CHYBA MĚŘENÍ A ODHAD PRAVÉHO SKÓRU: Připomenutí některých postupů Klasické testové teorie

HYNEK CÍGLER[1], MARTIN ŠMÍRA[1]

Abstrakt: *Práce s chybou měření patří k základním dovednostem při interpretaci výsledků psychologických výsledků. Bohužel, řada českých psychologických metod však neobsahuje veškeré informace o chybě měření, například intervaly spolehlivosti či odhad standardní chyby měření pro různá použití. I v případě, že tyto informace jsou dostupné, je často nutné zvážit i další okolnosti a způsob výpočtu přizpůsobit – ne vždy je přitom možné se spolehnout na informace poskytnuté distributorem testu. Ani v současné počítačové době navíc nejsou jednoduše dostupné příslušné aplikace a řadu základních výpočtů by si tak psycholog v ideálním případě měl umět provést sám. Článek v krátkosti shrne běžné postupy při interpretaci chyby měření s využitím intervalů spolehlivosti v rámci klasické testové teorie, a to včetně podrobných příkladů, aby text mohl sloužit jako návod pro psychology z praxe.*

Klíčová slova: *Klasická testová teorie, CTT, standardní chyba měření, SEM, interpretace testových výsledků*

Pro uskutečnění kvalitního diagnostického závěru je vždy nezbytné nějakým způsobem zvážit chybu měření. To lze provést řadou způsobů, ty však zpravidla nepatří do běžného repertoáru dovedností psychologa-diagnostika a konkrétní návod není součástí valné většiny metod používaných v České republice. Ovšem i v testech, kde jsou potřebné informace uvedeny, bývají některé údaje spočítány chybným či přinejmenším problematickým způsobem. Kritika těchto používaných postupů se navíc objevuje dlouhodobě (např. Dudek, 1979).

Konkrétních postupů vyjádření nejistoty měření bylo vyvinuto značné množství. Tento článek popisuje postupy založené na klasické testové teorii, tedy zejména tzv. „regresní model klasické testové teorie“ v podobě publikované např. Lordem a Novickem (1968), které se ale v obdobné podobě objevují ve většině zahraničních psychometrických i psychodiagnostických učebnic až do současnosti.

[1] Katedra psychologie, Fakulta sociálních studií MU, Joštova 10, 602 00, Brno

Tento článek proto nepřináší žádné zásadní nové informace, přesto však celá řada postupů popisovaných níže není dle našich zkušeností v České republice příliš rozšířená. Cílem textu je představit českému čtenáři některé statistické vlastnosti klasické testové teorie s přímými důsledky pro psychodiagnostickou praxi. Snažíme se o srozumitelnost i pro čtenáře bez výraznějších statistických znalostí, proto některé pasáže popisujeme velmi podrobně. V případě, kdy existuje více paralelních řešení s různými výsledky, volili jsme zpravidla první „dostatečně správné" řešení s nejjednodušší možností interpretace. Řada informací je proto podávána ve zjednodušené podobě.

Klasická testová teorie

Základním axiomem klasické testové teorie („classical test theory", CTT, označované někdy také jako „true-score theory") je, že pozorovaný skór X_i respondenta i je součtem tzv. pravého skóru τ_i a náhodné chyby měření e_i na pravém skóru nezávislé:

$$X_i = \tau_i + e_i \,. \qquad (1)$$

Pokud jsou všechny proměnné normálně rozložené, platí shodný vztah i pro jejich rozptyly[2]:

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2 \qquad (2)$$

Reliabilita je následně definovaná jako poměr rozptylu pravého skóre k pozorovanému (přičemž pravý skór lze vyjádřit jako rozdíl pozorovaného a chybového rozptylu):

$$r_{xx'} = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}\,, \qquad (3)$$

kde poslední úprava je shodná s tzv. „koeficientem determinace". Je tedy patrné, že (1.) reliabilita je rovna teoretické korelaci metody se sebou samou ($r_{xx'}$) a (2.) korelace pravého a pozorovaného je odmocnina z reliability, tedy

$$r_{x\tau} = \sqrt{r_{xx'}}\,.[3] \qquad (4)$$

Tyto vztahy jsou klíčové pro veškeré regresní operace popisované dále. Pokud ze vzorce 3 vyjádříme σ_e^2 a rovnici odmocníme, dostaneme běžný vzorec pro výpočet **standardní chyby měření** (která bývá v manuálech diagnostických metod zpravidla označována jako SE):

$$\sigma_e = \sigma_x\sqrt{1 - r_{xx'}} \qquad (5)$$

Pokud by bylo možné opakovaně měřit jediného respondenta i, aniž by se změnila úroveň měřeného rysu (což v praxi pochopitelně není možné z důvodu vlivu únavy, zácviku apod.), měly by naměřené hodnoty průměr shodný s jeho pravým skórem τ_i

[2] Odmocněný rozptyl, tedy „σ", je samozřejmě směrodatná odchylka.

[3] Reliabilita je tedy shodná s rozptylem měření vysvětleným úrovní pravého skóre. Což je ostatně důvod, proč v následujících vzorcích nefiguruje druhá mocnina reliability, ale přímo reliabilita samotná (reliabilita je již „umocněná korelace" pravého a pozorovaného skóre).

