WYKORZYSTYWANIE PROCEDURY SĘDZIÓW KOMPETENTNYCH W NAUKACH SPOŁECZNYCH I MOŻLIWOŚCI JEJ OCENY PSYCHOMETRYCZNEJ ZA POMOCĄ NARZĘDZI DOSTĘPNYCH W STATISTICA

Oleg Gorbaniuk, Uniwersytet Zielonogórski i Katolicki Uniwersytet Lubelski Jana Pawła II

Specyfika psychometryczna badań jakościowych

Badania jakościowe mają bardzo duże znaczenie w naukach społecznych, przede wszystkich z uwagi na ich dużą wartość ekologiczną i możliwość odkrycia zjawisk wykraczających poza dotychczasowe schematy i modele teoretyczne. Mogą one stanowić wartość samą w sobie lub być etapem w procesie badań jako faza wstępna, równoległa lub uzupełniająca badania ilościowe.

Analiza zebranego materiału jakościowego jest jednak podatna na tendencyjność osób dokonujących ich interpretacji lub oceny. Odchylenia mogą dotyczyć wyboru kryteriów oceny materiału lub stosowania tych kryteriów do oceny zebranych danych jakościowych. W zasadzie odchylenia są nieuniknione i wkalkulowane w ryzyko badań jakościowych. Zadaniem badacza jest jednak minimalizacja skali tych odchyleń w taki sposób, aby zapewnić powtarzalność wynikom realizowanych badań. Aby to osiągnąć, należy zadbać o trafność i rzetelność analizy danych jakościowych.

Jednym ze sposobów obiektywizacji ocen danych zgromadzonych w toku badań jakościowych jest zastosowanie metody sędziów kompetentnych, kiedy grupa osób w założeniu kompetentnych w danej dziedzinie ocenia materiał, a następnie oceny są uśredniane

i wykorzystywane w dalszych analizach statystycznych. Aby metoda sędziów kompetentnych była w pełni poprawnie zastosowana, należy przeprowadzić weryfikację kompetencji sędziów, ich sumienności w procesie analiz oraz ocenić jednomyślność ich opinii. Oznacza to, że sędziowie powinni być ocenieni w aspekcie trafności i rzetelności i takiej informacji należy oczekiwać w każdym sprawozdaniu z wyników badań jakościowych. Niestety w praktyce badacze z reguły poprzestają na podaniu ogólnikowej informacji o zastosowaniu metody sędziów kompetentnych bez jakiejkolwiek informacji na temat walidacji psychometrycznej ich decyzji klasyfikacyjnych. Z kolei tam, gdzie analiza psychometryczna ma miejsce, względnie częściej jest raportowana spójność decyzji sędziów, natomiast informacja na temat trafności należy do rzadkości.

Trafność decyzji sędziów kompetentnych i przygotowanie do roli sędziów

Trafność opinii sędziów – o ile w ogóle jest sygnalizowana – jest oceniana najczęściej w aspekcie zgodności kodowania materiału z przyjętym wzorcem poprawnych decyzji, uzyskanym we wcześniejszych badaniach lub na podstawie wyniku badań ekspertów, które pełnią funkcję zewnętrznego kryterium. Tak ujmowany wskaźnik dokładności jest odpowiednikiem trafności kryterialnej i jest obliczany osobno dla każdego z sędziów w celu oceny stopnia jego kompetencji, aby móc dopuścić go do analizy danych. Wartość oceny trafności kryterialnej zależy przede wszystkim od jakości wybranego kryterium zewnętrznego.

W zasadzie żadna osoba nie może być dopuszczona do oceny danych jakościowych bez odpowiedniego przygotowania, którego czas i zakres zależy z jednej strony od dotychczas posiadanych kompetencji przez badacza, a z drugiej od specyfiki zadania, które musi on realizować. Przygotowanie osób do roli sędziów kompetentnych powinno obejmować zdobycie i weryfikację wiedzy deklaratywnej („wiem, że”) i wiedzy proceduralnej („wiem, jak”).

W pierwszej kolejności badacz powinien przygotować trafne kategorie i kryteria klasyfikacji oraz ich szczegółowy opis wraz z definicją kluczowych terminów. Ocenę trafności klucza kodowego w skalach nominalnych można przeprowadzić na podstawie próbnych analiz zgromadzonego materiału jakościowego w aspekcie: (a) częstości wykorzystania kategorii (czy nie ma pustych?) oraz (b) częstości klasyfikowania do kategorii „inne" (czy lista kategorii jest wyczerpująca?). W przypadku pojawienia się pustych kategorii lub zbyt dużej frekwencji „inne" należy dokonać zmiany lub modyfikacji przygotowanych kategorii.

