SZACOWANIE POZIOMU ZGODNOŚCI OCEN W KONTROLI WIZUALNEJ – PROBLEMY W WYZNACZANIU WSPÓŁCZYNNIKÓW TYPU KAPPA

Magdalena DIERING, Krzysztof DYCZKOWSKI, Adam HAMROL

Streszczenie: Artykuł opisuje rozważania na temat sposobów wyznaczania i interpretacji współczynników typu Kappa w szacowaniu poziomu zgodności ocen operatorów i eksperta w kontroli wizualnej. Autorzy opisują swoje obserwacje i doświadczenia z badań prowadzonych w przedsiębiorstwach produkcyjnych, wskazują na problemy i wyzwania. Autorzy zauważają, że współczynniki typu Kappa skoncentrowane są na porównaniach między oceniającymi, a pomijają zdolność oceniających do powtarzania tych samych ocen. W artykule przedstawiono oryginalną koncepcję współczynnika typu Kappa, który uwzględnia zgodność wewnętrzną oceniających.

Słowa kluczowe: Współczynnik Kappa-Cohen'a, współczynnik AC_1 Gwet'a, metoda tabel krzyżowych, poziom zgodności oceniających, cecha niemierzalna.

1. Wprowadzenie

Analiza systemów pomiarowych dla cech niemierzalnych znajduje zastosowanie w procesach wytwarzania, w których ocena wyrobów odbywa się poprzez rozstrzyganie, czyli ocenę czy dany element jest zgodny czy nie w odniesieniu do stawianych mu wymagań. Uzyskany wynik pochodzi z określonego zbioru kategorii, np. ocena 0-1, wadliwy-dobry-do naprawy, Go/NoGo itd. Taka ocena jest charakterystyczna przede wszystkim w kontroli sprawdzianem/szablonem lub przy ocenie sensorycznej, opartej na kontroli wzrokowej (np. sprawdzanie barw na wydrukach), dotykowej , słuchowej, zapachowej czy smakowej (np. w branży spożywczej). Podejście to znajduje także szerokie zastosowanie do cech mierzalnych – na przykład gdy pomiar cechy jest utrudniony lub ekonomicznie nieuzasadniony (np. pomiaru średnicy wałka można dokonać suwmiarką lub ocenić sprawdzianem). Przykładem cech niemierzalnych mogą być:

- dokładność wykonania (w porównaniu ze wzorcem) zespołu zawiasu okna dachowego (m.in. czytelność kodu części, kolor powłoki galwanicznej, uszkodzenia, prawidłowość pozycji elementów zawiasu, płynność ruchu ślizgacza) – ocena organoleptyczna,
- dokładność wykonania bezobsługowych (samosmarownych) łożysk ślizgowych – ocena wizualna,
- dokładność wykonania poszczególnych elementów stosowanego w medycynie cewnika diagnostycznego (m.in. czytelność napisu, wtopienia, przebarwienia, nadlewy, wgłębienia) – ocena wizualna.

Nadzorowany powinien być nie tylko proces wytwarzania (wyroby w procesie) oraz stosowany w nim system pomiarowy (sposób oceny/kontroli) [1]. Metody i procedury do analizy systemów pomiarowych umożliwiają stwierdzenie czy badany system jest adekwatny (właściwy) do stawianych mu zadań i celów (ocena jakości wyrobu lub

sterowanie procesem).

W literaturze przedmiotu znaleźć można opisy procedur, metod i współczynników proponowanych do analizy systemów pomiarowych dla cech niemierzalnych [2,3,4]. W wyniku analizy otrzymuje się wskaźniki/współczynniki informujące o poziomie zgodności między oceniającymi, skuteczności prowadzonej przez nich kontroli oraz o poziomie popełnianych przez oceniających błędów I (odrzucenie wyrobu zgodnego) i II (akceptacja wyrobu niezgodnego) rodzaju. Ponadto, ocena systemów pomiarowych stwarza przedsiębiorstwu możliwość identyfikacji obszarów do doskonalenia w procesie produkcyjnym. Ocena uzyskanych wyników powinna także prowadzić do uzyskania odpowiedzi na pytania: czy oceniający mają odpowiednie kwalifikacje?, czy otoczenie (środowisko) przeprowadzania oceny jest odpowiednie?, czy ryzyko błędów jest akceptowalne dla klienta?