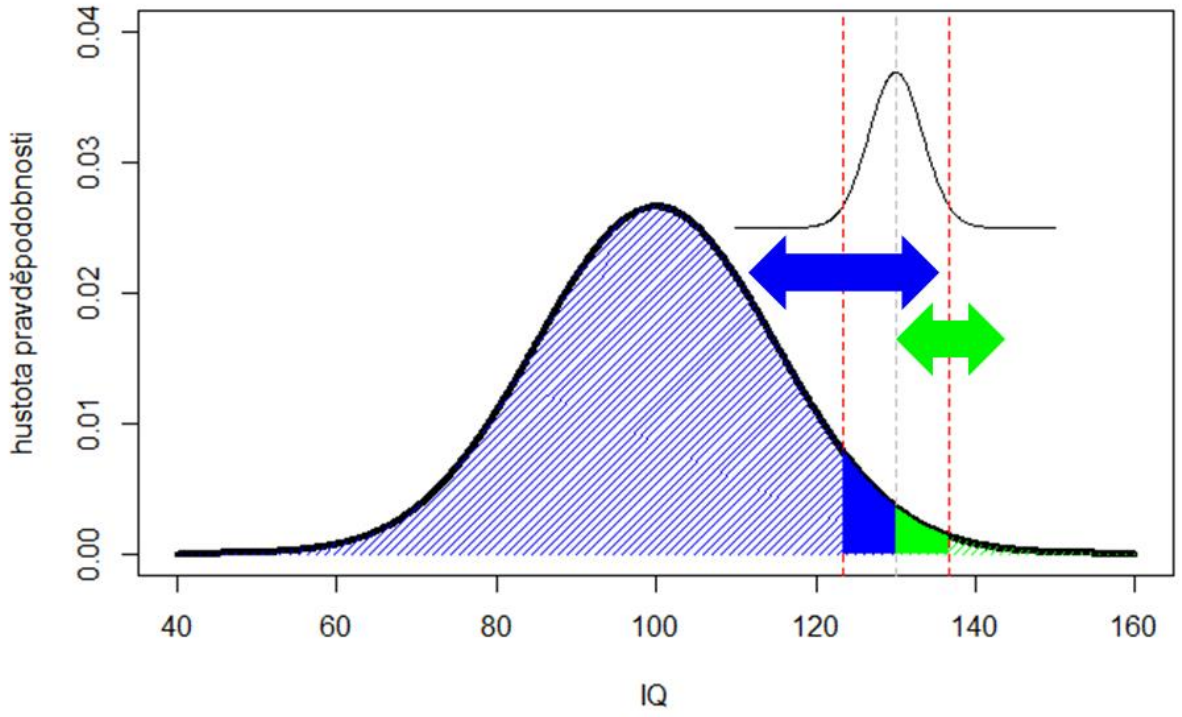
a směrodatnou odchylkou σ_i. Protože rozložení těchto „chybových" hodnot je normální a stejné pro všechny úrovně pravého rysu, je možné standardní chybu měření vynásobit příslušným kvantilem normálního rozložení[4] a sestavit intervaly spolehlivosti měření.

Regrese k průměru

Potenciálně pozorovatelné hodnoty jsou nicméně seskupeny kolem pravého skóru, nikoliv kolem aktuálně pozorovaného. Při pohledu na vzorec 2 je patrné, že pravé skóry budou mít nižší rozptyl (a tedy i směrodatnou odchylku) než skóry pozorované. Nejpravděpodobnější hodnota pravého skóru musí proto být o něco blíže k průměru než pozorovaný skór[5].

Tento příklad ilustruje Obrázek 1. Představme si situaci měření inteligenčním testem (M = 100, SD = 15) a reliabilitou 0,95, který má standardní chybu měření $\sigma_e = 15\sqrt{1 - 0{,}95} = 3{,}35$ (rozložení vpravo nahoře na grafu) a 95% interval spolehlivosti je tak 3,35 · 1,96 = ±6,67. Vidíme, že pokud bychom tento interval sestrojili kolem naměřené hodnoty 130 bodů IQ, směrem k průměru by obsahoval větší množství osob než směrem od průměru. Je tedy logické, že jsme spíše dotyčnou osobu „nadměřili" než „podměřili".

Obrázek 1: Rozložení chyby měření při pozorovaném skóre 130 (IQ) testem s reliabilitou 0,95. V polovině intervalu bližší průměru je více osob než v polovině vzdálenější.



[4] Pro 99% CI z = 2,58, 95% z = 1,96, pro 90% z = 1,64, pro 80 % z = 1,28 a pro 68% z = 1,0. Příslušný interval spolehlivosti je vhodné volit podle konkrétní diagnostické zakázky.

[5] Přesněji jde o průměr pravých skórů respondentů, u nichž jsme naměřili dané pozorované skóre.

Jinými slovy: nemůžeme si nikdy být jisti, zda je naměřená hodnota způsobena skutečnou úrovní rysu daného respondenta, nebo náhodnou chybou. Bez dalších informací proto lze předpokládat, že odchylku od průměru způsobila současně jak náhodná chyba, tak i skutečná úroveň měřeného rysu, a to v poměru definovaném reliabilitou.

Tento jev se obecně nazývá „regrese k průměru" a je patrný zejména při opakovaném měření: při zjištění extrémního (nadprůměrného či podprůměrného výkonu) v pretestu je pravděpodobné, že při retestu získáme hodnotu o něco blíže k průměru.

Zároveň je potřeba mít na paměti, že průměrem v tomto případě myslíme průměr populace, z níž je vybrán respondent[6].

Regresní model klasické testové teorie: Výpočet intervalů spolehlivosti

Zbývá tedy otázka: jaký je nejpravděpodobnější pravý skór při určitém pozorovaném skóru? Nejběžnější řešení využívá tzv. regresní model klasické testové teorie. Protože platí, že korelace pravého skóre a pozorovaného skóre je odmocnina z reliability (vzorec 4), pro test s průměrem 0 platí regresní vztah

$$6 \qquad \mathrm{E}[\tau] = r_{xx'} \cdot X,$$

tedy že očekávaný pravý skór $\mathrm{E}[\tau]$ je roven součinu odmocniny z reliability a naměřeného skóre[7]. Není už těžké odvodit vzorec pro test s průměrem odlišným od 0:

$$\mathrm{E}[\tau] = r_{xx'}(X - M_x) + M_x \,,$$

což lze upravit jako

$$7 \qquad \mathrm{E}[\tau] = r_{xx'}X + (1 - r_{xx'})M_x \,,$$

kde M_x je průměr pozorovaných skórů (shodný s průměrem pravých skórů). Z rovnice 7 je patrné (jak poukazují Lord a Novick, 1968), že čím více se reliabilita blíží jedné, tím vyšší váha je přikládaná pozorovaným hodnotám a nižší průměru; s reliabilitou klesající k nule naopak dostává na významnosti průměrný skór, kdežto pozorovaný skór hraje méně významnou roli.

