwestią zawiera 100% osiągalnej informacji badawczej, drugie zawiera z grubsza 50% itd. Tak więc, wczesne badania w dowolnym obszarze mają określony status. Ale pięćdziesiąte badanie zawiera jedynie 2% dostępnej informacji, setne około 1%. Czy będziemy mieli trudności motywując badaczy do przeprowadzenia pięćdziesiątego i setnego badania? Odpowiedź zależy od przyszłego systemu wynagrodzeń w naukach społecznych i behawioralnych. Jaka będzie struktura wynagrodzeń? Jaka być powinna?

Jedną z możliwości — niekoniecznie pożądaną — jest to, że będziemy mieli przedsięwzięcia naukowe o dwuwarstwowej (*two-tiered*) strukturze. Jedna grupa badaczy będzie specjalizowała się w prowadzeniu pojedynczych badań, druga zaś grupa będzie stosowała złożone i wyrefinowane metody metaanalizy do tych badań kumulatywnych oraz będzie dokonywała odkryć naukowych. Jak te dwie grupy powinny być wynagradzane? Jaki byłby ich względny status w ogólnych przedsięwzięciach naukowych? Czy byłoby to porównywalne z podziałem pracy w fizyce na fizyków teoretycznych oraz fizyków doświadczalnych? Fizycy doświadczalni prowadzą badania, a fizycy teoretyczni interpretują ich znaczenie. Ta analogia może być bardzo trafna: Hedges (1987) odkrył, że fizycy teoretyczni (i chemicy) używają metod „zasadniczo identycznych” z metaanalizą. W rzeczywistości podobna do tej struktura istnieje już w niektórych obszarach psychologii organizacji i pracy. A zatem, czy jest to powiew przyszłości?

Można by zapytać, dlaczego badacze przeprowadzający badania pierwotne nie mogą prowadzić też metaanalizywnych badań wtórnych? To może się zdarzyć i to już się zdarzyło w niektórych obszarach badawczych. Istnieją jednak niepokojące trendy, które mogą zniweczyć ten wynik w przyszłości. Opanowanie danego obszaru badań wymaga często poświęcenia całego czasu i wysiłku. Badania na poziomie „wstępnym”, często złożone, pochłaniają też mnóstwo czasu, pozostawiając niewiele czasu i energii na opanowanie nowych metod ilościowych, takich jak metaanaliza. Drugą sprawą jest fakt, iż w miarę ich udoskonalania, metody metaanalizy stają się co-

raz bardziej wymyślne i abstrakcyjne. Z własnego doświadczenia wiem, że wielu badaczy „wstępnych” odczuło już na początku lat osiemdziesiątych, że metody metaanalizy są dosyć złożone i nieco odstręczające. Patrząc wstecz, metody te były relatywnie proste. Nasza książka z 1982 roku o metaanalizie (Hunter i in., 1982) liczyła 172 strony; nasza książka z 1990 roku (Hunter i Schmidt, 1990b) ma 576 stron. Książka dotycząca metaanalizy Hedgesa i Olkina (1985) jest być może bardziej złożona pod względem statystycznym i liczy 325 stron. Zwiększenie dokładności metod metaanalizywnych jest uzasadnione naukową koniecznością precyzji. Jednakże zwiększenie dokładności można osiągnąć za cenę wzrostu złożoności oraz abstrakcyjności, co czyni te metody trudniejszymi do opanowania i stosowania.

Być może rozwiązaniem idealnym dla metaanalizy byłaby praca z jednym lub dwoma badaczami „wstępnymi” w przeprowadzaniu metaanaliz. Wiedza o obszarze badań pierwotnych jest pożądana, jeśli nie konieczna, po to, by jak najlepiej stosować metaanalizę. Ale to rozwiązanie unika przedsięwzięć naukowych o strukturze dwuwarstwowej dla tylko jednej małej grupy badaczy pierwotnych. Dla całej reszty rzeczywistością będzie struktura dwuwarstwowa.

WNIOSKI

Tradycyjne procedury analizy danych i interpretacji poszczególnych badań oraz przeglądów literatury hamują rozwój kumulatywnej wiedzy w psychologii. Procedury te, oparte na testach istotności statystycznej oraz hipotezie zerowej, w logiczny sposób prowadzą do błędnych wniosków, ponieważ przeceniają ilość informacji zawartej w pojedynczym badaniu, a także ignorują błędy drugiego rodzaju oraz moc statystyczną. Metaanaliza jest w stanie rozwiązać te problemy. Ale metaanaliza jest czymś więcej niż tylko nowym sposobem dokonywania przeglądów badań. Jest nowym sposobem patrzenia na znaczenie danych. Jako taka prowadzi do innego poglądu na pojedyncze badania oraz do zmienionej koncepcji odkrycia naukowego, może także prowadzić do zmiany podziału ról w przedsięwzięciach badawczych.