Następnie należy zadbać o teoretyczne przygotowanie przyszłych sędziów oraz weryfikację ich wiedzy teoretycznej. W zależności od poziomu przygotowania wyjściowego sędziów ta faza może ograniczyć się do przedstawienia definicji kluczowych pojęć lub obejmować bardziej szczegółową edukację. W kolejnym etapie należy przygotować klarowną i wyczerpującą instrukcję dla sędziego i zweryfikować jej rozumienie. Następnie należy wyjaśnić sposób rozumienia poszczególnych kryteriów i kategorii klasyfikacyjnych przyszłym sędziom oraz dokonać weryfikacji rozumienia tych kategorii. W zależności od złożoności zadania można przeprowadzić jedno lub więcej cykli szkoleń praktycznych na bazie podobnego materiału w celu opanowania stosowania przygotowanych kategorii do ocenianego materiału. Zaletą szkoleń praktycznych w grupie sędziów jest wymiana opinii pomiędzy sędziami, w ten sposób problemy z rozumieniem niezgłaszane przez jednych sędziów mogą być zgłoszone przez innych. Te problemy koniecznie jednak powinny być rozwiązane przez badacza w trakcie spotkań grupowych, inaczej może dojść do wyrobienia przez grupę błędnej interpretacji sposobu kodowania niezgodnego z kluczem, co oznacza wzrost spójności opinii grupy ze szkodą dla trafności. Po fazie grupowej należy przejść do fazy indywidualnego wykonania próbnych klasyfikacji na podstawie wcześniej przygotowanych materiałów, weryfikacji poprawności tych klasyfikacji poprzez obliczenie wskaźników zgodności z kluczem i indywidualne omówienie znaczących odstępstw w celu wyjaśnienia ich przyczyn. Po kilku cyklach treningowych można zweryfikować trafność kryterialną decyzji klasyfikacyjnych oraz ich stabilność na podstawie wcześniej przygotowanego materiału, które mogą być raportowane w sprawozdaniu z badań. Warto

zaznaczyć, że trafność ocen sędziów zależy nie tylko od procesu selekcji kandydatów, lecz przede wszystkim od stopnia, w jakim badacz zadbał o przygotowanie osób do roli sędziego kompetentnego. Nawet osoba z szerokimi kompetencjami w danej dziedzinie w przypadku niejasnej instrukcji lub braku precyzyjnych definicji kategorii ocen będzie podejmowała błędne decyzje klasyfikacyjne. Wiele cennych rekomendacji, jak należy zorganizować procedurę sędziowania danych jakościowych, można znaleźć u Krejtza i Krejtz (2009a).

Ocena rzetelności decyzji sędziów kompetentnych i organizacja procesu sędziowania

Rzetelność pracy sędziów należy oceniać co najmniej w aspekcie stabilności i zgodności ocen. Ocena rzetelności w aspekcie stabilności polega na co najmniej dwukrotnej ocenie tego samego materiału przez tego samego badacza. Wysoka powtarzalność ocen świadczy o dużej stabilności. Ten wskaźnik jest obliczany dla każdego z sędziów z osobna i jest miarą konsekwencji danego sędziego w ocenie zgromadzonego w badaniach jakościowych materiału. Warto zauważyć, że w miarę zdobywania doświadczenia sposób interpretacji kategorii ocen przez sędziego może się zmieniać, zarówno w stronę poprawną, jak również niepoprawną, w świetle założeń klucza kodowego. Stąd ważny jest wystarczająco długi etap szkoleń, zakończony weryfikacją stabilności ocen sędziego, ponieważ w początkowym etapie niestabilność ocen może wynikać z ewolucji rozumienia kategorii i nabierania wprawy w ich stosowaniu.