[6] Zpravidla jde o průměr standardizačního vzorku. Pokud však běžně testujeme například klinickou populaci, která setrvale vykazuje nadprůměrný či podprůměrný skór, skóry respondentů budou regredovat k průměru těchto klinických pacientů. Protože však „skutečnou" příslušnost k potenciální klinické skupině neznáme, není podle nás chybou počítat s průměrem standardizačního souboru.

[7] Nesmíme zapomenout, že se jedná o nejpravděpodobnější, očekávanou hodnotu pravého skóru (odhad), nikoliv pravý skór jako takový. Proto je notace „$\mathrm{E}[\tau]$", nikoliv přímo τ. Symbol se označuje jako expektance, a udává „očekávanou" hodnotu nějaké proměnné. Ve výše uvedeném případě by $\mathrm{E}[\tau]$ bylo průměrem pravých skórů všech osob, kterým bychom naměřili pozorovanou hodnotu X, případně průměrným skórem retestu osob, kterým při pretestu vyšel pozorovaný skór X.

Směrodatná odchylka rozdílu pravého skóru τ a jeho odhadu $\mathrm{E}[\tau]$, odvozeného prostřednictvím vzorce 7, je označována jako „**standardní chyba odhadu**“ a její velikost lze odvodit z předchozích vzorců:

$$\sigma_{e(\tau)} = \sigma_\tau \sqrt{1 - r_{xx'}} ,$$

kde σ_τ je směrodatná odchylka pravého skóru. Protože platí, že $\sigma_\tau = \sigma_x \sqrt{r_{xx'}}$, lze rovnici upravit na

8 $$\sigma_{e(\tau)} = \sigma_x \sqrt{r_{xx'}} \sqrt{1 - r_{xx'}} .$$

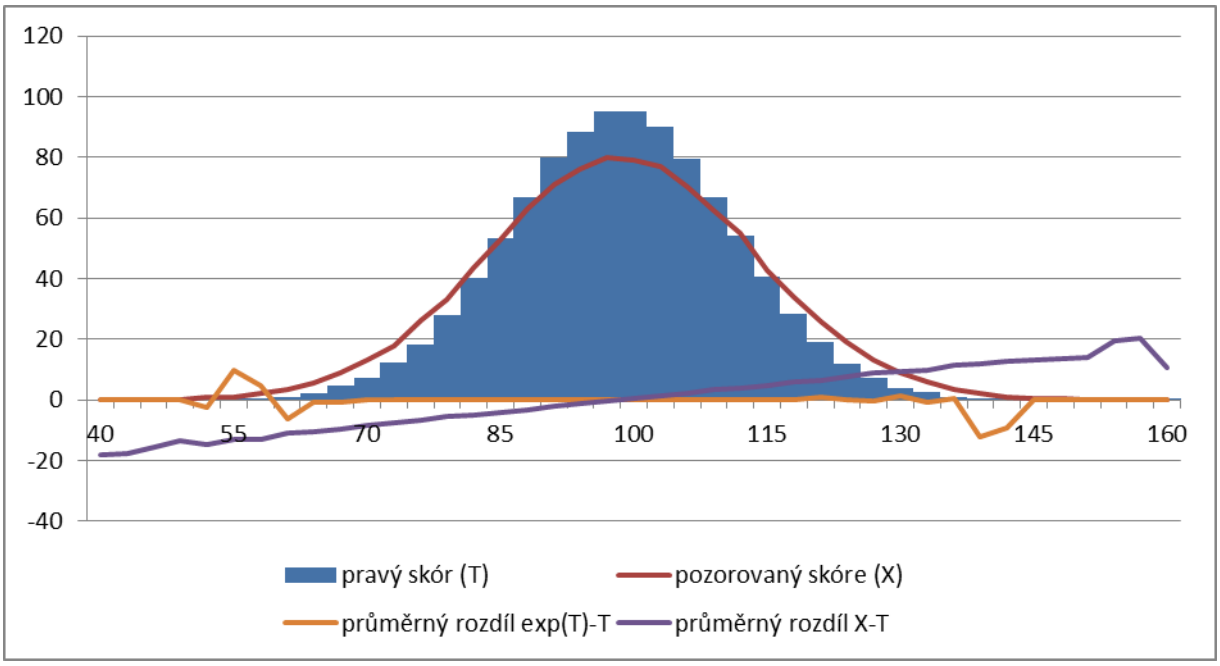
Tento vztah je prakticky totožný se vzorcem standardní chyby měření (5), obsahuje však navíc odmocninu z reliability.

Většinou autorit (např. Nunnally, 1978; Dudek, 1979; Lord a Novick, 1968; Revelle, 2015) doporučovaný postup výpočtu intervalu spolehlivosti naměřené hodnoty je proto odhadnout nejpravděpodobnější hodnotu pravého skóru pomocí vzorce 7 a následně kolem ní zkonstruovat interval spolehlivosti měření pomocí vzorce 5.[8] Protože pozorovaná hodnota je odlišná od středové hodnoty, kolem které interval spolehlivosti konstruujeme, je tento interval „asymetrický“ (či „posunutý“) vůči naměřené hodnotě.