2. Ocena poziomu zgodności – wybrane współczynniki typu Kappa

2.1. Zgodność ogólna i zgodność przypadkowa

Oszacowanie poziomu zgodności oceniających (w literaturze anglojęzycznej funkcjonuje określenie *inter-rater reliability*, czyli wiarygodność między oceniającymi) pozwala określić regularność uzyskiwanych klasyfikacji. Przy ocenie typu dobry/zły pojawiają się sytuacje, w których operator nie jest przekonany co do swojej oceny i „strzela". W zgodności ogólnej, czyli w liczbie przypadków, co do których oceniający zgodzili się w swojej ocenie (czyli we frakcji części o tych samych ocenach, oznaczanej jako p_a) są więc też takie, dla których oceniający podjęli decyzje przypadkowo („strzelali" w sposób losowy). Tą możliwość losowych decyzji nazywa się zgodnością przypadkową (oznaczana p_e; ang. *chance expected agreement*) [5]. Współczynniki typu Kappa uwzględniają poprawkę o wartość p_e [6]. Najbardziej znanymi współczynnikami tego typu są Kappa Scott'a κ_S (z roku 1955) dla par oceniających i będący jego modyfikacją Kappa Cohen'a κ_C (z roku 1960). Ten drugi znajduje szerokie zastosowanie w praktyce przedsiębiorstw produkcyjnych szczególnie przez rekomendacje grupy AIAG [3] do jego stosowania. Coraz bardziej znanym w statystyce inżynierskiej jest także współczynnik AC_1 Gwet'a (z roku 2001) (dla wielu oceniających). Różnica między tymi współczynnikami polega na sposobie wyznaczania wartości p_e. Scott przyjął, że przypadkowość ocen obu oceniających (w badanej parze oceniających) występuje z takim samym prawdopodobieństwem [7]. Cohen z kolei zakłada, że należy uwzględniać indywidualne preferencje poszczególnych oceniających [8]. Gwet zaproponował z kolei współczynnik dla dowolnej liczby oceniających i kategorii ocen oraz uwzględnił istnienie przypadków, dla których nie dokonano oceny [2].

Rozważania w dalszej części artykułu dotyczą współczynnika Kappa Cohen'a oraz AC_1.

2.2. Współczynnik Kappa Cohen'a

Współczynnik Kappa Cohen'a można zapisać (1) [8]:

$$\kappa_C = \frac{p_a - p_e}{1 - p_e}, \text{ gdzie } \; p_a = \sum_{k=1}^{q} p_{kk} \text{ i } \; p_e = \sum_{k=1}^{q} p_{k+} p_{+k}, \tag{1}$$

gdzie:

- κ_C – współczynnik Kappa Cohen'a, miara poziomu zgodności par operatorów lub par typu operator-ekspert (*inter-rater reliability*),
- p_a – suma zaobserwowanych zgodnych decyzji typu 0-0 i 1-1 w odniesieniu do liczby możliwych par decyzji,
- p_e – szansa (prowdopodobieństwo) na zgodność przypadkową dla decyzji typu 0-0 i 1-1 w odniesieniu do liczby możliwych par decyzji typu 0-0 i 1-1,
- q – liczba kategorii (najczęściej dwie: 0 – wyrób wadliwy, 1 – wyrób dobry),
- k – kategoria,
- p_{kk} – udział części, dla których oceniający A i B wskazali kategorię k,
- p_{k+} – udział części, dla których oceniający A wskazał kategorię k,
- p_{+k} – udział części, dla których oceniający B wskazał kategorię k.

Współczynnik Kappa Cohen'a może przyjmować wartości z przedziału <-1; 1>, przy czym dla wartości ujemnych uważa się, że zgodność oceniających jest niższa niż szansa ich zgodności przypadkowej, czyli brak zgodności, a dla wartości 0 zgodność oceniających jest na poziomie zgodności przypadkowej. W praktyce przyjęły się kryteria, które opracowano w oparciu o wytyczne Fleiss'a [9] i rekomendowane są przez AIAG [3]. Według AIAG, poziom zgodności operatorów badanego systemu pomiarowego jest dobry w zakresie od 0,4 do 0,75 i bardzo dobry, gdy wartość współczynnika Kappa przekracza 0,75 [3] (tabela 1).

Tab. 1. Ogólny Kryteria oceny poziomu zgodności w oparcu o wartość Kappa. Źródło: opracowanie własne na podstawie [2,3,9]

Wartość KAPPA	Interpretacja
Kappa min = -1,00	Brak zgodności
0 < Kappa < 0,40	Bardzo słaba zgodność
0,40 ≤ Kappa ≤ 0,75	Dobra zgodność
Kappa > 0,75 Kappa max = 1,00	Bardzo dobra zgodność Zgodność 100%

2.3. Problemy w wyznaczaniu i interpretacji współczynnika Kappa Cohen'a

Współczynnik Kappa Cohen'a jest wrażliwy na zmiany rozkładu wartości dla par zaobserwowanych zgodnych decyzji typu 0-0 i 1-1. Im większa koncentracja ocen w jednej z komórek dla par obserwacji 0-0 lub 1-1, tym mniejsza wartość współczynnika Kappa. Natomiast „przesuwanie" wyników obserwacji dla par 0-1 i 1-0 nie ma większego wpływu na jego warość [2,6,10]. Ilustrują to przykłady zawarte w tabeli 2.

