Možnosti, problémy a odporúčania pri meraní zhody medzi posudzovateľmi – deskriptívny prístup

LUCIA KOČIŠOVÁ[*]
Ústav experimentálnej psychológie, Slovenská akadémia vied, Bratislava

Options, Problems and Guidelines for Measuring Interrater Agreement – a Descriptive Approach

Abstract: Interrater agreement is one way to establish reliability (and also validity) in social science research. The traditionally preferred method of measuring interrater agreement is the descriptive approach owing to its simplicity. This approach is also associated with a number of different agreement indices, which makes it difficult to select the right index. This article summarises theoretical definition on the prevailing approach used to measure interrater agreement (in both quantitative and qualitative research). From a practical point of view, the article focuses on the possibilities of measuring agreement by using percent agreement, the kappa coefficient, and the AC_1 coefficient. A more detailed description of the indices explains how to define, calculate, and interpret them and the problems associated with their use. The indices are then discussed in comparison. Although underestimated and criticised, percent agreement may be a good indicator of interrater agreement. Several paradoxes accompany the use of the kappa coefficient, which is only possible under certain conditions. The appropriate alternative to it is the AC_1 coefficient. The article concludes with a summary of recommendations for improving the quantification of interrater agreement.

Keywords: interrater agreement, agreement index, percent agreement, kappa coefficient, AC_1 coefficient

Sociologický časopis / Czech Sociological Review, 2025, Vol. 61, No. 3: 277–300
Uverejnené online: 6. júna 2024, https://doi.org/10.13060/csr.2024.007

Úvod

Reliabilita merania je jednou zo základných charakteristík merania v kvantitatívnom sociálnovednom výskume. Dôkazy o reliabilite (spolu s dôkazmi o validite) sú predpokladom na zabezpečenie kvality meracieho nástroja, čo odkazuje na dôležitosť tejto témy v akomkoľvek výskume. Naše meranie môže byť skreslené rôznymi chybami, mieru ktorých sa snažíme zistiť. Kým validita je o chybách

* Všetku korešpondenciu posielajte na adresu: Lucia Kočišová, Ústav experimentálnej psychológie, Centrum spoločenských a psychologických vied, Slovenská akadémia vied, Dúbravská cesta 9, 841 04 Bratislava, Slovenská republika, e-mail: lucia.kocisova@savba.sk.

konštantných, reliabilita sa týka náhodných chýb merania a meranie považujeme za reliabilné, ak je miera náhodných chýb čo najnižšia[1].

Náhodnou chybou môže byť v podstate čokoľvek, keďže ju nevieme predpokladať a v rámci rôznych spôsobov odhadu reliability je aj jej definovanie rôzne. Reliabilita je následne označovaná ako stabilita testu v čase (odhad test-retestom), ako ekvivalencia (odhad paralelnými formami merania), či ako vnútorná konzistencia. Posledným odhadom reliability je použitie viacerých posudzovateľov[2] (inter-rater reliability), čo sa uplatňuje v procese skórovania, posudzovania či hodnotenia správania, vlastností alebo javov. Zdrojom náhodnej chyby sú posudzovatelia a reliabilita je označovaná ako zhoda medzi posudzovateľmi (inter-rater agreement) a konzistencia posudzovateľov (inter-rater consistency[3]).

Rozšíreným prístupom k meraniu zhody posudzovateľov je klasický deskriptívny prístup, ktorý zahŕňa všetky koeficienty/testy vhodné pre meranie súhlasu medzi dvomi či viacerými posudzovateľmi. Oproti modelovému prístupu je uprednostňovaný pre jeho jednoduchosť, ktorá podľa Xieho (2013) je však zdanlivá, pretože napriek jednoduchosti dochádza v niektorých prípadoch k nesprávnym postupom vo výpočtoch či k nesprávnej interpretácii uvádzaných hodnôt zhody. Zároveň tu vládne nekonzistentnosť v definovaní, vo vymedzení sa voči konzistencii či v používaných indexoch ako aj nekonzistentnosť v odporúčaných indexoch zhody.

Cieľom článku je sumarizovať informácie o deskriptívnom prístupe k meraniu zhody medzi posudzovateľmi. V úvode bude pozornosť venovaná teoretickému vymedzeniu zhody medzi posudzovateľmi s dôrazom na odlíšenie od konzistencie posudzovateľov. Následne sa pozornosť zameria na meranie zhody, pričom podrobnejšie budú popísané vybrané tri indexy zhody: *percentuálny súhlas* ako najjednoduchší index zhody, ktorý má aj napriek kritike svoj potenciál, *koeficient kappa* ako najčastejšie používaný index zhody, ktorý však podlieha skresleniam a tie nebývajú vo výskumoch reflektované a nakoniec *koeficient* AC_1,

[1] Tento predpoklad je spájaný s klasickou teóriou testov (classic test theory, CTT), kde skóre merania je sumou pravdivého skóre (true score) a chyby merania (error) a reliabilita je spájaná s relatívnou neprítomnosťou chýb merania, ktoré sú náhodné (pozri napr. Cígler a Šmíra, 2015; Řehák, 1998; Urbánek et al., 2011).

[2] Slovo posudzovateľ býva v rôznom kontexte nahrádzané iným pomenovaním ako napr. pozorovateľ, hodnotiteľ, tester, kóder, výskumník a pod. (pozri napr. aj Zhao et al., 2022). Aj keď rozdielnosť pojmov nemusí byť nutne kontraproduktívna, pretože konkrétne pomenovanie súvisí s konkrétnou výskumnou situáciou, problémom môže byť orientovanie sa v problematike a ťažkosti pri nachádzaní literatúry, a tak Kottner et al. (2011) navrhujú používať termín *rater*, pretože sa zdá byť vhodným pre charakterizovanie širokého spektra situácií, v ktorých posudzovatelia hodnotia iné osoby alebo objekty.

[3] Najčastejšie sa v angličtine stretneme s pojmom inter-rater reliability a inter-rater agreement. Pojem inter-rater consistency je menej častý, ale zdá sa nám vhodnejší pre rozlišovanie medzi meraním konzistencie ako jedného zo spôsobov zisťovania spoľahlivosti posudzovateľov od pojmu spoľahlivosť posudzovateľov, ktorú chápeme ako všeobecnejší pojem zahŕňajúci aj zhodu medzi posudzovateľmi.

ktorý je aktuálne najviac preferovaným indexom zhody. Po deskripcii uvedených indexov zhody bude nasledovať diskusia o porovnaní indexov.

Aj keď článok primárne neprináša zásadné nové informácie, na Slovensku a v Čechách nie sú veľmi rozšírené a tak je jeho poslanie zlepšovať informovanosť výskumníkov o zhode medzi posudzovateľmi.

Vymedzenie zhody medzi posudzovateľmi

Zhoda medzi posudzovateľmi predstavuje jeden zo spôsobov ako je zabezpečovaná reliabilita v prípade použitia dvoch alebo viacerých hodnotiteľov. Ak je posudzovateľ len jeden, potom hovoríme o intra- variante, podľa Gweta (2021) o seba-replikovateľnosti a v prípade ak máme namiesto posudzovateľov dve a viac opakovateľných meraní rozložených v čase, potom ide o replikovateľnosť merania, teda test-retest reliabilitu.

Predstavme si jednoduchý príklad. Dvaja posudzovatelia pozorujú rovnaký objekt. Ich správa o tom čo vidia by sa v ideálnom prípade mala zhodovať. Ak sa líši, predpokladá sa, že sú za to zodpovedné vlastnosti posudzovateľov. Tieto vlastnosti sú teda zdrojom chýb a výsledné hodnotenie je to určitej miery nepresné. Nedostatok rozdielov medzi posudzovateľmi predstavuje mieru spoľahlivosti a naopak, miera rozdielov medzi nimi prestavuje mieru nespoľahlivosti. Kvantifikácia týchto rozdielov sa tak stáva dôležitou súčasťou hodnotenia spoľahlivosti medzi hodnotiteľmi (DeVellis, 2005). Dosiahnuť dokonalú zhodu nie je vždy možné a dokonca ani potrebné. Extrémna nezhoda je štatisticky takmer rovnako neočakávaná ako dokonalá zhoda a nemala by sa vyskytovať, ak posudzovatelia používajú rovnaké inštrukcie na posudzovanie a pracujú nezávisle od seba (Krippendorff, 2004).

Vzhľadom k rozsiahlym možnostiam použitia sa zhoda medzi posudzovateľmi objavuje v mnohých vedných odboroch (najčastejšie je to v sociálnych vedách, v medicíne, v ošetrovateľstve, v mediálnom a komunikačnom odbore a ďalších) a existuje aj množstvo teoreticky a empiricky ladených článkov, kapitol v knihách, či priamo kníh, ktoré sa tejto téme venujú[4]. Tým sa sťažuje orientácia v tom, čo vlastne zhoda medzi posudzovateľmi vyjadruje a ktorý koeficient je najvhodnejší pre jej meranie. Nájsť konsenzus v takom množstve informácií je častokrát náročné a dá sa predpokladať že niekedy aj nemožné. Je však možné hľadať rovnaké či podobné vyjadrenia a empirické dôkazy či priznať nekonzistentnosť v tom, čo výskumníci tvrdia.

[4] Pre orientovanie sa vo vymedzení a rozdieloch medzi súhlasom posudzovateľov a konzistenciou posudzovateľov odporúčame články od Tinsleyho a Weissa (1975), LeBretona a Sentera (2008), Kottnera a Streinera (2011), Kottnera et al. (2011) či kapitolu od DeVellisa (2005). Popis viacerých indexov prinášajú knihy od Gwetta (2021) či od Von Eyeho a Muna (2005). Kritické zhodnotenie a porovnávanie viacerých indexov ponúkajú články od Fenga (2013, 2015) či Zhaoa a kolegov (2013, 2022).

