REPORTOVANIE SÚHLASU POSUDZOVATEĽOV A SPOĽAHLIVOSTI POSUDZOVATEĽOV

LUCIA KOČIŠOVÁ[1]
[1] *Ústav experimentálnej psychológie, SAV Bratislava*

Abstrakt: *V psychológii ale aj v mnohých iných oblastiach sa stretneme s použitím ďalšieho posudzovateľa pre potvrdenie validity a reliability našich záverov. Ide o súhlas posudzovateľov (inter-rater agreement), ktorý predstavuje zhodu v ich hodnotení a ak je zhoda dosiahnutá, hodnotitelia sú zameniteľní (Tinsley &Weiss, 1975) a spoľahlivosť posudzovateľov (inter-rater reliability) v zmysle konzistencie hodnotenia (LeBreton & Senter, 2008). Oba koncepty sa okrem definovania líšia aj v zodpovedaní rôznych výskumných otázok a spôsobu štatistickej analýzy. Cieľom príspevku je odpovedať na otázky, ktoré súvisia s praktickou potrebou reportovania súhlasu posudzovateľov a spoľahlivosti posudzovateľov. S tým sú spojené otázky, na ktoré príspevok hľadá odpovede: Aký počet posudzovateľov je vhodné zvoliť? Ako si vybrať vhodný index súhlasu a spoľahlivosti posudzovateľov? Existujú akceptované miery súhlasu a spoľahlivosti posudzovateľov? Ktoré faktory vplývajú na mieru súhlasu a spoľahlivosti posudzovateľov?*

Kľúčové slová: *súhlas posudzovateľov; spoľahlivosť posudzovateľov; výber indexu*

Úvod

Základ reportovania súhlasu posudzovateľov a spoľahlivosti posudzovateľov stojí a zároveň padá na výbere relevantného indexu[2], koeficientu alebo testu. Dôvodom je práve to, že nesprávne zvolený index môže nadhodnocovať alebo podhodnocovať výsledky, a tak budú naše zistenia skreslené či zavádzajúce. Preto je potrebné venovať výberu indexov súhlasu a spoľahlivosti posudzovateľov náležitú pozornosť a riadiť sa odporúčanými krokmi, ku ktorým sa zaraďuje určenie cieľa analýzy, počtu posudzovateľov, ujasnenie si úrovne merania a poznanie výpovednej hodnoty indexov (Stemler, 2004, Uebersax, 2008, von Eye & Mun, 2005, Gisev et al., 2013).

[1] Centrum spoločenských a psychologických vied SAV, Ústav experimentálnej psychológie SAV, Dúbravská cesta 9, 841 01 Bratislava
[2] Pojem index bude v článku používaný ako synonymum pre pojmy koeficient, test.
Korespondenční autor: Lucia Kočišová, e-mail: lucia.kocisova@savba.sk
Doručeno do redakce 17. 5. 2021

Cieľ analýzy

V prvom rade je potrebné si zodpovedať otázku: *Čo je cieľom analýzy?* Odpoveďou na ňu zistíme, či je pre nás dôležitá absolútna hodnota posudzovania alebo trend v hodnoteniach posudzovateľov (Gisev et al., 2013). Viacero výskumníkov sa zhoduje v tom, že existuje zásadný rozdiel, či ide o *súhlas posudzovateľov* (angl. inter-rater agreement), alebo o *spoľahlivosť posudzovateľov* (angl. inter-rater reliability)[3] (Stemler, 2004, LeBreton & Senter, 2008, Liao et al., 2010, Kottner & Streiner, 2011, Gisev et al., 2013, Stolarova et al., 2014 a iní). Na druhej strane je ešte stále väčšia skupina výskumníkov, ktorí medzi týmito dvoma pojmami nerozlišujú a používajú ich ako synonymá, čo následne vedie k nesprávnej interpretácii výsledkov.

Ak teda potrebujeme zistiť, či je hodnotenie posudzovateľov rovnaké, alebo sa líši (Kottner & Streiner, 2011), presnejšie ak chceme zistiť, aký je súhlas medzi posudzovateľmi (de Vet et al., 2006), potom cieľom analýzy bude zistenie absolútnej hodnoty zhody medzi posudzovateľmi. Ak nás ale zaujíma konzistencia posudzovateľov (Stemler, 2004) alebo chceme zistiť, aké spoľahlivé je meranie (de Vet et al., 2006), potom cieľom bude zistenie trendu v hodnoteniach.

Rozlišovanie medzi týmito dvoma pojmami je podstatné, pretože zásadne vplýva na výber vhodnej štatistickej analýzy. Iné indexy sa odporúčajú pre zisťovanie súhlasu posudzovateľov a iné zase pre zisťovanie spoľahlivosti posudzovateľov. Predstavte si, že chcete zistiť zhodu (súhlas) medzi dvoma posudzovateľmi, ktorí hodnotia prejavy agresívneho správania detí v škôlke na škále od 1 do 4 počas jednej hodiny. Keď použijete index pre zisťovanie spoľahlivosti, tak výsledok bude hovoriť o tom, ako sú posudzovatelia konzistentní vo svojich hodnoteniach na základe vlastného definovania tejto škály (napr. jeden z posudzovateľov môže byť nastavený prísnejšie pri danom hodnotení ako ten druhý), ale nebude to vypovedať o tom, či je ich hodnotenie rovnaké, alebo sa líši.

Ak v tomto momente nabádame vybrať si medzi súhlasom a spoľahlivosťou posudzovateľov, je tiež na mieste uviesť, že použitie oboch súčasne vo výskume je podľa niektorých výskumníkov vhodné, ba dokonca aj žiadúce. Tinsley a Weiss (1975) odporúčajú, aby v každej výskumnej štúdii boli súhlas aj spoľahlivosť posudzovateľov skúmané, a to pre získanie najlepšieho indikátora kvality hodnotenia. Wilhelm et al. (2018) uvádzajú, že poskytnutím len konsenzu (súhlasu) vynechávajú výskumníci dôležité informácie o presnosti procesu hodnotenia. Podľa zistení Hintzeho a Matthewsa (2004), miery založené na koreláciách, ktoré sú vyjadrením konzistencie (spoľahlivosti posudzovateľov, pozn. autora), lepšie zodpovedajú veľkosti nesúhlasu medzi posudzovateľmi a miery konsenzu lepšie zodpovedajú súhlasu medzi posudzovateľmi. Preto ak je to vhodné, odporúčame nastaviť ciele výskumu aj pre súhlas a aj pre

[3] Termín posudzovateľ (angl. rater) býva vzhľadom na kontext výskumu rôzny a tak sa stretávame s pojmami ako pozorovateľ (angl. observer), kóder (angl. coder), tester, examinátor, čitateľ, atď. Zároveň sa používa namiesto pojmu *intra* pojem *inter*, a to vtedy, ak je posudzovateľom jedna osoba a robí posudzovanie napr. viac krát v rôznych časoch.

spoľahlivosť posudzovateľov, pretože tak je možné získať viac informácií o tom, aké bolo hodnotenie posudzovateľov.