Co se stane, když sestrojíme interval spolehlivosti klasickým způsobem – tedy na základě pozorovaného skóre, nikoliv odhadu skóre pravého? Konkrétní příklad předkládá Obrázek 2, který obsahuje grafické znázornění simulace dat 100 000 virtuálních respondentů inteligenčního testu (M = 100, SD = 15) s nízkou reliabilitou $r_{xx'}$ = 0,7. Simulace je výhodná, protože v tomto případě známe pravý skór každého respondenta, zároveň jej ale můžeme zkusit zpětně „odhadnout“ na základě dostupných pozorovaných skórů a obě hodnoty srovnat.

[8] Použijeme zpravidla standardní chybu měření (vzorec 5), nikoliv odhadu (vzorec 8), a to z toho důvodu, že většinou používáme škálu pozorovaných skórů. Jinými slovy chceme mít interval spolehlivosti pro naměřenou hodnotu – chceme vědět, jaké jiné hodnoty jsme mohli pozorovat, když jsme naměřili určitý pozorovaný skór. Podrobné a jednoduché srovnání chyb měření, odhadu a predikce viz Dudek (1979). Této problematice se nicméně věnujeme ještě v poznámce pod čarou č. 12.

Obrázek 2: Simulace chyb měření a odhadu (*N* = 100 000)



V grafu vidíme, že simulované pravé skóre (modrá plocha) má skutečně o něco menší směrodatnou odchylku než pozorované skóre (červená čára), konkrétně je směrodatná odchylka odhadů pravých skóre $\sqrt{0,7}$ = 0,84násobek směrodatné odchylky pozorovaných skórů, což je rovno $0,84 \cdot 15 = 12,55$ bodů IQ namísto 15[9]. Fialová čára je průměrný rozdíl pozorovaného a pravého skóre ve zvoleném intervalu (širokém 3 body IQ). Vidíme, že vpravo nad průměrem jsme skutečně zpravidla „nadměřili“, a to o téměř 20 bodů při extrémních skórech; vlevo pod průměrem jsme analogicky „podměřili“. Oranžová značí průměrný rozdíl odhadu pravého skóru a skutečných pravých skórů, který je až na drobné náhodné odchylky v extrémních hodnotách (s nižším počtem pozorování) skutečně vždy blízký nule.

Příklad výpočtu

Ukažme si výpočet na konkrétním příkladu respondenta, který v inteligenčním testu (M = 100, SD = 15) s reliabilitou 0,7 dosáhl 130 bodů IQ – veškeré výpočty obsahuje Tabulka 1.

Vidíme, že nejpravděpodobnější skutečná hodnota respondentova pravého skóru, ačkoliv jsme „naměřili“ pozorovaný IQ skór 130, je jen 121 a s 95% pravděpodobností leží v rozmezí ±13,47, tedy 107,53–134,47.

Více nás však zajímá interval spolehlivosti měření, který je na škále pozorovaného skóru (která bývá zpravidla použitá ke konstrukci norem) a lze tedy snadno interpretovat. Při použití odhadu pravého skóru a standardní chyby měření sestrojíme 95% interval

[9] To je samozřejmě teoretický předpoklad – reálná hodnota bude nepatrně jiná pro každou simulaci stejně, jako pro každý opravdový sběr dat.

spolehlivosti měření 121±16,10 = 104,9–137,1 bodů IQ. Tento závěr lze použít pro účely diagnostické zprávy a může znít například: „Respondent dosáhl skóre 130 bodů IQ s 95% intervalem spolehlivosti 105–137“.

Pokud bychom znovu testovali všechny respondenty, kteří poprvé v testu dosáhli skóre 130 bodů, při opakovaném testování by jejich průměrný výkon byl trochu neintuitivně „jen“ 121 bodů IQ. Velký rozdíl devíti bodů je způsobený právě nepříjemně nízkou reliabilitou celého testu.

Tabulka 1: Příklad výpočtů regresního modelu CTT

Standardní chyba měření	$\sigma_e = \sigma_X\sqrt{1 - r_{xx'}} = 15\sqrt{1 - 0{,}7} = 8{,}22$
šířka 95% intervalu spolehlivosti měření	$CI_X = \pm z \cdot \sigma_e = \pm 1{,}96 \cdot 8{,}22 = \pm 16{,}10$
odhad pravého skóru	$\mathrm{E}[\tau] = r_{xx'}X + (1 - r_{xx'})M_x = 0{,}7 \cdot 130 + (1 - 0{,}7) \cdot 100 = 121$
standardní chyba odhadu	$\sigma_{e(\tau)} = \sigma_X\sqrt{r_{xx'}}\sqrt{1 - r_{xx'}} = 15\sqrt{0{,}7}\sqrt{1 - 0{,}7} = 6{,}87$
šířka 95% intervalu spolehlivosti odhadu	$\pm CI_{E(\tau)} = \pm z \cdot \sigma_{e(\tau)} = \pm 1{,}96 \cdot 6{,}87 = \pm 13{,}47$

Dvě a více měření: Standardní chyba rozdílu

Zatím jsme se věnovali pouze situaci, kdy máme jediné měření. Pokud chceme srovnat respondentův výkon s nějakou kritickou hodnotou, stačí zkonstruovat interval spolehlivosti na požadované hladině pravděpodobnosti a podívat se, zda hodnota v intervalu neleží. Co ale když jsou měření dvě (samozřejmě na stejné škále) a my chceme zjistit, zda se od sebe neliší[10]?

Už na konci šedesátých let Payne a Gwynne Jones (1959) publikovali studii o rozdílových skórech, ve kterých popsali některé dodnes používané postupy pro hlavní diagnostické otázky zahrnující do jisté míry i regresi k průměru. Postup byl nicméně časem vylepšován a dnes existuje více různých běžně používaných postupů, přičemž jejich závěry a možnosti interpretace se do jisté míry liší (Charter a Feldt, 2000).

