PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 4, 67–84
DOI: 10.31648/przegldpsychologiczny.10084

Alfa Cronbacha – co daje dobre wyniki?

Kilka uwag dotyczących budowania kwestionariuszy psychologicznych

Tomasz Rak[1,2]
Uniwersytet Papieski Jana Pawła II w Krakowie
https://orcid.org/0000-0002-3522-5176

Szymon Wrześniowski[2]
Uniwersytet Papieski Jana Pawła II w Krakowie
https://orcid.org/0000-0001-5553-4016

Streszczenie

Cokolwiek mierzy alfa Cronbacha – nie jest to spójność wewnętrzna, powszechnie błędnie rozumiana w psychologii jako średnia siła związków pomiędzy pozycjami kwestionariusza. W tym artykule badamy powody, dla których rozumienie alfa jako spójności wewnętrznej jest błędne i skupiamy się na działaniu inflacji (przeszacowania) współczynnika alfa w praktyce. Na bazie symulacji komputerowych określiliśmy dokładny (wspólny) wpływ na wartość alfa: liczby respondentów, zakresu skal pomiarowych (Likerta), liczby pytań w kwestionariuszu (itemów) oraz średniej korelacji między pozycjami. Wyniki potwierdzają występowanie inflacji poziomu alfa ze względu na liczbę pytań: alfa osiąga zadowalające wartości nawet przy minimalnej spójności wewnętrznej, jeśli w kwestionariuszu jest dużo itemów. Sugerujemy, że w przypadku słabych narzędzi pomiarowych rzetelność może być przeszacowywana ze względu na prezentowany tu krzywolinowy wzrost alfa. Liczba osób badanych i zakres skali nie miały wpływu na wartość alfa.

Słowa kluczowe: rzetelność, alfa Cronbacha, inflacja alfa, spójność wewnętrzna, symulacje

---

[1] Adres do korespondencji: tomasz.rak@o2.pl lub tomasz.rak@upjp2.edu.pl.

[2] Autorzy oświadczają, że oni sami, ani żaden członek ich rodziny, NIE mają powiązań ani nie są zaangażowani w jakąkolwiek organizację lub podmiot mający jakiekolwiek finansowe lub niefinansowe interesy związane z tematyką lub materiałami omawianymi w tym manuskrypcie, ani żadnymi innymi interesami lub działaniami, które mogą być postrzegane jako mające wpływ lub wzmagające stronniczość badania

W psychometrii analiza rzetelności pozwala określić precyzję pomiaru (wielkość błędu pomiaru) (Revelle i Condon, 2019) narzędzia pomiarowego (zazwyczaj kwestionariuszowego) (Sijtsma, 2020). Innymi słowy, rzetelność ma odpowiadać na pytanie, jak dobrze test mierzy to, co ma mierzyć, i jaka jest wielkość błędu pomiaru (Bajpai i Bajpai, 2014; Golafshani, 2003). Zgodnie z klasyczną teorią testów rzetelność jest to wielkość współczynnika korelacji między wynikiem obserwowanym a wynikiem prawdziwym lub pomiędzy wynikami wersji równoległych testów (Li i in., 1996; Metsämuuronen, 2022; Raykov i Marcoulides, 2011). W piśmiennictwie psychologicznym spośród wielu metod zdecydowanie najpopularniejszą miarą rzetelności jest obecnie współczynnik alfa Cronbacha (por. np. Flake i in., 2017; McNeish, 2018; Ponterotto. i Ruckdeschel, 2007; Taber, 2018) i to na nim chcemy skupić nasze rozważania.

Miara alfa Cronbacha wielokrotnie była poddawana krytyce, pod względem jej właściwości i możliwości interpretacji (np. Bonett i Wright, 2015; Dunn i in., 2014; Eisinga i in., 2013; Flora, 2020; Sijtsma i Pfadt, 2021). Chcemy natomiast już na tym etapie nadmienić, że w niniejszej pracy nie zamierzamy się mierzyć z teoretycznymi meandrami zjawiska pomiaru rzetelności (opisywanymi szerzej przez Borsbooma i Mellenbergha, 2002 lub Kane'a, 2013) czy z kwestią jakości samej miary alfa na tle innych miar rzetelności (Anselmi i in., 2019; Trizano-Hermosilla i Alvarado, 2016). Postaramy się też uniknąć złożonych dywagacji na temat struktury wzorów, żeby nie generować problemów z odbiorem prezentowanych tu treści wśród badaczy psychologów, którzy nie mają silnego ugruntowania w matematyce (por. Borsboom, 2006) – chcemy o psychometrii mówić w sposób prosty. Praca ta ma charakter czysto empiryczny i naszym celem jest przybliżenie pewnego praktycznego problemu z wykorzystywaniem współczynnika alfa tym psychologom, którzy po prostu w swojej pracy wykorzystują gotowe pakiety statystyczne. Mianowicie: współczynnik alfa nie powinien być interpretowany jako potocznie rozumiana spójność wewnętrzna (*internal consistency*).

Jak wskazują liczne publikacje, *rzetelność* w psychologii często rozumie się jako spójność wewnętrzna wyników uzyskanych w obrębie danego testu psychologicznego (Kalkbrenner, 2023; Revelle i Condon, 2019). Miara alfa jest często mylnie interpretowana jako *bezpośrednia miara spójności wewnętrznej* (Cho i Kim, 2015; Hayes i Coutts, 2020; Henson, 2001). Należy oczywiście zaznaczyć, że w niektórych przypadkach *spójność wewnętrzna* i *rzetelność* mogą być sobie równe (Lucke, 2005; Ten Berge i Sočan, 2004), jednak takie rozumienie w odniesieniu do współczynnika alfa jest po prostu błędne. W najlepszym wypadku współczynnik alfa zbliża się do tzw. największej dolnej granicy rzetelności[3] (GLB, *greatest lower bound to the reliability*), a na dodatek, nawet jeśli założymy, że pomiary są tau-ekwiwalentne (wyniki prawdziwe pomiarów mają tę samą lub porównywalną wariancję), GLB może nadal nie wyrażać rzetelności (Green i Yang, 2009; Sijtsma, 2009). Niemniej nawet pomijając tę złożoną

[3] Chcemy zauważyć, że obecnie nie ma jednoznacznego i dobrego tłumaczenia tego terminu na język polski.

kwestię, *spójność wewnętrzna* także jest sama w sobie przedmiotem debaty: jak ją rozumieć, czym właściwie jest i co tak naprawdę mierzy?