Počet posudzovateľov

V druhom kroku je podstatné určenie počtu posudzovateľov, nakoľko samotné indexy bývajú navrhované pre dvoch posudzovateľov, pre viac ako dvoch posudzovateľov, prípadne sú bez obmedzenia počtu posudzovateľov. V tejto fáze ide o jednoduchý krok, keďže počet posudzovateľov je známy a výskumník už pred samotnou realizáciou výskumu urobil rozhodnutie, koľko posudzovateľov zahrnie do hodnotenia. *Bolo však jeho rozhodnutie o počte posudzovateľov správne?* Odpoveď na túto otázku nie je jednoduchá, pretože neexistuje jednotné odporúčanie optimálneho počtu posudzovateľov pre všetky výskumy. Výber počtu posudzovateľov sa riadi najmä možnosťami výskumníka (personálne, ale aj časové či finančné), a mal by sa riadiť aj informáciami o tom, ako súvisí počet posudzovateľov s hodnotou výsledku či s cieľom výskumu.

Dôležitou informáciou je vplyv počtu posudzovateľov na hodnotu indexu súhlasu alebo spoľahlivosti posudzovateľov. Viacerí výskumníci uvádzajú zistenia, že ak budeme zvyšovať počet posudzovateľov, súhlas posudzovateľov bude klesať. Pri súhlase posudzovateľov vo všeobecnosti platí, že zvyšovaním počtu posudzovateľov sa zvyšuje aj variabilita hodnotení a tým sa hodnota indexu súhlasu bude znižovať. Súvisí to práve so spôsobom ako sú indexy súhlasu posudzovateľov počítané. A keďže výpočet indexov súhlasu sa líši, znižovanie hodnoty indexov zvýšením počtu posudzovateľov nepostupuje rovnakým tempom nadol.

Pre predstavu uvádzame zistenia Abediho et al. (1995), ktorí porovnávali viacero indexov súhlasu s Monte Carlo simuláciou dát a pozerali sa na počet posudzovateľov, počet prípadov hodnotenia a distribúciu dát. Prezentovaný príklad (tab.1) je pre dáta s normálnou distribúciou (okrem nej simulovali tiež dáta s uniformnou distribúciou, angl. rectangular distribution a dáta s distribúciou v tvare J). V týchto výsledkoch je možné sledovať trend, že ak je miera súhlasu medzi posudzovateľmi vysoká, potom pridanie ďalšieho posudzovateľa nezníži percentuálny súhlas tak výrazne, ako v prípade, ak je miera súhlasu medzi posudzovateľmi nižšia.

Tabuľka 1

Miera súhlasu pri rôznom počte posudzovateľov a rôznej veľkosti vzorky pri normálnom rozložení (Abedi et al., 1995)

miera zhody	nad 90%		50%-75%		menej ako 50%	
počet posudzovateľov	2	4	2	4	2	4
percentuálny súhlas (n = 500)	91,6%	90,4%	71,2%	61,8%	48,2%	31,6%
percentuálny súhlas (n = 50)	94%	92%	68%	64%	46%	34%

Obdobný pokles evidovali aj Bikmaz Bilgen & Doğan (2017) pri troch použitých indexoch súhlasu s počtami posudzovateľov 2, 5 a 10. Posudzovatelia hodnotili 50 výkonov žiakov 5. triedy. Najvyšší pokles bol evidovaný v prípade Kendallovho koeficientu.

Tabuľka 2

Miera súhlasu pri rôznom počte posudzovateľov (Bikmaz Bilgen & Doğan, 2017)

počet posudzovateľov	2	5	10
Cohennova kappa	0,38	0,24	0,27
Krippendorfova alfa	0,67	0,60	0,58
Kendallov koeficient	0,61	0,31	0,18

Nying (2004) nepotvrdil úplne tento klesajúci trend, nakoľko zistil, že so zvyšujúcim sa počtom posudzovateľov sa Kendallov koeficient nezmenil a koeficient kappa pre viacerých posudzovateľov kolísal. Ako bolo však už vyššie naznačené, do tohto procesu vstupujú aj ďalšie premenné, ktoré ho moderujú a určite nepredpokladáme medzi mierou súhlasu a počtom posudzovateľov perfektný lineárny negatívny vzťah.

Pri spoľahlivosti posudzovateľov sa podľa výsledkov niektorých štúdií (pozri nižšie) ukazuje, že zvyšovaním počtu posudzovateľov dochádza aj k zvyšovaniu reliability – konzistencie medzi posudzovateľmi. Predpokladáme, to súvisí s variabilitou hodnotení a s výpočtom niektorých indexov. Samotná reliabilita je definovaná cez variabilitu skóre[4] a ak je variabilita nízka, potom aj reliabilita bude nízka. Z toho sa dá usudzovať, že pridaním ďalšieho posudzovateľa sa zvýši aj variabilita v hodnoteniach, čo však ale nemusí nutne nastať.

Tabuľka 3

Miera spoľahlivosti pri rôznom počte posudzovateľov a rôznej veľkosti vzorky pri normálnom rozložení (Abedi et al., 1995)

miera zhody	nad 90%		50%-75%		menej ako 50%	
počet posudzovateľov	2	4	2	4	2	4
priemerné korelácie (n = 500)	0,88	0,89	0,57	0,63	0,29	0,30
priemerné korelácie (n = 50)	0,97	0,93	0,61	0,68	0,50	0,40
Cronbachova alfa (n = 500)	0,94	0,97	0,73	0,85	0,45	0,63
Cronbachova alfa (n = 50)	0,97	0,98	0,76	0,89	0,66	0,73

Pre zisťovanie spoľahlivosti medzi posudzovateľmi sa používa napríklad aj Cronbachova alfa (posudzovatelia predstavujú položky), ktorá je citlivá na počet položiek a zvyšuje sa

[4] Urbánek et al. (2011, s. 97) uvádzajú nasledovnú definíciu: Reliabilitu je možné „definovať teoreticky ako podiel variability pravých skóre k celkovej variabilite“.

po pridaní položiek (napr. Cortina, 1993; Marko, 2016). Takže zvýšením počtu posudzovateľov sa zvyšuje aj hodnota indexu Cronbachova alfa. Je to možné vidieť aj na výsledkoch od Abedi et al. (1995). Pri vysokej miere zhode medzi posudzovateľmi zistili na simulovaných dátach (Monte Carlo), že odhady spoľahlivosti posudzovateľov nie sú tak veľmi ovplyvnené počtom posudzovateľov ako je to pri nižšej miere zhody.