Definícia

Zhoda medzi posudzovateľmi je rôznymi výskumníkmi vzťahovaná rôzne k reliabilite: niektorí hovoria o dimenzii reliability (Wilhelm et al., 2018) a niektorí o indikátore reliability (Lombard et al., 2002). Krippendorff (2004) uvádza, že súhlas je to čo meriame a reliabilita je to, čo chceme z neho odvodiť a súhlas vníma ako index reliability. Zhoda medzi posudzovateľmi tak nie je priamo reliabilita, ale len spôsobom, akým sa dá reliabilita merania zisťovať – odhadnúť.

Zhoda medzi posudzovateľmi je vymedzovaná ako absolútny konsenzus vo viacnásobnom hodnotení (Gisev et al., 2013; LeBreton a Senter, 2008). Keďže ide o kvantitatívnu mieru, súhlas medzi posudzovateľmi znamená, že hodnotitelia používajú práve takú istú hodnotu pri hodnotení, takže sú zameniteľní (Kozlowski a Hattrup, 1992; Tinsley a Weiss, 1975) a zdieľajú všeobecnú interpretáciu hodnoteného konštruktu (Stemler, 2004).

Súhlas posudzovateľov smeruje k otázke, či je hodnotenie identické, podobné alebo do akej miery sa líši. V tejto situácii nás zaujíma absolútna miera chyby merania a akákoľvek variabilita medzi subjektmi alebo distribúcia hodnotenej vlastnosti v populácii nie je dôležitá (Kottner a Streiner, 2011). Ak chceme teda zistiť, aký je súhlas medzi rôznymi hodnotiteľmi, alebo opakovanými meraniami, odlišnými podmienkami, či medzi hodnotiteľom v rôznom čase, potom by sme sa mali vydať touto cestou (de Vet et al., 2006).

Súhlas nie je konzistencia

Súhlas posudzovateľov a konzistencia posudzovateľov[5] sa od seba líšia. Kým súhlas je o priraďovaní rovnakej hodnoty (ekvivalencia skóre), konzistencia je o zoraďovaní rovnakým spôsobom (ekvivalencia poradia) (LeBreton a Senter, 2008). Podľa Laury Goodwin (2001) predstavuje zhoda medzi posudzovateľmi absolútnu reliabilitu a konzistencia relatívnu reliabilitu. Pri odhadoch konzistencie nie je potrebné aby posudzovatelia zdieľali rovnakú interpretáciu posudzovacej škály, pokiaľ je každý hodnotiteľ konzistentný v tom ako posudzuje daný konštrukt či objekt/subjekt na základe vlastnej definície škály.

Keď je našou výskumnou otázkou zistiť, ako konzistentné je meranie, prípadne ako dobre je možné objekty/subjekty výskumu od seba odlíšiť aj napriek chybám v meraní, potom by sme si mali vybrať spoľahlivosť posudzovateľov (de Vet et al., 2006).

Spoľahlivosť a súhlas posudzovateľov môžu predstavovať dva prístupy k dátam výskumníka. Prvým z nich je vnímanie oboch pojmov cez prizmu kvality výskumu, s čím sú spojené spôsoby odhadu či stanovenia psychometrických kvalít ako je reliabilita a validita. Druhý spôsob použitia je spojený so štatistickým

[5] Pojem konzistencia používa napr. Stemler (2004), ale inak sa väčšinou uvádza pojem spoľahlivosť posudzovateľov (angl. inter-rater reliability).

spracovaním dát, keď primárnym záujmom výskumníka je zistiť, či sa posudzovatelia zhodnú, prípadne či sú konzistentní vo svojich hodnoteniach (Litwin, 1995).

Ešte na pripomenutie: súhlas a konzistencia posudzovateľov nie sú fixnou vlastnosťou výskumného nástroja, ale sú produktom interakcie medzi nástrojom, subjektom a kontextom hodnotenia (Kottner et al., 2011) alebo inak povedané sú vlastnosťou testovej situácie (Stemler, 2004). Podľa de Vet et al. (2006) parametre súhlasu budú stabilnejšie medzi odlišnými výskumnými vzorkami, kdežto parametre reliability sú veľmi závislé na variancii výskumnej vzorky a sú generalizovateľné len na vzorky s podobnou varianciou. Presnejšie povedané, spoľahlivosť posudzovateľov je charakteristikou použitého meracieho nástroja v konkrétnej výskumnej vzorke a súhlas posudzovateľov je viac charakteristikou samotného meracieho nástroja.

Takéto jasné rozlišovanie medzi súhlasom a konzistenciou nie je spoločné u všetkých výskumníkov a pojmy sa používajú zameniteľne. Napríklad Gwet (2021), ktorý sa tejto téme venuje už dlhšie a je autorom viacerých výskumných článkov, kníh ako aj autorom jedného z indexov zhody, používa oba pojmy zameniteľne a to z dôvodu, že dôvody od Tinsleyho a Weissa (1975) sa nevzťahujú na problém, ktorým sa on zaoberal (pozn. autora: problém reprodukovateľnosti) a iní autori túto problematiku moc neobjasňujú. K nerozlišovaniu pojmov prispieva aj nedostatočné terminologické rozlišovanie či už v anglickom jazyku ale aj na Slovensku či v Čechách. V tejto záležitosti sa prikláňame k dôslednému rozlišovaniu pojmov zhody medzi posudzovateľmi a konzistencii posudzovateľov, pretože oba koncepty sa líšia v tom, ako postupujú pri definovaní vzájomnej podobnosti, sú určené na zodpovedanie odlišnej výskumnej otázky a líšia sa aj v použitých štatistických postupoch, s čím súvisí aj adekvátna interpretácia (Kottner a Streiner, 2011; LeBreton a Senter, 2008).

Kvantitatívny výskum a kvalitatívny výskum

V rámci metodológie výskumu zasahuje zhoda medzi posudzovateľmi do viacerých oblastí. V kvantitatívnom výskume je okrem reliability jedným zo spôsobov kvantifikácie obsahových dôkazov validity (pozri napr. Martončik, 2019) ale zároveň je aj postupom ako je možné agregovať dáta pri viacúrovňovom výskume (napr. Bliese, 2000; Chan, 1998; Kozlowski a Klein, 2000; LeBreton a Senter, 2008).

V kvalitatívnom výskume sa zhoda medzi posudzovateľmi môže uplatňovať v situáciách pozorovania, hodnotenia, ale predovšetkým pri kódovaní kvalitatívnych dát. Postoj k jej používaniu je rôzny, čo súvisí aj s rôznym postojom výskumníkov ku zaisťovaniu kritérií kvalitatívneho výskumu. Na jednej strane sú tu výskumníci, ktorí úplne odmietajú používanie ďalšieho posudzovateľa/kódera, čo vysvetľujú práve odlišnými ontologickými a epistemologickými východiskami tohto výskumu. Na druhej strane sú výskumníci, ktorí vnímajú zaisťovanie kvality výskumu ako dôležité a aplikujú tak rôzne stratégie, ktoré zvyšujú dôveryhodnosť vo výskumné zistenia. Tieto stratégie sa nazývajú ako verifikačné stratégie a podľa delenia Lincolnovej a Guby (1985) ide o zaisťovanie štyroch kri-

térií: dôveryhodnosť (credibility), prenositeľnosť (transferability), spoľahlivosť (dependability) a potvrditeľnosť (confirmability). V takomto prípade je potom zhoda medzi posudzovateľmi používaná pri technike audit kolegov (peer auditing), ktorá slúži pre zaisťovanie dôveryhodnosti. O dôveryhodnosti hovoríme potom vtedy, ak sa nezávislí skúsení experti v téme zhodnú pri posudzovaní záverov výskumu (Švaříček, Šeďová a kol., 2007). Iné uplatnenie zhody medzi posudzovateľmi je pri kódovaní kvalitatívnych dát, čo je spôsob, akým sa zaisťuje spoľahlivosť. Použitie ďalších kóderov alebo pozorovateľov je vlastne spôsob triangulácie výskumníkov a táto stratégia sa používa pre zaistenie dôveryhodnosti, spoľahlivosti ale aj potvrditeľnosti (Krefting, 1991).

Síce v tejto skupine výskumníkov panuje zhoda, že by sa mal využívať ďalší posudzovateľ/kóder, nepanuje zhoda v tom, či je nutné ju aj kvantifikovať. Podľa niektorých výskumníkov môže byť uvádzanie hodnoty zhody mätúce, pretože kvantifikácia patrí do kvantitatívneho výskumu a dôležitým je predsa proces posudzovania, kódovania. Na druhej strane sú výskumníci, ktorí požadujú kvantifikovanie zhody za spôsob ochrany pred systematickým skreslením, čo nám môže pomôcť odhaliť nedostatky v definíciách, kódovaní, prekrývaní sa významu, či dokonca problémy pri dosahovaní zhody vzhľadom na povahu údajov (McDonald et al., 2019).

Rozhodnutie o tom, či vôbec použiť alebo nepoužiť zhodu v kvalitatívnom výskume vychádza najmä z prístupu akým získavame či analyzujeme dáta. Prizvanie ďalšieho výskumníka do týchto procesov a zisťovanie zhody môže byť vhodné, ak máme veľa dát, či potrebujeme potvrdiť konzistenciu nášho merania alebo identifikovať, kde skupina výskumníkov nesúhlasí v interpretácii. Ak však je naše kódovanie príliš jednoduché, či výskumník je expertom, alebo samotná zhoda nie je pre nás podstatná a kódy sú procesom nie produktom, potom prizvanie ďalšieho posudzovateľa nie je potrebné (McDonald et al., 2019).

