Problemy Zarządzania, vol. 12, nr 1 (45): 137–155
ISSN 1644-9584, © Wydział Zarządzania UW
DOI 10.7172/1644-9584.45.9

Ewaluacja projektów i abstraktów – wpływ indywidualnego stylu ewaluacji na oceny

Nadesłany: 30.10.13 | Zaakceptowany do druku: 20.12.13

Grzegorz Król*, Katarzyna Kinga Kowalczyk**

W artykule przedstawiamy zarys psychologicznego modelu ewaluacji oraz wynik analizy ocen 673 abstraktów zgłoszonych na międzynarodową konferencję i 323 projektów badawczych. W obu przypadkach jeden ewaluator (recenzent) dokonywał oceny wielu projektów lub abstraktów, co pozwoliło na wykazanie wpływu indywidualnych stylów ewaluacji (dla 18 ewaluatorów projektów badawczych i 33 recenzentów abstraktów), operacjonalizowanych przez poziom łagodności/surowości i stopień różnicowania ocen. Pokazano także związek kolejności oceniania z surowością ocen (pierwsze 15 abstraktów ocenianych było surowiej niż reszta), co zinterpretowano za pomocą dwóch alternatywnych wyjaśnień: stabilizacji utajonego wzorca porównawczego oraz zmęczenia.

Słowa kluczowe: ewaluacja, utajone wzorce, recenzje, efekt łagodności, różnicowanie ocen.

Evaluation of grant proposals and abstracts – the influence of individual evaluation style on ratings

Submitted: 30.10.13 | Accepted: 20.12.13

In this article we present an outline of psychological model of evaluation and results of ratings analysis for 673 abstracts submitted to international conference and 323 grant proposals. In both cases one evaluator (reviewer) evaluated many proposals or abstracts, which allowed us to demonstrate individual evaluation style (for 18 grant proposals evaluators and 33 abstract reviewers), operationalized by leniency/severity level and degree of ratings differentiation. We showed as well relation between rank order and severity rating (first 15 abstracts were rated more severely than others). It is suggested that this relation can be explained by two alternative explanations: stabilization of implicit comparative standard and fatigue.

Keywords: evaluation, implicit standards, reviews, leniency terror, ratings differentiation.

JEL: C18

* **Grzegorz Król** – dr, Wydział Zarządzania, Uniwersytet Warszawski.
** **Katarzyna Kinga Kowalczyk** – mgr, Wydział Zarządzania, Uniwersytet Warszawski.

Adres do korespondencji: Uniwersytet Warszawski, Wydział Zarządzania, ul. Szturmowa 1/3, 02-678 Warszawa; e-mail: gkrol@cjk.pl.

1. Wprowadzenie

Ocena jakości produktu materialnego (np. samochodu, układów scalonych) czy usług (np. biura podróży, banku) jest o wiele łatwiejsza niż ocena produktu niematerialnego, takiego jak np. projekt badawczy, abstrakt wystąpienia konferencyjnego czy dorobek naukowy. O ile w przypadku większości przedmiotów materialnych istnieją precyzyjnie określone normy (np. parametry techniczne), o tyle w ocenach produktów mentalnych proces obiektywizacji oceny jest dużym wyzwaniem.

Nic dziwnego, że powstała cała gałąź nauki zajmująca się problemami ewaluacji. W materiałach Polskiego Towarzystwa Ewaluacyjnego (2008, s. 3–4) możemy przeczytać, że „Ewaluacja koncentruje się na gromadzeniu wiedzy, która jest podstawą do formułowania sądów na temat wartości podejmowanych działań (…). Dla powodzenia ewaluacji podstawowe znaczenie ma: precyzyjnie i logicznie spójne określenie przedmiotu i celu ewaluacji, określenie kluczowych pytań, na które poszukuje się odpowiedzi oraz kryteriów, z zastosowaniem których przedmiot ewaluacji będzie ewaluowany". Ze względu na moment przeprowadzania badania wyróżniane są następujące rodzaje ewaluacji (Ciężka, 2005): (1) *ex-ante* (przed rozpoczęciem działania); (2) *mid-term* (w połowie trwania działania); (3) *ad hoc*/*on-going* (wykonywana na bieżąco, w dowolnym momencie) oraz (4) *ex-post* (po zakończeniu działania).

Narzędzia ewaluacyjne są wykorzystywane w niemal każdej dziedzinie zarządzania. Ich głównym celem jest zapewnienie (precyzyjniej: zwiększenie) obiektywizmu ocen poprzez skwantyfikowanie w ujęciu ilościowym proponowanych kryteriów. Jest to jednak zadanie trudne i obarczone wieloma potencjalnymi błędami. Ewaluatorzy dysponują najczęściej dostarczonymi przez prowadzącego ewaluację liczbowymi skalami odpowiedzi, za pomocą których mają wyrazić swoje opinie. W większości ankiet mogą również dołączyć słowny komentarz.

W artykule przedstawiamy analizy ewaluacji zbioru 673 abstraktów (nazywane dalej *zbiorem A*) zgłoszonych na międzynarodową konferencję i 323 projektów badawczych (nazywane dalej *zbiorem P*), więc rozważania teoretyczne będziemy ilustrować przede wszystkimi tymi dwoma przykładami. Są to ewaluacje ex-ante, których wyniki mają istotne konsekwencje. Odrzucone w jej wyniku projekty lub abstrakty nie uzyskują finansowania i nie są realizowane lub prezentowane. Zanim przejdziemy do psychologicznego modelu ewaluacji, przedstawiamy informacje o sposobie ewaluacji wykorzystanym w analizowanych danych.

Zbiór A zawiera oceny 673 abstraktów zgłoszonych na międzynarodową konferencję o tematyce zdrowia publicznego w 2011 r. Recenzentami byli przedstawiciele międzynarodowych organizacji pozarządowych, naukowcy oraz niezależni eksperci z całego świata. Autorzy najlepszych abstraktów mogli liczyć na stypendium konferencyjne, pokrywające m.in. koszty podróży. Proces

 DOI 10.7172/1644-9584.45.9

ewaluacji abstraktów był dwustopniowy, składał się z fazy recenzji on-line oraz fazy spotkania komitetu programowego, gdzie podejmowano ostateczne decyzje. Ocena abstraktów odbywała się za pośrednictwem narzędzi dostępnych on-line, poprzez przeglądarkę internetową, gdzie oceniający mieli dostęp do treści zgłoszeń. Abstrakty prezentowane były anonimowo i przypisywane przez system do recenzenta w sposób losowy, z zachowaniem kontroli konfliktu interesów. Ewaluatorzy oceniali prace na 6 wymiarach, opisanych na skali 0 do 10 każdy. Oceny cząstkowe zostały przekształcone w średnią ogólną z następującymi wagami: jakość (waga 0,1), wartość teoretyczna lub praktyczna (0,1), oryginalność (0,1), zgodność z tematem konferencji (0,1), jakość prezentacji (0,1), ocena podsumowująca i rekomendacja (0,5).

Zbiór P zawiera oceny 323 projektów badawczych dokonane przez ekspertów w ramach jednej z polskich organizacji pozarządowych. Zwycięskie projekty otrzymywały finansowanie w wysokości 28 tys. zł. Zgłoszenia pogrupowane były w 6 dziedzinach. Wszystkie projekty z danej dziedziny były oceniane niezależnie przez tych samych 3 recenzentów, którzy przyznawali im ocenę (na skali 1–5). Kolejny ekspert dokonywał osobnego rankingu, kierując się własną ocena wniosków oraz uzyskanymi przez nie wcześniejszymi recenzjami.

Celem przeprowadzonej ewaluacji zbioru A i P było liczbowe porównanie ocenianych obiektów. W obu przypadkach jeden ewaluator (recenzent) dokonywał ewaluacji wielu projektów lub abstraktów, co pozwala na analizę właściwego dla niego stylu ewaluacji, który omówimy w dalszej części tekstu.