Optimálny počet posudzovateľov bude vždy vzťahovaný ku konkrétnemu výskumu s konkrétne nastaveným cieľom. Napríklad v niektorých prípadoch nebude potrebné mať počet posudzovateľov vysoký, nakoľko cieľom výskumu nebude variabilita v hodnotení, ale zistenie, či je hodnotiaca škála aplikovaná podobne pri hodnotení správania. V kvalitatívnom výskume bude zase vhodné zapojenie viacerých posudzovateľov vtedy, keď bude potrebné zvýšiť kapacitu kódovania väčšieho množstva dát či pri formovaní kódov alebo pri redukovaní skreslení na strane jednotlivca. Na druhej strane, ak je výskumník expert s jedinečnými skúsenosťami alebo je posudzovanie jednoduché (napríklad binárne), tak nebude potrebné zisťovať súhlas či spoľahlivosť posudzovateľov a tak zapájať iných výskumníkov do posudzovania (McDonald et al., 2019).

Na základe uvedeného by sa dalo vo všeobecnosti povedať, že ak je našim cieľom zisťovať súhlas posudzovateľov, nie je potrebné mať vysoké počty hodnotiacich (odhliadnuc v tomto prípade od špecifických cieľov v kvalitatívnom výskume) a počet 2-3 posudzovateľov by mohol byť vhodný (s dobre nastaveným tréningom hodnotenia). Ak chceme zisťovať spoľahlivosť posudzovateľov, pre zvýšenie variability hodnotenia je vhodné zaradiť viacerých posudzovateľov ako dvoch. Bogartz (2005) na základe výsledkov simulácií pre zistenie optimálneho počtu posudzovateľov (okrem iných cieľov) odporúča použiť práve troch posudzovateľov v takomto prípade, čo by mohlo byť vodítkom aj pre nás. A ak by sme chceli zvýšiť spoľahlivosť posudzovateľov, tak zvýšime počet posudzovateľov.

V niektorých prípadoch môže byť počet posudzovateľov stanovený vopred, napr. pri zisťovaní súhlasu medzi žiakmi v triede pri hodnotení klímy triedy, kde chceme zahrnúť všetkých žiakov ako posudzovateľov (pozri napr. Gálová, 2010b). V takýchto prípadoch následne vieme využiť vyššie spomínané informácie o vplyve počtu posudzovateľov na mieru súhlasu a spoľahlivosti a tak lepšie interpretovať naše zistenia.

Výber indexu

Výber indexu pre súhlas a spoľahlivosť posudzovateľov usmerňuje (ako už bolo vyššie spomínané) typ výskumnej otázky, počet posudzovateľov a nakoniec aj úroveň merania. Samotný výber je však sťažený existenciou doslova desiatok[5] rôznych indexov, koeficientov či testov.

[5] Prehľady viacerých koeficientov ponúka napr. Popping (1988), ktorý identifikoval 38 indexov len pre nominálne dáta. Uebersax (2008) na svojej stránke ponúka prehľad mnohých indexov, ktoré delí podľa úrovne merania dát a podľa počtu meraní. Balík *irr* pre štatistický program R obsahuje 17 rozličných indexov (Gamer et al., 2012). Zhao et al. (2013) uvádza diskusiu o 22 rôznych koeficientoch, pričom zistenia vedú k tomu, že viaceré z nich sú matematicky ekvivalentné, čo vyústilo v 11 unikátnych indexov.

Ako si teda vybrať ten správny index? Zhao et al. (2013) uvádzajú, že potrebujeme index, ktorý by zodpovedal typickej výskumnej situácii, kde sa posudzovatelia snažia byť presní, ale niekedy nedobrovoľne pripustia určitú náhodnosť. Podľa nich existujúce indexy zatiaľ túto potrebu nespĺňajú a pokiaľ takýto index nebude vytvorený, budeme si musieť vybrať z existujúcich indexov[6].

V prvom rade sa indexy delia na tie, ktoré sú vhodné pre súhlas posudzovateľov a tie, ktoré sú vhodné pre spoľahlivosť posudzovateľov (i keď vo výskumoch sa častejšie stretnete s tým, že sa hovorí o spoľahlivosti posudzovateľov a používa sa na to index vhodný pre súhlas posudzovateľov). Výber konkrétneho indexu ovplyvní aj samotnú mieru zisťovaného súhlasu a spoľahlivosti, keďže výpočet jednotlivých indexov sa líši a tak môžeme pri použití viacerých indexov na rovnakých dátach získať rôzne výsledky[7].

V skupine indexov vhodných pre súhlas posudzovateľov je možné vyčleniť indexy, ktoré berú do úvahy náhodný súhlas tj. ktoré rôznym spôsobom kvantifikujú súhlas medzi posudzovateľmi na základe náhodnosti priraďovania odpovedí či doslova hádania odpovedí. Jedným z najpoužívanejších indexov z tejto skupiny je Cohenov koeficient kappa určený pre dvoch posudzovateľov a pre nominálne dáta[8]. Jeho popularita je doteraz veľmi veľká aj napriek preukázaným problémom, ktoré však bývajú často opomínané[9]. Keďže paradoxy[10], s ktorými koeficient kappa bojuje, robia z neho index vhodný pre špecifické situácie, určite nie je vhodné používať ho za akýchkoľvek podmienok. Ak ho však aj napriek tomu použiť chceme (napr. chceme naše výsledky porovnať so zisteniami iných, ktorí použili koeficient kappa), je vhodné doplniť ho o ďalšie informácie ako je pozitívny súhlas, negatívny súhlas, index prevalencie a index skreslenia (pozri napr. Gálová, 2010a).

Ktorý index je však vhodnejší než koeficient kappa? Percentuálny súhlas, ktorý je jednoduchý a intuitívny, býva kritizovaný práve preto, že mu chýba korekcia náhodnosti. Podľa Fenga (2013) keď je hodnotenie náročné (v zmysle obtiažnosti úlohy), percentuálny súhlas nadhodnocuje zhodu medzi posudzovateľmi (pretože náročnosť ide ruka v ruke s náhodnosťou odpovedania, pozn. autora). Ale ak je hodnotenie jednoduché, môže byť dobrým ukazovateľom súhlasu posudzovateľov. To z neho robí vhodný index

[6] Zhao et al. (2013) nerozlišujú medzi súhlasom a spoľahlivosťou posudzovateľov a tak tu hovoria len o indexoch súhlasu posudzovateľov.

[7] Napr. ten Hove et al. (2018) publikovali rozdiely medzi 20 rôznymi koeficientmi pre rôzne úrovne merania na rovnakom dátovom sete. Aj mnohé iné štúdie prinášajú rozdiely v jednotlivých hodnotách indexov súhlasu a spoľahlivosti, o čom sme sa presvedčili aj my pri výskume hodnotenia klímy školskej triedy žiakmi (Gálová, 2010b), alebo pri hodnotení interakčného štýlu učiteľa žiakmi (Gálová, 2014).

[8] Existujú viaceré verzie tohto koeficientu odvodené od pôvodného a zároveň vhodné aj pre viac posudzovateľov či pre iné ako nominálne dáta.