Samo sformułowanie *spójność wewnętrzna* (*internal consistency*) było dotychczas różnie definiowane i autorzy prac w omawianym obszarze wskazują, że już na poziomie definicyjnym istnieje spore zamieszanie sprawiające problemy interpretacyjne (Bentler, 2009; Streiner, 2003; Tang i in., 2014). Na przykład rozróżnienie pomiędzy spójnością wewnętrzną (*internal consistency*), jednorodnością (*homogenicity*), rzetelnością (*reliability*), nasyceniem czynnika (*general factor saturation*) itp. jest problematyczne, gdyż terminów tych często (i zazwyczaj błędnie) używa się zamiennie. Badacze wskazują również, że psychologia na ogół utożsamia tę wewnętrzną spójność z pojęciem tego, czy seria pozycji (np. pytań kwestionariusza) mierzy „mniej-więcej-to-samo" (por. McCrae i in., 2011; Sijtsma, 2009). Takie sformułowanie jest również nieprecyzyjne i co za tym idzie mylące. Można jednak z dużą dozą odpowiedzialności założyć, iż dla psychologów spójność wewnętrzna często oznacza, że pozycje kwestionariusza w ramach jednej skali po prostu odpowiednio silnie korelują ze sobą przy pomiarze określonej właściwości psychologicznej (por. Tang i in., 2014; Thigpen i in., 2017; Vaske i in., 2017). A więc psychologia zdaje się mylić rzetelność wyrażoną miarą alfa z przeciętną korelacją między pozycjami kwestionariusza. Czasem w odniesieniu do wyników współczynników spójności wewnętrznej pojawia się założenie istnienia jednowymiarowej struktury zjawiska, tzn. w wypadku użycia narzędzi takich jak analiza czynnikowa (PCA) lub analiza konfirmacyjna (CFA) uda nam się wykazać, iż badana struktura jest niepodzielna (zob. np. Bentler, 2009; Gignac i in., 2007). To założenie do pewnego stopnia jest zgodne z ideą wysokiej korelacji pozycji w obrębie założonej struktury i choć różne metody redukcyjne dadzą nieco inne rezultaty, gdy część itemów koreluje mocniej, a część słabiej (zob. Jolliffe i Cadima, 2016; McDonald, 2013), to nadal można założyć, że obracamy się w obrębie zagadnienia związków między itemami (pytaniami) danego kwestionariusza. Dlatego też, aby wykazać interesujący nas efekt, na potrzeby niniejszego tekstu przyjmijmy, że *spójność wewnętrzna*, jak powszechnie się ją rozumie, jest średnią korelacją pomiędzy pozycjami w ramach danego zjawiska psychologicznego. W związku z tym pierwszy obszar zainteresowania naszego artykułu to precyzyjne określenie nieliniowej zależności między średnią korelacją pozycji kwestionariusza a rzetelnością mierzoną współczynnikiem alfa.

W obszarze pomiaru rzetelności metodą alfa Cronbacha narosły też pewne nieścisłości lub mity. Niektórzy badacze wskazywali, iż poziom współczynników alfa może zależeć od liczby itemów kwestionariusza, a większa ich liczba zazwyczaj doprowadza do inflacji alfa (Cortina, 1993; DeVellis, 2006; Duhachek i in., 2005; Dunn i in., 2014). Pośrednio można założyć, biorąc pod uwagę zarówno wzór samego współczynnika alfa, skalujący ten współczynnik względem liczby itemów (Cortina, 1993; Thompson, 2003), jak i podstawę proroczego wzoru Spearmana-Browna (de Vet i in., 2017), że inflacja współczynnika alfa jest faktem. Nie było jednak dotychczas jasne, jak duże przeszacowanie poziomu rzetelności następuje wraz z dokładnym wzrostem liczby itemów mierzących dany konstrukt. Ponadto uważamy, że dla psychologów, którzy nie posługują się

biegle matematyką, wyobrażenie sobie na podstawie wzorów, jak zachowa się pewna krzywoliniowa zależność w przestrzeni, może być niezwykle trudne. Drugim celem niniejszego artykułu jest więc przejrzyste przedstawienie tego, w jaki sposób liczba itemów realnie rzutuje na przeszacowanie współczynnika alfa, oraz prezentacja wizualna kluczowych progów przeszacowania.

Kolejnym interesującym nas zagadnieniem było określenie zależności między wartością alfa a długością skali odpowiedzi, którą mają do dyspozycji osoby badane. Dotychczas sugerowano, że rzetelność uzyskanych wskaźników (wymiarów latentnych kwestionariusza) zwiększa się wraz z długością użytej w badaniu skali Likerta (Leung, 2011; Preston i Colman, 2000; Taherdoost, 2022), choć nie zawsze zwiększa to *spójność wewnętrzną*. Wydaje się to mało możliwe w kontekście wspomnianych wzorów obliczania alfa, nieuwzględniających przecież długości użytej skali, ponieważ w uproszczeniu wartość alfa powinna opierać się głównie na związkach pomiędzy poszczególnymi pytaniami kwestionariusza. Wariancja wyników może być (i zazwyczaj w warunkach naturalnych jest) różna dla wyników zebranych z użyciem tych samych pytań, ale o różnych rozpiętościach skal odpowiedzi, jednak choć wariancja wyników sama w sobie gra pośrednią rolę w obliczeniu współczynnika alfa, to z pewnością nie jest jasne, jak długość skali ma się przekładać na spójność wewnętrzną (*internal consistency*) (np. Chang, 1994; Matell i Jacoby, 1972). Interesujące było więc dla nas również to, aby określić, jak duże różnice wartości alfa uzyskamy np. między skalami dychotomicznymi (badany udziela odpowiedzi „tak" lub „nie") a skalami wielopunktowymi (siedmiostopniowymi skalami Likerta), dla zbliżonych związków między itemami kwestionariusza. Należy tu nadmienić, iż w piśmiennictwie sugeruje się, że współczynnik alfa może nie nadawać się do określania *rzetelności* skal dychotomicznych (Barbaranelli i in., 2015; Pastore i Lombardi, 2014), niemniej debata ta nie jest przedmiotem zainteresowania niniejszego artykułu. Skupimy się tutaj na *spójności wewnętrznej*, dla której uzyskaliśmy ciekawe wyniki i przedstawimy je w dalszej części pracy.