2. Psychologiczny model procesu ewaluacji[1]

Każdy sąd ewaluatywny, będący np. odpowiedzią na pytanie, czy projekt badawczy zasługuje na finansowanie, czy abstrakt zasługuje na publiczną prezentację w trakcie konferencji oraz przyznanie stypendium, wymaga porównania ocenianego obiektu z jakimś standardem lub wzorcem. Zwraca na to uwagę także Wolfe, w swoim modelu oceny esejów (Wolfe, 1997; za: Eckes, 2012).

2.1. Utajone wzorce porównawcze

Jedną z podstawowych zdolności naszego umysłu jest kategoryzowanie doświadczeń, łączenie informacji i odkrywanie zależności, które krystalizują się w postaci wzorców. Wzorce zależą od uprzednich kontaktów z podobnymi obiektami (a więc doświadczenia) i zestawienia informacji (a więc aktywności poznawczej).

Wzorcem używanym do porównań kolejnych obiektów może być:

1) wyabstrahowana ze szczegółów reprezentacja będąca uśrednieniem cech napotkanych wcześniej obiektów,
2) teoretyczna wizja (nieistniejącego w rzeczywistości) egzemplarza,
3) jeden ze spotkanych wcześniej egzemplarzy danej kategorii.

Jeżeli oceniamy naszego szefa to wzorcem #1 będzie uśredniona reprezentacja naszej wiedzy na temat tego, jacy szefowie bywają, wzorcem #2 wizja, jaki powinien być szef, wzorcem #3 może stać się np. nasz poprzedni szef, do którego będziemy porównywać aktualnego.

Proces wykorzystania różnych typów wzorców w ewaluacji wyjaśnimy na przykładzie oceny trudno kwantyfikowalnego dorobku naukowego. W przypadku postępowania o nadanie tytułu profesora, recenzent – a więc profesor z tytułem – musi określić, *jak dobry* jest kandydat do tytułu, trzeba więc określić jakość jego dorobku.

Problem definiowania jakości staje się jasny, gdy przypomnimy opowieść Pirsiga (2005). Wykładowca poprosił studentów, aby uszeregowali cztery eseje według ich jakości. Wykonali oni to zadanie bez najmniejszego problemu. Kiedy zapytano ich, jak należy zdefiniować jakość, studenci zgodnie wskazywali m.in. na takie aspekty, jak: spójność, żywy styl, siła przekonywania, zaangażowanie emocjonalne, utrzymywanie czytelnika w napięciu, ale nie byli w stanie podać jej definicji. Musieli więc przyznać, że chociaż w ich głowach istnieje *porównawczy wzorzec jakości*, do którego odwoływali się, rangując eseje, to nie potrafią go zwerbalizować. Skoro wzorzec nie poddaje się werbalizacji można go nazwać *utajonym*.

Profesor oceniający dorobek kandydata do tytułu staje przed podobnym problemem. Musi dokonać porównania. I choć może sobie tego nie uświadamiać, najczęściej „ma do dyspozycji" trzy wzorce porównawcze: (1) wyabstrahowaną ze szczegółów reprezentację będącą uśrednieniem cech recenzowanych wcześniej „profesur", (2) idealną wizję (teoretycznie możliwą, ale niekoniecznie istniejącą w rzeczywistości) dorobku profesora, (3) dorobek innego profesora lub własny. Reprezentacja poznawcza wzorca ma budowę prototypową, tzn. składa się z prototypu (wzorca) i zakresu dopuszczalnych transformacji. Metaforycznie można powiedzieć, że oceniając podobieństwo dostarczonej dokumentacji do wzorca, dokonuje się mentalnej transformacji jednej reprezentacji w drugą. Choć ostateczna decyzja ma charakter binarny (kandydat na profesora spełnia wymagania lub nie), to jednak recenzenci są w stanie określić „odległość" dorobku kandydata od prototypu kategorii „dorobek profesorski". Jeśli odległość przekracza zakres dopuszczalnych transformacji prototypu (a więc w zbyt dużym stopniu odbiega od wzorca), recenzent odmawia swojego poparcia kandydaturze.

Kategoria „dorobek profesora" może mieć parę prototypów (np. prototyp profesora empirysty – publikującego przede wszystkim artykuły empiryczne, prototyp profesora teoretyka – piszącego głównie monografie; prototyp profesora praktyka – odnoszącego wielkie sukcesy w zastosowaniach).

To, czy dorobek zostanie uznany za wystarczający do przyznania tytułu, nie zależy tylko od jego zawartości, ale też od typu wzorca, z jakim będzie porównywany (wzorzec typowego, idealnego lub konkretnego obiektu). To, który wzorzec zostanie zaktywizowany, zależy od nawykowych wyborów recenzenta (niektórzy zawsze koncentrują się na wzorcu idealnym)

 DOI 10.7172/1644-9584.45.9

i zmiennych czynników sytuacyjnych (np. ostatnio pisana recenzja). Może wystarczyć, aby ktoś przypomniał recenzentowi podobny przypadek oceniany 2 lata wcześniej, i wtedy ślad pamięciowy dotyczący tego kandydata staje się wzorcem porównawczym.

Analogicznie podczas oceny projektu badawczego/abstraktu utajonym poznawczym wzorcem porównawczym dla ewaluatora może być np.:

1) wzorzec #1 – reprezentacja będąca uśrednieniem cech abstraktów/projektów wcześniej ocenianych, powstająca jako odpowiedź na pytanie: „Jak jest najczęściej?";
2) wzorzec #2 – reprezentacja „idealnego" projektu/abstraktu, powstająca jako odpowiedź na pytanie „Jak być powinno?" (taki obiekt jest wytworem myślenia abstrakcyjnego i może nie istnieć w świecie realnym);
3) wzorzec #3 – reprezentacja konkretnego obiektu (ostatnio ocenianego lub takiego, który z jakiś względów „zapadł w pamięć").

Może się zdarzyć, że przystępując do oceny, recenzenci nie mają gotowego wzorca porównawczego – np. po raz pierwszy recenzują projekty badawcze w tym konkursie – wtedy wzorzec zaczyna się wyłaniać w trakcie kolejnych ocen. Co ważne, to pierwsze abstrakty i projekty wywierają większy wpływ na tworzenie się tych prototypów niż pozostałe. W psychologii prawidłowość ta określana jest jako *efekt pierwszeństwa*. Przykładowo, te same błędy umieszczone w pierwszej części wypracowania mają większy wpływ na ocenę, niż gdy były umieszczone w jego dalszej części. Student odpowiadający dobrze na pierwsze pytania, popełniający parę błędów na końcu, jest oceniany wyżej niż ten, który popełnia te same błędy na początku, a potem odpowiada bezbłędnie, mimo że w obu przypadkach liczba i jakość błędów były identyczne (Jones, Rock, Shaver, Goethals i Ward, 1968). Badania nad efektem pierwszeństwa potwierdzają obiegową mądrość, że liczy się… pierwsze wrażenie.