Rzetelność w aspekcie spójności polega na ocenie zgodności ocen tego samego materiału przez grupę sędziów. Ten wskaźnik jest obliczany dla grupy sędziów. Bez osiągnięcia satysfakcjonującego stopnia zgodności nie można przeprowadzać agregacji ocen sędziów. Ponadto należy odpowiedzieć na pytanie, czy wszyscy sędziowie w jednakowym stopniu przyłożyli się do wykonywania zadań klasyfikacyjnych i czy nie należy któregoś z nich usunąć z uwagi na niską wiarygodność jego ocen. Zabieg „oczyszczania” grupy sędziów z sędziów nierzetelnych jest jednak zbyt rzadko stosowany, w przeciwieństwie do

usuwania pozycji skal obniżających ich zgodność wewnętrzną, co wynika najczęściej z nieświadomości możliwości zaistnienia tego problemu i nieznajomości narzędzi służących do jego wykrycia.

W przypadku dłuższego materiału do kodowania, wymagającego wielu godzin, a nawet miesięcy pracy, może pojawić się problem niesystematyczności pracy sędziego i związany z nim problem utraty wprawy, reinterpretacji kategorii ocen, zmęczenie i pośpiech w wykonywaniu zadań w końcowej fazie kodowania, a więc też powierzchowność ocen. Jednym ze sposobów sprawdzenia, czy takie zjawisko ma miejsce, jest podzielenie całego zakodowanego materiału przez sędziów kompetentnych na części i sprawdzenie stopnia zgodności ocen sędziów w ramach tych części, a następnie określenie stopnia, w jakim opinia każdego z sędziów z osobna odbiega od opinii pozostałych sędziów (analogia do *item-to-total correlation*). Pojawienie się dużych wahań może świadczyć o braku konsekwencji danego sędziego w ocenie danych lub też błędów technicznych popełnionych przez niego w którymś etapie kodowania (np. przesunięcie danych, przypadkowe lub niepotrzebne intencjonalne przesortowanie danych itd.). W tym ostatnim przypadku, aby zachować wyniki pracy sędziego dla agregacji, należałoby ponownie ocenić wadliwą część klasyfikacji lub przywrócić zaburzoną kolejność danych, rozpoznając jej przyczynę. Częściowa ocena zgodności pozwala także sprawdzić, czy zgodność (spójność) sędziów jako grupy przez cały materiał utrzymywała się na podobnym poziomie. W ten sposób można uzyskać dodatkowy wskaźnik wiarygodności pracy sędziów.

W przypadku, gdy klasyfikacja danych jakościowych obejmuje wiele tygodni, dobrym rozwiązaniem jest wyznaczenie z góry harmonogramu wysyłania kolejnych porcji poklasyfikowanego materiału. W ten sposób: (a) minimalizuje się problem utraty kompetencji wskutek dużych przerw (zapominanie), (b) minimalizuje się problem kumulacji wykonywania zadań przed końcowym terminem, (c) zmniejsza się prawdopodobieństwo utraty całości pracy sędziego wskutek nieprzewidzianych okoliczności (np. uszkodzenie komputera lub dysku z danymi).

Przygotowanie danych do walidacji psychometrycznej

Aby móc przeprowadzić ilościową ocenę trafności i rzetelności decyzji sędziów kompetentnych, należy spełnić wiele warunków. Obiekty/jednostki stanowiące przedmiot oceny powinny być niezależne od siebie (np. różne osoby, z których każda jest poddawane diagnozie klinicznej). Kategorie w kluczu kodowym powinny być dobrze zdefiniowane, rozłączne i wyczerpujące. Sędziowie powinni podejmować decyzje klasyfikacyjne niezależnie od siebie. Niezależność opinii sędziów można ocenić na poszczególnych etapach lub na koniec sędziowania, korelując opinie sędziów między sobą. Zbyt wysokie korelacje wyraźnie odstające od pozostałych par sędziów mogą wskazywać np. na wspólne sędziowanie, okresowe uzgadnianie opinii lub ściąganie. Wzajemna znajomość sędziów, trudność zadania, jego czasochłonność i warunki mogą zwiększać prawdopodobieństwo odstępstwa od warunku niezależności ocen sędziów.

W większości wskaźników zgodności sędziów kodujemy jako kolejne zmienne (czyli jedna kolumna = jeden sędzia). Przedmiot oceny sędziego kodujemy jako kolejne wiersze, przy czym w jednym wierszu umieszczone są oceny różnych sędziów, ale tego samego obiektu oceny. Tylko w ten sposób możliwe jest skonfrontowanie decyzji różnych sędziów odnośnie tego samego obiektu. Jeżeli ten sam obiekt przez tego samego sędziego musi być oceniany na większej liczbie wymiarów/kryteriów, to zmiennych (kolumn w pliku danych) związanych z danym sędzią będzie tyle, ile jest wymiarów/kryteriów, oraz dla każdego z tych wymiarów trzeba niezależnie policzyć wskaźnik zgodności.