[9] Xie (2013) uvádza, že výskumníci začali na problémy s koeficientom kappa upozorňovať už pred viac ako 40-timi rokmi.

[10] Paradoxy spájané s indexom kappa boli "preslávené" v sérii prác Feinsteina a Cicchettiho (Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990). Nízka hodnota koeficientu kappa sa môže vyskytnúť pri vysokej zhode a nesymetrické marginálne rozdelenia produkujú vyššie hodnoty koeficientu kappa než viac symetrické marginálne rozdelenia. Inak povedané, koeficient kappa je ovplyvnený zošikmením rozdelenia kategórií (problém prevalencie) a mierou nezhody medzi posudzovateľmi (problém skreslenia, *bias problem*; DiEugenio & Glass, 2004).

použiteľný za „špeciálnych okolností“ (Zhao, 2013). Ak si vezmeme učebnicový príklad, kde je hodnotenie nenáročné a vychádza z dobre vypracovaného protokolu pre posudzovanie a neočakávame tak náhodný súhlas, vtedy môže byť percentuálny súhlas vhodný, ale zároveň by nemal byť použiteľný samostatne (Feng, 2015).

Ostatné indexy, ktoré do úvahy náhodný súhlas berú, do svojich výpočtov zahŕňajú (tak ako aj koeficient kappa) pozorovateľný súhlas a očakávaný súhlas na základe náhody. Jednotlivé indexy sa líšia potom práve v kvantifikácii náhodného súhlasu. Feng (2013) konštatuje, že náhodný súhlas u rôznych koeficientov býva ovplyvnený počtom kategórií (koeficient S), marginálnym rozdelením[11], náročnosťou hodnotiacich úloh a interakciou medzi nimi a tiež uvádza, že náročnosť úloh abnormálne vplýva na náhodný súhlas (ide o koeficient kappa, koeficient π, či Krippendorffov koeficient α). Abnormálnosťou má na mysli to, že čím sú úlohy hodnotenia ťažšie, tým sú šance na zhodu nižšie. Pravdepodobnejšie je však, že posudzovatelia budú pri náročnejších úlohách viac hádať, čo znamená, že náhodný súhlas by mal pozitívne korelovať s náročnosťou úlohy, čo však v prípade uvedených koeficientov neplatí.

Gwetov koeficient AC_1 je tiež ovplyvnený počtom kategórií a interakciou s marginálnym rozdelením, ale v porovnaní s koeficientom kappa je to v menšej miere (Wongpakaran et al., 2013). Zároveň náhodný súhlas tohto indexu pozitívne koreluje s náročnosťou úloh a preto je podľa Fenga (2013) lepším indexom súhlasu než koeficient kappa. Koeficient AC_1 preferujú všetci výskumníci, ktorý ho v rámci svojich výskumov porovnávali s koeficientom kappa (napr. Gwet, 2002a; Gwet, 2002b; Xie, 2013; Wongpakaran et al., 2013; Dettori & Norvell, 2020). Aktuálne je teda vhodné používať pri nominálnych premenných koeficient AC_1 a pri ordinálnych a intervalových premenných koeficient AC_2. Zároveň je dobrou praxou použiť viac rôznych indexov v prípade zisťovania súhlasu posudzovateľov (Xie, 2013), pretože výsledky môžu lepšie popisovať skutočnosť.

Spoľahlivosť posudzovateľov je definovaná cez konzistenciu a nie je spájaná s toľkými indexami ako súhlas posudzovateľov. Potom aj výber vhodného indexu je uľahčený. Vo všeobecnosti sa odporúča používať korelačné koeficienty s ohľadom na povahu dát (napr. Pearsonov korelačný koeficient, Spearmanov koeficient poradovej korelácie a pod.). Najčastejšie sa používajú vnútrotriedne korelácie (angl. *intraclass correlation*).

Uvedené odporúčania určite nie sú paušálne vhodné pre všetky výskumné situácie, kde sú zahrnutí viacerí posudzovatelia. Podľa Uebersaxa (2008) je len jedno všeobecné pravidlo, ktoré hovorí o tom, že žiadny index nie je tým najlepším pre všetky výskumy. Dôvodmi použitia rôznych indexov sú zosúladenie s výskumnou otázkou, znalosť indexov či dostupnosť výpočtových postupov a logistika používania nástrojov (Wilhelm et al., 2018). V Prílohe 1 prinášame zoznam 35 indexov s informáciami o type výskumnej otázky, počte posudzovateľov a úrovni merania dát, ktorý však ani v tomto počte nepovažujeme za konečný vzhľadom na množstvo rôznych variácií uvedených indexov

[11] Marginálne (okrajové) rozdelenie (distribúcia) popisuje v kontingenčnej tabuľke riadkové alebo stĺpcové početnosti premennej/premenných.

a taktiež neustále vznikajúce nové indexy. Zoznam môže byť využitý ako odrazový mostík pri rozhodovaní sa o výbere vhodného indexu, ktorý nám zúži naše zameranie, čo určite pri prvom stretnutí sa s touto problematikou šetrí čas.

Akceptovaná miera súhlasu/spoľahlivosti posudzovateľov

Podľa LeBretona a Sentera (2008) sa dá vo všeobecnosti povedať, že čím väčšie sú dôsledky, ktoré vyplývajú z hodnotenia posudzovateľmi, tým väčšia je potreba dosiahnuť vysokú mieru súhlasu a spoľahlivosti posudzovateľov. Avšak aj napriek rozdielom medzi súhlasom a spoľahlivosťou posudzovateľov sa zdá, že existuje štandardná prax považovať v oboch prípadoch hodnotu 0,7 alebo 0,8 za dostatočnú. Vypátrať históriu tejto hodnoty je podľa Wilhelma et al. (2018) nemožné a tak sa usudzuje, že stanovenie a používanie hodnoty 0,7 alebo 0,8 ako akceptovanej hodnoty pre súhlas a spoľahlivosť posudzovateľov vzniklo pravdepodobne nedopatrením či nesprávnou interpretáciou generalizovania akceptovanej hodnoty z klasickej teórie testov pre vnútornú konzistenciu. Toto jednotné pravidlo však nie je opodstatnené, nakoľko hodnoty súhlasu a spoľahlivosti je možné interpretovať mnohými spôsobmi a skôr by sa mal robiť úsudok týkajúci sa interpretácie hodnôt súhlasu a spoľahlivosti posudzovateľov s ohľadom na povahu štúdie a možné dôsledky výsledkov (Gisev et al., 2013). Základom je uvedomiť si, o čo nám vlastne ide. Je pre nás nutné aby súhlas medzi posudzovateľmi (spoľahlivosť posudzovateľov) dosiahol konkrétnu hodnotu? Alebo potrebujeme zistiť, ako vybraní posudzovatelia hodnotia určený objekt/subjekt (prípadne ako sú konzistentní)? Dôsledkom hodnotenia môžu byť napríklad závažné rozhodnutia ako je určenie diagnózy a v takom prípade bude pre nás podstatnejšie, aby bol súhlas posudzovateľov vysoký než to požadujeme pri hodnotení eseje študentov. Môžeme mať však úplne iný cieľ výskumu, kde stanovená hranica súhlasu nebude pre nás dôležitá a budeme sa primárne zaujímať o to, ako posudzovatelia vnímajú hodnotený objekt/subjekt a prípadne v čom tkvie ich nesúhlas.