Ostatnio oceniane obiekty (projekty/abstrakty) stanowią kontekst dla kolejnych, mogą więc wpływać na zmianę pierwszego typu wzorca porównawczego („Jak jest najczęściej") i zmieniać oceny. Wyniki dotychczasowych badań sugerują, że surowość ocen pierwszych obiektów w serii jest wyższa niż obiektów ostatnich. Wykazano na przykład, że wyższe noty w Międzynarodowym Konkursie Muzycznym im. Królowej Elżbiety Belgijskiej otrzymywali muzycy występujący w późniejszych dniach (Flores i Ginsburgh, 1996), w konkursie gastronomicznym „Das Perfekte Dinner" częściej wygrywały osoby gotujące w piątki niż poniedziałki (Haigner, Jenewein, Müller i Wakolbinger, 2010), w zawodach łyżwiarzy figurowych uczestnicy, którzy występowali później, otrzymywali lepsze noty zarówno w pierwszej, jak i drugiej rundzie (Bruine de Bruin, 2006). Co ważne, efekt ten występował niezależnie od stopnia znajomości ocenianego obiektu i kryterium oceny (Wedell, Parducci i Lane, 1990; za: Skład i Wieczorkowska, 2001). Jeżeli założymy, że wzorzec (prototyp oceny) wyłania się podczas oceny pierwszych obiektów, to można sądzić, że w pewnym momencie serii dokona się jego *krystalizacja*, co może się przejawiać w zmianie wartości średniej.

W procesie ewaluacji podstawową rolę odgrywa *efekt zakotwiczenia* (*anchoring effect*). W wielu badaniach pokazano, że podana liczba stanowi punkt odniesienia (kotwicę) dla formułowania dalszych ocen. Na przykład uczestników pytano, ile byliby skłonni przeznaczyć na wsparcie programu ochrony ptaków morskich zagrożonych przez tankowce; grupy badanych różniła suma wymienioną w pytaniu „Czy byliby Państwo skłonni przekazać datek w wysokości x dolarów". W grupie 1., w której nie wymieniono sumy, badani deklarowali gotowość wpłacenia 64 dol., w grupie 2., pytanej o 5 dol. średnia z deklaracji wyniosła 20 dol., w grupie 3. pytanej o 400 dol. średnia wzrosła do 143 dol. (Kahneman, 2011). Wprowadzona przez pytanie wartość (5 dol. vs 400 dol.) stała się w umysłach badanych porównawczym wzorcem (kotwicą) i wpłynęła na udzielane odpowiedzi.

W analogiczny sposób zaktywizowany utajony wzorzec porównawczy dla projektu badawczego czy abstraktu staje się kotwicą i wpływa na jego ocenę.

Podsumowując, w procesie oceny – oprócz wyartykułowanych kryteriów oceny dostarczonych przez zamawiającego ewaluację – ważną rolę odgrywają utajone wzorce porównawcze, będące efektem indywidualnego doświadczenia recenzenta z tą kategorią obiektów oraz indywidualna strategia oceny ewaluatora.

2.2. Strategie w formułowaniu sądów ewaluatywnych

Wpływ na ocenę ma także strategia, jaką zastosuje ewaluator. Przykładowo, w czasie głosowania nad poparciem wniosku o nadanie tytułu profesora członkowie Rady Wydziału (ewaluatorzy) mogą użyć jednej z czterech strategii formułowania sądu (Forgas i Vargas, 2005):

1. *Strategii odtwarzania gotowych ocen* – polegającej na wyszukiwaniu w pamięci istniejącej już wcześniej odpowiedzi. Taką gotową oceną jest opinia środowiska o kandydacie, konkluzje w recenzjach.
2. *Strategii przetwarzania zmotywowanego* – stosowanej wtedy, gdy mamy silne preferencje dotyczące oceny, jaką mamy sformułować. Jeżeli lubimy kandydata, to będziemy głosować na „tak", nie zważając na argumenty zawarte w recenzjach.
3. *Strategii heurystycznego (uproszczonego) przetwarzania informacji* – wykorzystywanej wtedy, gdy nie zależy nam bardzo na trafności lub jesteśmy przeciążeni. Oceny heurystyczne opierają się na skojarzeniach – głosujący może decydować impulsywnie, łatwo poddawać się efektowi zakotwiczenia.
4. *Strategii przetwarzania analitycznego (szczegółowego)* – wymagającej wyselekcjonowania i zinterpretowania nowej informacji, a następnie powiązania jej z wiedzą. Jest ona stosowana, gdy głosujący ma wystarczające zdolności/umiejętności, aby zrozumieć np. źródło sprzecznych konkluzji w recenzjach, oraz zasoby poznawcze (nie jest zmęczony) i jest zainteresowany trafnością swojej oceny.

W psychologii panuje zgoda co do istnienia dwóch funkcjonalnie różnych systemów (por. Kahneman, 2011) przetwarzania informacji: (1) szyb-

 DOI 10.7172/1644-9584.45.9

kiego – impulsywnego, działającego w sposób automatyczny i intuicyjny oraz (2) wolnego – refleksyjnego, odpowiadającego za opanowanie automatycznych reakcji systemu pierwszego. System refleksyjny – związany ze strategiami przetwarzania analitycznego – wymaga zasobów psychoenergetycznych (np. Gailliot, Baumeister, DeWall, Maner, Plant, Tice i Schmeichel, 2007; Gailliot i Baumeister, 2007), które w warunkach presji czasowej lub zmęczenia są wyczerpywane. Powoduje to przełączenie przetwarzania na system pierwszy, związany ze strategiami heurystycznego (a więc uproszczonego) przetwarzania informacji, co często prowadzi do powierzchownych ocen. Zniekształcenia z tym związane są „coraz większe w warunkach nacisku czasu i konieczności szybkiego podejmowania decyzji oraz natychmiastowego działania. Powstaje wtedy stan stresu informacyjnego, stanowiącego dodatkowe źródło błędów" (Nosal, 2001, s. 36).

Miło byłoby twierdzić, że ewaluatorzy zawsze stosują strategie przetwarzania analitycznego, ale jeżeli sobie wyobrazimy, że niektórzy z nich dokonują ewaluacji 61 abstraktów w ciągu 1 dnia, to można założyć, że zmęczenie, znużenie może narastać w czasie ewaluacji.

Podsumowując, o wyborze strategii przetwarzania informacji decydują zarówno *możliwości i preferencje poznawcze recenzenta* (np. poziom refleksyjności), jego *stan energetyczny* (przeciążenie, nastrój), *cele* wyznaczające stopień zaangażowania, jak i *cechy obiektu oceny* (stopień znajomości i złożoności) oraz *cechy sytuacji* (presja czasowa, aprobata społeczna, konsekwencje w przypadku popełnienia błędu).

Na ewaluację wpływa także łatwość dostępu do informacji, ich wyrazistość, świeżości i formy zakodowania (Nosal, 2001). Na tworzenie adekwatnej reprezentacji sytuacji problemowej może mieć także wpływ lękowe nastawienie, nadmierna koncentracja na szczegółach, dogmatyczność, podejrzliwość (Wieczorkowska-Siarkiewicz, 1992).

3. Indywidualny styl ewaluacji[2]

Zadaniem osoby oceniającej zbiór obiektów (projektów, abstraktów) jest dokonanie ich kategoryzacji na te, które są warte poparcia, i te, które należy odrzucić (skala oceny dwuwartościowa). Najczęściej ewaluator proszony jest o dokonanie oceny na skalach wielowartościowych. Nawet gdy skala oceny jest precyzyjnie określona (np. od 1 do 5), recenzenci mogą w różny sposób dokonywać transformacji swojej skali w skalę odpowiedzi (Wieczorkowska-Siarkiewicz, 1992). Jeden recenzent może różnicować obiekty tylko na części skali, np. 3–5, inny może używać tylko dwóch ocen, np. 2 i 5. Jest to przejawem specyficznego dla danej osoby *stylu ewaluacji*, który może manifestować poprzez: (1) zbyt surowe (lub łagodne) oceny (Hoyt, 2000), (2) brak różnicowania cząstkowych wymiarów oceny – efekt halo (Landy, Vance, Barnes-Farrell i Steele, 1980) – i/lub ocenianych obiektów. Można go wykryć i oszacować tylko wtedy, gdy dysponujemy

ocenami serii obiektów dokonanymi przez tego samego ewaluatora. Najczęściej jednak mamy do czynienia z pojedynczymi ocenami dokonywanymi przez różnych recenzentów, więc wpływ indywidualnego stylu ewaluatora nie był przedmiotem wielu badań, ze względu na brak odpowiednich danych.