Meranie zhody medzi posudzovateľmi

Zhoda medzi posudzovateľmi je merateľná cez indexy zhody. Prvé indexy zhody sú dokumentovateľné už viac ako 120 rokov[6] a odvtedy vzniklo množstvo indexov, ktoré sa väčšinou snažili prekonať problémy už existujúceho indexu. Zaradiť sem môžeme podiel celkového súhlasu (%), podiel pozitívneho súhlasu, podiel negatívneho súhlasu, Cohenov koeficient kappa κ, vážený koeficient kappa κ_w, Fleissov koeficient kappa κ_{Fleiss}, Congerov exaktný koeficient kappa κ_{Exact}, Lightov koeficient kappa κ_{Light}, Brennan-Predigerov koeficient kappa κ_{BP}, Krippendorffov koeficient alfa α, Scottov koeficient π, koeficient S Benneta, Alperta a Goldsteina, Jaccardov koeficient J, Stuart-Maxwellov test χ^2, McNemarov test χ^2, Bhapkar test χ^2, Gwetov koeficient AC_1 a AC_2, Kendallov koeficient τ, Kendallov koeficient W, r_{wg} koeficient Jamesa, Demareeho a Wolfa ale aj t-test pre dva závislé výbery, priemernú od-

[6] Benini už v roku 1901 dokumentoval percentuálny súhlas a predstavil koeficient β (in Zhao et al., 2018).

chýlku AD, pozorovateľnú variabilitu dát (napr. štandardná odchýlka) alebo test ANOVA (two-way). Indexy sa líšia v tom, pre aký typ dát a pre aký počet posudzovateľov sú vhodné (pozri napr. Kočišová, 2022). Zároveň sa dajú veľmi nahrubo rozdeliť na tie, ktoré sú bez korekcie a tie, ktoré berú do úvahy náhodný súhlas[7].

V súčasnosti sa preferujú práve indexy zhody, ktoré korigujú náhodný súhlas, a aj keď sa výskumníci nezhodnú na tom, ako je definovaná zhoda medzi posudzovateľmi, či na výbere merania zhody, zhodnú sa, že pri meraní je potrebné korigovať súhlas o náhodu, ktorá sa pri meraní môže vyskytnúť (napr. Feng, 2015; Zhao et al., 2018). Vo vyššie uvedenom jednoduchom príklade môže jeden z posudzovateľov alebo aj dokonca obaja udeliť náhodné hodnotenie v dôsledku čoho sa vyskytne náhodný súhlas. Takéto náhodné hodnotenie nastáva, ak si hodnotiteľ nie je istý ako klasifikovať hodnotený objekt, k čomu môže dôjsť vtedy, keď charakteristiky nezodpovedajú pokynom hodnotenia. A aj keď posudzovateľ uvedie náhodné hodnotenie, stále je tu šanca, že bude v hodnotení dosiahnutý súhlas. Tento typ zhody je však nepredvídateľný a ťažko odôvodniteľný a je zrejmé, že to nie je spôsob, akým by výskumník chcel, aby sa posudzovatelia zhodli. Preto je náhodná zhoda nežiadúca, pretože ju nemožno považovať za dôkaz toho, že posudzovatelia zvládajú hodnotiaci proces (Gwet, 2021).

Pre vysvetlenie náhodného súhlasu je možné použiť analógiu dvoch posudzovateľov, ktorí losujú z urny plnej bielych a čiernych guličiek. Povedzme že posudzujú diagnózu pacienta. Ak si vytiahnu obaja čiernu guličku, zhodnú sa, že pacient je pozitívny, ak si vytiahnu bielu, zhodnú sa, že pacient je negatívny. V oboch prípadoch je to náhodná zhoda, pretože je to bez toho, aby sa pozreli na pacienta alebo aby nahliadli do jeho záznamov (Zhao, 2011).

Korigovanie náhodného súhlasu je v súčasnosti preferované, avšak vhodným indexom[8] môže byť aj ten najjednoduchší index zhody – percentuálny súhlas, čo sa ukazuje vo výskumoch porovnávajúcich rôzne indexy zhody (Zhao et al., 2013; Zhao et al., 2022). Práve z tohto dôvodu je mu venovaný priestor aj v tomto článku. Následne sa venujeme Cohenovmu koeficientu kappa, ktorý je najznámejším a najpoužívanejším indexom zhody, ale ktorý zároveň podlieha aj viacerým skresleniam. Nakoniec popisujeme Gwetov koeficient AC_1, ktorý v súčasnosti mnohí uprednostňujú ako lepšiu voľbu než kritizovaný koeficient kappa.

Percentuálny súhlas

Percentuálny súhlas[9] je najjednoduchším spôsobom ako vyjadriť zhodu medzi posudzovateľmi. Medzi jeho výhody patrí ľahký výpočet, intuitívnosť a zrozumiteľnosť výsledného indexu (Stemler, 2004). Zároveň je flexibilný a je možné

[7] Pozri napr. Zhao et al. (2013).

[8] Vhodným v zmysle, že menej podlieha skresleniam ako by sme očakávali na základe toho, že nekoriguje náhodný súhlas.

[9] Autor percentuálneho súhlasu nie je známy a niekedy je možné stretnúť sa s ním pod pojmom Osgoodov koeficient alebo Holstiho koeficient CR (Krippendorff, 2004).

ho použiť pri nominálnych, ordinálnych či kvantitatívnych premenných (Syed a Nelson, 2015). Vzorec pre výpočet vyzerá nasledovne:

$$P_A^{10} = \frac{N_A}{N_A + N_D} * 100 \quad (1),^{10}$$

kde N_A je celkovým súhlasom (agreement) a N_D je celkovým nesúhlasom (disagreement). V podstate ide o jednoduchý pomer počtu položiek, v ktorých posudzovatelia súhlasili, voči celkovému počtu položiek, z čoho plynie aj interpretácia tohto koeficientu. Ak zistíme zhodu nižšiu ako 100 %, vieme z nej odvodiť množstvo prípadov či pozorovaní v ktorých sa hodnotitelia nezhodli[11]. Rozpätie hodnôt pri percentuálnej zhode je od 0 % (ak nie je zaznamenaný žiadny súhlas) do 100 % (perfektný súhlas).

Ako referenčná hodnota pre posúdenie dostatočnosti miery percentuálnej zhody (Hartmann et al., 2004) v prípade komplexnejších meraní je úroveň 70 %[12] a inak je dostatočná zhoda medzi posudzovateľmi na úrovni 80–90 %.

Celá táto úvaha je aplikovateľná ak je hodnotenie binárne (v zmysle hodnôt 0–1) a počet posudzovateľov je dva. Ak sa zvýši počet posudzovateľov, potom aj postup výpočtu sa stáva zložitejší, nie však nemožný. Keď zostaneme pri binárnom hodnotení, potom je možné vypočítať pre každé hodnotenie percentuálny súhlas a tieto hodnoty následne spriemerovať. McHugh (2012) uvádza, že výhodou takto zobrazenej matice (Tabuľka 1) je, že sa môžeme pozrieť na to, či sú chyby náhodné a teda pomerne rovnomerne rozložené medzi všetkých hodnotiteľov a medzi hodnotenia, alebo konkrétny posudzovateľ udáva iné hodnotenia ako zvyšní posudzovatelia. Taktiež môžeme identifikovať premenné, ktoré môžu byť problematické, t.j. ktoré dosahujú nižšiu zhodu.

Ak by sme pridali viac kategórií hodnotenia, sťaží sa interpretovateľnosť percentuálneho súhlasu ako aj výpočet (Graham et al., 2012) a zároveň bude obtiažnejšie dosiahnuť perfektný súhlas. V takomto prípade je možné použiť modifikáciu odhadu zhody rozšírením definície zhody o susedné bodové kategórie na hodnotiacej škále[13] (Graham et al., 2012; Stemler, 2004). Predstavme si, že máme sedembodovú hodnotiacu škálu, kde posudzovatelia hodnotia od 1 do 7. Naše definovanie zhody môžeme rozšíriť tak, že za zhodu budeme považovať aj rozdiely v hodnotení o jeden bod nadol či smerom nahor (napr. zhoda bude aj hodnotenie 2–2, ale aj hodnotenie 1–2 či 2–3). Na jednej strane ide o zmierne-

[10] Označenie percentuálneho súhlasu môže byť rôzne, napr. a0 či p0.

[11] Nezhoda nemusí nutne znamenať že sa jeden z posudzovateľov mýli, ale podľa definície súhlasu posudzovateľov ide o odlišné nazeranie na to, čo je posudzované.

[12] Všeobecne hodnota 0,70 pri reliabilite sa používa skôr už tak „zo zotrvačnosti", pričom nevieme vypátrať históriu tejto hodnoty. Wilhelm et al. (2018) uvádzajú, že to vzniklo pravdepodobne nedopatrením či nesprávnou interpretáciou. Podnetným čítaním na túto tému je napríklad aj článok od Lance et al. (2006), ktorí rozoberajú všeobecne reportovanie cut-off kritérií a medzi nimi aj hodnotu 0,70 pri indexe r_{wg}.

[13] Angl. adjacent agreement.

Tabuľka 1. Percentuálny súhlas pri viacerých posudzovateľov (adaptované podľa McHugh, 2012)

Hodnotenie	P1	P2	P3	P4	P5	% súhlas
1	1	1	1	1	1	100
2	0	1	1	1	1	80
3	0	1	0	0	0	80
4	0	0	0	0	0	100
5	1	1	1	0	0	60
Priemerná zhoda medzi posudzovateľmi						84

Poznámka: P – posudzovateľ.

nie kritéria zhody, na druhej strane to môže viesť k nadhodnoteným odhadom súhlasu medzi posudzovateľmi a to najmä vtedy, ak je počet kategórií nižší, kde takmer všetky body budú susedné. Pri nižšom počte kategórií hodnotenia tak tento postup nie je vhodný. Zároveň si je potrebné uvedomiť, že táto technika používania susedných bodov vedie k situácii, kde percentuálna zhoda bude na krajných koncoch hodnotiacej škály takmer vždy nižšia ako v strede škály.