Dôležité je tiež poznamenať, že existujú dva prístupy pri hodnotení súhlasu. Prístup, ktorý kladie dôraz na to, či súhlas dosiahol alebo nedosiahol stanovenú hodnotu, hovorí o praktických štandardoch. Na druhej strane prístup štatistických štandardov nie je zameraný na absolútny súhlas ale na to, či pozorovaný súhlas vo vzorke je väčší ako by sa očakávalo na základe náhody, tj. ide o štatistický test hypotézy. Štatistický prístup sa používa menej, čo môže byť spôsobené neznalosťou tejto metódy, prípadne ide aj o to, že štatistický súhlas môže byť obtiažne dosiahnuť v bežne sa vyskytujúcich praktických situáciách (O'Neill, 2017).

Graham et al. (2012) popisujú tri zásady, podľa ktorých je možné sa riadiť pri určení toho, aká miera súhlasu je dostatočná. Prvá zásada hovorí, že jednotlivé indexy majú stanovenú minimálnu hodnotu, ktorá je akceptovateľná. Ak tomu tak nie je, je možné využiť porovnanie miery súhlasu/spoľahlivosti posudzovateľov s tým, čo uvádzajú iné výskumy na danú tému. Treťou zásadou je potom zameranie sa na proporciu nesúhlasu a zváženie, nakoľko je táto proporcia podstatná pre výsledky, čo nás vracia k názoru LeBretona a Sentera (2008) uvedeného vyššie.

De Vet s kolegami (2006) uvádzajú, že parametre súhlasu sú stabilnejšie medzi rôznymi populáciami než parametre spoľahlivosti, pričom tie sú viac závislejšie na variancii v populačnom výbere a sú zovšeobecniteľné len pre výbery s podobnou varianciou. Spoľahlivosť je charakteristikou použitia nástroja v konkrétnom populačnom výbere a súhlas posudzovateľov je viac charakteristikou samotného meracieho nástroja. Je tu zhoda s tým, čo uvádzajú Gisev et al. (2013) a to, že výsledky súhlasu a spoľahlivosti sú jedinečné pre konkrétnu výskumnú štúdiu, nakoľko sú funkciou populácie a sú závislé od posudzovateľov, odpovedí a hodnotiacej škály. Preto by nemali byť generalizované na iné štúdie. Takéto tvrdenie vyznieva v rozpore so zásadou Grahama et al. (2012), aby sme pri určení akceptovateľnej hodnoty porovnali mieru súhlasu/spoľahlivosti posudzovateľov s inými výskumami. Podľa nás to však skôr upozorňuje na to, že súhlas aj spoľahlivosť sú značne závislé od variability hodnoteného subjektu/objektu v skupine posudzovateľov a tak je generalizovanie výsledkov v jednej štúdii na iné štúdie náročnejšie (i keď nie nemožné). A aj pri porovnávaní mier súhlasu/spoľahlivosti posudzovateľov vo viacerých štúdiách treba pamätať práve na tú variabilitu.

Faktory vplývajúce na súhlasu/spoľahlivosti posudzovateľov

Ak skúmame súhlas posudzovateľov, v každom prípade sa objaví určitá miera nesúhlasu. Bez vysvetlenia tejto miery budú závery, ktoré budú chcieť generalizovať výsledky výskumu, slabé. Posúdenie toho, čo je možné považovať za zdroj nesúhlasu by zlepšilo výpovednú hodnotu výskumu (Slaug et al., 2012). To platí v prípade, že nesúhlas predstavuje niečo, čo chceme eliminovať. Stretneme sa totiž aj so situáciami, kedy nesúhlas hľadáme (napríklad pri kvalitatívnom kódovaní pri definovaní kódov) a v tomto prípade faktory, ktoré ho môžu spôsobovať by mali byť tiež kontrolované len v opačnom význame.

Na jednej strane môže byť nesúhlas zapríčinený samotným posudzovateľom, jeho osobnostnými charakteristikami (napr. prísnosť-zhovievavosť, haló-efekt, centrálna tendencia), skúsenosťami či očakávaním (Fradenburg et al., 1995, Slaug et al., 2012). Väčšia pozornosť posudzovateľovi sa venovala najmä pri hodnotení osobnosti druhých. Funder (2012) označuje posudzovateľa vo svojom modeli RAM[12] ako moderátora zhody medzi posudzovateľmi. Posudzovanie osobnosti je komplexným problémom, na čo vplývajú nielen explicitné a implicitné znalosti posudzovateľa o osobnosti, ale taktiež jeho percepčné a kognitívne schopnosti a aj motivácia (Hrebíčková, 2003).

Spôsob, ktorý môže byť nápomocný v tejto chvíli je tréning posudzovateľov. Graham et al. (2012) konštatujú, že tréning je jeden z najdôležitejších nástrojov ako zlepšiť súhlas posudzovateľov. Je to aj napriek tomu, že viaceré štúdie (napr. Lumley & McNamara, 1995, Hoyt & Kerns, 1999; Wang, Wong & Kwong, 2010) zistili, že určitá miera variability medzi posudzovateľmi pretrváva aj po veľmi dlhých tréningoch a tréning posudzovateľov redukuje najmä extrémne rozdiely medzi posudzovateľmi a náhodne chyby v

[12] Model RAM (Realistic Accuracy Model) definuje moderátory presného posudzovania osobnosti: a) posudzovateľ, b) cieľová osoba, c) charakteristika, ktorá je posudzovaná, d) informácie ako dlho sa posudzovateľ a posudzovaný poznajú a v akom sú vzťahu (Funder, 2012).

hodnoteniach. Pri tréningu posudzovateľov je dôležitá jeho dĺžka ako aj pretestovanie schopnosti posudzovateľa hodnotiť/pozorovať/kódovať. A ak ani rozšírený tréning nezabezpečí, že posudzovateľ bude viac menej v zhode s ostatnými, potom je namieste ho nezahrnúť do hodnotenia. Niektorí výskumníci v tomto prípade navrhujú rekrutáciu väčšieho počtu posudzovateľov a tých, ktorí neprejdú cez skríning súhlasu a spoľahlivosti vylúčiť.