3.1. Styl ewaluatora: poziom łagodności

Wszyscy wiedzą, że egzaminatorzy różnią się poziomem surowości. U jednych trudno zdać egzamin, u drugich dużo łatwiej, mimo że teoretycznie wymagania formalne są takie same. Jeśli ewaluator wykazuje systematyczne odchylenie od średniej w jednym kierunku (lub brak tej inklinacji), to można powiedzieć, że charakteryzuje go indywidualny styl ewaluacji (Baron, 1985; 1986; za: Nosal, 1990). Wynika to z różnic w nawykowo aktywizowanych wzorcach porównawczych. Ci bardziej surowi porównują egzaminowanego do teoretycznego wzorca opisującego „Jak być powinno", ci łagodniejsi zapewne porównują z wzorcem typowym, opisującym „Jak jest". Można to wykryć, licząc średnią z wszystkich ocen wystawionych przez tego samego recenzenta. W badaniach pokazano, że niektórzy ewaluatorzy konsekwentnie oceniali eseje surowiej, podczas gdy inni charakteryzowali się niezmiennie bardziej pozytywną oceną tych samych prac. Jest to określane jako *błąd łagodności/surowości* (Holzbach, 1978). Jeżeli oceniający stosuje strategie przetwarzania zmotywowanego i lubi stawiać dobre oceny, będzie bardziej wyrozumiały w ocenach, jeżeli zaś lubi wytykać błędy, negatywne kategorie są bardziej dostępne i sądy będą surowsze (Feldman, 1981).

3.2. Styl ewaluatora: stopień różnicowania

Tę samą średnią ocenę, np. 3 na pięciostopniowej skali, mogą uzyskać zarówno ewalutorzy, którzy słabo różnicują obiekty, przyznając połowie projektów 4, a połowie 2, jak i ci, którzy każdą z ocen 1, 2, 4, 5 przyznali tyle samo razy. Nie różnią się poziomem łagodności wskaźnikowanym przez średnią, ale różnią się istotnie stopniem różnicowania.

Wtedy, gdy recenzenci oceniają obiekty na wymiarach cząstkowych, tak jak to było w zbiorze abstraktów, stopień skorelowania ocen cząstkowych może być wynikiem opisywanego w psychologii *efektu halo* (nazywanego także *efektem aureoli*), polegającego na tym, że przypisanie obiektowi jednej ważnej pozytywnej lub negatywnej cechy wpływa na skłonność do przypisywania innych cech zgodnych ewaluatywnie (Brzezińska, Brzeziński i Eliasz, 2004). Można więc mówić o funkcjonalnym zróżnicowaniu oceniającego „*differential rater functioning*" (Eckes, 2012).

4. Analizowane dane

W tabeli 1 przedstawiono różnice i podobieństwa między dwoma zbiorami danych opisanymi we wprowadzeniu.

 DOI 10.7172/1644-9584.45.9

Zbiór A. Oceny abstraktów konferencyjnych	Zbiór P. Oceny projektów badawczych
– zawiera oceny 673 abstraktów zgłoszonych na konferencję – wszystkie abstrakty były oceniane niezależnie przez 3 recenzentów – każdy z 33 recenzentów oceniał średnio 61 abstraktów (od 17 do 88) – dla każdego abstraktu recenzenci byli losowani przez system, więc różne abstrakty mogły być oceniane przez różne zestawy recenzentów	– zawiera oceny 323 projektów badawczych – projekty dotyczyły 6 dziedzin badawczych – wszystkie projekty z danej dziedziny były oceniane niezależnie przez tych samych 3 recenzentów – analizowano styl ewaluacji 18 recenzentów

Tab. 1. Analizowane dane. Źródło: opracowanie własne.

5. Wskaźniki stylu ewaluacji

5.1. Poziom łagodności/surowości

W zbiorze projektów przedmiotem analizy był indywidualny styl ewaluacji 18 recenzentów. Dla każdego z nich tendencja centralna wzorca porównawczego była wskaźnikowana przez średnią ocenianych przez niego projektów (liczba projektów dla jednego recenzenta zmieniała się od 16 do 49 w zależności od dziedziny).

W zbiorze abstraktów przedmiotem analizy był indywidualny styl ewaluacji 33 recenzentów. Dla każdego z nich tendencja centralna wzorca porównawczego była wskaźnikowana przez średnią z dokonanych przez niego ocen około 61 abstraktów (liczba abstraktów ocenionych przez jednego recenzenta zmieniała się od 17 do 88).

5.2. Stopień różnicowania

Różnicowanie przez recenzenta ocenianych projektów/abstraktów. W obu zbiorach stopień różnicowania przez recenzenta ocenianych projektów/abstraktów był wskaźnikowany przez odchylenie standardowe wystawionych przez niego ocen, rozrzut ocen (różnica pomiędzy oceną maksymalną a minimalną) oraz współczynnik zmienności. W odróżnieniu od odchylenia standardowego, które określa bezwzględne zróżnicowanie cechy, współczynnik zmienności jest miarą względną, czyli zależną od wielkości średniej arytmetycznej. Współczynniki zmienności obliczono, dzieląc odchylenia standardowe przez średnią ocen dla danego recenzenta.

Różnicowanie przez recenzenta ocen na wymiarach cząstkowych. Przypomnijmy, że w zbiorze A ewaluatorzy oceniali (na skali 0 do 10) abstrakty na 6 wymiarach, z których 4 – tj. (I) wartość teoretyczna lub praktyczna, (II) oryginalność, (III) zgodność z tematem konferencji, (IV) jakość prezentacji – teoretycznie powinny być niezależne. Aby to sprawdzić, dla każdego

z 33 ewaluatorów wykonano analizę czynnikową ocenianych przez niego abstraktów na 4 wymienionych wyżej wymiarach. Procent zmienności wyjaśniony przez pierwszy czynnik był miarą skorelowania w umyśle recenzenta tych teoretycznie niezależnych wymiarów.

5.3. Kolejność oceniania

Wysyłając abstrakt na konferencję, nie tylko nie wiemy, kto będzie go recenzował, ale nie wiemy także, czy będzie on oceniany przez wylosowanego przez system recenzenta jako pierwszy, drugi, a może ostatni. Aby sprawdzić, czy to, którą pozycję w serii ocen miał nasz abstrakt, wpływa na jego ocenę, dokonano agregacji ocen po abstraktach, co oznacza, że każdemu abstraktowi przypisano *średnią kolejność*, w jakiej był oceniany. Jeśli np. był oceniany jako 15. przez recenzenta #1, jako 25. przez recenzenta #2, jako 50. przez recenzenta #3, to średnia kolejność dla tego abstraktu wynosiła 30. W analogiczny sposób obliczono dla każdego abstraktu średnią pozycję z ocen 3 recenzentów.

W zbiorze P kolejność, w jakiej recenzenci oceniali projekty, nie była znana, więc analizy były przeprowadzone wyłącznie na zbiorze A. Prawdą jest, że w zbiorze A oceny poszczególnych abstraktów nie są od siebie niezależne, ponieważ każdy z nich był oceniany przez różny zestaw recenzentów, ale ze względu na losowanie ewaluatorów przez system z 5456 możliwych zestawów, wpływ ewaluatora w tej analizie można zaniedbać.