Tylko dla układu 50% części jednej kategorii i 50% części kategorii drugiej możliwe jest uzyskanie wartości maksymalnej współczynnika Kappa, czyli 1 (przy założeniu, że oceniający nie popełnili ani jednego błędu). Brak zgodności w każdej kolejnej ocenie powoduje, że wartość Kappa „szybko" maleje. Jeśli relacja części dobrych do części złych będzie się znacznie różniła, wartość Kappa zostanie „zaburzona" (ilustrują to przykłady 2 i 3 oraz 4 i 5 w tabeli 2 – ogólna zgodność oceniających A i B w obu przypadkach wynosi 80 na 100 wskazań, ale wartość współczynnika Kappa dla przypadku pierwszego jest znacznie wyższa). W takim przypadku uzyskanie w drodze badania informacji o faktycznej zgodności oceniających będzie bardzo trudne, a czasem wręcz niemożliwe.

2.4. Współczynnik AC_1 Gwet'a

Gwet zaproponował nowe podejście do szacowania współczynnika Kappa – AC_1. Współczynnik ten jest rozwinięciem Kappy Fleiss'a [9], może być stosowany dla wielu oceniających i wielu kategorii ocen. AC_1 rozwiązuje także problem wrażliwości Kappy na nierównomierny rozkład częstości występowania par ocen typu 1-1 i 0-0. Poza tym, AC_1 można oszacować w przypadku, gdy uczestnicy badania oceniali różną (względem siebie) liczbę części (uwzględnia tzw. missing rating, czyli brak oceny).

Współczynnik Gwet'a można zapisać (2) [2]:

Tab. 2. Wrażliwość współczynnika Kappa Cohen'a na rozkład wartości par obserwacji. Źródło: opracowanie własne

oceniający / kategoria		B 1	B 0	suma:	Kappa
A	1	**50**	**0**	50	1,00
	0	**0**	**50**	50	
suma:		50	50	100	

oceniający / kategoria		B 1	B 0	suma:	Kappa
A	1	**40**	**10**	50	0,60
	0	**10**	**40**	50	
suma:		50	50	100	

oceniający / kategoria		B 1	B 0	suma:	Kappa
A	1	**70**	**10**	80	0,38
	0	**10**	**10**	20	
suma:		80	20	100	

oceniający / kategoria		B 1	B 0	suma:	Kappa
A	1	**40**	**4**	44	0,61
	0	**16**	**40**	56	
suma:		56	44	100	

oceniający / kategoria		B 1	B 0	suma:	Kappa
A	1	**70**	**16**	86	0,39
	0	**4**	**10**	14	
suma:		74	26	100	

oceniający / kategoria		B 1	B 0	suma:	Kappa
A	1	**10**	**40**	50	-0,60
	0	**40**	**10**	50	
suma:		50	50	100	

oceniający / kategoria		B 1	B 0	suma:	Kappa
A	1	**0**	**50**	50	-1,00
	0	**50**	**0**	50	
suma:		50	50	100	

$$AC_1 = \frac{P_a - P_e}{1 - P_e}, \tag{2}$$

gdzie

$$p_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r_i(r_i-1)},\ p_e = \frac{1}{q-1}\sum_{k=1}^{q}\pi_k(1-\pi_k) \text{ i } \pi_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r_i},$$

- AC_1 – współczynnik typu Kappa, zaproponowany przez Gwet'a, miara wiarygodności (poziomu zgodności) między oceniającymi,
- p_a – udział zaobserwowanych zgodnych decyzji w odniesieniu do liczby możliwych decyzji zespołu oceniających,
- p_e – szansa (prowdopodobieństwo) na zgodność przypadkową,
- r – liczba oceniających,
- r_{ik} – liczba oceniających, którzy zaklasyfikowali obiekt *i* do kategorii *k*,
- π_k – prawdopodobieństwo wskazania kategorii *k* przez dowolnego oceniającego dla dowolnej części,
- q – liczba kategorii (najczęściej dwie: 0 – wyrób wadliwy, 1 – wyrób dobry, ale możliwych jest więcej kategorii ze skali nominalnej),
- k – kategoria,
- n – liczba wszystkich obiektów w badaniu.

Gwet zaproponował także modyfikację AC_1 – AC_2, który uwzględnia dodatkowo czy ocenienie danej części przez operatora sprawia mu trudności (H-subject) czy jest to dla niego łatwe zadanie (E-subject) [2]. Rozważania na temat obszaru zastosowań AC_2 i możliwych modyfikacji tego współczynnika są przedmiotem obecnych dyskusji autorów.