Wskaźniki zgodności

Pomiar decyzji sędziów kompetentnych odbywa się na skali porządkowej lub nominalnej. Przedmiotem pomiaru jest identyczność podjętych decyzji, która z kolei zależy wprost od stopnia: (a) podobieństwa rozumienia znaczenia ocenianego materiału, (b) podobieństwa rozumienia kategorii klasyfikacyjnych (klucza kodowego) oraz (c) podobieństwa kodowania. Jako miary zgodności opinii sędziów czasami wykorzystuje się współczynniki

korelacji oparte na statystyce chi^2 lub współczynniki korelacji porządkowej. Nie jest to jednak postępowanie poprawne, ponieważ współczynniki korelacji, operując wystandaryzowanymi danymi, mówią nam o ich współzmienności, a nie o stopniu podobieństwa porównywanych par wartości. Współczynniki korelacji informują nas o relatywnym podobieństwie uporządkowania wartości, a nie stopniu ich identyczności, ponieważ oceniają względne podobieństwo pomiędzy profilami wyników sędziów, nie uwzględniają natomiast absolutnej różnicy pomiędzy profilami. O ile silnie ujemna korelacja świadczy o niskim stopniu zgodności pomiędzy sędziami, o tyle silnie dodatnia korelacja nie pozwala wnioskować o podobieństwie/identyczności decyzji klasyfikacyjnych sędziów.

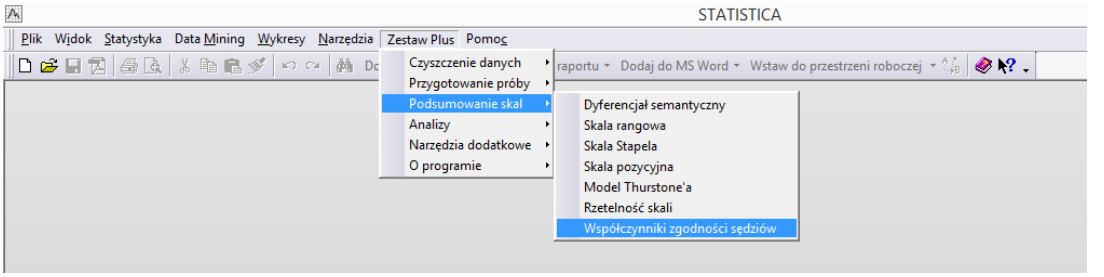
Najprostszym, a zarazem najmniej doskonałym wskaźnikiem zgodności opinii sędziów, jest odsetek zgodnych kategoryzacji, obliczany jako liczba identycznych decyzji klasyfikacyjnych na tle wszystkich możliwych decyzji. Ten wskaźnik nie uwzględnia prawdopodobieństwa przypadkowych zbieżnych decyzji sędziów, które wzrasta wraz ze zmniejszeniem się liczby kategorii, osiągając maksimum przy klasyfikacji typu Tak/Nie.

Najbardziej odpowiednią miarą podobieństwa/zgodności opinii sędziów z kryterium zewnętrznym (np. współczynnik trafności kryterialnej) lub w parach między sobą (np. współczynnik zgodności sędziów lub współczynnik stabilności opinii pojedynczego sędziego) są współczynniki zgodności uwzględniające prawdopodobieństwo przypadkowych zbieżności decyzji sędziów. Wybór właściwego współczynnika zależy od skali pomiarowej decyzji sędziego, liczby sędziów oraz od akceptowanego przez badacza stopnia liberalizmu vs. konserwatyzmu miary zgodności (zob. tabela 1).

Tabela 1. Wybór współczynnika zgodności.