Za limit percentuálneho súhlasu považuje Uebersax (2018) nerozlišovanie medzi pozitívnym a negatívnym súhlasom. V Tabuľke 1 je to aj vidieť, kde je 100% súhlas pripísaný aj keď označili posudzovatelia všetci zhodne odpovede hodnotou 1 a aj keď ich označili hodnotu 0. Hlavná kritika percentuálneho súhlasu je zameraná na skutočnosť, že neberie do úvahy náhodný súhlas a je častokrát spájaná s vytváraním nového indexu (napr. Bennet et al., 1954; Cohen, 1960 a pod.). Predpoklad žiadneho náhodného súhlasu vedie podľa Zhaoa et al. (2013) k dôležitému paradoxu: aj náhodné hádanie môže byť reliabilné. To však nemusí z percentuálneho súhlasu robiť nepoužiteľný index, ale skôr index so špeciálnym účelom ako je napríklad použitie v jednoduchých prípadoch, v ktorých práve neočakávame žiadny náhodný súhlas (Feng, 2013; Zhao et al., 2013).[14]

Koeficient kappa

V roku 1960 vytvoril Jacob Cohen koeficient kappa (κ), ktorý patrí medzi najpoužívanejšie koeficienty pre vyjadrenie súhlasu medzi posudzovateľmi. Dôvodom vytvorenia bola práve kritika percentuálneho súhlasu, ktorý (ako už bolo vyššie spomínané) neberie do úvahy náhodný súhlas. Už pred Cohenom bol takto vy-

[14] Situáciu, keď súhlas v jednej kategórii má inú váhu než súhlas v druhej kategórii (napr. pri stanovovaní diagnózy sú kategóriami: prítomnosť diagnózy vs. neprítomnosť diagnózy) je možné riešiť cez koeficient A_1 od Rogota a Goldberga z roku 1966, avšak ani tento koeficient nie je priamo vylepšením percentuálneho súhlasu a taktiež neberie do úvahy náhodný súhlas (Zhao et al., 2013).

tvorený koeficient S[15] (Bennet et al., 1954) či koeficient π[16] (Scott, 1955), ktoré svojim vlastným spôsobom definovali náhodný súhlas. Keďže koeficient S vznikol ako kritika percentuálneho súhlasu a koeficient π ako kritika aj percentuálneho aj koeficientu S, koeficient κ vznikol ako vhodnejší postup (pozn. podľa Cohena) pre počítanie náhodného súhlasu než boli už existujúce koeficienty.

Kappa ako pôvodný Cohenov koeficient je vhodný pre nominálne premenné a pre dvoch posudzovateľov. Predpoklady pre jeho použitie zahŕňajú nezávislosť posudzovateľov, nezávislosť kategórií, ktoré sa navzájom vylučujú a nezávislosť hodnotených položiek (Cohen, 1960).

Výpočet pre koeficient kappa je nasledovný:

$$\kappa = \frac{(p_0 - p_c)}{(1 - p_c)} \quad (2),$$

kde p_0 predstavuje percentuálny súhlas a p_c je súhlasom náhodným[17],[18],[19].

Náhodný súhlas je počítaný násobením marginálnych súčtov v kontingenčnej Tabuľke 2, ktoré delíme N. Pre lepšie pochopenie výpočtu náhodného súhlasu si môžeme pomôcť kontingenčnou tabuľkou 2 × 2, kde sú sumarizované údaje z hodnotenia dvomi posudzovateľmi používajúcimi dve odpoveďové kategórie (áno, nie).

Marginálny súčet (a + c) predstavuje celkový počet prípadov, kedy posudzovateľ 2 uviedol áno a súčet (a + b) zase celkový počet prípadov, kedy posudzovateľ 1 odpovedal áno. Ak tieto súčty vydelíme celkovým počtom odpovedí N, potom zistíme, aká je pravdepodobnosť odpovedí pri súhlasných odpovediach u každého posudzovateľa zvlášť. Rovnako to je aj v prípade súčtov (b + d) a (c + d), ktoré predstavujú prípady, kedy posudzovateľ 2 aj posudzovateľ 1 odpovedali kategóriou nie.

Náhodný súhlas potom podľa Cohena predstavuje vynásobenie súčtu súhlasných odpovedí oboch posudzovateľov (pravdepodobnosť, že obaja posudzo-

---

[15] Pri koeficiente S sa pozerá na počet príslušných kategórií a miera náhodného súhlasu je definovaná ako inverzum počtu odpoveďových kategórií ($p_c = 1/k$, k – počet kategórií) (pozri napr. Gálová, 2010).

[16] Koeficient π predpokladá, že distribúcie proporcií hodnotení medzi kategóriami sú známe a rovnaké medzi oboma hodnotiteľmi a náhodný súhlas je závislý na počte odpoveďových kategórií, ako aj na frekvencii, s ktorou posudzovatelia využívajú tieto kategórie.

[17] Takto vyzerá vzorec aj koeficientu S a koeficientu π, líšia sa len v tom, ako definujú náhodný súhlas.

[18] Pre lepšie pochopenie základného vzorca najmä ohľadom menovateľa odporúčame časť textu od Zhao (2013).

[19] Kappa ako funkcia podielu pozorovaného a očakávaného súhlasu je podľa Warrensa (2015) len jedným zo spôsobov ako nazerať na tento koeficient. Podľa neho (a matematických výpočtov) môže byť interpretovaná aj ako podiel súhlasu korigovaný o náhodu, či ako priemerná spoľahlivosť kategórie alebo aj ako vnútroskupinová korelácia.

Tabuľka 2. Kontingenčná tabuľka pre 2 posudzovateľov

		Posudzovateľ 2		
		áno	nie	spolu
Posudzovateľ 1	áno	a	b	a + b
	nie	c	d	c + d
	spolu	a + c	b + d	N (a + b + c + d)

vatelia povedia áno) a vynásobenie súčtu nesúhlasných odpovedí oboch posudzovateľov (pravdepodobnosť, že obaja posudzovatelia povedia nie) a ich následné sčítanie, čo je celková pravdepodobnosť náhodného súhlasu. V matematickom zápise s ohľadom na príklad v Tabuľke 2 by to vyzeralo nasledovne:

$$p_c = \left(\frac{a+c}{N} \times \frac{a+b}{N}\right) + \left(\frac{b+d}{N} \times \frac{c+d}{N}\right) \quad (3)[20].$$

Porozumenie výpočtu koeficientu kappa nám pomôže si uvedomiť, že náhodný súhlas je závislý od distribúcie odpovedí posudzovateľov na rozdiel napr. od koeficientu S, kde náhodný súhlas priamo ovplyvňuje počet kategórií, alebo od koeficientu π, kde sú marginálne súčty zrátavané a nie násobené.

Kappa môže nadobúdať ľubovoľnú hodnotu z intervalu od −1 do +1, kde hodnota +1 predstavuje perfektnú zhodu. Z matematického hľadiska je však obtiažne dosiahnuť perfektnú zhodu a tá je zaznamenaná len za extrémnych okolností. Záporná hodnota naznačuje, že pozorovaný súhlas je nižší ako by sa očakávalo na základe náhody a opačne, ak sú hodnoty vyššie ako 0, tak je pozorovaná hodnota väčšia ako by sa očakávalo na základe náhody. Hodnotu 0 nadobúda vtedy, ak je miera súhlasu rovná miere náhodného súhlasu (Cohen, 1960). Zriedkavo sa stretneme s tým, že hodnota koeficientu kappa je nižšia ako 0 pokiaľ nie sú medzi posudzovateľmi veľké nezhody. Hypotéza nulovej hodnoty koeficientu[21] sa skúma málokedy a častejšie je uvádzanie konfidenčného intervalu (CI), ktorý môže byť počítaný pomocou vzorca (Cohen):

$$95\%\ CI = \kappa \pm 1{,}96 \times SE_\kappa \quad (4),$$
$$99\%\ CI = \kappa \pm 2{,}58 \times SE_\kappa \quad (5),$$

[20] Matematický zápis náhodného súhlasu pri koeficiente kappa môže vyzerať aj inak napr: $p_c = ((a + c)(a + b)) + ((b + d)(c + d)) / N^2$.

[21] Pre viac informácií o štatistickom testovaní koeficientu kappa pozri napr. Sim a Wright (2005).

kde pre štandardnú chybu kappy je výpočet nasledujúci:

$$SE_{\kappa} = \sqrt{\frac{p_0\,(1-p_0)}{N(1-p_c)^2}} \qquad (6).$$

Tu je potrebné si všimnúť, že štandardná chyba čiastočne závisí od veľkosti vzorky a čím vyšší je počet meraní, tým menšia je štandardná chyba. Aj keď sa dá κ vypočítať aj pre malé vzorky (menej ako 10), potom konfidenčný interval bude dosť široký a preto sa podľa McHughovej (2012) odporúča, že veľkosť vzorky by mala byť viac ako 30 hodnotení/meraní.

Veľkosť koeficientu κ predstavuje podiel zhody väčší ako sa očakáva náhodne (Sim a Wright, 2005) a pre tento koeficient boli vytvorené viaceré slovné interpretácie[22], ktoré sú uvedené v Tabuľke 3. Medzi najznámejšie patrí interpretácia podľa Landisa a Kocha (1977), ktorá však nemá podľa Ludbrooka (2002) teoretický základ a tak môže byť pre výskumníkov zavádzajúca. Fleiss et al. (2003) označujú ako slabú zhodu hodnotu kappa nižšiu ako 0,40 (0,41–0,75: priemerná až dobrá, 0,75–1: veľmi dobrá). Krippendorff (1980) poskytuje konzervatívnejšiu interpretáciu, ktorá navrhuje, aby sa závery pre hodnoty koeficientu κ nižšie než 0,67 vyvodzovali opatrne, pre hodnoty medzi 0,67 a 0,80 predbežne a pre hodnoty nad 0,80 ako definitívne závery. McHugh (2012) je tiež v interpretácii prísnejšia, nakoľko hodnota 0,41 ako akceptovateľná môže byť pre medicínsky výskum príliš mierna.