W strukturze wzorów dotyczących alfa nie uwzględnia się bezpośrednio rozmiaru skali, na której odpowiada badany, i tak samo nie pojawiają się tam (bezpośrednio) elementy odnoszące się do liczebności osób badanych. Psychologowie niezajmujący się statystyką matematyczną nie zadają sobie zwykle pytania, czy istnieje jakiś pośredni związek między macierzą kowariancji a liczebnością (można wskazać wzory wykazane np. u Brannicka, 1995; Brecklera, 1990 czy Thalla i Vaila, 1990), ale na potrzeby badań zachęca się ich do zwiększania wielkości próby, na której waliduje się dane narzędzie pomiarowe (Chan, 2014; Zumbo i Chan, 2014). O ile ma to z pewnością znaczenie dla określenia jego struktury (Comrey i Lee, 1992; Tabachnick i Fidell, 2007) czy jego normalizacji (Guidroz i in., 2009; Macey i Eldridge, 2006), o tyle nie było dotychczas jasne, jaki dokładnie wpływ ma zwiększenie liczebności próby zarówno na poziom alfa, jak i na wspominaną wcześniej spójność wewnętrzną (*internal consistency*).

Podsumowując powyższe rozważania, nasz artykuł podejmuje zagadnienie wpływu poziomu średniej korelacji pomiędzy itemami narzędzia pomiarowego, liczby itemów, długości skali odpowiedzi oraz liczebności osób badanych na wielkość współczynnika alfa Cronbacha.

Metoda

Na potrzeby badania napisano oprogramowanie w języku Delphi[4], którego zadaniem było wygenerowanie na podstawie metody Monte Carlo (Dimov, 2008; Dunn i Shultis, 2011) serii losowych setów danych o różnych parametrach. Pojedyncze sety danych symulowały odpowiedzi badanych uzyskiwane z badań kwestionariuszowych i różniły się: liczbą itemów (*NI*), liczbą przypadków (*NN*), zakresem skal odpowiedzi (*NL*) oraz średnią korelacją pomiędzy pozycjami kwestionariusza (*Mr*). Każdy pojedynczy set danych mógł zawierać od 2 (minimalna liczba do określenia rzetelności) do 20 itemów (pytań) kwestionariusza (*NI* = {2,3,4…20}) oraz zawierać od 100 do 1000 przypadków (fikcyjnych osób badanych), przygotowanych w odstępie co 100 (*NN* = {100,200,300…1000}). Następnie ustalono różne zakresy skal odpowiedzi w obrębie symulowanego instrumentu pomiarowego – od skali dychotomicznej (dwupunktowej, reprezentującej odpowiedzi „tak" vs „nie"), po skale siedmiopunktowe. Rozpiętość skal mieściła się więc w zakresie od 1–2 do 1–7 punktów (*NL* = {2,3,4…7}). Ostatnim ustalanym warunkiem była średnia korelacja między pozycjami kwestionariusza mierzona współczynnikiem r-Pearsona, a jej zakres ustalono w przedziale od 0,10 do 0,90 skokowo, co 0,10 (*Mr* = {0,10, 0,20, 0,30…0,90}). Wymienione warunki wejściowe mogły więc dać 19 × 10 × 6 × 9 = 10 260 możliwych kombinacji takich symulowanych narzędzi kwestionariuszowych (setów danych). Dla każdej pojedynczej kombinacji warunków wygenerowano serię 100 setów kwestionariuszy danego typu, co dało łącznie 1 026 000 pojedynczych symulowanych wyników kwestionariuszowych o różnych parametrach (lub inaczej 10 260 serii setów danych).

Na potrzeby niniejszej pracy i dla uproszczenia analizowano jedynie macierze dodatnie; zakłada się, że ujemne korelacje wskazują na pozycje, które przed pomiarem rzetelności powinny zostać odwrócone (poruszają to pośrednio Bland i Altman, 1997). W celu uzyskania wyników maksymalnie zbliżonych do wyników realnych, generator wprowadzał odpowiedni poziom szumu do danych. Algorytm generował takie zestawienia odpowiedzi w których średnia korelacja między pozycjami mieściła się w obrębie +/– 0,02 ustalonego progu, np. dla progu 0,50 dawało to realnie zakres $0{,}48 < rM <= 0{,}52$. Zatem mimo ustalonych wartości skokowych ostatecznie uzyskiwano dość zróżnicowany wynik średniej korelacji, który wiarygodnie oddawał to, co uzyskuje się w realnych badaniach. W obrębie 100 różnych setów o tych samych parametrach wejściowych generatora (jednej serii setów) nieznacznie różniły się również wartości korelacji pomiędzy poszczególnymi pozycjami, ponieważ zakładano, że np. dwie pozycje mogą korelować silniej ze sobą, a z trzecią z nich korelacja mogła być słabsza niż założony próg itd. W efekcie, mimo zbliżonych finalnie średnich korelacji między dwoma różnymi setami, dawało to duży rozrzut różnych możliwych do uzyskania danych o pożądanych właściwościach finalnych.

[4] Kod lub pseudokod tego oprogramowania do generowania symulowanych danych może być potencjalnie niebezpieczny (np. pozwolić na sfingowanie badań eksperymentalnych), dlatego może zostać udostępniony tylko po uprzednim porozumieniu z autorami.

Kombinacje liczby itemów (*NI*), liczby przypadków (*NN*), zakresu skal odpowiedzi (*NL*) oraz średniej korelacji pomiędzy pozycjami (*Mr*) potraktowano jako zmienne niezależne. Zmienną zależną były obliczone dla danego setu wartości alfa-Cronbacha. Surowe sety danych[5] zostały przetworzone z użyciem skryptu w języku R. Dla każdego setu danych obliczono dokładną średnią korelację między pozycjami, a następnie współczynnik alfa-Cronbacha, co pozwoliło uzyskać plik danych wyjściowych zawierający 1 026 000 finalnych obserwacji. W tych finalnych danych mogły się znaleźć takie „przypadki", dla których obliczona średnia korelacja była nieco słabsza niż założony próg skokowy korelacji. Pozwoliło to na większe zagęszczenie danych na wykresach i opracowanie wyników w sposób bliższy funkcjom ciągłym niż skokowym. W prezentacji dla większej czytelności wykresów użyto uśrednionych wartości dla skoków korelacji. Plik z danymi jest dostępny jako materiał uzupełniający do niniejszego artykułu[6].

Przed przystąpieniem do analiz przeprowadzono kontrolę równoważności warunków planu badawczego. Przyjęto, że poprawnie wygenerowane dane będą miały dokładnie takie same średnie poziomy korelacji w każdej kombinacji warunków ze względu na liczbę itemów (*NI*), liczbę przypadków (*NN*) oraz rozstęp maksymalny skali Likerta (*NL*). Wynik ten oczywiście został uzyskany – sety można uznać za równoważne, ponieważ średni poziom korelacji w każdym z tych warunków jest właściwie taki sam (bliski 0,51).