6. Wyniki analiz

6.1. Prawidłowość 1. Ewaluatorzy różnią się poziomem łagodności i stopniem różnicowania projektów/abstraktów.

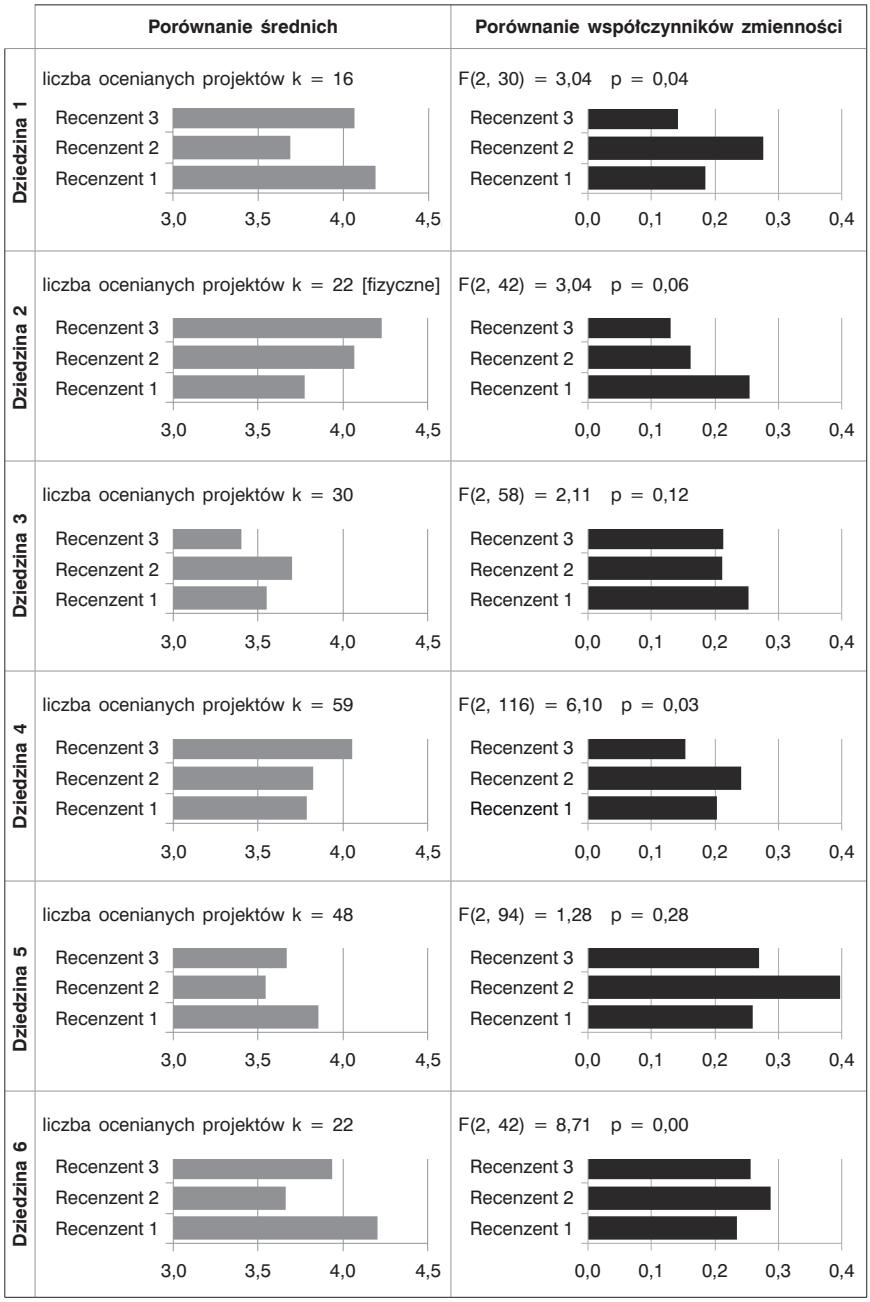
W zbiorze P, ze względu na fakt, że każdy projekt był oceniany przez tę samą trójkę recenzentów, można było policzyć efekt łagodności ewaluatora.

Na rysunku 1 przedstawiono charakterystykę ocen (średnia, współczynnik zróżnicowania wyrażony w proporcjach) trzech recenzentów dla 6 dziedzin.

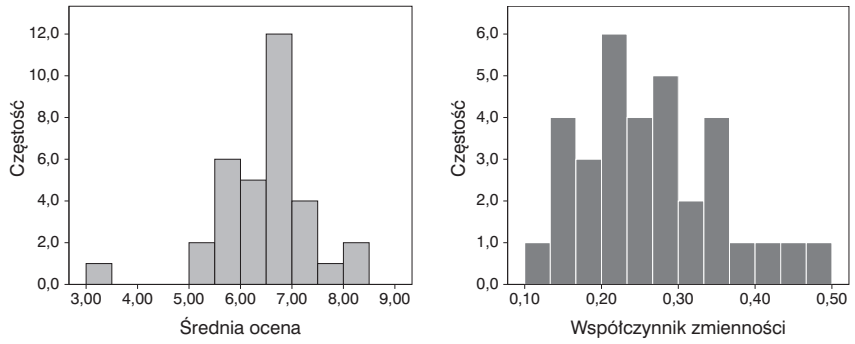
Analiza wariancji wykazała istotny wpływ ewaluatora w 4 z 6 dziedzin. Stopień skorelowania średniej ocen ewaluatorów był różny w różnych zestawach. Analiza korelacji pozwala określić, że np. w dziedzinie 3 jeden z recenzentów oceniał „inaczej" niż dwóch pozostałych (zob. rysunek 1). Wniosek: ewaluatorzy mogą różnić się zarówno surowością (operacjonalizowaną przez średnią), jak i stopniem różnicowania projektów badawczych.

Takich analiz nie można przeprowadzić dla zbioru A, ale na wykresach 2 i 3 widać, że także w tym przypadku recenzenci różnią się zarówno poziomem łagodności (dla 33 recenzentów średnia zmieniała się od 3,44 do 8,40), jak i stopniem różnicowania (współczynnik zmienności od 0,12 do 0,46).

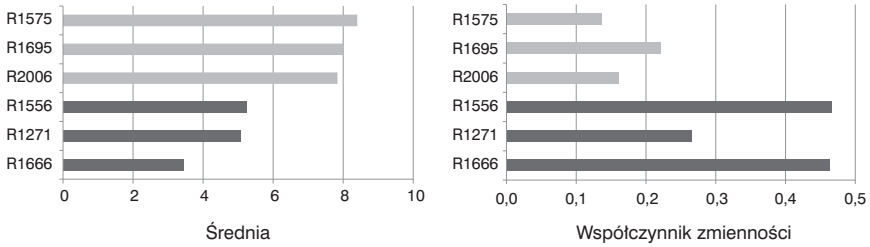
 DOI 10.7172/1644-9584.45.9



Rys. 1. Charakterystyka ocen (średnia, współczynnik zróżnicowania wyrażony w proporcjach) trzech recenzentów dla sześciu dziedzin. Źródło: opracowanie własne.



Rys 2. Charakterystyka ocen (średnia, współczynnik zróżnicowania wyrażony w proporcjach) dla 33 recenzentów. Źródło: opracowanie własne.