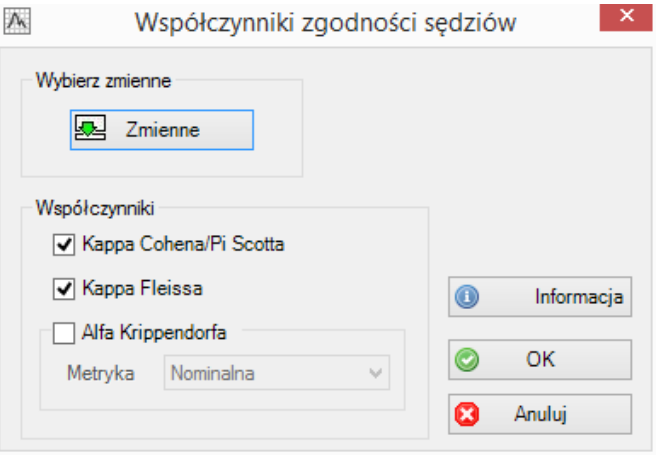
Skala pomiarowa	Liczba sędziów	
	2	więcej niż 2
nominalna	κ Cohena π Scotta	π Scotta κ Fleissa
porządkowa	α Krippendorfa	α Krippendorfa W Kendalla

W programie Statistica wyboru właściwego współczynnika korelacji dokonujemy wybierając opcję Współczynniki zgodności w ramach *Zestawu Plus* (zob. rys. 1).



Rys. 1. Narzędzia dostępne w *Zestawie Plus*.

W oknie dialogowym należy wybrać właściwy współczynnik zgodności, definiując jednocześnie liczbę sędziów i skalę pomiaru decyzji klasyfikacyjnych (zob. rys. 2).



Rys. 2. Okno definiowania liczby sędziów, skali pomiaru i współczynnika zgodności.

Wartości współczynników świadczące o różnym stopniu zgodności porównywanych opinii sędziów wahają się w granicach od 0 do 1, gdzie 0 oznacza zgodność na poziomie przypadku, natomiast 1,0 oznacza identyczność decyzji sędziów. Problemem w interpretacji współczynników zgodności jest jednak brak wyrobionej i powszechnie akceptowanej opinii w środowisku badaczy na temat tego, jak oceniać uzyskane współczynniki zgodności. Sugestie różnych autorów odnośnie interpretacji współczynników zgodności można

próbować uogólnić w postaci następujących przedziałów i odpowiadających im ocen (zob. tabela 2):

Tabela 2. Ocena stopnia zgodności sędziów na podstawie wartości współczynnika.

współczynnik	ocena zgodności
> 0,80	bardzo dobra
0,60-0,80	dobra
0,40-0,59	zadowalająca
< 0,40	niewystarczająca

Niżej zostanie omówiony pokrótce każdy z najczęściej wykorzystywanych w badaniach naukowych współczynników zgodności wraz z opisem sposobu ich obliczenia za pomocą programu Statistica oferującego najpełniejszy dostęp do różnych współczynników zgodności w ramach *Zestawu Plus*. Pominięte zostanie natomiast omówienie wzorów, ponieważ są one szczegółowo omówione w polskojęzycznej literaturze (np. Krejtz i Krejtz, 2009b), gdzie zainteresowany czytelnik może się z nimi zapoznać.

π Scotta i κ Fleissa

Współczynniki π Scotta (Scott, 1955) i κ Fleissa (Fleiss, 1971) mogą być interpretowane jako odsetek zgodnych kategoryzacji między sędziami skorygowany o prawdopodobieństwo przypadkowej zgodności. Mają one zastosowanie w sytuacji, kiedy pomiar decyzji klasyfikacyjnej jest dokonywany na skali nominalnej, a kategorii jest dwie lub więcej. Pierwotnie współczynnik π Scotta został opracowany na potrzeby pomiaru zgodności opinii dwóch sędziów. Natomiast w *Zestawie Plus* programu Statistica ten współczynnik jest dostępny w wersji z poprawką, umożliwiając pomiar zgodności trzech lub większej liczby sędziów pod nazwą κ Fleissa. Wartości współczynnika π Scotta i κ Fleissa wahają się w granicach od -1 (całkowita niezgodność) do +1 (całkowita zgodność), a 0 – oznacza zgodność na poziomie przypadku. Możliwy do uzyskania zakres wartości zależy od liczby

sędziów (większa liczba sędziów powoduje zawyżenie wskaźnika zgodności) oraz liczby kategorii w ramach decyzji klasyfikacyjnej. Wadą współczynnika π Scotta jest jego konserwatywność, co oznacza, że bardzo trudno uzyskać maksymalny wskaźnik zgodności. Ponadto ten współczynnik nie uwzględnia różnic w rozkładzie wykorzystywania poszczególnych kategorii decyzyjnych przez różnych sędziów. Jeżeli różnice są duże, lepszym jest współczynnik kappa Cohena.