Používanie uvedených referenčných stupníc však môže byť zavádzajúce a to z nasledovných dôvodov (Dettori a Norvell, 2020): vypočítaný koeficient κ je špecifický pre skupinu subjektov a zmení sa, ak sa zmenia subjekty a veľkosť koeficientu κ závisí od veľkosti vzorky, počtu kategórií či rozdelenia subjektov medzi kategóriami. Napríklad hodnota κ = 0,54 na základe 200 posudzovaní je oveľa presvedčivejšia než hodnota κ = 0,6 na základe 10 posudzovaní. Zároveň zavádzajúce môže byť aj porovnávanie hodnoty koeficientov, pretože nespája interpretáciu miery zhody so stupňom neistoty a neumožňuje porovnať rozsah zhody medzi rôznymi štúdiami, pokiaľ sa nevykonávajú za rovnakých experimentálnych podmienok (Vanacore a Pellegrino, 2021).

Sim a Wright (2005) uvádzajú, že existujú aj ďalšie faktory (ako prevalencia, skreslenie a nezávislosť hodnotení), ktoré môžu ovplyvniť veľkosť koeficientu alebo priamo interpretáciu. S týmito faktormi sa spája kritika koeficientu κ a viacerí výskumníci (napr. Byrt et al. 1993; Cicchetti et al., 1990; Di Eugenio a Glass, 2004; Feinstein a Cicchetti, 1990; Gwet, 2002; Zhao 2011 a iní) vrátane samotného autora koeficientu upozorňujú na to, že hodnota κ môže byť za určitých okolností

[22] Nie všetci výskumníci uznávajú takýto postup interpretácie. Napr. Zhao et al. (2013) tvrdia, že jedinou interpretáciou je, že hodnota 1 znamená dokonalú zhodu a akýkoľvek výsledok menší než 1 znamená, že zhoda nie je dokonalá. Iný pohľad na interpretáciu má aj Gwet (2021) a o tom budeme písať nižšie pri koeficient AC_1.

Tabuľka 3. Interpretácia koeficientu kappa*

kappa	Landis a Koch (1977)	Altman (1991)	kappa	McHugh (2012)
<0	žiadna		0–0,2	žiadna
0–0,20	mierna	slabá	0,21–0,39	minimálna
0,21–0,40	dostatočná	dostatočná	0,40–0,59	slabá
0,41–0,60	priemerná	priemerná	0,60–0,79	priemerná
0,61–0,80	významná	dobrá	0,80–0,90	silná
0,81–1	vynikajúca	veľmi dobrá	nad 0,9	takmer perfektná

* Tieto interpretácie sa často používajú aj pre iné indexy zhody, ktoré korigujú náhodný súhlas.

neadekvátna a to konkrétne vtedy, ak ide o nerovnomerné rozdelenie distribúcie (ak sa marginálne distribúcie jednej kategórie veľmi líšia od marginálnych distribúcií druhej kategórie). Pozrime sa na nasledujúce dva príklady v Tabuľke 4 a 5 (Di Eugenio a Glass, 2004).

Prvý príklad (Tabuľka 4) predstavuje vyváženú distribúciu, kde percentuálny súhlas bude mať hodnotu $p_0 = 0,9$, náhodný súhlas $p_c = 0,5$ a koeficient $\kappa = 0,8$. V druhom príklade je distribúcia zošikmená, pričom percentuálny súhlas bude mať rovnakú hodnotu ako v príklade s rovnomerným rozložením a to $p_0 = 0,9$, náhodný súhlas sa zvýši $p_c = 0,905$ a koeficient kappa sa zníži $\kappa = -0,0526$. Je to celkom neočakávaný výsledok, najmä ak uvážime, že zhoda vyjadrená cez súhlasné odpovede je v oboch prípadoch zhodná.

To, čo sa v tomto príklade prejavilo je tzv. kappa paradox, na ktorý upozornili už Feinstein a Cicchetti (1990) a ktorý vlastne spochybňuje predpoklad, že hodnota koeficientu κ sa zvyšuje so zhodou údajov. Štúdie senzitivity (Gwet, 2002) ukázali, že paradox vzniká vtedy, keď posudzované subjekty majú tendenciu byť zaradené do jednej z možných kategórií. To môže byť spôsobené ak povaha samot-

Tabuľka 4. Príklad vyváženej distribúcie

		P2		
		áno	nie	spolu
P1	áno	45	5	50
	nie	5	45	50
	spolu	50	50	100

Poznámka: P – posudzovateľ.

Tabuľka 5. Príklad zošikmenej distribúcie

		P2		
		áno	nie	spolu
P1	áno	90	5	95
	nie	5	0	5
		95	5	100

ného výsledku zahŕňa vysokú prevalenciu jednej kategórie, alebo tým, že jeden z hodnotiteľov má tendenciu častejšie používať jednu z kategórií (Zec et al., 2017).

Prvý paradox súvisí teda s prevalenciou atribútu a efekt prevalencie sa prejavuje vtedy keď sa podiel zhody pri pozitívnej klasifikácii líši od podielu zhody pri negatívnej klasifikácii. Matematicky sa to dá vyjadriť ako (pri použití zápisu z Tabuľky 2):

$$index\ prevalencie = \frac{|a - d|}{N} \tag{7}.$$

Ak je index vysoký, náhodný súhlas bude tiež vysoký a κ bude primerane znížená ako keď je index prevalencie nízky alebo nulový (Sim a Wright, 2015). V štúdii, ktorú uskutočnili Zec et al. (2017) sa zistilo, že paradox sa začína objavovať pri prevalencii vyššej ako 60 % (v našom počítanom príklade zošikmenej distribúcie v Tabuľke 5 je index prevalencie = 0,9).

Druhým paradoxom je skreslenie (predpojatosť, angl. bias), čo predstavuje mieru, v akej sa posudzovatelia nezhodujú na podiele pozitívnych (alebo negatívnych) prípadov a matematicky sa to dá vyjadriť nasledovne:

$$index\ skreslenia = \frac{|b - c|}{N} \tag{8}.$$

Keď je index skreslenia vyšší, koeficient kappa je vyšší ako keď je index skreslenia nízky či nulový. Na rozdiel od prevalencie je účinok skreslenia väčší, keď je κ nízka ako keď je vysoká (Sim a Wright, 2015). Pre lepšie porozumenie je možné zobrazenie v príklade (Di Eugenio a Glass, 2004):

V oboch prípadoch vychádza percentuálny súhlas p_0 = 65 %. V Tabuľke 6 sú nezhody medzi posudzovateľmi 1 a 2 vcelku symetrické a náhodný súhlas má hodnotu p_c = 0,52 a celkovo koeficient κ = 0,27. V Tabuľke 7 sú tieto nezhody medzi posudzovateľmi asymetrické a v dôsledku väčšieho skreslenia sú výsledné hodnoty koeficientu kappa odlišné – náhodný súhlas má hodnotu p_c = 0,45 a celkovo κ = 0,418.

Podľa Zhaoa (2011) má κ ďaleko viac paradoxov ako uvedené dva: napríklad nedefinovateľnosť koeficientu (ak sa posudzovatelia zhodnú, že percentuál-

na frekvencia jednej kategórie je 100% a druhej 0, tak koeficient nie je možné definovať), maximalizáciu náhodnosti (t.j. vylúčenie, že posudzovatelia môžu byť aj čestní), porovnávanie jabĺk s hruškami (koeficient kappa porovnáva objektívnu zhodu a náhodnú nezhodu) a mnohé ďalšie.

Keďže koeficient kappa predstavený Cohenom je vhodný len pre nominálne premenné a dvoch posudzovateľov, tento koeficient sa dočkal sa aj rôznych „rozšírení": vážený koeficient kappa κ_w (použiteľný pre dvoch posudzovateľov pri nominálnych či ordinálnych premenných), exaktný koeficient kappa κ_{Exact} (generalizácia koeficientu kappa pre *m* posudzovateľov; Conger, 1980), kappa κ_{Light} (použiteľný pre viacerých ako dvoch posudzovateľov a kategorické dáta; je priemerom všetkých možných kombinácií bivariačného koeficientu kappa medzi posudzovateľmi; Light, 1971)[23]. Aj keď sa tieto koeficienty snažili prekonať problémy, ktoré kappa má, nakoľko vychádzajú z jej základu, je zrejmé, že sa ich úplne nezbavili.

Ak sa teraz pýtate, že či je správne používať naďalej tento koeficient, tak je to úplne legitímna otázka. Jednou z možností je pokračovať v jeho používaní, avšak okrem uvádzania hodnoty koeficientu by sme mali doplniť ďalšie hodnoty ako je pozitívny súhlas, negatívny súhlas, index prevalencie a index skreslenia (napr. Cunningham, 2009). Oveľa častejšie sa však vyzýva na jeho nahradenie inými paradoxom nepodliehajúcimi indexami zhody.

Tabuľka 6. Príklad „symetrického" nesúhlasu

		P2		
		áno	nie	spolu
P1	áno	40	15	55
	nie	20	25	45
	spolu	60	40	100

Tabuľka 7. Príklad asymetrického nesúhlasu

		P2		
		áno	nie	spolu
P1	áno	40	35	75
	nie	0	25	25
		40	60	100

[23] Upozornenie: Fleissov koeficient kappa je generalizáciou Scottovho koeficientu π a Brennan-Predigerov koeficient kappa je zase generalizáciou G-indexu.