Okrem posudzovateľa vstupuje do procesu hodnotenia aj samotný nástroj pre hodnotenie a jeho charakteristiky (napr. presnosť/znenie položiek, stupnica hodnotenia) ako aj kontext, v ktorom sa hodnotenie uskutočňuje (napr. časový rámec hodnotenia, obdobie v roku, teplota) (Slaug et al., 2012). Celkom to pripomína distraktory v experimente, ktoré je potrebné sledovať a kontrolovať.

V prípade spoľahlivosti posudzovateľov sa nedá hovoriť o ich nesúhlase, ale keďže sme na poli reliability, tak ide o náhodné chyby[13]. Náhodné chyby sú nekonzistentné chyby, ktoré nie je možné predpovedať a sú nepriamo úmerné spoľahlivosti meracieho nástroja (Volchok, 2015), v našom prípade spoľahlivosti posudzovateľov. Keďže ide o chyby, ktoré sú ťažko predpovedateľné a v podstate môže ísť o čokoľvek, čo sa vyskytne počas merania/testovania či hodnotenia, je náročné zostaviť zoznam náhodných chýb. Podľa nášho názoru môže ísť tak o tie isté chyby, ktoré sa objavujú aj pri súhlase posudzovateľov. Nezahrnuli by sme sem ale nástroj posudzovania a chyby v presnom definovaní položiek, či chyby v stupnici, pretože v tomto prípade by išlo o chyby systematické (viac o náhodných a systematických chybách merania je napríklad v publikácii od Urbánka et al., 2011, alebo v Štandardoch pre pedagogické a psychologické testovanie, 2014).

Reportovanie súhlasu/spoľahlivosti posudzovateľov

Po realizácii výskumu so získaním dát, analýzou výsledkov a interpretáciou zistení nasleduje fáza reportovania súhlasu a/alebo spoľahlivosti posudzovateľov. V tejto chvíli je nápomocné využiť usmernenia pre reportovanie (príloha), ktoré vytvorili Kottner et al. (2011), ktorí majú skúsenosti s vývojom nových nástrojov, ich hodnotením, s odhadmi reliability a súhlasu alebo s recenziami štúdií o reliabilite. Usmernenia vedú výskumníka od začiatku až po koniec výskumu a sú mu zdrojom kladenia otázok pri jednotlivých fázach výskumu.

K podstatným informáciám, ktoré by si mal čitateľ z článku odniesť, patrí postupnosť výberu vhodného indexu pre zisťovanie súhlasu a spoľahlivosti posudzovateľov. Na začiatku je potrebné rozlišovať medzi súhlasom a spoľahlivosťou a nastaviť ciele výskumu konkrétnejšie tak, aby zodpovedali naozaj tomu, čo chceme zisťovať. Ak je to v našom výskume vhodné, odporúča sa zisťovať súčasne súhlas aj spoľahlivosť posudzovateľov kvôli podrobnejším informáciám o hodnotení posudzovateľov.

[13] Reliabilita býva vymedzovaná ako relatívna neprítomnosť náhodných chýb (Urbánek et al., 2011).

Na výber konkrétneho indexu vplýva už spomínaný cieľ výskumu, počet posudzovateľov a úroveň merania. Rozhodnutie o počte posudzovateľov prichádza ešte pred samotnou realizáciou výskumu a výskumník by mal do úvahy pri ňom brať nielen svoje možnosti, ale taktiež informácie o jeho vplyve na mieru súhlasu a mieru spoľahlivosti posudzovateľov ako aj informácie o náročnosti úlohy, prípadne či je potrebné z hľadiska cieľov výskumu mať vyššiu variabilitu v hodnotení alebo nie.

V oblasti súhlasu posudzovateľov je z množstva indexov aktuálne preferovaný Gwetov koeficient AC_1 (nominálne premenné) a AC_2 (ordinálne, intervalové premenné) a pri spoľahlivosti posudzovateľov sú to korelačné koeficienty s preferovanými vnútrotriednymi koreláciami. Keďže ale nejde o indexy, ktoré sú vhodné pre všetky výskumné situácie, je potrebné sa oboznámiť aj s inými možnosťami. Odporúčané je aj použitie viacerých indexov súhlasu (pri spoľahlivosti sa s viacerými indexmi skôr nestretávame), pričom je potrebné mať na pamäti, že následne jednotná hranica akceptovanej miery súhlasu (a aj miery spoľahlivosti) nie je relevantná. Zároveň je nutné si uvedomiť, že tak ako súhlas tak aj spoľahlivosť posudzovateľov sú vzťahované ku konkrétnej populácii, ku konkrétnej hodnotiacej škále a ku konkrétnym posudzovateľom a preto nie vždy je možné generalizovať dosiahnutú zhodu medzi posudzovateľmi a ich konzistentnosť.

V súčasnej dobe Open Science a čoraz častejšieho využívania preregistrácie je vhodné do nej zahrnúť aj rozhodnutie o počte posudzovateľov, počte objektov/subjektov, ako aj rozhodnutie o vybranom indexe. Mali by sme poskytnúť taktiež informácie o nástroji hodnotenia, jeho stupnici, o tom, čo bude hodnotené (napr. či pôjde o zúčastnené pozorovanie alebo sa budú hodnotiť videozáznamy či rôzne nahrávky, prepisy a pod.) a tiež o celej schéme hodnotenia (či všetci posudzovatelia budú hodnotiť všetko, alebo či necháme časť posudzovateľov hodnotiť polovicu a druhú časť druhú polovicu a pod.). Ide o kroky, ktoré sa dejú ešte pred realizáciou výskumu, takže z tohto pohľadu by to nemalo byť problematické. Uvedomujeme si ale, že to môže byť rovnako náročné ako napríklad nastavenie experimentu, takže najmä dôkladné premyslenie a nastavenie celého procesu hodnotenia je veľmi dôležité.

Predkladaný článok rieši základné problémy, ktoré sa vyskytnú pri reportovaní súhlasu a spoľahlivosti posudzovateľov. Okrem nich existujú ďalšie témy, ktoré s danou problematikou súvisia a ktorým sa v článku nevenujeme, prípadne iba okrajovo. V prvom rade ide o teoretické vymedzenie a zarámcovanie súhlasu a spoľahlivosti posudzovateľov v kvantitatívnom aj kvalitatívnom výskume. Téme sa venujú napríklad Tinsley & Weiss (1975), Stemler (2004), LeBreton & Senter (2008), či Kottner & Streiner (2011).

Dôležitou témou je aj veľkosť vzorky, ktorá vplýva na mieru súhlasu a spoľahlivosti posudzovateľov a o jej determinovaní sa viac dočítate napríklad v článkoch od Temela & Erdogana (2017) alebo od Shana (2018). A nakoniec ide o vizualizáciu súhlasu posudzovateľov, kde sa používa Bland-Altmanov graf, ktorý bol pôvodne vyvinutý pre

meranie súhlasu v štúdiách porovnávajúcich metódy, ale môže byť použitý tiež v tomto prípade (Bland & Altman, 1999). V posledných rokoch sa vytvárajú aj nové techniky pre zobrazenie napr. pre koeficient kappa (Eubanks, 2017) a zároveň sa viete dopracovať aj ku kódom na vytvorenie grafu v programe R (napr. https://www.datanovia.com/en/lessons/inter-rater-agreement-chart-in-r/).