3. Metoda tabel krzyżowych

W przewodniku AIAG dla branży motoryzacyjnej [3] do szacowania poziomu zgodności oceniających wskazana jest metoda tabel krzyżowych (ang. *Cross-Tab Method*). W ramach tej metody zalecane jest szacowanie współczynnika Kappa Cohen'a wraz z oceną skuteczności. Standardowy układ takiego badania to: od 30 do 50 części, 2 lub 3 oceniających i 2-3 serie oceny, czyli do 450 ocen oraz decyzje eksperta. Ważnym jest, żeby części do badania były tak dobrane, żeby około 50% z nich były częściami jednej kategorii (części dobre), druga połowa – drugiej kategorii (wyroby wadliwe). W obu grupach powinny znajdować się takie części, dla których decyzja jest trudna do podjęcia (części na granicy akceptowalności/odrzucenia).

Skuteczność systemu pomiarowego szacuje się dla zespołu operatorów (ocena między-operatorami) oraz dla zespołu operatorów i eksperta i w obu przypadkach w odniesieniu do liczby części w badaniu. Uzyskuje się także wyniki skuteczności dla poszczególnych operatorów. Ocena skuteczności proponowanej w przewodniku MSA4 dostarcza, między innymi, informacji na temat:

- udziału części (w odniesieniu do liczby części w badaniu), dla których poszczególni oceniający trzykrotnie (czyli za każdym razem) wskazali tą samą kategorię,
- liczby części, dla których oceniający popełnili błąd I rodzaju (część zgodna była trzykrotnie/zawsze niesłusznie odrzucana) lub błąd II rodzaju (cześć niezgodna trzykrotnie /zawsze niesłusznie akceptowana),
- liczby części, dla których oceniający popełniali sprzeczne decyzje (podejmowali różne decyzje co do tej samej części w trzech seriach ocen).

Kryteria akceptacji przy ocenie skuteczności przedstawiono w tabeli 3.

Tab. 3. Kryteria akceptacji przy ocenie skuteczności systemu pomiarowego dla cech niemierzalnych. Źródło: [3]

DECYZJA	Skuteczność	Udział złych części uznanych za dobre (ang. Miss Rate)	Udział dobrych części uznanych za złe (ang. False Alarm Rate)
System pomiarowy jest akceptowalny	≥ 90%	≤ 2%	≤ 5%
System pomiarowy jest warunkowo akceptowalny	≥ 80%	≤ 5%	≤ 10%
System pomiarowy jest nieakceptowalny	< 80%	> 5%	> 10%

4. Szacowanie poziomu zgodności między oceniającymi w praktyce przedsiębiorstw produkcyjnych – obserwacje i doświadczenia autorów

Autorzy prowadzili obserwacje w kilku przedsiębiorstwach produkcyjnych, w których realizowane były badania systemów pomiarowych: w firmie produkującej okucia okien dachowych – badania wizualnej oceny jakości elementu części ościeżnicowej i ramowej okna dachowego; w firmie produkującej łożyska dla samochodów osobowych, mających zastosowanie w kilkunastu częściach i mechanizmach w pojeździe oraz w firmie produkującej narzędzia i urządzenia medyczne, między innymi cewniki diagnostyczne. Oberwacje dotyczyły stosowanej w firmach metodyki wyznaczania i interpretacji wyników z analizy systemów pomiarowych.

Z doświadczeń autorów wynika, że w praktyce przedsiębiorstw produkcyjnych ocena poziomu zgodności oceniających najczęściej polega na wyznaczeniu wartości współczynnika Kappa Cohoen'a dla par typu operator-ekspert, często także dla par operator-operator. Ocena skuteczności dokonywana jest rzadko. Dla właścicieli procesów poziom zgodności poszczególnych operatorów z ekspertem jest informacją najważniejszą i na tej się koncentrują. Informacja o poziomie zgodności między operatorami okazuje się mniej istotna. Wynika to stąd, że zwykle ekspertem jest inżynier, specjalista ds. jakości lub kontroler jakości i to on przekłada wymagania klienta na specyfikację wyrobu i warunki jego odbioru. Ekspert tworzy katalog błędów i wad dla produktu, a ten jest podstawą instrukcji stanowiskowych, które obowiązują poszczególnych operatorów. Poziom zgodności operator-ekspert pozwala uzyskać odpowiedź na pytanie czy zapewnienie wiedzy (m.in. w postaci instrukcji stanowiskowych) danemu operatorowi przekłada się na jego umiejętności oceny i podejmowania decyzji o wyrobie.