Koeficient AC_1

Gwetov koeficient zhody prvého rádu – AC_1 (2001) bol predstavený ako alternatívny náhodu korigujúci index k všetkým doposiaľ existujúcim indexom. Gwet matematicky dokázal, že jeho index je menej zaťažený skresleniami než iné indexy (pozn. autora: aj ako koeficient kappa). Jeho definícia náhodného súhlasu je založená na premise, že náhodný súhlas sa vyskytuje vtedy, keď aspoň jeden z posudzovateľov pri svojom hodnotení háda, alebo hodnotí náhodne, ale tieto hodnotenia nie sú úplne náhodné, náhodná je len časť hodnotení[24]. Inak povedané, náhodný súhlas sa vyskytuje len vtedy, ak sa dvaja hodnotitelia zhodujú, avšak aspoň jeden z nich vykonal náhodné hodnotenie (Xu a Lorber, 2014).

V porovnaní s ostatnými indexami, ktoré vznikli pred koeficientom AC_1 je Gwetove chápanie náhodného súhlasu jedinečné (Zhao et al., 2013). Vráťme sa k príkladu s guličkami (spomínaný pri definovaní náhodného súhlasu). Kým ostatné indexy (ako κ, π, S a pod.) predpokladajú, že posudzovatelia hodnotia náhodne, keď sa guličky zhodujú a hodnotiac poctivo, ak sa nezhodujú, tak Gwet predpokladá, že posudzovatelia hodnotia náhodne, keď sa guličky nezhodujú a poctivo, ak sa zhodujú.

Na základe simulácií Gwet (2002) uvádza, že pravdepodobnosť zhody posudzovateľov za predpokladu náhodného hodnotenia je 0,5. To znamená, že ak jeden z hodnotiteľov náhodne vyberie kategóriu odpovede, dosiahnu zhodu v 50 % prípadov.

Základný výpočet indexu AC_1 je v podstate rovnaký ako je výpočet kappy (vzorec 2):

$$AC_1 = \frac{(p_0 - p_c)}{(1 - p_c)} \tag{7}.$$

Rozdiel je už potom v tom, ako je definovaný náhodný súhlas (podľa kontingenčnej Tabuľky 2):

$$p_c = 2\left(\frac{\frac{a+b}{N} + \frac{a+c}{N}}{2}\right)\left(1 - \frac{\frac{a+b}{N} + \frac{a+c}{N}}{2}\right) = \frac{1}{2}\left(1 - \left(\frac{a-d}{N}\right)^2\right) \tag{8}.$$

Aj keď je tento koeficient stále závislý od marginálneho rozdelenia, teda od prevalencie, vôbec nesúvisí so skreslením (bias) a to preto, že skreslenie je vymazané už v definovaní náhodnej zhody použitím priemerných marginálnych rozdelení. Závislosť od prevalencie je žiaducou vlastnosťou miery zhody, pretože je v súlade zo zdravým rozumom, že vyššiu úroveň spoľahlivosti očakávame vtedy,

[24] Iné indexy zhody, ktoré korigujú náhodu (a sem patrí aj koeficient kappa) maximalizujú náhodný súhlas, teda predpokladajú, že celé hodnotenie je náhodné.

keď dvaja odborne kvalifikovaní posudzovatelia dosahujú podobnú alebo vyššiu úroveň zhody (Xie, 2013).

Vráťme sa teraz k príkladom vyššie, kde sme preberali problém prevalencie a problém skreslenia u koeficientu kappa. V prípade normálnej distribúcie (Tabuľka 4) vychádza hodnota AC_1 rovnako ako koeficient kappa ($p_0 = 0,9$, $\kappa = 0,8$, $AC_1 = 0,8$). V príklade so zošikmenou distribúciou (Tabuľka 5), kde bola hodnota kappa nezmyselná, vychádza hodnota AC_1 porovnateľne ako v prípade nezošikmenej distribúcie ($p_0 = 0,9$, $\kappa = -0,0526$, $AC_1 = 0,89$).

V ďalších dvoch príkladoch (Tabuľka 6 a 7), kde bol ukázaný problém skreslenia pri koeficiente kappa, vychádza pre oba príklady náhodný súhlas aj koeficient AC_1 rovnako: $p_c = 0,49$, $AC_1 = 0,31$ (pre pripomenutie: koeficient κ vychádza rozdielne: $\kappa = 0,27$ a $\kappa = 0,418$)

Ohľadom interpretácie koeficientu AC_1 ale aj ostatných indexov zhody navrhuje Gwet (2014, 2021) štandardizovanejšiu metódu, ktorá prekonáva viaceré problémy s referenčnými hodnotami. Na základe hodnôt koeficientu zhody a jeho štandardnej chyby je možné vypočítať pravdepodobnosť, že koeficient zhody bude patriť každej z referenčných kategórií (uvedených v Tabuľke 3). Následne postupne cez kumulatívnu pravdepodobnosť budeme zisťovať rozsah, do ktorého najpravdepodobnejší odhad zhody patrí, pričom to bude vtedy, ak prekročí určitú hranicu (napríklad 95 %). Tento postup bol porovnávaný s postupom referenčných hodnôt cez Monte Carlo simulácie a porovnávala sa výkonnosť každého postupu z hľadiska váženej miery chybnej klasifikácie vypočítanej pre všetky kategórie zhody ako aj pre každú kategóriu odpovede samostatne (Vanacore a Pellegrino, 2021). V prípade malých vzoriek simulácie ukazujú uspokojivé a porovnateľné charakteristiky váženej miery chybnej klasifikácie u oboch postupov a pri väčších vzorkách má postup porovnávania navrhnutý Gwetom vo všeobecnosti horšie výsledky, takže sa v tomto prípade sa odporúča použiť referenčné porovnávanie dolnej hranice či už percentilového rozptylu alebo parametrického intervalu spoľahlivosti.

Gwettov koeficient AC_1 sa síce úspešne popasoval s dvomi problémami, ktoré sú spojené s koeficientom kappa, ale to neznamená, že sa s ním nespájajú iné paradoxy. Podľa Zhao et al. (2013) je u neho badateľný klasický paradox súvisiaci s prázdnymi kategóriami odpovedí či paradox miešania poctivých súhlasov s náhodnými nesúhlasmi a pod.

Okrem koeficientu AC_1, ktorý je vhodný pre dvoch a viac posudzovateľov a nominálne kategorické premenné, Gwet (2021) vytvoril aj koeficient druhého rádu AC_2, ktorý je vhodný pre ordinálne, intervalové či pomerové premenné a taktiež pre dvoch a viacerých posudzovateľov.

Porovnanie percentuálneho súhlasu, koeficientu κ a koeficientu AC_1

Použitie rôznych indexov zhody vedie k rôznym hodnotám súhlasu medzi posudzovateľmi, keďže sa od seba líšia v samotných výpočtoch a preto by výber indexu mal byť dobre premyslený a jeho interpretácia adekvátna. Pozrime sa

Tabuľka 8. Kontingenčná tabuľka pre empirické dáta (Madysová et al., 2012)

		Posudzovateľ 2		
		áno	nie	spolu
Posudzovateľ 1	áno	3	3	6
	nie	1	13	14
	spolu	4	16	20

na porovnanie nami popisovaných indexov na empirických dátach. Zvolili sme dáta z článku od Mandysovej et al. (2012), ktoré sú vhodné pre demonštrovanie rozdielov medzi indexami na rovnakých dátach. Dáta pochádzajú od dvoch posudzovateľov, ktorí hodnotili 20 prípadov (pacientov) podľa rovnakej škály, kde hodnota 1 znamenala prítomnosť rizika vzniku dekubitov a hodnota 0 absenciu rizika vznikov dekubitov[25]. Dáta uverejnené v článku sme zapísali do kontingenčnej Tabuľky 8.

Už z rozloženia dát v tabuľke vidíme, že nie symetrické, čo podľa vyššie uvedených informácií o indexoch napovedá, že aj keď bude percentuálny súhlas vyšší, koeficient kappa bude skreslený práve nesymetrickosťou prevalencie dát a koeficient AC_1 by mal toto skreslenie prekonať smerom k vyššej hodnote zhody.

Na základe výpočtov (uvedených v Apendixe 1[26]) zisťujeme, že náš predpoklad je správny. Podľa percentuálneho súhlasu je zhoda medzi posudzovateľmi v hodnote $P_A = 0,8$, čo znamená, že posudzovatelia sa zhodli pri svojom hodnotení v 80 % prípadov. Výsledná hodnota koeficientu kappa v hodnote $\kappa = 0,474$ CI [0,013; 0,9346] hovorí, že zhoda medzi posudzovateľmi podľa kritérií McHughovej (ktorá je prísnejšia najmä ak ide o medicínsky výskum) je slabá a zároveň že tento súhlas je odlíšiteľný od náhody. Po výpočte koeficientu $AC_1 = 0,68$ zisťujeme, že je naozaj vyšší ako koeficient kappa a tak je bližší k hodnote percentuálneho súhlasu, pričom berie do úvahy aj náhodný súhlas[27].