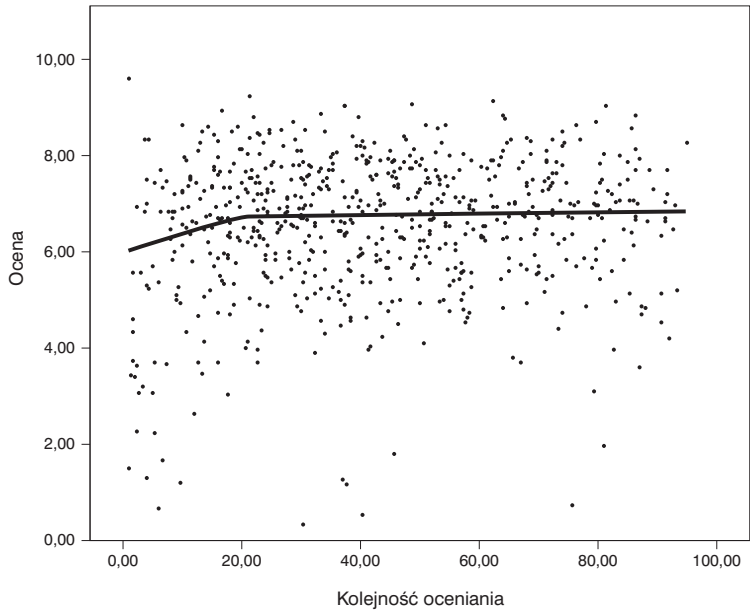


Rys 3. Charakterystyka ocen (średnia, współczynnik zróżnicowania wyrażony w proporcjach) dla trzech najniżej i trzech najwyżej oceniających recenzentów abstraktów. Źródło: opracowanie własne.

6.2. Prawidłowość 2. Ocena abstraktu zależy od kolejności, w której abstrakt był oceniany (zbiór A)

W celu sprawdzenia związku między średnią oceną abstraktu oraz średnią kolejnością, w jakiej był on oceniany, przeprowadzono dla 673 abstraktów analizę regresji, w której wykazano liniowy związek ($a = 6{,}24$; $b = 0{,}007$; beta = 0,118; $r^2 = 0{,}014$, $F(1{,}671) = 9{,}51$, $p < 0{,}002$;) obu zmiennych. Niski współczynnik dopasowania linii prostej r^2 oraz wykres zależności średniej oceny od średniej kolejności dla 673 abstraktów (rysunek 4) pokazuje, iż warto szukać nieliniowego związku pomiędzy kolejnością oceny i średnim poziomem oceny. Próba dopasowania związku krzywoliniowego za pomocą krzywej *inverse* dała 3-krotnie wyższy poziom wyjaśnionej zmienności wysokości ocen. O ile dla początkowych 15–30 ocen model liniowy jest skutecznym narzędziem wyjaśniania związku ocen i kolejności, dla całości danych dopasowanie modelu krzywoliniowego jest wyższe ($r^2 = 0{,}014$ oraz odpowiednio $r^2 = 0{,}048$).

 DOI 10.7172/1644-9584.45.9



Rys 4. Średnia ocena abstraktu (oś OY) w zależności od średniej kolejności oceny (oś OX). Dopasowana krzywa wskazuje na zmianę w średnich wynikach oceny po około 15 ocenianych abstraktach. Źródło: opracowanie własne.

W kolejnym kroku określono punkt, w którym następuje zmiana relacji pomiędzy oceną i kolejnością. Punkt ten wyznaczono poprzez maksymalizację wielkości wyjaśnionej wariancji dla wybranej liczby początkowych ocen. Powyżej tego punktu związek pomiędzy tymi zmiennymi staje się statystycznie nieistotny. Na podstawie wartości r^2 w regresji liniowej oraz krzywoliniowej (*inverse*) określono, iż zmiana następuje po około 15 ocenach.

Liczba początkowych ocen	r^2 w regresji liniowej	b	r^2 w regresji krzywoliniowej	B
10	0,09	0,21*	0,05	–2,20***
15	0,14	0,17***	0,08	–3,30***
20	0,12	0,11***	0,09	–3,35***
25	0,11	0,08***	0,10	–3,64***
Całość (685)	0,014	0,01**	0,05	–3,64***

* $p<0,05$; ** $p<0,01$; *** $p<0,001$.

Tab. 2. Wartości r2 w analizie regresji liniowej oraz krzywoliniowej (krzywa „odwrotna") dla różnych podzbiorów danych – zmienna zależna: łączna ocena. Źródło: opracowanie własne.

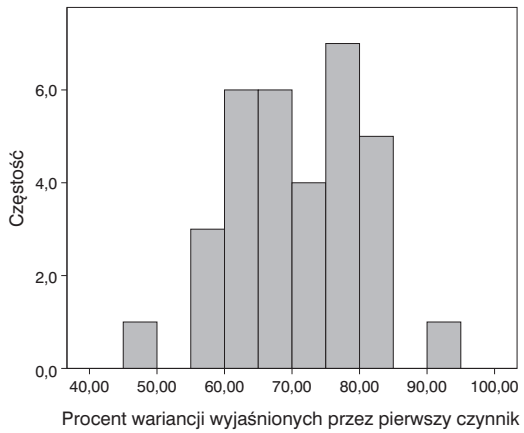
Przypomnijmy, że każdy recenzent oceniał około 61 abstraktów. W następnym kroku dokonano podziału zbioru na 2 grupy: (1) abstrakty, które należały do początku (pierwsze 15) serii ocen recenzentów, i (2) te, które były oceniane później (kolejność wyższa niż 15). Grupy różnią się istotnie średnią (M1 = 5,81, M2 = 6,65) oraz wariancją ocen (SD1 = 0,15 i SD2 = 0,06).

Można więc powiedzieć, że abstrakty oceniane na początku otrzymują niższe oceny aż o 0,84 punktu. Większe jest też zróżnicowanie ocen w grupie ocen początkowych niż w grupie ocen dalszych, co można przewidzieć istotnością liniowego związku między kolejnością a oceną w grupie abstraktów początkowych.

6.3. Prawidłowość 3. Występowanie efektu halo w ocenach recenzentów (zbiór A)

Przypomnijmy, że w celu określenia poziomu skorelowania 4 ocen cząstkowych dla każdego recenzenta przeprowadzono analizę czynnikową ocen średnio 61 abstraktów na poszczególnych wymiarach.

Jeżeli byłyby to niezależne wymiary percepcji, to nie powinny się dać łatwo zredukować do 1 czynnika. Okazało się, że dla wszystkich 33 recenzentów oceny cząstkowe można przekształcić w jeden czynnik, który wyjaśniał od 47 do 91% zmienności ocen (rysunek 5). Ten procent jest *wskaźnikiem różnicowania wymiarów cząstkowych* oceny przez recenzenta. Bardzo wysoki procent wyjaśnionej wariancji przez jeden czynnik oznacza, że recenzent oceniał abstrakty prawie identycznie we wszystkich 4 wymiarach cząstkowych (np. 1, 2, 1, 1 vs 7, 8, 7, 7), a więc słabo różnicował wymiary.



Rys 5. Rozkład wielkości wariancji wyjaśnionej przez pierwszy czynnik w analizie czynnikowej czterech wymiarów cząstkowych oceny. Źródło: opracowanie własne.

 DOI 10.7172/1644-9584.45.9

7. Dyskusja wyników i podsumowanie

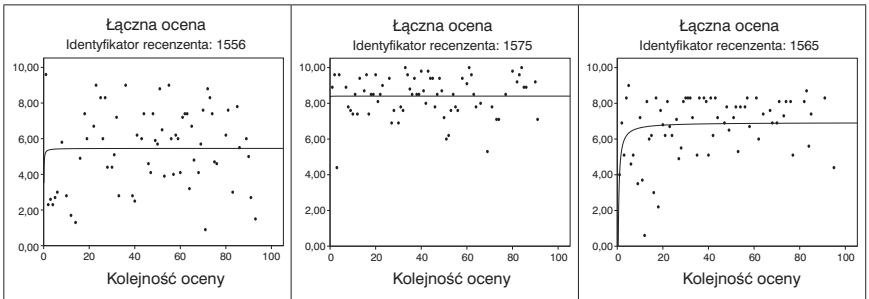
Pokazaliśmy, że recenzenci projektów/abstraktów różnią się indywidualnym stylem ewaluacji, operacjonalizowanym przez stopień łagodności/surowości i stopień różnicowania.