Na podstawie obserwacji codziennej pracy operatorów i ekspertów, pogłębionych wywiadów z uczestnikami badań i uzyskanych z nich wyników sformułowano wnioski i spostrzeżenia:

- Zgodności dla poszczególnych par operatorów mogą być na wysokim poziomie, podczas gdy poziomy zgodności operatorów z ekspertem mogą nie być zadowalające. Oceny eksperta traktowane są jako wartości referencyjne i nadrzędne w stosunku do ocen operatorów, a tymczasem w praktyce okazuje się, że ekspert (zwłaszcza jeśli nie jest nim klient czy specjalista ds. jakości) też może popełnić błąd (i czasami popełnia!) w ocenie wyrobów.
- Ekspert rzadko (a czasem wcale) pracuje bezpośrednio na stanowisku, dla którego przygotowuje instrukcje stanowiskowe i stąd może nie uwzględniać kilku aspektów, na przykład: szczegółowość i obszerność instrukcji (za dużo treści),

lokalizacja instrukcji, brak precyzyjnych opisów wad i zdjęć (np. zdjęcie wady w powiększeniu ale bez informacji o zastosowanej skali) itp.

- Doświadczony operator nierzadko zna lepiej proces niż ekspert – ma większą wiedzę w odniesieniu do możliwych błędów i wad – umiejętne korzystanie przez eksperta z wiedzy i doświadczenia operatorów na potrzeby tworzenia katalogu wad i błędów przyczynia się do wzrostu w zespole świadomości roli jakości ocen i zaangażowania pracowników w działania doskonalące.

Ponadto:

- Operatorzy odbierają badanie systemu pomiarowego jako kontolę ich umiejętności i – mimo wyjaśnień ze strony nadzorującego badanie – nierzadko bardzo stresują się podczas podejmowania decyzji.
- W ramach badania metodą tabel krzyżowych operatorzy oceniają kilkadziesiąt części kilkukrotnie, co zajmuje sporo czasu i wiąże się ze wzmożoną koncentracją (operatorzy przecież wiedzą, że w ocenianym zbiorze wiele części jest wadliwych), inną niż podczas ich codziennej pracy (w przypadku stabilnych i zdolnych statystycznie procesów wadliwe wyroby pojawiają się bardzo rzadko). Wykonanie więcej niż jednej serii ocen w ciągu zmiany roboczej stanowi dla operatorów ogromny wysiłek i obniża ich zaangażowanie i powoduje zwątpienie w celowość i sensowność takich analiz, obniża także „wiarę" we własne umiejętności.
- Operatorom łatwo jest szczegółowo opisać zaobserwowane niedoskonałości na ocenianej części i rzadko nie dostrzegają odstępstw od wzorca dobrej części, jednakże często w takich przypadkach trudno jest im podjąć jednoznaczną decyzję o tym czy dany wyrób jest dobry czy wadliwy. Jak sami twierdzą (w pogłębionych wywiadach), przyczyną tego często jest możliwość konsultacji oceny z przełożonym (ekspertem) lub kontrolerem jakości, co „zwalnia" ich z odpowiedzialności za decyzję, a podczas badania nie mogą się z nikim konsultować i w takich przypadkach dokonują oceny przypadkowo, losowo.
- Operatorzy nierzadko posługują się innym językiem (innymi sformułowaniami) w odniesieniu do nazw wad i błędów, co może prowadzić do nieporozumień w komunikacji między ekspertem nadzorującym badanie a operatorami.
- Katalogi błędów i wad często nie są dobrze przygotowane, między innymi: wady opisane są nieprecyzyjnie (np. „kilka małych plamek" – nie wiadomo czy dwie plamki to już kilka czy dopiero pojawienie się czwartej plamki skutkować powinno odrzuceniem; nie wiadomo też co to znaczy „mała plamka"), brakuje zdjęć z niedoskonałościami na granicy akceptacji czy odrzucenia, zamieszczone w katalogu zdjęcia wad są często w powiększeniu, ale bez informacji o zastosowanej skali (i wtedy żadna plamka nie jest tak duża, jak ta „mała plamka" opisana w katalogu) i inne.
- Osoby przygotowujące badanie często nie uwzględniają w arkuszu danych innych kategorii niż „dobry" i „wadliwy", tymczasem w praktyce dla procesów produkcyjnych, w których dokonuje się oceny cech niemierzalnych, często funkcjonuje trzecia kategoria – „do poprawy", „do konsultacji" lub inna. Brak tej kategorii podczas wyznaczania poziomu zgodności oceniających może mieć ogromny wpływ na jakość i efekt całego badania.
- W wielu przypadkach autorzy zaobserwowali także tylko podstawową wiedzę inżynierów (osób nadzorujących badanie) z zakresu analizy systemów pomiarowych dla cech niemierzalnych.