[10] Nejjednodušší možnost, přímé srovnání dvou intervalů spolehlivosti zkonstruovaných separátně pro obě změřené hodnoty, je jen orientační řešení pro ty případy, kdy nemáme k dispozici kalkulačku nebo počítač. Tento postup je bez dalších korekcí „přísnější“ než přesný výpočet, a navíc je i tak nutné řešit metodologické zádrhely popsané v následujících odstavcích. Konkrétní rozdíl oproti přímému výpočtu prezentují např. Charter a Feldt (2000). Protože každý má dnes kalkulačku v mobilu, srovnávání dvou intervalů spolehlivosti v diagnostické situaci nepovažujeme za vhodné.

Obecný postup pro testování rozdílu dvou skórů počítá s tím, že výsledkem součtu či rozdílu dvou náhodných normálně rozložených proměnných je opět normální rozložení s rozptylem $\sigma^2_{A\pm B}$, který má hodnotu

9 $$\sigma^2_{A\pm B} = \sigma^2_A + \sigma^2_B \pm 2r_{AB}\sigma_A\sigma_B \,,$$

kde σ_A a σ_B jsou směrodatné odchylky proměnných A a B a r_{AB} je jejich korelace[11]. Protože chyby měření jsou na sobě (teoreticky) nezávislé a jejich vzájemná korelace je tedy nulová, v případě rozdílu dvou proměnných se známými chybami měření $\sigma^2_{e(A)}$ a $\sigma^2_{e(b)}$ lze ze vzorce 9 odvodit tzv. standardní chybu rozdílu

10 $$\sigma_{e(A-B)} = \sqrt{\sigma^2_{e(A)} + \sigma^2_{e(B)}} \,.$$

Situaci nicméně komplikuje několik rozhodnutí – můžeme totiž srovnávat

1. přímo pozorované skóry;
2. odhady pravých skórů;
3. odhad pravého skóre z jednoho měření s pozorovaným skóre měření druhého,

a to s využitím chyb měření (vzorec 5), chyb odhadu (vzorec 8), nebo nějaké jejich kombinace. Zároveň obě měření mohou, ale nemusejí mít stejnou reliabilitu, mohou být provedena tím stejným, nebo jiným testem, mohou na sobě být statisticky závislá či nezávislá atp.[12] Každá diagnostická otázka předpokládá odlišnou volbu mezi těmito rozhodnutími.

V následující části článku proto představíme základní diagnostické otázky a vybereme vždy jedno preferované řešení.

Rozdíl dvou opakovaných měření stejným testem u jednoho respondenta

Zde uvedený postup využijeme při opakovaném měření. Příkladem může být diagnostika organického poškození mozku, pokud známe původní míru schopností, měření efektu terapie, efektu učení[13] atd.

[11] Podotýkáme, že $r_{AB}\sigma_A\sigma_B$ je jejich kovariance.

[12] Situaci navíc komplikuje fakt, že některé testy – jmenovitě například WAIS-III^{UK} – používají škálu pravých skórů, nikoliv skórů pozorovaných (Charter a Feldt, 2000). Je však otázkou, nakolik tento postup využívá i česká standardizace WAIS-III, manuál v tomto směru mlčí. Jiné testy, například WJ-IE-II, jsou konstruovány s využitím teorie odpovědi na položku a výsledné skóry tak mohou být již regresním odhadem „latentního rysu“. I ve všech těchto případech však použití zde uvedeného základního postupu není chybné.

[13] Zejména v posledních dvou příkladech může být kompenzovat efekt zlepšení či zhoršení společného pro celý vzorek osob. Touto problematikou se zda nezabýváme.

Protože reliabilita je „korelace metody se sebou samou“, lze využít jednoduchého vzorce pro standardní chybu odhadu, která bývá v tomto případě označována (např. Lord a Novick, 1968) jako „**standardní chyba predikce**“:

$$11 \qquad \sigma_{pred} = \sigma_x\sqrt{1 - r_{xx'}^2}$$

kde σ_x je standardní odchylka pozorovaných skórů a $r_{xx'}^2$, druhá mocnina reliability. Výsledkem je očekávaná směrodatná odchylka rozdílu pozorovaného skóre v retestu a odhadu pravého skóre v pretestu – z důvodu regrese k průměru při druhém měření je nevhodné srovnávat přímo naměřené hodnoty (Lord a Novick, 1968; Dudek, 1979; Nunnally, 1978).

Dudek (1979) upozorňuje na zajímavý fakt, že standardní chyba predikce je kombinací chyby odhadu a chyby měření podle výše uvedeného vzorce 10 (čtenář může do následujícího vzorce 12 dosadit rovnice 5 a 8, výsledkem úprav je právě vzorec 11):

$$12 \qquad \sigma_{pred} = \sqrt{\sigma_e^2 + \sigma_{e(\tau)}^2}\,.$$

Příklad výpočtu
Respondenta z předchozího příkladu (IQ = 130) jsme testovali ještě jednou tím stejným testem a naměřili jsme hodnotu 105. Chceme vědět, zda se změnila úroveň latentního rysu – tedy zda je podaný výkon statisticky významně odlišný od výkonu podaného v pretestu.

Predikovaný skór na základě pretestu je 121 (viz předchozí příklad), testovaný rozdíl je tedy „jen“ $121 - 105 = 16$ bodů. Následně použijeme k výpočtu standardní chyby predikce vzorec 11: $\sigma_{pred} = 15\sqrt{1 - 0{,}7^2} = 10{,}71$. 95% interval spolehlivosti tedy je $CI_{95\%} = \pm 10{,}71 \cdot 1{,}96 = \pm 21{,}00$. Rozdíl 16 bodů je menší než kritická hodnota 21, proto můžeme říct, že na 5% hladině významnosti se výsledky neliší. Na tomto místě je důležité zmínit, že pokud bychom zanedbali regresi k průměru, rozdíl by byl 25 bodů, a tedy na 5% hladině významnosti signifikantní (což by však byl chybný závěr).