Zdroje

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing: National Council on Measurement in Education.* Washington DC: American Educational Research Association.

Abedi, J., Baker, E. L., & Herl, H. (1995). *Comparing reliability indices obtained by different approaches for performance assessments* (CSE Report 401). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bikmaz Bilgen, Ö., & Doğan, N. (2017). The comparison of interrater reliability estimating techniques. *Journal of Measurement and Evaluation in Education and Psychology*, 8(1). 63–78. https://doi.org/10.21031/epod.294847

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical methods in medical research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Bogartz, R. S. (2005). *1 Interrater Agreement and Combining Ratings*. http://people.umass.edu/~bogartz/Interrater%20Agreement.pdf

Cicchetti, D. V., & 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

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* Vol. 20, No. 1, 1960. 37–46. https://doi.org/10.1177/001316446002000104

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1) 1, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Dettori, J. R., & Norvell, D. C. (2020). Kappa and Beyond: Is There Agreement? *Global spine journal*, *10*(4), 499–501. https://doi.org/10.1177/2192568220911648

Eubanks, D.A. (2017). (Re)Visualizing Rater Agreement: Beyond Single-Parameter Measures. *Journal of Writing Analytics,* (1). 276-310. https://doi.org/10.37514/JWA-J.2017.1.1.10

von Eye, A., & Mun, E. Y. (2005). *Analyzing Rater Agreement: Manifest Variable Methods* (1st ed.). Psychology Press. p. 202 ISBN 0-8058-4967-X https://doi.org/10.4324/9781410611024

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, *43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-l

Feng, G. C. (2013). Underlying determinants driving agreement among coders. *Quality & Quantity: International Journal of Methodology, 47*(5), 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

Fradenburg, L. A., Harrison, R. J., & Baer, D. M. (1995). The effect of some environmental factors on interobserver agreement. *Research in Developmental Disabilities, 16*(6), 425–437. https://doi.org/10.1016/0891-4222(95)00028-3

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science,* 21(3), 177–182. https://doi.org/10.1177/0963721412445309

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *irr: Various coefficients of interrater reliability and agreement [computer software].* https://CRAN.R-project.org/package=irr

Gálová, L. (2010a). *Koeficient kappa -aplikačné možnosti, výhody a nevýhody.* In: 2. Česko-slovenská konference doktorandů oborů pomáhajících profesí: sborník z vědecké konference konané v Ostravě 3. února 2010. Ostravská univerzita.

Gálová, L. (2010b). *Konsenzus a konzistencia hodnotenia klímy školskej triedy.* In: Sociálne procesy a osobnosť 2010: zborník z medzinárodnej vedeckej konferencie. Košice 20. - 22. september 2010. - Bratislava: SAV.

Gálová, L. (2014). Výber indexu súhlasu žiakov pri hodnotení interakčného štýlu učiteľov. *Forum statistikum slovacum* 2/2014. ISSN 1336-7420 31-35.

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

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

Gwet, K. L. (2002a). *Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity*. Statistical Methods for Inter-Rater Reliability Assessment No. 2, May 2002 1–9.

Gwet, K. L. (2002b). *Kappa statistic is not satisfactory for assessing the extent of agreement between raters.* Statistical Methods for Inter-Rater Reliability Assessment, No. 1, April 2002 1–5.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. School Psychology Review, 33(2), 258–270. https://doi.org/10.1080/02796015.2004.12086247

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Hřebíčková, M. (2003). Metodologické souvislosti výzkumu shody mezi sebeposouzením a posouzením druhými. *Československá Psychologie: Časopis Pro Psychologickou Teorii a Praxi, 47*(6), 533–547.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & 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

Kottner, J., & 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

LeBreton, J. M., & 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

Liao, S. C., Hunt, E. A., & Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Annals of the Academy of Medicine*, 39(8), 613–618.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. Language Testing, 12, 54–72. https://doi.org/10.1177/026553229501200104

Marko, M. (2016). Využitie a zneužitie Cronbachovej alfy pri hodnotení psychodiagnostických nástrojov. *Testfórum* 5(7), 99–107. https://doi.org/10.5817/TF2016-7-90

McDonald, N., Schoenebeck, S., & 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*, November 2019 Article No.: 72 https://doi.org/10.1145/3359174

Nying, E. (2004). *A Comparative Study of Interrater Reliability Coefficients Obtained from Different Statistical Procedures Using Monte Carlo Simulation Techniques*. *Dissertations*. 1267. https://scholarworks.wmich.edu/dissertations/1267

O'Neill, T. A. (2017). An Overview of Interrater Agreement on Likert Scales for Researchers and Practitioners. *Frontiers Psychology* 8:777. 1–15. https://doi.org/10.3389/fpsyg.2017.00777

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris I. N. Gallhofer (Eds.), *Sociometric research* (pp. 90–105). London, UK: Palgrave Macmillan.

Shan, G. (2018). Sample size calculation for agreement between two raters with binary endpoints using exact tests. *Statistical Methods in Medical Research*, *27*(7), 2132–2141. https://doi.org/10.1177/0962280216676854

Slaug, B., Schilling, O., Helle, T., Iwarsson, S., Carlsson, G., & Brandt, Å. (2012). Unfolding the phenomenon of interrater agreement: a multicomponent approach for in-depth examination was proposed. *Journal of clinical epidemiology*, *65*(9), 1016–1025. https://doi.org/10.1016/j.jclinepi.2012.02.016

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. In: A peer-reviewed electronic journal. *Practical Assessment, Research & Evaluation*, 9(4). https://doi.org/10.7275/96jp-xz07

Stolarova, M., Wolf, C., Rinker, T., & Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in psychology*, *5*, 509. https://doi.org/10.3389/fpsyg.2014.00509

Temel G., & Erdogan, S. (2017). Determining the sample size in agreement studies. *Marmara Medical Journal*. 30: 101–112. https://doi.org/10.5472/marumj.344822

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2018). On the usefulness of interrater reliability coefficients. In Wiberg, M., Culpepper, S., Janssen, R., González, J., & Molenaar, D. (Eds.), *Quantitative Psychology*: The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017 (pp. 67-75). (Springer Proceedings in Mathematics & Statistics; Vol. 233). Springer.