5. Współczynnik typu Kappa z uwzględnieniem zgodności wewnętrznej oceniającego – koncepcja

5.1. Zgodność wewnętrzna oceniającego – dyskusja

Zdaniem autorów, ocena skuteczności zaproponowana w ramach metody tabel krzyżowych stanowić może uzupełnienie w badaniu poziomu zgodności oceniających, niemniej jednak łączenie jej z wyznaczaniem współczynnika Kappa w ramach jednej metody powoduje trudności w interpretacji i problemy z ostateczną decyzją analityka o wiarygodności i przydatności stosowanego systemu pomiarowego. Dla każdej pary oceniających osoba prowadząca analizę otrzyma po dwa trudno-porównywalne ze sobą wyniki. Pierwszy z nich to poziom zgodności ocen danej pary (reprezentowany przez wartość współczynnika Kappa, wartość od -1 do 1) i drugi – skuteczność tej pary (w %). Poza tym, w ocenie skuteczności nie uwzględnia się podstawowego założenia, które przyjmowane jest przy szacowaniu współczynnika typu Kappa – nie uwzględnia się, że w liczbie przypadków, co do których oceniający zgodzili się w swojej ocenie są też takie, dla których oceniający podjęli decyzje losowo. Stąd, porównywanie wyników obu tych analiz nie jest łatwe i autorzy zalecają co najmniej rozwagę w formułowaniu wniosków na temat poziomu zgodności oceniających w oparciu o zestawienie wyników metody tabel krzyżowych.

Pomijając ocenę skuteczności, można by zrezygnować z kolejnych serii powtórzeń ocen dla tych samych części (przy zapewnieniu odpowiedniej liczby części do wyznaczenia współczynnika Kappa; o liczbie części do badania piszą, między innymi, Feliks i Lichota [11] czy Gwet [2]), ponieważ tych współczynnik Kappa nie uwzględnia w swojej formule. Jednakże, proponowana w [3] ocena skuteczności zawiera w sobie próbę oszacowania powtarzalości poszczególnych operatorów (poprzez wyznaczanie frakcji trzykrotnych takich samych wskazań w odniesieniu do liczby części). Autorzy podjęli dyskusję na temat znaczenia zdolności oceniającego do powtarzania tych samych ocen w kolejnych seriach. Uważają, że przy szacowaniu poziomu zgodności oceniających – bez względu na to który ze współczynników typu Kappa zostanie użyty do obliczeń – należy także uwzględnić zgodność wewnętrzną, indywidualną każdego z badanych, także osoby eksperta (zwłaszcza, jeśli wartości referencyjne są nadawane przez kontrolera jakości, inżyniera jakości czy przełożonego operatorów-oceniających). Współczynniki typu Kappa jej nie uwzględniają, są skoncentrowane na porównaniach między oceniającymi, a tymczasem wątpliwą będzie analiza zgodności pary jeśli poszczególni oceniający mogą nie być zgodni sami ze sobą. Stąd, w metodzie tabel krzyżowych przy interpretowaniu wyniku współczynnika Kappa należy uwzględnić wynik skuteczności. Autorzy proponują jednak, by zgodność wewnętrzną oceniającego uwzględnić we współczynniku typu Kappa. Zgodność wewnętrzną można oszacować, korzystając na przykład ze współczynnika AC_1, ale z uwagą, że zamiast liczby oceniających jest liczba serii ocen tego samego oceniającego. Przykładowo, współczynnik taki może mieć postać (3):

$$\kappa_{win} = \frac{\sum_{r=1}^{m} \kappa_w \kappa_r}{\sum_{r=1}^{m} \kappa_w}, \tag{3}$$

gdzie:

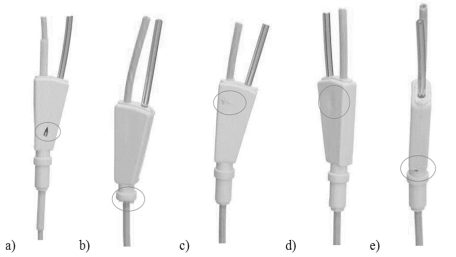
- κ_{win} – współczynnik typu Kappa, łączna miara poziomu zgodności wewnętrznej oceniających (***w**ithin-rater reliability*) i zgodności par oceniających (***in**ter-rater reliability*),

- κ_r – wartość współczynnika typu Kappa (np. Kappa Cohen'a, AC_1 lub inny) dla badanych oceniających,
- κ_w – współczynnik typu Kappa, miara poziomu zgodności wewnętrznej oceniających (***w**ithin-rater reliability*), wyznaczany analogicznie jak κ, przy czym liczbę oceniających zastępuje liczba serii powtórzeń ocen tego samego oceniającego,
- m – liczba oceniających,
- r – oceniający.

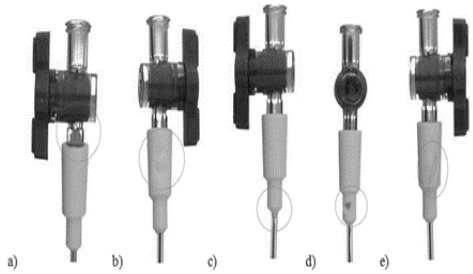
Koncepcja ujęta w formułę (3) stanowi dla autorów punkt wyjścia do szczegółowego zamodelowania takiego współczynnika oceny zgodności, w którym ujęta zostanie zgodność przypadkowa oraz zgodność wewnętrzną poszczególnych oceniających.