Uvedené rozdiely boli očakávané, pretože rôzne indexy zhody majú rôzny teoretický základ a rôzne predpoklady. Zhao et al. (2013) uvádzajú, že hlavný rozdiel medzi indexami zhody je v ich predpoklade o tom, ako sa správa posudzovateľ. Percentuálny súhlas predpokladá, že posudzovatelia nikdy nehádajú, t.j. nepodávajú náhodné hodnotenie, zatiaľ čo indexy, ktoré náhodu korigujú (kam radíme aj koeficient kappa a AC_1) predpokladajú, že posudzovatelia maximalizujú náhodné hodnotenie.[28] Podľa Krippendorffa (2004) korigovanie náhody v zho-

[25] V tomto prípade ide o dichotomizáciu priradených bodov v rámci škály.
[26] Apendix je dostupný online na https://doi.org/10.13060/csr.2024.007.
[27] V Apendixe 2 sme sumarizovali možnosti výpočtu týchto troch koeficientov v najčastejšie používaných štatistických programoch.
[28] Zhao et al. (2013) rozlišujú indexy korigujúce náhodu na indexy založené na kategóriách a indexy založené na distribúcii, pričom tie, ktoré sú založené na kategóriách predpokladajú rovnaké rozdelenie a indexy založené na distribúcii predpokladajú kvóty pri posudzovaní.

de medzi posudzovateľmi nie je všeliekom. Nakoľko náhoda môže znamenať rôzne veci, aj korekcie náhody sú rôzne, čím je vlastne meraná zhoda na mierne odlišných stupňoch. Tým pádom porovnávanie indexov v rôznych štúdiách prestáva dávať zmysel, pretože už dopredu vieme, že indexy sa budú líšiť.

Môžeme však povedať, že ktorý index je pre kvantifikovanie najlepší? Vráťme sa naspäť k nášmu príkladu. Pri výbere indexu zhody je dôležitým krokom poznanie vybraného indexu a identifikovanie jeho predností ako aj slabín, k čomu mám môžu pomôcť rôzne prehľadové články[29]. Najnovšia štúdia Zhao et al. (2022) je jednou z nich. Títo výskumníci svojimi závermi potvrdili, že náhodný súhlas existuje a ak ho budeme prehliadať, tak percentuálny súhlas bude zhodu medzi posudzovateľmi nadhodnocovať. Zároveň na základe svojich zistení uvádzajú, že náhodná zhoda nie je až taká veľká, ako sa pôvodne predpokladalo. Vo svojom výskume (Zhao et al., 2022) v medzisubjektovom experimente[30] testovali vplyv troch faktorov (počet kategórií hodnotenia, skreslenie distribúcie a náročnosť úlohy) na výsledok 7 najznámejších indexov zhody medzi nimi aj na nami bližšie popísané indexy. Zistili, že každý z indexov zhody implikuje jeden alebo viac nesprávnych predpokladov o náhodnej zhode. Percentuálny súhlas ju prehliada, koeficienty S, l_r a AC_1 sa nevhodne spoliehajú na kategórie a koeficienty π, κ a α[31] zase na skreslenie distribúcie. Obtiažnosť úlohy mala silný a pozitívny vplyv na náhodnú zhodu a všetky zahrnuté indexy, ktoré náhodnú zhodu korigujú, vôbec s obtiažnosťou úlohy neoperujú, dokonca sa spoliehajú na jej ľahkosť. Z testovaných indexov voči pravému skóre reliability vyšiel najlepšie percentuálny súhlas ako najpresnejší prediktor spoľahlivosti a ako tretí najlepší aproximátor. Koeficient AC_1 bol druhým najlepším prediktorom a zároveň najlepším aproximátorom. No a zvyšné miesta v prvej top trojke patrili koeficientu S. Indexy π, κ a α si viedli horšie a to čiastočne preto, lebo zahŕňajú viac nesprávnych predpokladov.

Koeficient AC_1 odporúčajú uprednostňovať aj Konstantinidis et al. (2022), ktorí na základe simulácií uvádzajú, že pri nesymetrickej prevalencii je tento koeficient lepší než Fleissov koeficientom κ, Lightov koeficient κ a Congerov koeficient κ. Ak je však prevalencia symetrická, potom sú rozdiely medzi indexami zhody nízke, ale nakoľko nevieme, aká bude prevalencia v populácii, lepšie je použiť už spomínaný koeficient AC_1. To je jedno z potvrdení toho, na čo upozorňujú kritici koeficientu kappa.

---

[29] Napríklad: Zhao (2011) sa vo svojom príspevku venuje koeficientu kappa a popisuje 14 paradoxov, ktoré ukazujú, že κ nie je všeobecným indikátorom a to cez matematické výpočtu a základnú logiku; Zhao et al. (2013) vo svojej štúdii vysvetľujú rôzne nedostatky náhodne korigovaných koeficientov a to cez popísanie problémov a paradoxov s nimi spojených; van Oest (2022) matematicky porovnáva vážené koeficienty súhlasu a to vplyvom parametra sily, ako reálne mocninového parametra, ktorý zachytáva bežné váhové schémy (lineárne, kvadratické, identifikačné a radikálne); Zhao et al. (2022) poskytujú zistenia z experimentu ohľadom 7 indexov zhody voči pozorovanej reliabilite; Konstantinidis et al. (2022) porovnávajú odpovede viacerých posudzovateľov (2–5) v 20, 50, 300 a 500 pozorovaniach cez štyri základné indexy zhody.

[30] Experiment zahŕňal posudzovanie dĺžky tyčí, čo bolo navoliteľné programovo a tak existoval „zlatý štandard". Išlo o medzisubjektový experiment $4 \times 8 \times 3$ so 4 subjektami v každej bunke.

[31] Máme na mysli Krippendorffov koeficient α.

Koeficient AC_1 má aktuálne širokú podporu ale aj napriek tomu je ešte stále popularita koeficientu κ či percentuálneho súhlasu veľká. Pri percentuálnom súhlase to ale nemusí byť nakoniec taký veľký faux pas keď ho použijeme, ako sme už spomínali vyššie. Jeho popularitu je možné pripísať viacerým faktorom (Button et al., 2020). Za prvé, rady odborníkov sú zmätočné (ak medzi samými výskumníkmi – metodológmi panuje nezhoda v tom, aký index je najlepší, aj praktický výskumníci majú v tom zmätok). Za druhé, percentuálny súhlas je ľahko vypočítateľný a je všeobecne zrozumiteľný a tak je celkom logické, že výskumníci po ňom siahnu, pretože ho poznajú. A za tretie, percentuálna zhoda zodpovedá nášmu intuitívnemu chápaniu toho, čo znamená zhodnúť sa na niečom a to spôsobom, akým to pravdepodobnostné myslenie nedokáže. Popularitu pri koeficiente κ skôr pripisujeme tomu, že dlho nebol dostupný index zhody, ktorý by prekonal jeho spomínané paradoxy a tento koeficient je rozšírený všade: spomínajú ho učebnice metodológie, jeho výpočet obsahujú mnohé štatistické programy a ak ho výskumníci vidia použitý v iných štúdiách, jednoducho uveria „autorite" a použijú ho ďalej bez jeho dôkladného poznania.

Podľa Zhao et al. (2013) však potrebujeme index založený na predpokladoch variabilnej náhodnosti a poctivých posudzovateľov a tiež potrebujeme, aby tento index ako hlavný faktor využíval stupeň náročnosti (obtiažnosť úlohy) a nie kategórie odpovedí či distribúcie odpovedí. Kým sa ale takého indexu dočkáme, musíme si vybrať to, čo najlepšie máme k dispozícii a to je aktuálne percentuálny súhlas spolu s koeficientom AC_1.

Záver

Zhoda medzi posudzovateľmi ako téma zastáva svoju pozíciu v metodológii kvantitatívneho výskumu ako aj pri získavaní a analyzovaní kvalitatívnych dát. Zaradenie ďalšieho posudzovateľa do výskumu sa objavuje v mnohých oblastiach. S tým rastie aj množstvo výskumníkov, ktorí sa téme venujú a tak rastie aj množstvo informácií, ktoré sú k téme publikované, čo sťažuje orientovanie sa v nich. Cieľom tohto článku bolo sumarizovať základné teoretické vymedzenie zhody a zároveň popísanie troch indexov (percentuálny súhlas, koeficient kappa a koeficient AC_1), ktoré slúžia pre jej kvantifikáciu. Pri následnom porovnaní týchto indexov boli potvrdené predpokladané rozdiely, na základe ktorých odporúčame v súčasnosti používať koeficient AC_1 doplnený percentuálnym súhlasom. Aj keď oba koeficienty majú svoje skreslenia, musíme si zatiaľ vystačiť s tým čo máme. Stále je tu ale potreba v tom, aby nové indexy vznikali, čo sa vlastne aj naďalej deje (napr. nový index od van Oesta, 2019).

Okrem výskytu nových indexov predpokladáme, že ďalšou témou, ktorú bude potrebné riešiť je rozmach umelej inteligencie a s ním prichádzajú na scénu aj e-posudzovatelia, pri ktorých je možné naprogramovať ich tak, že ich rozhodnutia či očakávania budú jasné a konzistentné. E-posudzovanie nie je však nová téma[32]

[32] Automatické skórovanie je používané často pri hodnotení položiek s konštruovanou odpoveďou vo vzdelávacích testoch.

a už dlhšiu dobu sa objavujú články, ktoré e-posudzovateľov zapájajú do výskumu. Predpokladáme, že táto téma naberie na intenzite a bude sa riešiť aj to, či sú existujúce indexy zhody vhodné pre e-posudzovateľov a či nebude potrebné vyvíjať nové indexy s inými predpokladmi náhodnosti posudzovania.

Lucia Kočišová *pôsobí na Ústave experimentálnej psychológie Slovenskej akadémie vied, kde sa zameriava na výskum konceptov dôchodku. Dlhodobejšie sa zameriava na reliabilitu posudzovateľov v kvantitatívnom aj kvalitatívnom výskume.*
ORCID: 0000-0002-8747-1651

Literatúra

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman & Hall. https://doi.org/10.1201/9780429258589

Bennett, E. M., Alpert, R. a Goldstein, A. C. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, *18*, 303–308. https://doi.org/10.1086/266520

Bliese, P. D. (2000). *Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis.* In K. J. Klein a S. W. J. Kozlowski (eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (s. 349–381). Jossey-Bass.