Przykładem zastosowania współczynnika π Scotta do oceny trafności klasyfikacji mogą badania psycholeksykalne języków naturalnych. W ramach tych badań przymiotniki opisujące ludzkie cechy wyselekcjonowane ze słownika (np. w języku polskim jest ok. 6-7 tys. przymiotników-deskryptorów osobowych) są klasyfikowane do 12 kategorii na skali dychotomicznej typu Tak/Nie (Angleitner, Ostendorf i John, 1990), przy czym ten sam przymiotnik może jednocześnie być sklasyfikowany do kilku kategorii, np. przymiotnik *spokojny* może jednocześnie opisywać cechy temperamentu i stan emocjonalny. Każdy sędzia po kilkumiesięcznym przeszkoleniu teoretycznym i praktycznym dostał zestaw 240 słów, które we wcześniejszych badaniach zostały jednomyślnie przez sędziów przypisane do odpowiednich kategorii, a ich decyzje zostały skonfrontowane z wynikami badań leksykalnych innych języków. W ten sposób badacz dysponował listą kryterialną, z którą można było porównać decyzje klasyfikacyjne każdego sędziego w ramach każdego kryterium klasyfikacji.

Tabela 3. Współczynniki zgodności sędziów z kryterium zewnętrznym π Scotta.

Kryterium klasyfikacji	Sędzia	
	S1	S2
Temperament i charakter	0,69	0,83
Zdolności	0,82	0,85
Stan emocjonalny	0,93	0,79
Role i relacje	0,93	0,43
Reakcje społeczne	0,89	0,89
Anatomia i morfologia	0,87	0,86

Tabela 3 przedstawia wyniki oceny trafności klasyfikacji dwóch sędziów, przy czym współczynniki π Scotta zostały obliczone dla każdej kategorii i dla każdego sędziego osobno. Porównując uzyskane współczynniki zgodności z tabelą 2, możemy stwierdzić, że trafność klasyfikacji jest dobra (0,60-0,80 lub bardzo dobra (0,81-1,00) lub wręcz doskonała dla sędziego S1 w kategorii *Stan emocjonalny* (π = 0,93) i *Reakcje społeczne* (π = 0,93). Z kolei dla sędziego S2 w kategorii *Role i relacje* trafność klasyfikacji była niska (π = 0,43). Gdyby okazało się, że wszyscy sędziowie mają niski wskaźnik trafności przy jednocześnie wysokich wskaźnikach według innych kryteriów, oznaczałoby to niewłaściwe rozumienie definicji kryterium lub obiektywną trudność rozpoznania przynależności słowa do danej kategorii. Jeżeli dla któregoś sędziego wszystkie wskaźniki trafności byłyby wyraźnie niższe niż u pozostałych sędziów, to mogłoby wskazywać na niskie kompetencje sędziego i jego nieprzydatność do wykonania zadania. Z kolei jeżeli niska trafność dotyczy jednej kategorii u pojedynczego sędziego, oznacza to konieczność ponownego przeszkolenia sędziego w zakresie stosowania danego kryterium do analizy danych jakościowych.

κ Cohena

Współczynnik zgodności sędziów κ Cohena (Cohen, 1960) ma zastosowanie wtedy, kiedy sędziów jest tylko dwóch, a decyzja klasyfikacyjna jest mierzona na skali nominalnej i przyjmuje dwie lub więcej wartości (kategorii). Podobnie jak poprzedni współczynnik zgodności κ Cohena uwzględnia prawdopodobieństwo przypadkowej zbieżności, a zakres wahań wartości wynosi od -1 (całkowita niezgodność) do +1 (doskonała zgodność), a 0 oznacza zgodność na poziomie przypadku. Cechą specyficzną tego współczynnika jest uwzględnianie rozkładu częstotliwości wykorzystywania przez poszczególnych sędziów różnych kategorii klasyfikacyjnych.

Tabela 4. Współczynniki zgodności dwóch sędziów κ Cohena.