5.2. Badanie poziomu zgodności ocen w kontroli wizualnej z uwzględnieniem zgodności wewnętrznej oceniającego – przykład

Dla przykładu oszacowano wartość współczynnika κ_{wi} dla zbioru ocen, które uzyskano podczas badania systemu pomiarowego dla cech niemierzalnych przy produkcji cewników diagnostycznych. Do badania – po konsultacjach ze specjalistą ds. jakości – wybrano dwa elementy cewnika: rozgałęzienie (oznaczane w obliczeniach jako RO) oraz przedłużenie balonowe (PB). Na podstawie instrukcji stanowiskowych, katalogu błędów i wad oraz specyfikacji technicznych zdefiniowano listę możliwych wad, które mogą pojawić się w produkcji, a które wykrywać powinni operatorzy tych stanowisk, na których elementy te są wykonywane. Wybrane z wad przedstawiono na rysunkach 1 i 2.



Rys. 1. Rozgałęzienie – przykłady wad: a) przebicie, b) niedolanie, c) wtopienie materiału, d) zapadnięcie, e) pęknięcie. Źródło: opracowanie własne



Rys. 2. Przedłużenie balonowe – przykłady wad części: a) wystająca opaska, b) zapadnięcie, c) flashing, d) dziura na wyprasce, e) przebarwienia. Źródło: opracowanie własne

Ponieważ serie produkcyjne są krótkie, a wyroby wadliwe to rzadkość, zbieranie części do badań trwało wiele tygodni. Przygotowano próbkę 50 sztuk rozgałęzienia oraz 50 sztuk przedłużenia balonowego. W przygotowaniu próbek uczestniczyli Ekspert i specjalista ds. jakości. W tym przypadku nie policzono zgodności wewnętrznej Eksperta, przyjmując udzielone przez niego i skonsultowane ze specjalistą ds. jakości oceny jako referencyjne. O dokonanie trzech serii ocen poproszono operatorów: A, B i C. Zestawienie wskazań oceniających zestawiono w tabelach 4 i 5.

Tab. 4. Podsumowanie wskazań operatorów A, B i C w badaniu poziomu zgodności oceniających dla elementów cewnika diagnostycznego. Źródło: opracowanie własne

Produkt **RO**

Operator A

Seria ocen	Ocena		Razem
	0	1	
ocena 1	29	21	50
ocena 2	28	22	50
ocena 3	27	23	50
Srednia	28	22	50

Operator B

Seria ocen	Ocena		Razem
	0	1	
ocena 1	30	20	50
ocena 2	29	21	50
ocena 3	25	25	50
Srednia	28	22	50

Operator C

Seria ocen	Ocena		Razem
	0	1	
ocena 1	30	20	50
ocena 2	31	19	50
ocena 3	31	19	50
Srednia	30,7	19,3	50

Produkt **PB**

Operator A

Seria ocen	Ocena		Razem
	0	1	
ocena 1	40	10	50
ocena 2	41	9	50
ocena 3	42	8	50
Srednia	41	9	50

Operator B

Seria ocen	Ocena		Razem
	0	1	
ocena 1	36	14	50
ocena 2	34	16	50
ocena 3	33	17	50
Srednia	34,3	15,7	50

Operator C

Seria ocen	Ocena		Razem
	0	1	
ocena 1	31	19	50
ocena 2	29	21	50
ocena 3	30	20	50
Srednia	30	20	50

Tab. 5. Zestawienie wskazań typu 0-0, 1-1, 0-1 i 1-0 dla par Operator A-Ekspert, Operator B-Ekspert i Operator C-Ekspert w badaniu poziomu zgodności oceniających dla elementów cewnika diagnostycznego. Źródło: opracowanie własne

Produkt **RO**

Operator A

	0	1	Suma
0	**69**	6	75
1	15	**60**	75
Suma	84	66	150

Operator B

	0	1	Suma
0	**63**	12	75
1	21	**54**	75
Suma	84	66	150

Operator C

	0	1	Suma
0	**75**	0	75
1	17	**58**	75
Suma	92	58	150

Produkt **PB**

Operator A

	0	1	Suma
0	**72**	3	75
1	51	**24**	75
Suma	123	27	150

Operator B

	0	1	Suma
0	**65**	10	75
1	38	**37**	75
Suma	103	47	150

Operator C

	0	1	Suma
0	**62**	13	75
1	28	**47**	75
Suma	90	60	150

By oszacować poziom zgodności między oceniającymi, w pierwszej kolejności oszacowano ich zgodność wewnętrzną κ_w. W tym celu posłużono się współczynnikiem AC_1 (zgodnie ze wzorem (2)) dla 3 oceniających z tą uwagą, że zamiast 3 oceniających w obliczeniach uwzględniono 3 serie ocen tego samego oceniającego. Następnie oszacowano poziom zgodności poszczególnych operatorów z ekspertem (zgodność między operatorem a ekspertem), także w oparciu o współczynnik AC_1. W ostatnim etapie oszacowano współczynnik zgodności łącznej κ_{win} (zgodnie z formułą (3)), czyli miarę zgodności wewnętrznej oceniających (***w**ithin-rater reliability*) i zgodności między oceniającymi (***in**ter-rater reliability*). Uzyskano wysoką zgodność wewnętrzną (od 0,68 do 0,97) oraz

wysoką zgodność łączną dla RO – 0,70 i dosyć niską zgodność dla PB – 0,39. Uzyskane wyniki zestawiono w tabeli 6.