Pewną trudność w dalszych analizach wyników stanowiło wskazanie „dobrej" miary alfa. Mimo coraz bardziej restrykcyjnych zaleceń akceptowalne są nadal raportowane wartości powyżej 0,80, a nawet powyżej 0,70 (Taber, 2018). Stąd podczas prezentacji wyników odnosimy się raczej do wartości surowej alfa, pozostawiając czytelnikowi samodzielną interpretację rzetelności na podstawie wartości współczynnika.

Wyniki

Nie uzyskano wpływu liczebności grupy (*NN*) na poziom alfa w serii setów. Nie odnotowano także wpływu długości skali Likerta na poziom alfa (*NL*), zarówno wyrażonej w postaci równania regresji (tabela 1, s. 73), jak i ogólnego modelu liniowego (tabela 2, s. 73). Natomiast liczba pozycji kwestionariusza (*NI*) i poziom korelacji między nimi (*Mr*) okazały się dodatnio związane z poziomem alfa. Te trzy czynniki pozostawały ze sobą w nieliniowej zależności i należy ją tutaj nieco dokładniej opisać.

[5] Ich rozmiar przekracza 12 GB, więc nie są one załączone jako materiał uzupełniający, ale mogą być dostępne po uprzednim ustaleniu z autorami.

[6] Rozmiar modelu zawierający taką liczbę „przypadków" nie nadaje się do standardowych technik modelowania poprzez ogólne modele liniowe. Na potrzeby prezentacji modeli istotnościowych opracowano także bazę danych z uśrednionymi w obrębie stu setów danych wynikami.

Tabela 1

Predyktory przeciętnego poziomu alfa dla uproszczonego zestawu danych (regresja liniowa bez interakcji)

Predyktor	*B (SE)*	β
Stała	0,45 (0,00)	---**
NL	0,00 (0,00)	0,00
NN	0,00 (0,00)	0,01
NI	0,01 (0,00)	0,42**
Mr	0,05 (0,00)	0,78**
F	93350,89**	
R2	0,784	

* $p < 0,05$, ** $p < 0,01$; *NI* – liczba itemów/pytań kwestionariusza, *NN* – liczba przypadków/ osób badanych, *NL* – rozstęp maksymalny skali Likerta, *Mr* – średnia korelacja pomiędzy pozycjami kwestionariusza

Tabela 2

Predyktory przeciętnego poziomu alfa dla uproszczonego zestawu danych (ogólny model liniowy z interakcjami)

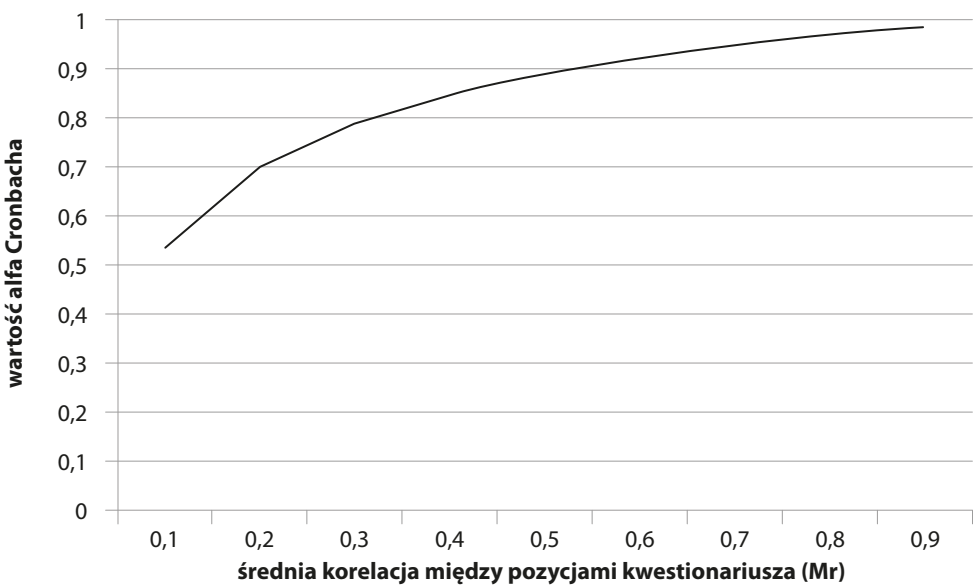
źródło zmienności	*F*	*p*	*η2*
NL	0,10	0,992	0,000
NN	0,31	0,973	0,000
NI	9740,78	< 0,001	0,214
Mr	497350,05	< 0,001	0,608
NL × *NI*	< 0,01	1,000	0,000
NL × *NI*	< 0,01	1,000	0,000
NI × *Mr*	3640,24	< 0,001	0,080
NL × *NN*	0,01	1,000	0,000
NL × *Mr*	0,03	1,000	0,000
NN × *Mr*	0,11	0,999	0,000
NL × *NN* × *NI*	< 0,01	1,000	0,000
NL × *NI* × *Mr*	< 0,01	1,000	0,000
NN × *NI* × *Mr*	< 0,01	1,000	0,000
NL × *NN* × *Mr*	< 0,01	1,000	0,000
NL × *NN* × *NI* × *Mr*	< 0,01	1,000	0,000

NI – liczba itemów/pytań kwestionariusza, *NN* – liczba przypadków/ osób badanych, *NL* – rozstęp maksymalny skali Likerta, *Mr* – średnia korelacja pomiędzy pozycjami kwestionariusza

Wzrost poziomu alfa obserwowany ze względu na wartość średniej korelacji między pozycjami (*Mr* w przedziale od 0,1 do 0,9) został przestawiony na wykresie (rysunek 1). Jest to nieliniowa zależność, która może być wyrażona dla wygenerowanych danych jako logarytm (*alfa* = 1,02 + 0,21* LN(*Mr*)). Można sformułować ogólny wniosek – wraz ze wzrostem średniej korelacji między pozycjami rosła wartość alfa, ale ten wzrost był najsilniejszy (najbardziej stromy) do poziomu przeciętnej korelacji bliskiej 0,40 – wówczas alfa przekraczała „akceptowalną" wartość 0,80. Ponadto wzrost poziomu alfa następował wraz ze wzrostem liczby itemów kwestionariusza (rysunek 2, s. 75). Jest to również zależność logarytmiczna (*alfa* = 0,58 + 0,12 × LN(*NI*)). Dość silny przyrost alfa był obserwowany do 5 itemów/pytań kwestionariusza, gdzie alfa również była bliska 0,80. Oba te obserwowane wyniki są jednak jedynie uproszczonym rzutem interakcji średniej korelacji i liczby itemów na dwuwymiarową przestrzeń. Przyjrzyjmy się więc tej interakcji w ujęciu trójwymiarowym (rysunek 3, s. 75).