W zbiorze A, w którym znana była kolejność, w jakiej recenzenci oceniali abstrakty, stwierdziliśmy *tendencję do niższego oceniania początkowych kilkunastu prac*. Może to być interpretowane w terminach kształtowania się w umysłach recenzentów w czasie recenzowania *utajonego wzorca porównawczego*. Po ocenie kilkunastu prac wzorzec się już stabilizuje – co w analizowanym zbiorze owocowało wzrostem średniej ocen dla abstraktów ocenianych później. Można spekulować, że recenzenci na początku aktywizują wzorzec #2, będący zapisem ich wyobrażenia teoretycznie idealnego abstraktu (odpowiedź na pytanie: „Jak być powinno?"), pod wpływem czytania kolejnych prac tworzy się wzorzec porównawczy #1, będący odpowiedzią na pytanie „Jak jest?" i oczekiwania się obniżają, a więc oceny rosną. Wpływ kolejności oceniania powinien być najsilniejszy u recenzentów, którzy nie mieli wcześniejszych doświadczeń. Część recenzentów wykonywała tę samą pracę rok wcześniej i warto byłoby w kolejnych analizach spróbować dotrzeć do tego typu danych.

W badaniach (Marsh, Jayasinghe i Bond, 2008) porównywano ewaluacje osób oceniających jeden lub więcej projektów. Ze względu na to, że większość recenzentów oceniała tylko jeden projekt, gdy ocenili go niżej niż pozostali recenzenci, nie wiemy, czy był to efekt słabego projektu i łagodności pozostałych recenzentów, czy też indywidualnego stylu ewaluacji. Stwierdzono, że oceny tych, którzy oceniali co najmniej 3 projekty, były bardziej spójne z ocenami innych recenzentów tego samego projektu i bardziej trafne (zgodne z ocenami finałowymi). Można więc sądzić, że ocena więcej niż 1 projektu pozwalała na stworzenie w umyśle recenzenta wzorca porównawczego. Stwierdzono także, że ich oceny były też średnio bardziej surowe niż tych, którzy oceniali tylko jeden projekt. Badacze podkreślają jednak, że w grupie 15 ewaluatorów, którzy ocenili 10 lub więcej projektów, można wskazać osoby, które konsekwentnie są bardziej łagodne w swoich ocenach niż reszta, co potwierdza wpływ indywidualnego stylu ewaluacji.

Alternatywna do stabilizacji wzorca porównawczego interpretacja podkreśla rolę zmęczenia recenzenta i może zostać zweryfikowana w planowanych badaniach eksperymentalnych, w których różni ewaluatorzy będą oceniali w tej samej kolejności ten sam zestaw abstraktów.

Podsumowując, przeprowadzona przez nas analiza pokazała, że na ocenę abstraktów i projektów badawczych wpływ miały: *indywidualny styl ewaluacji* (poziom łagodności/surowości, skłonność do różnicowania) oraz *kolejność abstraktów* (pierwsze abstrakty oceniane były surowiej). Podsumowującą ilustracją różnic w stylach ewaluacji mogą być przedstawione na rysunku 6 zależności między kolejnością oceny a poziomem łagodności i stopniem różnicowania pochodzące od 3 typów recenzentów ze zbioru A.



6.a. Typ #1: Niska średnia ocena i wysokie różnicowanie abstraktów (M = 5,24; s = 2,44, N = 69).	6.b. Typ #2: Wysoka średnia ocen i małe różnicowanie abstraktów (M = 8,40; s = 1,14, N = 70).	6.c. Typ #3: Istotny wzrost ($p < 0,03$) wysokości oceny wraz z kolejnością oceniania (M = 6,69; s = 1,73, N = 65).

Rys. 6. Przykłady stylów ewaluacji (średnia, zróżnicowanie, zmiana oceny wraz z kolejnością ocenianego abstraktu) w zbiorze A dotyczącym recenzji abstraktów konferencyjnych. Źródło: opracowanie własne.

Czytelnik sam może odpowiedzieć na pytanie, z której kategorii chciałby wylosować recenzenta do oceny jego abstraktu, bo trzeba pamiętać, że w typowych warunkach konkursowych dla każdego projektu dobierani są inni recenzenci. Może się więc zdarzyć, że wylosujemy 3 recenzentów z jednej kategorii.

Indywidualny styl ewaluacji recenzenta nie miał wpływu na decyzje dotyczące finansowania projektów (zbiór P), ponieważ ci sami recenzenci oceniali wszystkie projekty. Porównanie ocen może być utrudnione w najczęściej spotykanym przypadku, gdy część projektów jest oceniana przez bardziej surowego, część przez bardziej łagodnego recenzenta. Tak było w przypadku innego konkursu projektów badawczych (Michałowicz, 2013), w którym dla aż 48 projektów (co stanowi 38%) różnica między oceną maksymalną a minimalną wyniosła powyżej 30 punktów na 100-stopniowej skali. Aż 13 recenzentów wystawiło projektom oceny poniżej 20 punktów (na 100-punktowej skali), 54 recenzentów wystawiło oceny powyżej 80 punktów. W tym wypadku nie można stwierdzić, czy recenzent, który ocenił projekt na 2,5 (jeden przypadek), jest nadmiernie surowy, czy też trafił na bardzo słaby projekt i bardzo łagodnych współrecenzentów. W przypadku gdy ewaluator ocenia tylko jeden projekt, nie można określić jego indywidualnego stylu ewaluacji (a wiec poziomu łagodności, stopnia różnicowania).

Aby zminimalizować wpływ indywidualnego stylu ewaluatorów, podejmuje się różnego typu środki zaradcze (Raymond i Viswervaran, 1991; Raymond i Huston, 1990). Z jednej strony są to programy szkoleniowe dla oceniających, z drugiej korekty statystyczne, np. podejmowanie decy-

 DOI 10.7172/1644-9584.45.9

zji na podstawie wyników standaryzowanych. Niestety pierwsze rozwiązanie wymaga sporo czasu i jest kosztowne, a jego skuteczność również podlega dyskusji (Landy, Vance, Barnes-Farrell i Steel, 1980; Bernardin i Pence, 1980). W przypadku ocen pracowników zaleca się (Landy i in., 1980):

– zwiększenie liczby obserwacji pracownika lub liczby oceniających,
– standaryzowanie ocen wewnątrz grup pracowników (np. sprzedawców, informatyków) i przyznawanie np. premii na podstawie ocen standaryzowanych, a nie surowych, co pozwoli na porównywanie ocen pracowników w różnych działów i na różnych stanowiskach,
– kontrolowanie w analizach zależności wpływu zmiennych potencjalnie zniekształcających relacje, np. staż pracy, wysokość wynagrodzenia

Uważamy, że w przypadku ocen projektów rozwiązaniem mogło by być „kotwiczenie" osób oceniających, a więc ro*zpoczynanie przez nich procesu oceny od kilku obiektów o znanej wartości* (np. 3 najlepsze projekty vs 3 najgorsze projekty badawcze z poprzedniego konkursu)[3]. Powinno to zapewnić ewaluatorom zbliżone wzorce porównawcze i minimalizować wpływ utajonych wzorców porównawczym specyficznych dla każdego recenzenta. Przedstawianie ewaluatorom jedynie ogólnych wytycznych zmusza do samodzielnego przekładania ich na wzorce. Wynik tej transformacji podlega silnemu wpływowi różnic indywidualnych w stylach ewaluacji, które chcielibyśmy wyeliminować. Jeżeli dysponowalibyśmy dla każdego recenzenta jego ocenami projektów/abstraktów o znanej a priori wartości, moglibyśmy policzyć i uwzględniać w ocenach jego styl ewaluacji. Jeśli recenzent X oceniłby 3 bardzo dobre projekty jedynie na 4 na skali 7-stopniowej, to moglibyśmy uznać, że jest on surowym recenzentem i ważyć jego oceny projektów o nieznanych wcześniej wartościach wyżej, niż recenzenta, który wzorcowe obiekty ocenił na 7. Mając na uwadze częstotliwość wykorzystywania narzędzi ewaluacyjnych i ich znaczenie (od oceny ewaluatora zależy np. przyznanie dofinansowania), bardzo ważne są dalsze badania psychologicznych uwarunkowań procesu ewaluacji.