Kryterium klasyfikacji	κ Cohena
Temperament i charakter	0,62
Zdolności	0,73
Stan emocjonalny	0,80
Role i relacje	0,46
Reakcje społeczne	0,85
Anatomia i morfologia	0,78

Przykładem zastosowania współczynnika κ Cohena może być ocena zgodności opinii dwóch sędziów, którzy poklasyfikowali listę 480 przymiotników osobowych. Ocenę zgodności przeprowadzono osobno dla każdej kategorii, stanowiącej kryterium oceny. Przedstawione w tabeli 4 współczynniki wskazują na różną zgodność sędziów w zależności od kryterium oceny. W przypadku klasyfikacji słów do kategorii *Temperament i charakter, Zdolności* oraz *Anatomia* i *morfologia* zgodność można ocenić jako dobrą, w przypadku kategorii *Stan emocjonalny* i *Reakcje społeczne* – jako bardzo dobrą, natomiast w przypadku *Role i relacje* – jako bliską niezadowalającej (κ = 0,46).

α Krippendorfa

Współczynnik α Krippendorfa (Krippendorf, 2012) jest najbardziej uniwersalnym współczynnikiem zgodności. Ten współczynnik nie ma ograniczeń ani co do skali pomiarowej decyzji klasyfikacyjnej (nominalna, porządkowa, ilościowa), ani co do liczby wartości/kategorii w ramach skali pomiarowej, ani co liczby sędziów kompetentnych (dwóch lub więcej), ani co do minimalnej liczby ocenianych obiektów. Poza tym jest on odporny na braki danych. Zakres wartości waha się, podobnie jak w dotychczas omówionych współczynnikach, w granicach od -1 do +1 o podobnej jak poprzednio interpretacji.

Tabela 5. Współczynniki zgodności siedmiu sędziów α Krippendorfa.

	α Krippendorfa
Temperament i charakter	0,68
Zdolności	0,59
Stan emocjonalny	0,83
Role i relacje	0,72
Reakcje społeczne	0,86
Anatomia i morfologia	0,76

W taksonomii psycholeksykalnej bierze udział z reguły od 5 do 10 sędziów, co zwiększa rzetelność zagregowanej opinii sędziów pod warunkiem osiągnięcia akceptowalnej zgodności wewnętrznej. Tabela 5 przedstawia wyniki analizy zgodności opinii 7 sędziów, którzy oceniali ten sam materiał składający się z 480 słów. Ponieważ sędziowie oceniali każde słowo z uwagi na 6 kryteriów, współczynnik zgodności α Krippendorfa został obliczony osobno w ramach każdej kategorii. Wyniki analiz wskazują, że sędziowie osiągnęli bardzo dobrą zgodność w zakresie dwóch kategorii: *Stan emocjonalny* i *Reakcje społeczne*. Dla kategorii *Temperament i charakter, Role i relacje* oraz *Anatomia i morfologia* zgodność była dobra. Natomiast w kategorii *Zdolności* ($\alpha = 0{,}59$) była ona przeciętna. Ogólnie w odniesieniu do analizowanych kategorii zgodność sędziów można uznać za satysfakcjonującą i można polegać na zagregowanej opinii. W przypadku pomiaru na skali typu Tak/Nie uznaje się, że oceniana jednostka należy do kategorii, jeżeli została do niej zakwalifikowana przez większość sędziów, czyli w analizowanym przypadku przez co najmniej 4 spośród 7 sędziów.

W Kendalla

W przypadku kiedy zadanie sędziów polega na porządkowaniu materiału poprzez nadanie kolejnym obiektom rang/hierarchii według określonego kryterium, idealną miarą zgodności

sędziów jest współczynnik W Kendalla (Ferguson i Takane, 2004). Całkowita zgodność sędziów oznaczałaby w tym przypadku przypisanie danemu obiektowi takiej samej rangi przez wszystkich sędziów. Na przykład selekcja kandydatów na stanowisko pracy przez grupy sędziów jest dobrą okazją do weryfikacji zgodności opinii komisji rekrutacyjnej za pomocą współczynnika W Kendalla (pytanie, jak często taka weryfikacja ma miejsce i jakie wnioski są wyciągane w przypadku braku takiej zgodności odnośnie kompetencji komisji?). Ogólnie należy wskazać na dwa kluczowe warunki, które powinny być spełnione, aby móc zastosować współczynnik W Kendalla: (a) sędziów powinno być trzech lub więcej, (b) pomiar decyzji sędziów dokonywany jest na skali porządkowej. Wartości współczynnika wahają się w przedziale od 0 (zupełny brak zgodności) do 1 (całkowita zgodność).