Rysunek 1

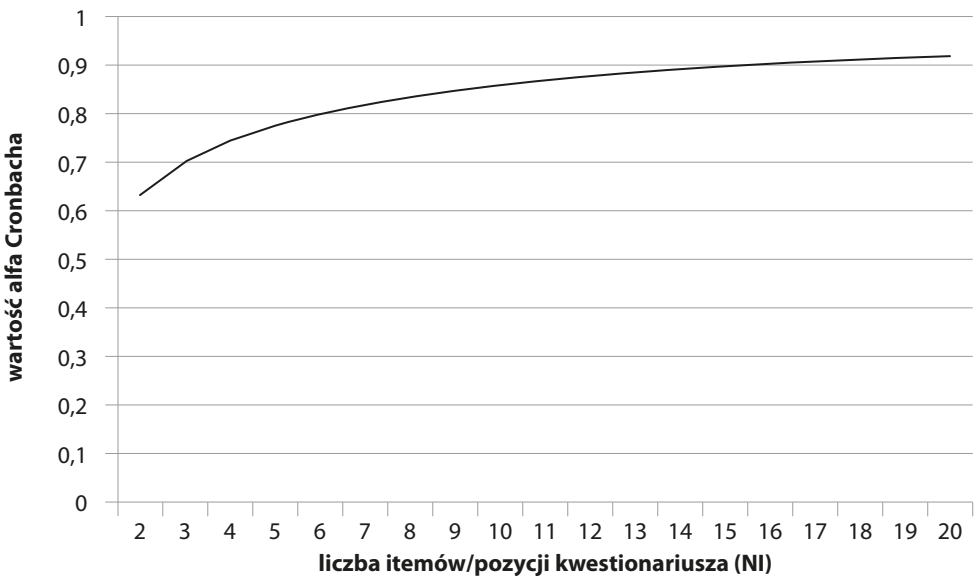
Wartość alfa Cronbacha jako przeciętny wynik uzyskany dla kwestionariusza o założonym poziomie korelacji między pozycjami



Dla dwóch itemów (pytań) kwestionariusza (*NI* = 2) zależność między średnią korelacją (*Mr*) a wartością alfa była bliska liniowej – wartość alfa była niemal równa przeciętnej korelacji między pozycjami. Natomiast wraz ze wzrostem liczby itemów (pytań) kwestionariusza zależność ta była coraz silniej krzywoliniowa (por. rysunek 3, tabela 3) – przy 20 itemach kwestionariusza (*NI* = 20)

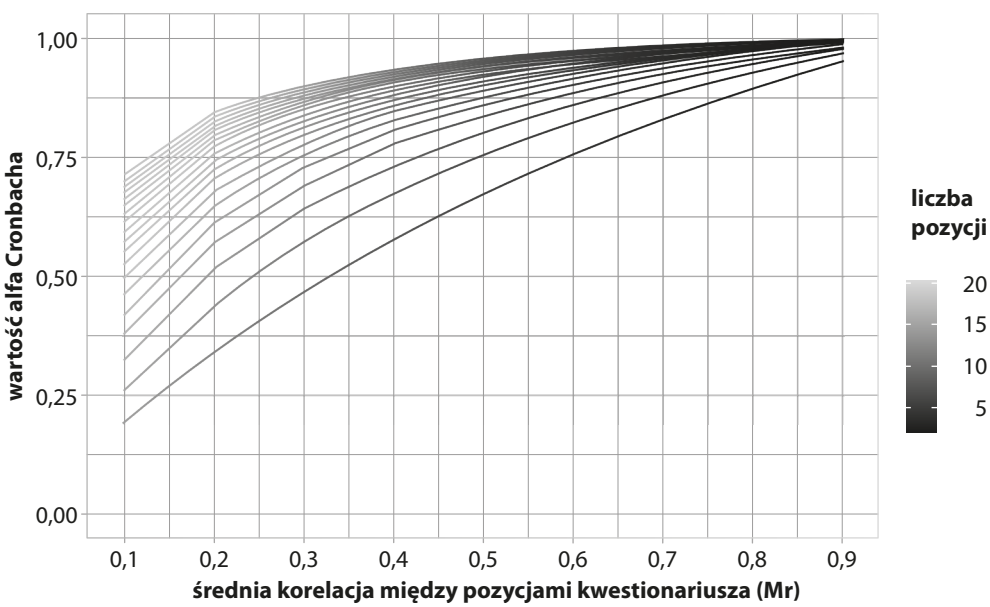
Rysunek 2

Wartość alfa Cronbacha jako przeciętny wynik uzyskany dla kwestionariusza o założonej liczbie pytań



Rysunek 3

Zależność między średnim poziomem korelacji składowych kwestionariusza a średnim poziomem alfa Cronbacha; linie reprezentują różną liczbę itemów kwestionariusza – od 2 pytań (najciemniejsze linie, niżej położone) do 20 pytań (najjaśniejsze linie, wyżej położone)



wystarczyły bardzo słabe związki między itemami ($Mr \approx 0{,}20$-$0{,}30$), by uzyskać poziom alfa raportowany zazwyczaj jako bardzo dobry (0,85). Warto też zauważyć, że nawet przy minimalnej założonej średniej korelacji (0,10) wraz z liczbą itemów wzrastała wartość średnia alfa, aby przekroczyć poziom 0,70 przy 20 itemach kwestionariusza. Innymi słowy, nawet jeśli kwestionariusz ma bardzo słabe przeciętne związki pomiędzy poszczególnymi pytaniami, a więc ma bardzo złą potocznie rozumianą spójność wewnętrzną, to duża liczba pytań tego kwestionariusza podwyższa silnie wartość współczynnika alfa do poziomu, który wydaje się już wysoki bądź bardzo wysoki. Zależność tę można wyrazić wzorem $alfa = 0{,}0211 + 0{,}1678 \times Mr + 0{,}0531 \times Mr - 0{,}0082 \times Mr$ ^2–$0{,}0032 \times Mr \times NI - 0{,}0011 \times NI$ ^2.