Tab. 6. Poziomy zgodności wewnętrznej operatorów A, B i C oraz miara łączna w badaniu poziomu zgodności oceniających dla elementów cewnika diagnostycznego. Źródło: opracowanie własne

		Produkt	
		RO	**PB**
Zgodność κ_w	Operator **A**	0,89	0,92
	Operator **B**	0,68	0,79
	Operator **C**	0,97	0,82
Zgodność κ_r (zgodnie z AC_1)	Operator **A**	0,72	0,35
	Operator **B**	0,56	0,38
	Operator **C**	0,78	0,46
Wskaźnik κ_{win}		0,70	0,39

6. Podsumowanie

Reasumując, autorzy dostrzegają wiele problemów i wyzwań w wyznaczaniu i interpretacji współczynników typu Kappa przy szacowaniu poziomu zgodności oceniających w kontoli wizualnej. Przeprowadzone obserwacje, opisane w artykule, stanowią wstęp do badań nad rozwojem metod i metodyki wyznaczania poziomu zgodności ocen operatorów i eksperta. Tym samym autorzy podjęli się opracowania modelu nowego współczynnika typu Kappa, który będzie uwzględniał zgodność wewnętrzną poszczególnych oceniających oraz będzie rozróżniał dobrane do badania części ze względu na poziom trudności oceny. Podstawą do rozważań nad sformułowaniem nowego współczynnika będą współczynniki Gwet'a – AC_1 i AC_2 oraz wnioski z kolejnych badań autorów realizowanych w przedsiębiorstwach produkcyjnych.

Bibliografia

1. Diering M., Kujawińska A., Dyczkowski K., Rogalewicz M.: Logika rozmyta w ocenie alternatywnych systemów pomiarowych jako jeden z kierunków rozwoju MSA, w: Innowacje w zarządzaniu i inżynierii produkcji, Oficyna Wydawnicza Polskiego Towarzystwa Zarządzania Produkcją, Opole 2014, tom 2, str. 348-359.
2. Gwet K. L.: Handbook of Inter-Rater Reliability. The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters, 3rd ed., 2012.
3. Measurement Systems Analysis, 4th ed., Reference manual, AIAG-Work Group, Daimler Chrysler Corporation, Ford Motor Company, General Motors Corporation, 2010.
4. Hamrol A.: Zarządzanie jakością z przykładami, Wyd. PWN, Warszawa 2008.
5. Carletta J.: Assessing agreement on classification tasks: the kappa statistic, Computational Linguistics, vol. 22/2, str. 1-6
6. Jarosz-Nowak J.: Modele oceny stopnia zgody pomiędzy dwoma ekspertami z wykorzystaniem współczynników kappa, Matematyka Stosowana, nr 8, Wrocław 2007, s. 126-154.

7. Scott W.A.: Reliability of Content Analysis: The Case of Nominal Scale Coding, The Public Opinion Quarterly, 19(3):321-325, 1955.
8. Cohen J.: A coefficient of agreement for nominal scales, Educational and Psychological Measurement 1960, vol. 20, pp. 37-46.
9. Fleiss J., Cohen J.: The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement 1973, vol. 33, pp. 613-619.
10. Viera Anthony J., Garrett Joanne M.: Understanding Interobserver Agreement: The Kappa Statistic, Family Medicine, Vol. 37, No. 5, pp. 360-363, 0742-3225
11. Feliks J., Lichota A.: Wspomaganie analizy systemów pomiarowych (MSA), Archives of Foundry Engineering, Volume 10, Special Issue 3/2012, s. 169-174.

Dr inż. Magdalena DIERING
Prof. dr hab. inż. Adam HAMROL
Katedra Zarządzania i Inżynierii Produkcji
Wydział Budowy Maszyn i Zarządzania
Politechnika Poznańska
Pl. M. Skłodowskiej-Curie 5, 60-965 Poznań
Tel./fax.: (061) 665 2738 / (061) 665 2774
e-mail: magdalena.diering@put.poznan.pl

Dr Krzysztof DYCZKOWSKI
Zakład Metod Przetwarzania Informacji Nieprecyzyjnej
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza
ul. Umultowska 87, 61-614 Poznań