Aby poprawnie obliczyć współczynnik W Kendalla, dane należy przygotować w innym układzie niż wcześniej omówione współczynniki (zob. tabela 6): sędziowie powinni być w wierszach (wszystkie dane uzyskane od tego samego sędziego powinny znajdować się w tym samym wierszu), natomiast obiekty oceny umieszczamy w kolumnach/zmiennych (w jednej kolumnie/zmiennej znajdują się oceny różnych sędziów dotyczących tego samego obiektu). Często spotykanym w praktyce błędem jest odwrotna orientacja danych i wskutek tego uzyskiwanie błędnych wartości współczynników W Kendalla.

Jeśli jest więcej niż jeden wymiar oceny tego samego obiektu, współczynnik W Kendalla należy obliczyć osobno dla każdego wymiaru. Wówczas należy odpowiednio przygotować dane. Możliwe są dwa rozwiązania. Pierwsze polega na zakodowaniu opinii sędziów osobno dla każdego wymiaru w osobnych plikach. Drugie rozwiązanie polega na zakodowaniu wszystkich danych w tym samym pliku. Oznacza to, że opinie tego samego sędziego zajmują tyle wierszy, ile jest wymiarów oceny obiektu (np. kandydat na stanowisko pracy może być osobno oceniany przez sędziów z uwagi na wykształcenie, doświadczenie zawodowe, komunikatywność, prezencję itd.). Za pomocą osobnej zmiennej oznaczamy wymiar, którego opinia dotyczy, a w procesie analiz poprzez opcję „podziel dane na podzbiory" wybieramy jako kryterium podziału zmienną, w której zakodowaliśmy oceniane

wymiary. Wówczas współczynnik W Kendalla będzie policzony jednocześnie dla wszystkich wymiarów osobno.

Tabela 6. Wyniki uporządkowania kandydatów na stanowisko pracy przez członków komisji.

Sędzia	Kandydat					
	a	b	c	d	e	f
A	6	4	1	2	3	5
B	3	5	1	2	4	6
C	6	4	1	2	3	5
D	5	4	2	1	3	6

Jeżeli zastosujemy współczynnik korelacji do wyników oceny 6 kandydatów na stanowisko pracy przez 4 członków komisji (1 – oznaczało najlepszy kandydat, 6 – oznaczało najgorszy kandydat), to uzyskamy współczynnik korelacji W = 0,86, co świadczyłoby o dużej zgodności komisji w ocenach kandydatów, a uśrednione opinie można wykorzystać do podjęcia ostatecznej decyzji o zatrudnieniu.

Podsumowanie

Z uwagi na to, że nie ma jednomyślności wśród badaczy w ocenie wysokości współczynników zgodności i tego, który spośród nich jest najlepszy w sytuacji, kiedy warunki zastosowania pozwalają użyć kilku spośród dostępnych współczynników, warto podawać je równolegle w raportach z badań. Poza tym w niektórych obszarach badań praktykuje się wykorzystanie współczynników korelacji (np. ρ Spearmana lub α Cronbacha) do badania zgodności opinii sędziów. Wówczas dla porównania z wynikami wcześniejszych badań można obliczyć je, uzupełniając jednocześnie współczynnikami, które uwzględniają prawdopodobieństwo przypadkowości decyzji sędziów.

Literatura

1. Angleitner, A., Ostendorf, F., & John, O. P. (1990). Towards a taxonomy of personality descriptors in German: a psycho-lexical study. *European Journal of Personality*, 4(2), 89-118.
2. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
3. Ferguson, G. A., Takane, Y. (2004). *Analiza statystyczna w psychologii i pedagogice*. Wydawawnictwo Naukowe PWN.
4. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
5. Krejtz, K., Krejtz, I. (2009a). Rzetelność w analizie treści. (W:) K. Stemplewska-Żakowicz, K. Kreitz (red.) *Wywiad psychologiczny* (t.1, s.217-230). Warszawa: Pracownia Testów Psychologicznych.
6. Krejtz, K., Krejtz, I. (2009b). Wybrane statystyki zgodności miedzy sędziami w analizie treści. (W:) K. Stemplewska-Żakowicz, K. Krejtz (red.) *Wywiad psychologiczny* (t.1, s. 231-249). Warszawa: Pracownia Testów Psychologicznych.
7. Krippendorff, K. (2012). Content analysis: *An introduction to its methodology*. Newbury Park: Sage.
8. Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3), 321-325.