Tabela 3

Zależność między średnim poziomem korelacji składowych kwestionariusza (rM) i liczbą itemów/pytań kwestionariusza (NI) a średnim poziomem alfa Cronbacha – dokładne wyniki

					rM				
NI	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
I2	0,192	0,342	0,468	0,577	0,672	0,755	0,828	0,893	0,951
I3	0,265	0,439	0,571	0,673	0,755	0,822	0,878	0,926	0,967
I4	0,326	0,512	0,640	0,733	0,804	0,861	0,906	0,943	0,975
I5	0,378	0,568	0,690	0,775	0,837	0,885	0,923	0,954	0,980
I6	0,422	0,612	0,727	0,805	0,861	0,903	0,935	0,962	0,983
I7	0,461	0,648	0,757	0,828	0,878	0,915	0,944	0,967	0,986
I8	0,494	0,678	0,781	0,846	0,892	0,925	0,951	0,971	0,987
I9	0,524	0,703	0,800	0,861	0,903	0,933	0,956	0,974	0,989
I10	0,550	0,725	0,817	0,873	0,912	0,939	0,960	0,977	0,990
I11	0,574	0,744	0,831	0,883	0,919	0,944	0,964	0,979	0,991
I12	0,595	0,760	0,842	0,892	0,925	0,949	0,967	0,980	0,991
I13	0,614	0,774	0,853	0,900	0,931	0,953	0,969	0,982	0,992
I14	0,632	0,787	0,862	0,906	0,935	0,956	0,971	0,983	0,993
I15	0,648	0,798	0,870	0,912	0,939	0,959	0,973	0,984	0,993
I16	0,662	0,808	0,877	0,917	0,943	0,961	0,975	0,985	0,994
I17	0,676	0,818	0,883	0,921	0,946	0,963	0,976	0,986	0,994
I18	0,688	0,826	0,889	0,925	0,949	0,965	0,977	0,987	0,994
I19	0,700	0,834	0,894	0,929	0,951	0,967	0,979	0,988	0,995
I20	0,710	0,841	0,899	0,932	0,954	0,969	0,980	0,988	0,995

Dyskusja i wnioski

Przeprowadzone analizy wykazały, że na poziom alfa ma wpływ średnia korelacja między pozycjami oraz liczba itemów, a liczba osób badanych i długość użytej skali odpowiedzi ma znaczenie marginalne (właściwie żadne).

Brak wpływu liczebności grupy na poziom alfa można pośrednio przewidywać, biorąc pod uwagę przytoczone wcześniej wzory opisujące współczynnik alfa (Cortina, 1993; de Vet i in., 2017; Thompson, 2003). Ciekawe wydaje się, że uzyskane wyniki z naszego badania opartego na symulacjach rozkładów odpowiedzi badanych nie są zgodne z teoretycznym wywodem przeprowadzonym przez Bujang i in. (2018), którzy postulują, by określać wielkość próby (*sample size*) potrzebną do uzyskania satysfakcjonującej rzetelności badanego narzędzia. Nie wchodząc jednak w szczegóły, ich wywód teoretyczny dotyczy przede wszystkim małych prób ($N <= 10$), my natomiast skupiliśmy się tu na wielkościach próby większych od $N = 100$[7]. Podobny raport dla małych prób można znaleźć też u Šerbetara i Sedlar (2016).

Obserwowano brak wpływu długości skal odpowiedzi, ale należy także odnieść się do kwestii, czy alfa nadaje się do określenia spójności wewnętrznej (*internal consistency*) dla skal dychotomicznych (Barbaranelli i in., 2015; Pastore i Lombardi, 2014). Po raz kolejny należy zwrócić uwagę na fakt, że nie uzyskujemy tutaj odpowiedzi na pytanie, czy i jak długość skali oddziałuje z rozmaicie rozumianą w psychometrii *rzetelnością*, natomiast w prezentowanych wynikach dla poziomu alfa nie miało żadnego znaczenia to, czy skala była dychotomiczna, trzy-, pięcio- czy siedmiostopniowa, ponieważ podstawą tego wskaźnika jest po prostu (w uproszczeniu) średnia korelacja między pozycjami. Można z dużą dozą prawdopodobieństwa wskazać, że to związki między pozycjami determinują poziom współczynnika alfa, natomiast z perspektywy metodologicznej można co najwyżej zastanawiać się, czy po prostu w przypadku skal dychotomicznych nie jest trudniej uzyskać korelacje między pozycjami niż w przypadku skal szerszych, ale to już kwestia, której niniejszy artykuł nie obejmuje.

Naturalnie literatura zaleca użycie innych miar, jak Omega McDonalda, ale trudno powiedzieć, czy i kiedy zdobędą popularność, skoro ich użycie jest trudne bądź niemożliwe w obrębie popularnych pakietów statystycznych (Hayes i Coutts, 2020). Tak jak wcześniej wskazywaliśmy, naszym celem nie jest tutaj porównywanie alternatyw dla alfa. Czytelnik, który chciałby wybrać dla siebie najlepszy z kilkudziesięciu dostępnych współczynników, powinien raczej zapoznać się z pracą Cho (2022) lub Trizano-Hermosilla i Alvarado (2016).

Wydaje się ciekawe, że zależność między średnią korelacją a poziomem alfa jest tak silnie zakrzywiana przez liczbę itemów. Z praktycznego punktu widzenia rzetelność musi odnosić się do jakiegoś konstruktu psychologicznego o jednolitej definicji. Oznacza to, że pytania kwestionariuszowe powinny dotyczyć podobnego konstruktu lub jednolitego obszaru – wtedy możemy mówić ogólnie

[7] Należy oczywiście nadmienić, że w walidacji narzędzi badawczych zalecane są zwykle duże grupy, np. u Chartera (1999).

o *rzetelności*. Jednak w przypadku, gdy pytania właściwie nie są ze sobą powiązane (tj. odpowiedzi respondentów nie są ze sobą skorelowane), nadal można uzyskać wysoki poziom rzetelności dla całkowicie niepotwierdzonego konstruktu o nieznanym charakterze. Na podstawie przeprowadzonych obliczeń można powiedzieć, że przy związkach, które nawet w psychologii uznaje się często za pomijalne, współczynnik alfa jest na poziomie akceptowalnym już przy 20 itemach, a zwiększając liczbę pytań, nawet przy związkach bliskich zeru, można uzyskać niemal idealną rzetelność dla 90 itemów. Przykładem może być tutaj popularna w Polsce adaptacja kwestionariusza *Cor-Evaluation Hobfolla* (Gruszczyńska, 2012) złożonego z 90 pytań, które pozwalają uzyskać wynik ogólny i ten wskaźnik ogólny czytelnik mógłby uznać za *rzetelny*.