Rozdíl dvou skórů dvou respondentů v jednom testu

Odlišná situace může nastat, pokud potřebujeme srovnat výkon dvou respondentů v jednom testu. Protože jsou obě měření nezávislá, lze k tomu použít přímo standardní chybu rozdílu uvedenou ve vzorci 10. Protože jsou navíc oba respondenti měřeni tím samým testem, jsou obě směrodatné chyby pod odmocninou stejné, $\sigma_{e(A)} = \sigma_{e(B)} = \sigma_x$, a lze dosazením vzorce 5 výpočet zjednodušit na

$$\sigma_{e(A-B)} = \sqrt{\sigma_{e(A)}^2 + \sigma_{e(B)}^2} = \sqrt{2\left(\sigma_x\sqrt{1 - r_{xx'}}\right)^2}\,,$$

což lze upravit jako

$$13 \qquad \sigma_{e(A-B)} = \sigma_x\sqrt{2}\sqrt{1 - r_{xx'}}\,,$$

kde σ_x je směrodatná odchylka testu a $r_{xx'}$ jeho reliabilita. Protože testujeme nulovou hypotézu, že rozdíl pravých skórů je roven nule ($\tau_A - \tau_B = 0$), předpokládáme, že rozdíl mezi dvěma pozorováními X_A a X_B je způsoben pouze chybami měření: $X_A - X_b = \tau_A - \tau_B + e_A - e_B = 0 + e_A - e_B$. Pro směrodatnou odchylku rozdílu $e_A - e_B$ pak platí přímo vztah uvedený ve vzorci 13, a proto také lze v tomto případě srovnat přímo naměřené hodnoty bez regrese k průměru.

Příklad výpočtu

V případě testu s reliabilitou $r_{xx'} = 0{,}7$ bude 95% interval spolehlivosti pro rozdíl pozorovaných skórů dvou respondentů roven $CI_{95\%} = \pm z \cdot \sigma_X \sqrt{2}\sqrt{1 - r_{xx'}} = \pm 1{,}96 \cdot 15\sqrt{2}\sqrt{1 - 0{,}7} = 22{,}77$ bodů IQ.

Teprve při rozdílu větším než 22,77 bodů můžeme tvrdit, že na hladině pravděpodobnosti $p < 0{,}05$ má jeden respondent vyšší skór než ten druhý.

Rozdíl ve dvou různých testech u jediného respondenta

Poslední situací je případ „ipsativní“ diagnostiky, tedy když srovnáváme výkon jediného respondenta ve dvou různých testech a zjišťujeme „strukturu“ jeho schopností. Příkladem může být test rozdílu skóre v subtestu a celkového skóru inteligenční baterie, nebo výkonu ve dvou různých subtestech. Tato situace je velmi podobná předchozím, lze se však na ni dívat z více úhlů pohledu.

Rozdíl ve dvou rovnocenných testech

Zaprvé a nejjednodušeji se můžeme ptát, zda respondent získal v jednom testu vyšší či nižší skór než v testu druhém. V literatuře (např. Furr a Bacharach, 2014; Harvill, 1991; také i Payne a Gwynne Jones, 1959) je nejčastěji uváděn následující postup.

Protože odhadujeme oba skóry nezávisle na sobě a žádný test není „primární“, je situace shodná s předchozím příkladem a použijeme vzorec 10 s tím rozdílem, že chyby měření $\sigma_{e(x)}$ a $\sigma_{e(y)}$ se týkají odlišných testů s různou reliabilitou. Po dosazení tedy platí, že

$$\sigma_{\Delta xy} = \sqrt{\sigma_{e(x)}^2 + \sigma_{e(y)}^2} = \sqrt{\sigma_{xy}^2(1 - r_{xx'}) + \sigma_{xy}^2\left(1 - r_{yy'}\right)},$$

což lze upravit jako

$$\sigma_{\Delta xy} = \sigma_{xy}\sqrt{2 - r_{xx'} - r_{yy'}}, \qquad 14$$

kde $\sigma_{e(X)}^2$ a $\sigma_{e(Y)}^2$ jsou standardní chyby měření testů X, Y, $r_{xx'}$ a $r_{yy'}$ jsou jejich reliability a σ_{xy} je společná směrodatná odchylka obou testů (oba testy musí být pochopitelně převedeny na shodné jednotky).

Tento postup opět předpokládá, že rozdíl pozorovaných skórů byl způsoben pouze chybou měření, pravé skóry jsou tedy shodné a nepoužívá regresi k průměru. Chceme však upozornit, že takový předpoklad není zcela adekvátní a může vést k dezinterpretaci zjištěných rozdílů. Protože směrodatná odchylka pravých skórů testů s různou

reliabilitou je též různá, znamená to, že tentýž pravý skór může v každém testu znamenat odlišnou úroveň výkonu. Například pravý skór 121 v testu s reliabilitou 0,7 odpovídá 95. percentilu; tentýž pravý skór v testu s reliabilitou 0,9 odpovídá jen 93. percentilu (za předpokladu normálního rozložení pravého skóru).

Proto nelze předpokládat, že stejné „úrovni schopností" odpovídá ve dvou testech s různou reliabilitou též shodný pravý skór – ačkoliv právě ověření shodnosti schopností (nikoliv pravých skórů) je naším cílem[14]. Je tedy vhodnější vzít v úvahu regresi pozorovaných skórů k průměru, a navíc je převést na jednotky se stejnou směrodatnou odchylkou. Předchozí postup testoval, zda jsou dva pravé skóry shodné – následující postup ověřuje, zda se neliší úroveň schopností respondenta.