Button, C. M., Snook, B. a Grant, M. J. (2020). Inter-rater agreement, data Reliability, and the crisis of confidence in psychological research. *The Quantitative Methods for Psychology*, *16*(5), 467–471. https://doi.org/10.20982/tqmp.16.5.p467

Byrt, T., Bishop, J. a Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429. https://doi.org/10.1016/0895-4356(93)90018-v

Cicchetti, D. V. a Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-m

Cígler, H. a Šmíra, M. (2015). Chyba měření a odhad pravého skóru. *Testfórum*, *6*, 67–84. https://doi.org/10.5817/TF2015-6-104

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322–328. https://doi.org/10.1037/0033-2909.88.2.322

Cunningham, M. 2009. More than just the kappa coefficient: A program to fully characterize inter-rater reliability between two raters. *SAS Global Forum*, Paper 242-2009. https://support.sas.com/resources/papers/proceedings09/242-2009.pdf

de Vet, H. C. W., Terwee, C. B., Knol, D. L. a Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

Dettori, J. R. a Norvell, D. C. (2020). Kappa and beyond: Is there agreement? *Global Spine Journal*, *10*(4), 499–501. https://doi.org/10.1177/2192568220911648

DeVellis, R. F. (2005). Inter-rater reliability. In K. Kempf-Leonard (ed.), *Encyclopedia of social measurement* (s. 317–322). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00095-5

Di Eugenio, B. a Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101. https://doi.org/10.1162/089120104773633402

Feinstein, A. R. a D. V. Cicchetti. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Feng, G. C. (2013). Underlying determinants driving agreement among coders. *Quality and Quantity*, *47*, 2983–2997. https://doi.org/10.1007/s11135-012-9807-z

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(1), 13–22. https://doi.org/10.1027/1614-2241/a000086

Fleiss, J. L., Levin, B. a Paik, M. C. (2003). *Statistical methods for rates & proportions*. Wiley & Sons. https://doi.org/10.1002/0471445428

Gálová, L. (2010, 3. február). *Koeficient kappa – aplikačné možnosti, výhody a nevýhody* [príspevok prednesený na konferencii]. 2. česko-slovenská konference doktorandů oborů pomáhajících profesí, Ostrava.

Gisev, N., Bell, J. S. a Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social & Administrative Pharmacy*, *9*(3), 330–338. https://doi.org/10.1016/j.sapharm.2012.04.004

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science*, *5*(1), 13–34. https://doi.org/10.1207/S15327841MPEE0501_2

Graham, M., Milanowski, A. a Miller, J. (2012). *Measuring and Promoting Inter-rater Agreement of Teacher and Principal Performance Ratings.* Center for Educator Compensation and Reform. Stiahnuté 10. 2. 2023 z http://es.eric.ed.gov/fulltext/ED532068.pdf

Gwet, K. L. (2002). Kappa Statistic Is Not Satisfactory for Assessing the Extent of Agreement between Raters. *Statistical Methods for Inter-Rater Reliability Assessment*, *1*, 1–6. https://agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf

Gwet, K. L. (2014, 12. december). Benchmarking agreement coefficients. *Inter-rater Reliability Blog*. https://inter-rater-reliability.blogspot.com/2014/

Gwet, K. (2021). *Handbook of Inter-rater reliability. The definite guide to measuring the extent of agreement. Analysis of categorical rating*. Gaithersburg: AgreeStat Analytics.

Hartmann, D. P., Barrios, B. A. a Wood, D. D. (2004). Principles of behavioral observation. In M. Hersen (ed.), *Comprehensive handbook of psychological assessment* (s. 108–137). John Wiley & Sons.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*(2), 234–246. https://doi.org/10.1037/0021-9010.83.2.234

Kočišová, L. (2022). Reportovanie súhlasu posudzovateľov a spoľahlivosti posudzovateľov. *Testfórum*, *15*, 41–57. https://doi.org/10.5817/TF2022-15-14647

Konstantinidis, M., Le, L. W. a Gao, X. (2022). An empirical comparative assessment of inter-rater agreement of binary outcomes and multiple raters. *Symmetry, 14*(2), 262. https://doi.org/10.3390/sym14020262

Kottner, J. a Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology*, *64*(6), 701–702. https://doi.org/10.1016/j.jclinepi.2010.12.001

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M. a Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*(1), 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Kozlowski, S. W. J. a Klein, K. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. in K. J. Klein a S. W. J. Kozlowski (eds.). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (s. 3–90). San Francisco: Jossey-Bass.

Kozlowski, S. W. a Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, *77*(2), 161–167. https://doi.org/10.1037/0021-9010.77.2.161
Krefting, L. (1991). Rigor in qualitative research: the assessment of trustworthiness. *American Journal of Occupational Therapy*, *45*, 214–222. https://doi.org/10.5014/ajot.45.3.214
Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills: Sage Publications.
Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x
Lance, C. E., Butts, M. M. a Michels L. C. (2006). The sources of four commonly reported cutoff criteria: what did they really say? *Organizational Research Methods, 9*(2), 202–220. https://doi.org/10.1177/1094428105284919
Landis, J. R. a Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310
LeBreton, J. M. a Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852. https://doi.org/10.1177/1094428106296642
Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365–377. https://doi.org/10.1037/h0031643
Lincoln, Y. a Guba, E. G. (1985). *Naturalistic inquiry*. Sage. https://doi.org/10.1016/0147-1767(85)90062-8
Litwin, M. S. (1995). *How to measure survey reliability and validity*. Sage Publications. https://doi.org/10.4135/9781483348957
Lombard, M., Snyder-Duch, J. a Bracken, C. C. (2002). Content analysis in mass communication research: An assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x
Ludbrook, J. (2002). Statistical techniques for comparing measures and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology*, *29*(7), 527–536. https://doi.org/10.1046/j.1440-1681.2002.03686.x
Mandysová, P., Ehler, E. a Trejbalová, L. (2012). Česká verze Škály Bradenové: metodika prekladu a shoda mezi posuzovateli. *Ošetrovateľstvo*, *2*(4), 137–142.
Martončik, M. (2019). *Validita merania v sociálnych vedách.* Prešovská univerzita.
McDonald, N., Schoenebeck, S. a Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, *72*. https://doi.org/10.1145/3359174
McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031
Řehák, J. (1998). Kvalita dat I. Klasický model měření reliability a jeho praktický aplikační význam. *Sociologický časopis*, *34*(1), 51–60. https://doi.org/10.13060/00380288.1998.34.1.07
Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325. https://doi.org/10.1086/266577
Sim, J. a Wright, Ch. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257
Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). https://doi.org/10.7275/96jp-xz07

Syed, M. a Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, *3*(6), 375–387. https://doi.org/10.1177/2167696815587648

Švaříček, R. a Šeďová, K. a kol. (2007). *Kvalitativní výzkum v pedagogických vědách.* Portál.

Tinsley, H. E. a Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*(4), 358–376. https://doi.org/10.1037/h0076640

Uebersax J. (2018, 19. september). *Raw Agreement Indices*. https://www.john-uebersax.com/stat/raw.htm

Urbánek, T., Denglerová, D. a Širůček, J. (2011). *Psychometrika: měření v psychologii.* Portál.

van Oest, R. (2019). A new coefficient of interrater agreement: The challenge of highly unequal category proportions. *Psychological Methods*, *24*(4), 439–451. https://doi.org/10.1037/met0000183

Vanacore, A. a Pellegrino, M. S. (2021). Benchmarking procedures for characterizing the extent of rater agreement: a comparative study. *Quality and Reliability Engineering International*, *38*(3), 1404–1415. https://doi.org/10.1002/qre.2982

Von Eye, A. a Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Lawrence Erlbaum.

Warrens, M. J. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, *5*(4), 1. https://doi.org/10.4172/2161-0487.1000197

Wilhelm, A. G., Rouse, A. G. a Jones, F. (2018). Exploring differences in measurement and reporting of classroom observation inter-rater reliability. *Practical Assessment, Research, and Evaluation*, *23*(4). https://doi.org/10.7275/at67-md25

Xie, Q. (2013, 4.–6. november). *Agree or disagree? A demonstration of an alternative statistic to Cohen's kappa for measuring the extent and reliability of agreement between observers* [príspevok prednesený na konferencii]. Federal Committee on Statistical Methodology Research Conference, The Council of Professional Associations on Federal Statistics, Washington, DC. https://nces.ed.gov/FCSM/pdf/J4_Xie_2013FCSM.pdf

Xu, S. a Lorber, M. F. (2014). Interrater agreement statistics with skewed data: evaluation of alternatives to Cohen's kappa. *Journal of Consulting and Clinical Psychology*, *82*(6), 1219–1227. https://doi.org/10.1037/a0037489

Zec, S., Soriani, N., Comoretto, R. a Baldi, I. (2017). High agreement and high prevalence: The paradox of Cohen's kappa. *The Open Nursing Journal*, *11*, 211–218. https://doi.org/10.2174/1874434601711010211

Zhao, X. 2011 (11.–30. máj). *When to use Cohen's* κ*, if ever?* [príspevok prednesený na konferencii]. 61st annual conference of International Communication Association, Boston.

Zhao, X., Feng, G. C., Ao, S. H. a Liu, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology*, *22*(1), 232. https://doi.org/10.1186/s12874-022-01707-5

Zhao, X., Feng, G. C., Liu, K. a Deng, J. S. (2018). We agreed to measure agreement – redefining reliability de-justifies Krippendorff's alpha. *China Media Research*, *14*(2), 1–15. https://repository.um.edu.mo/handle/10692/25978

Zhao, X., Liu, J. S. a Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association*, *36*(1), 419–480. https://doi.org/10.1080/23808985.2013.11679142