To, że liczba itemów jest tak istotna dla współczynnika alfa, wskazuje jego poważną wadę w pomiarze potocznie rozumianej *spójności wewnętrznej*. Może zdarzyć się, że kwestionariusz ma bliskie zeru związki w obrębie poszczególnych pozycji, ale uzyska on wysoki współczynnik alfa wyłącznie z powodu dużej liczby pytań. Dodać należy, iż w literaturze często zaleca się zwiększanie liczby pytań w kwestionariuszu badawczym, aby zwiększyć rzetelność pomiaru (stąd istnienie i powszechne stosowanie formuły proroczej Spearmana-Browna, która notabene miała korygować rzetelność długich testów). Niemniej jednak, jak pokazaliśmy, przy (podkreślmy to jeszcze raz) pomijalnych korelacjach pomiędzy pozycjami kwestionariusza nadal można uzyskać wysoką rzetelność, która będzie po prostu przeszacowana i błędna.

Zależność między potocznie rozumianą *spójnością wewnętrzną* a wynikiem alfa jest w zasadzie prosta do określenia – jest to zależność krzywoliniowa, która wskazuje, że już przy słabych korelacjach w obrębie narzędzia pomiarowego (ok. 0,30) można uzyskać dość satysfakcjonującą rzetelność (bliską alfa = 0,80), a przy korelacjach 0,50 bardzo wysoki współczynnik rzetelności (bliski alfa = 0,90). Naturalnie nie chcemy tu dyskutować kiedy korelacje są „silne", a rzetelność „satysfakcjonująca" – chcemy tylko zwrócić uwagę, że alfa nie jest dobrą miarą *spójności wewnętrznej*, bo wyraźnie ją przeszacowuje. Wydaje się równie ciekawe, że alfa jako miara rzetelności rozumiana jako *spójność wewnętrzna* (*internal consistency*) przeszacowuje tę właśnie spójność wewnętrzną. Być może popularność tej metody nie wynika z jej prostoty użycia czy dostępności w SPSS (co sugeruje Borsboom, 2006)? Można tu także zadać pytanie, czy popularność alfa nie wynika z tego, iż bardzo łatwo wykazać, że tworzone narzędzie badawcze jest „rzetelne" (cokolwiek by to nie znaczyło), że ma wysoką miarę *alfa*, nawet jeśli *spójność wewnętrzna* jest słaba lub właściwie zerowa/pomijalna?

Przedstawione przez nas rezultaty mogą przyczynić się do bardziej starannego określania rzetelności i bardziej świadomego podejścia do szacowania błędu pomiaru. Na podstawie załączonego ostatniego wykresu lub tabeli można dokonać w miarę prostej predykcji podczas planowania i konstrukcji własnego narzędzia pomiarowego – dzięki niemu znając przeciętny poziom korelacji, łatwo można określić liczbę pozycji kwestionariusza potrzebną, aby uzyskać pożądany poziom alfa. Może być on także użyty do oceny wstecznej jakości danego narzędzia badawczego – znając poziom alfa i liczbę itemów, można wstecznie określić przeciętny poziom korelacji między pozycjami kwestionariusza (oczywiście z pewnym przybliżeniem).

Podsumowując, długość skali oraz liczba osób badanych (powyżej 100) nie ma specjalnego znaczenia dla określenia poziomu rzetelności rozumianej jako *internal consistency*, a współczynnik alfa jest wyraźnie przeszacowywany na skutek większej liczby pytań kwestionariusza. Jest to uwaga krytyczna względem samego współczynnika *alfa*, ale może być wykorzystana w planowaniu tworzenia narzędzi pomiarowych, które z powodu trudno obserwowalnych cech latentnych mogą wymagać po prostu zwiększenia liczby pytań w celu poprawy parametrów kwestionariusza (to zresztą sugeruje np. praca Hoyt i in., 2006). Sugerujemy jednak ostrożność w przypadku przedstawiania w piśmiennictwie tzw. dobrych miar alfa dla tych narzędzi, które po prostu mają dużo itemów.

Dlatego też jako rozwiązanie problemu chcemy zaproponować, aby w praktyce razem z miarami rzetelności raportować średnie korelacje. Niska średnia korelacja między pozycjami, wysoka rzetelność i duża liczba pytań mogą sugerować wykazany tu błąd inflacji alfa. Do oceny *spójności wewnętrznej* zalecamy raczej średni poziom korelacji (co jest zgodne z zaleceniami Cho i Kim, 2015). Biorąc pod uwagę jego popularność, sugerujemy pewną ostrożność w określaniu rzetelności za pomocą współczynnika alfa, w przypadku więcej niż 8 pozycji używanie alternatywnych miar, lub po prostu przyjęcie perspektywy wskazującej, że alfa Cronbacha nie jest przeznaczona do określania *wewnętrznej spójności*. Wreszcie w przypadku narzędzi składających się z więcej niż 20 pozycji w jednej skali zalecamy zastosowanie wieloaspektowego i ostrożniejszego podejścia do testowania rzetelności niż przeprowadzanie wyłącznie analizy z użyciem alfa Cronbacha.

W duchu uprawiania otwartej nauki autorzy artykułu zachęcają do zapoznania się z otwartymi danymi badawczymi dostępnymi w repozytorium cyfrowym pod tym adresem: https://tiny.pl/dttmk

Bibliografia

Anselmi, P., Colledani, D., Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, *10*, artykuł 2714. https://doi.org/10.3389/fpsyg.2019.02714

Bajpai, S., Bajpai, R. (2014). Goodness of measurement: Reliability and validity. *International Journal of Medical Science and Public Health*, *3*(2), 112–115. https://doi.org/10.5455/ijmsph.2013.191120133

Barbaranelli, C., Lee, C. S., Vellone, E., Riegel, B. (2015). The problem with Cronbach's alpha: comment on Sijtsma and van der Ark (2015). *Nursing Research*, *64*(2), 140–145. https://doi.org/10.1097/NNR.0000000000000079

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137–143. https://doi.org/10.1007/s11336-008-9100-1

Bland, J. M., Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *British Medical Journal*, *314*(7080), artykuł 572. https://doi.org/10.1136/bmj.314.7080.572

Bonett, D. G., Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, *36*(1), 3–15. https://doi.org/10.1002/job.1960

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D., Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505–514. https://doi.org/10.1016/S0160-2896(02)00082-X

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, *16*(3), 201–213. https://doi.org/10.1002/job.4030160303

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, *107*(2), 260–273. https://doi.org/10.1037/0033-2909.107.2.260

Bujang, M. A., Omar, E. D., Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: a simple guide for researchers. *The Malaysian Journal Of Medical Sciences: MJMS*, *25*(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Chan, E. K. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. W: *Validity and validation in social, behavioral, and health sciences* (s. 9–24). Springer.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, *18*(3), 205–215. https://doi.org/10.1177/014662169401800302