Z předchozích vzorců lze snadno odvodit řešení (podrobně viz Přílohu 1). Rozdíl pravých skórů po převedení na shodnou škálu (pozorovaných skórů) má hodnotu

$$15 \qquad E\left(\tau'_x - \tau'_y\right) = \sqrt{r_{xx'}}(X - M) - \sqrt{r_{yy'}}(Y - M)$$

se standardní chybou

$$16 \qquad \sigma_{\Delta\tau'_x\tau'_y} = \sigma_{xy}\sqrt{2 - r_{xx'} - r_{yy'}}\,,$$

která je shodná se vzorcem 14. Standardní chyba je proto při použití obou postupů shodná, liší se ale testovaný rozdíl.

Je patrné, že pokud jsou obě reliability stejně vysoké, platí $\sqrt{r_{xx'}} = \sqrt{r_{yy'}}$, a testovaný rozdíl podle vzorce 15 bude roven $\sqrt{r_{xx'}}(X - Y)$, což je menší než rozdíl pozorovaných skórů $X - Y$ (je nutný vyšší rozdíl naměřených skórů pro dosažení shodné statistické významnosti). Lze nicméně snadno dokázat, že rozdíl $E\left(\tau'_x - \tau'_y\right)$ může být nejen nižší, ale i vyšší[15] než rozdíl $X - Y$, záleží na vzájemných vztazích obou reliabilit i pozorovaných skórů. Odlišnosti obou postupů budou zpravidla tím větší, čím extrémnějších hodnot budou nabývat pozorované skóry obou testů, a čím vyšší bude rozdíl reliabilit.

Náš postup tedy předpokládá shodnou míru latentního rysu v obou testech, v literatuře (např. Furr a Bacharach, 2014; Harvill, 1991; Payne a Gwynne Jones, 1959) častěji uváděný postup předpokládá shodné pravé skóry za předpokladu, že tyto pravé skóry byly standardizované na stejné jednotky (například během standardizace testu)[16]. Pokud je tento předpoklad porušen a pro tvorbu norem byly použity pozorované skóry, což je v českém prostředí zřejmě pravidlem, je vhodnější použít náš postup, který podává nezkreslené výsledky. Reálný rozdíl však bude spíše zanedbatelný.

[14] Což je zřejmě důvodem, proč některé testy standardizují nikoliv pozorovaný, ale pravý skór (více viz poznámku pod čarou [12]).

[15] Například pokud $X = Y$ a zároveň $r_{xx'} > r_{yy'}$.

[16] Jak jsme uvedli výše, za určitých podmínek stejná úroveň měřeného rysu může být reprezentována odlišným pravým skórem (pokud byly pro tvorbu norem použity pozorované skóry namísto pravých a oba testy mají odlišnou reliabilitu).

Rozdíl v testu a subtestu (primárním a sekundárním testu)

Jiná situace nastává, pokud „odhadujeme“ dosažený skór v jednom testu na základě testu jiného za předpokladu, že oba měří to samé (nebo téměř to samé). Příkladem je otázka, zda se liší výkon respondenta v subtestu oproti výsledku celého testu – jinými slovy usuzujeme, zda se pozorované skóre subtestu neliší proti očekávanému pravému skóre celé škály[17]. V tomto případě předpokládáme kauzalitu, např. že obecná inteligence (g-faktor) „ovlivňuje“ některé specifické schopnosti, například kvantitativní usuzování. Upozorňujeme, že ne vždy je tento předpoklad na místě. Jiným příkladem může být retest po určité době, avšak jiným testem, měřícím opět tentýž rys (vyšetření pomocí WISC-III, testu B, rok po administraci WJ-II-IE, testu A)[18].

Tento příklad se nevyskytuje v nám známé literatuře, podle našeho názoru je nicméně nejvhodnější srovnávat odhad pravého skóre celkové (nezávislé, příp. pretestové) škály Y a pozorované skóre v subtestu X (závislé, resp. posttestové škále). Protože však stejně jako v předchozím příkladu shodný pravý skór v obou testech neznamená stejný výkon, je nutné odhad pravého skóre nezávislé škály převést na pravé skóre škály závislé. Rozdíl má pak hodnotu

17 $$\mathrm{E}[Y - \tau'_X] = \sqrt{r_{xx'} r_{yy'}}\,Y + \left(1 - \sqrt{r_{xx'} r_{yy'}}\right)M - X$$

a jeho standardní chyba je

18 $$\sigma_{\mathrm{e}(Y-\tau'_X)} = \sigma_{xy}\sqrt{1 - r_{xx'} r_{yy'}}\,,$$

kde $r_{xx'}$ a X je reliabilita a pozorovaný skór závislé škály a $r_{yy'}$ reliabilita a Y pozorovaný skór škály nezávislé. M a σ_{xy} je jejich společný průměr a směrodatná odchylka.

Vidíme, že pokud jsou reliability obou testů shodné, jsou rovnice 11 a 18 identické a neliší se ani vzorec 17 od rozdílu pretestového a posttestového skóre respondenta v tom stejném testu. Tento výpočet je proto zobecněním výše uvedeného postupu.

Abnormalita rozdílu ve dvou testech

Někdy nás nezajímá, zda respondent dosáhl v jednom testu vyššího výkonu než ve druhém, ale chceme vědět, „jak často se podobně velký rozdíl objeví v populaci“. Tento případ popisují jako primární a nejdůležitější už Payne a Gwynne Jones (1959), ačkoliv v pozdější literatuře se vyskytuje jen zřídka a důraz je obvykle kladen na existenci rozdílu jako takového (Furr a Bacharach, 2014; Revelle, 2015).

V tomto případě nás tedy nezajímá ani tak samotná existence rozdílu, ale spíše jeho „klinická významnost“. Zejména v případě, že testy spolu korelují velmi málo, nemusí být

[17] Tento příklad je nicméně komplikován tím, že celkový skór je součtem všech subtestů. Chyba posuzovaného subtestu tak nebude nekorelovaná s chybou celkového skóre. Velikost této korelace je sice možné odhadnout a odečíst ve vzorci 18, podle našeho názoru je však možné toto zanedbat – reálný vztah bude velmi slabý.