Przypisy

[1] Model jest kompilacją rożnych tez zawartych w publikacjach Wieczorkowskiej (1992; 1998; 2011). Autorzy dziękują prof. G. Wieczorkowskiej za pomoc w opisaniu modelu.

[2] Wpływ indywidualnego stylu ewaluacji na oceny zajęć akademickich jest przedmiotem intensywnych badań B. Michałowicza: Michałowicz, B. (2013). Koncepcja rozprawy doktorskiej pt. „Ankiety ewaluacyjne w szkolnictwie wyższym: wpływ wyboru ewaluatorów". Warszawa: Wydział Zarządzania Uniwersytetu Warszawskiego.

[3] Na przykład podobne rozwiązania stosowane są przy ocenie aplikacji w ramach amerykańskiego programu stypendialnego Graduate Research Fellowships w NSF (GRFP).

Bibliografia

Bernardin, H.J. i Pence, E.C. (1980). Effects of Rater Training. *Journal of Applied Psychology*, *65*, 60–66, http://dx.doi.org/10.1037/0021-9010.65.1.60.

Bruine de Bruin, W. (2006). Save the Last Dance II: Unwanted Serial Position Effects in Figure Skating Judgments. *Acta Psychologica*, *123*, 299–311, http://dx.doi.org/10.1016/j.actpsy.2006.01.009.

Brzezińska, A., Brzeziński, J. i Eliasz, A. (red.). (2004). *Ewaluacja a jakość kształcenia w szkole wyższej*. Warszawa: Wydawnictwo SWPS „Academica".

Ciężka, B. (2005). *Ewaluacja – kwestie ogólne*. Warszawa: Polskie Towarzystwo Ewaluacyjne.

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, *9* (3), 270–292, http://dx.doi.org/10.1080/15434303.2011.649381.

Feldman, J.M. (1981). Beyond Attribution Theory: Cognitive Processes in Performance Appraisal. *Journal or Applied Psychology*, *66* (2), 127–148, http://dx.doi.org/10.1037/0021-9010.66.2.127.

Flôres Jr., R.G., i Ginsburgh, V.A. (1996). The Queen Elisabeth Musical Competition: How Fair Is the Wnal Ranking? *The Statistician*, *45*, 97–104, http://dx.doi.org/10.2307/2348415.

Forgas, J.P. i Vargas, P.T. (2005). Wpływ nastroju na społeczne oceny i rozumowanie. W: M. Lewis i J.M. Haviland-Jones (red.), *Psychologia emocji*. Gdańsk: GWP.

Gailliot, M.T. i Baumeister, R.F. (2007). The Physiology of Willpower: Linking Blood Glucose to Self-Control. *Personality and Social Psychology Review*, *11* (4), 303–327, http://dx.doi.org/10.1177/1088868307303030.

Gailliot, M.T, Baumeister, R.F, DeWall, C., Maner, J., Plant, E., Tice, D. i Schmeichel, B. (2007). Self-Control Relies on Glucose as a Limited Energy Source: Willpower Is More Than a Metaphor. *Journal Of Personality And Social Psychology*, *92* (2), 325–336, http://dx.doi.org/ 10.1037/0022-3514.92.2.325.

Hainger, S., Jenewein, S., Müller, H.-C. i Wakolbinger, F. (2010). The First Shall Be Last: Serial Position Effects in the Case Contestants Evaluate Each Other. *Economics Bulletin*, *30* (4), 3170–3176.

Holzbach, R.L. (1978). Rater Bias in Performance Ratings: Superior, Self-, and Peer Ratings. *Journal of Applied Psychology*, *63* (5), 579–588, http://dx.doi.org/10.2307/2348415.

Hoyt, W.T. (2000). Rater Bias in Psychological Research: When Is It a Problem and What Can We Do about It? *Psychological Methods*, *5* (1), 64–86, http://dx.doi.org/IO.I037//I082-9S9X.5.1.64.

Johnson, J.S. i Lim, G.S. (2009). The Influence of Rater Language Background on Writing Performance Assessment. *Language Testing*, *26*, 485–505, http://dx.doi.org/10.1177/0265532209340186.

Jones, E.E., Rock, L., Shaver, K.G., Goethals, G.R. i Ward, L.M. (1968). Pattern of Performance and Ability Attribution: An Unexpected Primacy Effect. *Journal of Personality and Social Psychology*, *10*, 317–349, http://dx.doi.org/10.1037/h0026818.

Kahneman, D. (2011). *Pułapki myślenia*. Poznań: Wydawnictwo Media Rodzina.

Landy, F.J., Vance, R.J., Barnes-Farrell, J.L. i Steele, J.W. (1980). Statistical Control of Halo Error in Performance Ratings. *Journal of Applied Psychology*, *65* (5), 501–506, http://dx.doi.org/10.1037/0021-9010.75.3.290.

Marsh, H.W., Jayasinghe, U.W. i Bond, N.W. (2008). Improving the Peer-Review Process for Grant Applications. Reliability, Validity, Bias, and Generalizability. *American Psychologist*, *63* (3), 160–168, http://dx.doi.org/10.1037/0003-066X.63.3.160.

Michałowicz, B. (2013). Koncepcja rozprawy doktorskiej pt. Ankiety ewaluacyjne w szkolnictwie wyższym: wpływ wyboru ewaluatorów. Warszawa: Wydział Zarządzania Uniwersytetu Warszawskiego.

Nosal, C. (1990). *Psychologiczne modele umysłu*. Warszawa: PWN.

 DOI 10.7172/1644-9584.45.9

Nosal, C. (2001). *Psychologia myślenia i działania menedżera. Rozwiązywanie problemów. Podejmowanie decyzji. Kreowanie strategii*. Wrocław: Wydawnictwo AKADE.

Pirsig, R.M. (2005). *Zen i sztuka oporządzania motocykla*. Poznań: Dom Wydawniczy Rebis.

Polskie Towarzystwo Ewaluacyjne. (2008). Standardy ewaluacji. Pozyskano z: http://www.ewaluacja.org.pl/download/Standardy_ewaluacji_PTE.pdf (08.12.2013).

Raymond, M.R. i Huston, W.M. (1990). Detecting and Correcting for Rater Effects in Performance Assessment. *Act Research Report Series*, (December), 90–14.

Raymond, M.R. i Viswesvaran, C. (1991). Least Squares Models to Correct for Rater Effects in Performance Assessment. *Journal of Educational Measurement*, *30* (3), 253–268, http://dx.doi.org/10.1111/j.1745-3984.1993.tb00426.x.

Skład, M. i Wieczorkowska, G. (2001). Sztuka układania ankiet ewaluacyjnych. W: *Psychologia społeczna: Jednostka – społeczeństwo – państwo* (s. 250–266). Gdańsk: GWP.

Wieczorkowska-Nejtardt, G. (1998). *Inteligencja motywacyjna: mądre strategie wyboru celu i sposobu działania*. Warszawa: WISS.

Wieczorkowska-Siarkiewicz, G. (1992). *Punktowe i przedziałowe reprezentacje celu. Uwarunkowania i konsekwencje*. Warszawa: Oficyna Wydawnicza Wydziału Psychologii Uniwersytetu Warszawskiego.

Wieczorkowska-Wierzbińska, G. (2011). *Psychologiczne ograniczenia*. Warszawa: Wydawnictwo Naukowe Wydziału Zarządzania UW.