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566. https://doi.org/10.1007/978-3-319-07794-9_2

Cho, E. (2022). The accuracy of reliability coefficients: A reanalysis of existing simulations. *Psychological Methods*. Online first. https://doi.org/10.1037/met0000475

Cho, E., Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Comrey, A. L., Lee, H. B. (1992). *A first course in factor analysis* (wyd. 2). Erlbaum.

de Vet, H. C., Mokkink, L. B., Mosmuller, D. G., Terwee, C. B. (2017). Spearman–Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*, 45–49. https://doi.org/10.1016/j.jclinepi.2017.01.013

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), 50–59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Dimov, I. T. (2008). *Monte Carlo methods for applied scientists*. World Scientific. https://doi.org/10.1142/9789812779892

Duhachek, A., Coughlan, A. T., Iacobucci, D. (2005). Results on the standard error of the coefficient alpha index of reliability. *Marketing Science*, *24*(2), 294–301. https://doi.org/10.1287/mksc.1040.0097

Dunn, T. J., Baguley, T., Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Dunn, W. L., Shultis, J. K. (2011). *Exploring Monte Carlo methods*. Elsevier. https://doi.org/10.1016/B978-0-444-51575-9.00007-5

Eisinga, R., Grotenhuis, M. T., Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*(4), 637–642. https://doi.org/10.1007/s00038-012-0416-3

Flake, J. K., Pek, J., Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Gignac, G. E., Bates, T. C., Jang, K. L. (2007). Implications relevant to CFA model misfit, reliability, and the five-factor model as measured by the NEO-FFI. *Personality and Individual Differences*, *43*(5), 1051–1062. https://doi.org/10.1016/j.paid.2007.02.024

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597–607.

Green, S. B., Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Gruszczyńska, E. (2012). Kwestionariusz Samooceny Zysków i Strat – polska adaptacja Cor-Evaluation Se Hobfolla i jej podstawowe właściwości psychometryczne. W: E. Bielawska-Batorowicz i B. Dudek (red.), *Teoria zachowania zasobów Stevana E. Hobfolla. Polskie doświadczenia*. Wydawnictwo Uniwersytetu Łódzkiego.

Guidroz, A. M., Yankelevich, M., Barger, P., Gillespie, M. A., Zickar, M. J. (2009). Practical considerations for creating and using organizational survey norms: Lessons from two long-term projects. *Consulting Psychology Journal: Practice and Research*, *61*(2), 85–102. https://doi.org/10.1037/a0015969

Hayes, A. F., Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But… *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189. https://doi.org/10.1080/07481756.2002.12069034

Hoyt, W. T., Warbasse, R. E., Chu, E. Y. (2006). Construct validation in counseling psychology research. *The Counseling Psychologist*, *34*(6), 769–805. https://doi.org/10.1177/0011000006287389

Jolliffe, I. T., Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), artykuł 20150202. https://doi.org/10.1098/rsta.2015.0202

Kalkbrenner, M. T. (2023). Alpha, Omega, and *H* internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, *14*(1), 77–88. https://doi.org/10.1080/21501378.2021.1940118

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research*, *37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Li, H., Rosenthal, R., Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1*(1), 98–107. https://doi.org/10.1037/1082-989X.1.1.98

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, *29*(1), 65–81. https://doi.org/10.1177/0146621604270882

Macey, W. H., Eldridge, L. D. (2006). National norms versus consortium data: What do they tell us. W: A. I. Kraut (red.), *Getting action from organizational surveys: New concepts, technologies, and applications* (s. 352–376). Jossey-Bass.

Matell, M. S., Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology*, *56*(6), 506–509. https://doi.org/10.1037/h0033601

McCrae, R. R., Kurtz, J. E., Yamagata, S., Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Metsämuuronen, J. (2022). The effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability: Seeking the best options of correlation for deflation-corrected reliability. *Behaviormetrika*, *49*(1), 91–130. https://doi.org/10.1007/s41237-022-00158-y

Pastore, M., Lombardi, L. (2014). The impact of faking on Cronbach's alpha for dichotomous and ordered rating scores. *Quality & Quantity*, *48*(3), 1191–1211. https://doi.org/10.1007/s11135-013-9829-1

Ponterotto, J. G., Ruckdeschel, D. E. (2007). An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual and Motor Skills*, *105*(3), 997–1014. https://doi.org/10.2466/pms.105.3.997-1014

Preston, C. C., Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Raykov, T., Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. https://doi.org/10.4324/9780203841624

Revelle, W., Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Šerbetar, I., Sedlar, I. (2016). Assessing reliability of a multi-dimensional scale by coefficient alpha. *Journal of Elementary Education*, *9*(1/2), 189–196.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K. (2020). *Measurement Models for Psychological Attributes: Classical Test Theory, Factor Analysis, Item Response Theory, and Latent Class Models*. CRC Press. https://doi.org/10.1201/9780429112447-2

Sijtsma, K., Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*(4), 843–860. https://doi.org/10.1007/s11336-021-09789-8

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Tabachnick, B. G., Fidell, L. S. (2007). *Using multivariate statistics* (wyd. 5). Allyn & Bacon.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Taherdoost, H. (2022). What is the best response scale for survey and questionnaire design; review of different lengths of rating scale / attitude scale / Likert scale. *International Journal of Academic Research in Management*, *8*(1), 1–10.

Tang, W., Cui, Y., Babenko, O. (2014). Internal consistency: Do we really know what it is and how to assess it. *Journal of Psychology and Behavioral Science*, *2*(2), 205–220.

Ten Berge, J. M., Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–625. https://doi.org/10.1007/BF02289858

Thall, P. F., Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, *46*(3), 657–671. https://doi.org/10.2307/2532086

Thigpen, N. N., Kappenman, E. S., Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123–138. https://doi.org/10.1111/psyp.12629

Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues* (wyd. 1). Sage Publications, Inc. https://doi.org/10.4135/9781412985789.n1

Trizano-Hermosilla, I., Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*, artykuł 769. https://doi.org/10.3389/fpsyg.2016.00769

Vaske, J. J., Beaman, J., Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences*, *39*(2), 163–173. https://doi.org/10.1080/01490400.2015.1127189

Vehkalahti, K., Puntanen, S., Tarkkonen, L. (2006). *Estimation of reliability: a better alternative for Cronbach's alpha*. Department of Mathematics and Statistics, University of Helsinki.

Zumbo, B. D., Chan, E. K. (2014). *Validity and validation in social, behavioral, and health sciences*. Social Indicators Research Series, Vol. 54. Springer. https://doi.org/10.1007/978-3-319-07794-9