LITERATURA

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Psychological Association, Division of Industrial and Organizational Psychology (Division 14). (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.
- Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, 99, 388-399.
- Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology*, 65, 543-558.
- Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer generated sampling distributions of the multiplicative model variance estimates: Results for petroleum industry validation research. *Journal of Applied Psychology*, 66, 274-281.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1, 98-101.
- Fisher, C. D., & Gittelson, R. (1983). A meta-analysis of the correlates of role conflict and ambiguity. *Journal of Applied Psychology*, 68, 320-333.
- Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Edinburgh, Scotland: Oliver & Boyd.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Glymour, C. (1980). The good theories do. In E. B. Williams (Ed.), *Construct validity in psychological measurement* (pp. 13-21). Princeton, NJ: Educational Testing Service.
- Guzzo, R. A., Jackson, S. E., & Katzell, R. A. (1986). Meta-analysis analysis. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 9). Greenwich, CT: JAI Press.
- Guzzo, R. A., Jette, R. D., & Katzell, R. A. (1986). The effects of psychologically based intervention programs on worker productivity: A meta-analysis. *Personnel Psychology*, 38, 276-292.
- Hackett, R. D., & Guion, R. M. (1985). A re-evaluation of the absenteeism job satisfaction relationship. *Organizational Behavior and Human Decision Processes*, 35, 340-381.
- Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.
- Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist*, 42, 443-455.
- Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin*, 88, 359-369.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Hirsh, H. R., Northrop, L., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology*, 39, 399-420.
- Hunter, J. E. (1988). A path analytic approach to analysis of covariance. Unpublished manuscript, Michigan State University, Department of Psychology.
- Hunter, J. E., & Gerbing, D. W. (1982). Unidimensional measurement, second order factor analysis and causal models. [w:] Staw, B. M. & Cummings, L. L. (Eds.), *Research in organizational behavior* (Vol. 4). Greenwich, CT: JAI Press.
- Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper & I. I. Robertson (Eds.), *International review of industrial and organizational psychology 1987* (pp. 321-357). New York: Wiley.
- Hunter, J. E., & Schmidt, F. L. (1990a). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, 75, 334-349.
- Hunter, J. E., & Schmidt, F. L. (1990b). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Iaffaldone, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin*, 97, 251-273.
- Jackson, S. E., & Schuler, R. S. (1985). A meta-analysis and conceptual critique of research on role ambiguity and role conflict in work settings. *Organizational Behavior and Human Decision Processes*, 36, 16-78.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.
- Koch, S. (1964). Psychology and emerging conceptions of knowledge as unitary. [w:] Wann, T. W. (Ed.), *Behaviorism and Phenomenology*. Chicago: University of Chicago Press.
- Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 41, 429-471.
- Loehlin, J. C. (1987). *Latent variable models: An introduction to factor path, and structural analysis*. Hillsdale, NJ: Erlbaum.
- Mabe, P. A., III, & West, S. G. (1982). Validity of self evaluations of ability: A review and meta-analysis. *Journal of Applied Psychology*, 67, 280-296.
- Mackenzie, B. D. (1977). *Behaviorism and the limits of scientific method*. Atlantic Highlands, NJ: Humanities Press.
- McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods of rating training and experience in personnel selection. *Personnel Psychology*, 41, 283-314.
- McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Mauer, S. (1991). The validity of employment interviews: A review and meta-analysis. Manuscript submitted for publication.
- McEvoy, G. M., & Cascio, W. E. (1985). Strategies for reducing employee turnover: A meta-analysis. *Journal of Applied Psychology*, 70, 342-353.
- McEvoy, G. M., & Cascio, W. F. (1987). Do poor performers leave? A meta-analysis of the relation between performance and turnover. *Academy of Management Journal*, 30, 744-762.

- McKeachie, W. J. (1976). Psychology in America's bicentennial year. *American Psychologist*, 31, 819-833.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
- Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, 65, 373-406.
- Peters, L. H., Harthe, D., & Pohlman, J. (1985). Fiedler's contingency theory of leadership: An application of the meta-analysis procedures of Schmidt and Hunter. *Psychological Bulletin*, 97, 274-285.
- Petty, M. M., McGee, G. W., & Cavander, J. W. (1984). A meta-analysis of the relationship between individual job satisfaction and individual performance. *Academy of Management Review*, 9, 712-721.
- Premack, S., & Wanous, J. P. (1985). Meta-analysis of realistic job preview experiments. *Journal of Applied Psychology*, 70, 706-719.
- Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. *Journal of Applied Psychology*, 68, 382-395.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.
- Rothstein, H. R., Schmidt, E. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology*, 75, 175-184.
- Schlagel, R. H. (1979, September). *Evaluation in the philosophy of science: Implications for method and theory in psychology*. Invited address at the 87th Annual Convention of the American Psychological Association, New York.
- Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology*, 65, 643-681.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529-540.
- Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, 36, 1128-1137.
- Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, 71, 432-439.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, 66, 166-185.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology*, 38, 697-798.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology*, 32, 257-381.
- Schmidt, F. L., Hunter, J. E., & Urry, V. E. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, 61, 473-485.
- Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. A. (in press). Refinements in validity generalization procedures: Implications for the situational specificity hypothesis. *Journal of Applied Psychology*.
- Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology*, 38, 509-524.
- Sedlmeier P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-316.
- Suppe, E. (Ed.). (1977). *The structure of scientific theories*. Urbana: University of Illinois Press.
- Terborg, J. R., & Lee, I. W. (1982). Extension of the Schmidt-Hunter validity generalization procedure to the prediction of absenteeism behavior from knowledge of job satisfaction and organizational commitment. *Journal of Applied Psychology*, 67, 280-286.
- Toulmin, S. E. (1979, September). The cult of empiricism. In P. E. Secord (Chair), *Psychology and Philosophy*. Symposium conducted at the 87th Annual Convention of the American Psychological Association, New York.

Tłumaczył: Grzegorz Mizera