Tinsley, H. E., & 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. (2008). *Statistical methods for rater agreement*. http://www.john-uebersax.com/stat/agree.htm

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

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

Volchok, E. (2015). *Measurement & Measurement scales.* http://media.acc.qcc.cuny.edu/faculty/volchok/Measurement Volchok/index.html

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology*, 95(3), 546–561. https://doi.org/10.1037/a0018866

Wilhelm, A. G., Rouse, A. G., & Jones, F. (2018). Exploring Differences in Measurement and Reporting of Classroom Observation Inter-Rater Reliability. *Practical Assessment, Research, and Evaluation:* Vol. 23 , Article 4. https://doi.org/10.7275/at67-md25

Wongpakaran, N., Wongpakaran, T., & Wedding, D. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology* 13, 61 https://doi.org/10.1186/1471-2288-13-6

Xie, Q. (2013). *Agree or disagree? A demonstration of an alternative statistic to Cohens kappa for measuring the extent and reliability of agreement between observer.* In Proceedings of the Federal Committee on Statistical Methodology Research

Conference, The Council of Professional Associations on Federal Statistics, Washington, DC, USA, 2013. https://nces.ed.gov/FCSM/pdf/J4_Xie_2013FCSM.pdf

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

Lucia Kočišová (2022): Reporting of inter-rater agreement and inter-rater reliability

Abstract: *In psychology as well as in many other areas we use multiple raters to assess the validity and reliability of our results. It is the inter-rater agreement, which represents the raters consensus, and if it reaches a certain level, the raters are interchangeable (Tinsley & Weiss, 1975), and the inter-rater reliability in terms of consistency of assessment (LeBreton & Senter, 2008). Both concepts differ in definition, answer different research questions and differ in methodology. The aim of this article is to answer questions related to the practical need to report the inter-rater agreement and the inter-rater reliability. The article is looking for answers to these questions: How many raters should be used? How to choose a suitable index of inter-rater agreement and inter-rater reliability? Is there an accepted level of agreement and reliability of the raters? What factors affect inter-rater agreement and inter-rater reliability?*

Key words: *inter-rater agreement, inter-rater reliability, index selection*

Príloha 1: Prehľad indexov pre zisťovanie súhlasu a spoľahlivosti posudzovateľov vzhľadom na typ výskumnej otázky, počet posudzovateľov a úroveň merania dát

názov indexu, označenie, autor/i	**súhlas/spoľahlivosť posudzovateľov**	**počet posudzovateľov**	**úroveň merania dát**
podiel celkového súhlasu (%)	súhlas	neobmedzene	nominálna, ordinálna, intervalová
podiel pozitívneho súhlasu	súhlas	2	nominálna
podiel negatívneho súhlasu	súhlas	2	nominálna
Cohenov koeficient kappa κ	súhlas	2	nominálna
vážený koeficient kappa κ_w	súhlas	2	nominálna
Fleissov koeficient kappa κ_{Fleiss}	súhlas	>2	nominálna
exaktný koeficient kappa κ_{Exact} (Conger)	súhlas	>2	nominálna
Koeficient kappa κ_{Light} (Light)	súhlas	>2	nominálna
Brennan-Predigerov koeficient kappa κ_{BP}	súhlas	>2	nominálna
Krippendorffov koeficient alfa α	súhlas	>2	nominálna, ordinálna, intervalová
koeficient π (Scott)	súhlas	2	nominálna
koeficient S (Bennet, Alpert, Goldstein)	súhlas	2	nominálna
Jaccardov koeficient J	súhlas	>2	nominálna
Stuart-Maxwellov test X^2	súhlas	2	nominálna, ordinálna
McNemarov test X^2	súhlas	2	nominálna, ordinálna
Bhapkar test X^2	súhlas	2	nominálna
koeficient AC_1 (Gwet)	súhlas	neobmedzene	nominálna
Kendallov koeficient τ	súhlas	2	ordinálna
Kendallov koeficient W	súhlas	>2	ordinálna
koeficient AC_2 (Gwet)	súhlas	neobmedzene	ordinálna, intervalová
r_{wg} koeficient (James, Demaree, Wolf)	súhlas	neobmedzene	ordinálna, intervalová
t-test pre dva závislé výbery	súhlas	2	intervalová

priemerná odchýlka AD	súhlas	neobmedzene	intervalová
pozorovateľná variabilita dát (napr. štandardná odchýlka)	súhlas	neobmedzene	intervalová
ANOVA (two-way)	súhlas	>2	intervalová
„latent class model“	spoľahlivosť	neobmedzene	nominálna
longlineárne, asociačné alebo kvázi symetrické modely	spoľahlivosť	2	nominálna
tetrachorický korelačný koeficient	spoľahlivosť	2	nominálna
„latent trait“ model	spoľahlivosť	>2	nominálna
vnútrotriedne korelácie ICC	spoľahlivosť	neobmedzene	nominálna, ordinálna, intervalová
polychorické korelácie	spoľahlivosť	2	ordinálna
Spearmanov koeficient rho	spoľahlivosť	2	ordinálna
Pearsonov korelačný koeficient r	spoľahlivosť	2	intervalová
Cronbachov koeficient alfa α	spoľahlivosť	>2	intervalová

Príloha 2:

Usmernenia pre reportovanie súhlasu a spoľahlivosti posudzovateľov (Kottner et al., 2011)[1]

Názov a abstrakt:

1. Identifikujte v názve alebo abstrakte, či bude skúmaná spoľahlivosť a/alebo súhlas.

Úvod

2. Pomenujte a explicitne popíšte diagnostický alebo merací nástroj.

3. Špecifikujte populáciu výberu.

4. Špecifikujte populáciu posudzovateľov (ak je to možné).

5. Popíšte čo je známe o reliabilite a súhlase a poskytnite odôvodnenie výskumu (ak je to možné).

[1] Tieto usmernenia sú vhodné, ak je súhlas alebo spoľahlivosť posudzovateľov primárnym záujmom výskumníka. V prípade, že to tak nie je, napr. zisťovanie súhlasu/spoľahlivosti posudzovateľov je súčasťou overovania validity či reliability, aj tak je vhodné riadiť sa odporúčaniami a uvádzať dôležité informácie. Nemusí to však byť v prezentovanej štruktúre.

Metódy

6. Popíšte, ako bol uskutočnený výber vzorky. Uveďte určený počet posudzovateľov, subjektov/objektov a počet opakovaných pozorovaní.

7. Popíšte metódu zberu dát.

8. Popíšte proces merania/hodnotenia (napr. časový interval medzi opakovanými meraniami).

9. Uveďte, či merania/hodnotenia boli vykonávané nezávisle.

10. Popíšte štatistickú analýzu dát.

Výsledky

11. Uveďte aktuálny počet posudzovateľov a subjektov/objektov, ktoré boli zahrnuté a počet replikácií pozorovaní, ktoré boli vykonané.

12. Popíšte charakteristiky výskumnej vzorky, posudzovateľov a subjektov (napr. tréning, skúsenosti).

13. Reportujte odhady spoľahlivosti a súhlasu zahŕňajúce meranie štatistickej neistoty.

Diskusia

14. Diskutujete o praktickej relevantnosti výsledkov.

Pomocný materiál

15. Ak je to možné, poskytnite podrobné výsledky (napr. online).