[18] V tomto případě již chyby měření korelované nebudou (při dodržení předpokladů CTT, zejména ekvivalence položek – výsledek respondenta tedy nesmí být ovlivněn další skrytou latentní proměnnou, např. příslušností k minoritě).

i velký (a statistický významný rozdíl) mezi skóry obou testů významný klinicky – mohlo k němu dojít náhodou, podobného rozdílu třeba dosahuje značné množství lidí[19]. Chceme upozornit, že statistická významnost tohoto rozdílu bývá zpravidla nečekaně nízká.

Řešení se nijak neliší od obecného postupu, který popisuje vzorec 9. Pokud jsou oba testy X a Y ve stejných standardních jednotkách ($\sigma_x = \sigma_y = \sigma_{xy}$), platí

$$\sigma_{x-y} = \sqrt{\sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y}\,,$$

což lze upravit jako

19 $$\sigma_{x-y} = \sigma_{xy}\sqrt{2}\sqrt{1 - r_{xy}}\,,$$

kde r_{xy} a je korelace obou testů a σ_{xy} jejich standardní odchylka.

Protože korelace popisuje přímo vztah pozorovaných skórů, není v tomto případě nutné pracovat s regresí k průměru[20].

Příklady

1. Respondent vyplnil dva různé testy měřící potřebu kognitivního uzavření. V testu X získal T-skór 65 (reliabilita r_x = 0,7), v testu Y T-skór 50 (reliabilita r_y = 0,9). Jde o statisticky významný rozdíl?
   V první řadě spočítáme standardní chybu rozdílu, tedy $\sigma_{\Delta xy} = \sigma_{xy}\sqrt{2 - r_x - r_y} = 10\sqrt{2 - 0{,}7 - 0{,}9} = 6{,}32$. 95% interval spolehlivosti je $CI = 0 \pm 1{,}96 \cdot 6{,}32 = \pm 12{,}40$. Rozdíl obou testů je větší, $65 - 50 = 15$, a závěr tedy zní, že respondent dosáhl v testu X statisticky významně vyššího pravého skóre než v testu Y.
   Protože uvedený postup však neumožňuje testovat, zda se liší skutečná míra latentního rysu, spočítáme i druhou předloženou (a tedy vhodnější) variantu výpočtu. V tomto případě má testovaný rozdíl hodnotu $E(\tau'_x - \tau'_y) = \sqrt{0{,}7}(65 - 50) - \sqrt{0{,}9}(50 - 50) = 12{,}55$. Protože chyba měření (a tedy i interval spolehlivosti) je shodná s předchozím postupem, i tentokrát je rozdíl signifikantní na p < 0,05, byť výsledek nevypadá již zdaleka tak jistě. Můžeme

[19] Příkladem může být rozdíl hmotnosti a výšky člověka. Protože metr i běžná váha měří velmi přesně, i malý rozdíl (je-li výška v centimetrech a váha v kilogramech standardizována např. na T-skóry) je statisticky významný. Je však evidentní, že lidé se v tomto ohledu liší, ostatně i pásmo normy BMI indexu obsahuje určité rozmezí. Člověk musí být extrémně vysoký a současně štíhlý, nebo naopak spíše nižší a obézní, aby rozdíl hmotnosti a výšky byl významný „klinicky“. Z psychologické oblasti můžeme jako příklad zvolit rozdíl ve skóru neuroticismu a otevřenosti v dotazníku NEO – protože oba rysy spolu prakticky nesouvisí, rozdíl např. 20 T-skórů (jakkoliv je významný statisticky) není významný klinicky.

[20] Na základě předchozího textu je však patrné, že srovnávání pozorovaných skórů s sebou i v tomto případě nese případě jisté interpretační obtíže. Z důvodu jednoduchosti článku a nepříliš častému používání „klinicky významného rozdílu“ (což je podle našeho názoru škoda) jsme se rozhodli situaci dále nekomplikovat a použít jednodušší, byť ne zcela správné řešení. Pozorný čtenář si sám může odvodit správnější řešení, tj. srovnání jednoho pozorování s regresním odhadem na základě pozorování druhého.

však uzavřít, že na 5% hladině pravděpodobnosti se liší míra schopností respondenta změřená oběma testy.

2. Respondent absolvoval inteligenční test. Celkový výsledek po převedení na T-skóry byl $X = 65$ (reliabilita r_x = 0,9), T-skóre jednoho ze subtestů bylo $Y = 50$ (r_y = 0,7) – čísla jsou stejná jako v předchozím případě.
Prvně si spočítáme očekávaný rozdíl, tedy $\mathrm{E}[Y - \tau'_X] = 65\sqrt{0{,}7 \cdot 0{,}9} + 50\left(1 - \sqrt{0{,}7 \cdot 0{,}9}\right) - 50 = 11{,}91$, a chybu měření, $\sigma_{\mathrm{e}(Y-\tau'_X)} = 10\sqrt{1 - 0{,}7 \cdot 0{,}9} = 6{,}08$. 95% interval spolehlivosti je potom $CI = 0 \pm 1{,}96 \cdot 6{,}08 = \pm 11{,}92$. V tomto případě tedy respondent neskóroval na hladině pravděpodobnosti p < 0,05 v subtestu (závislém testu) hůře, než bychom předpokládali na základě celkového skóre (nezávislého testu).

3. Víme, že oba testy z předchozího příkladu spolu korelují středně silně, $r_{xy} = 0{,}45$. Standardní chyba rozdílu je tedy $\sigma_{x-y} = 10\sqrt{2}\sqrt{1 - 0{,}45} = 10{,}49$. 95% interval spolehlivosti je $CI = 0 \pm 1{,}96 \cdot 10{,}49 = \pm 20{,}56$. Výsledek respondenta v obou testech se tedy neliší „abnormálním způsobem“.

Shrnutí

Výše uvedený text předkládá řešení většiny otázek, které si běžně klade praktický psycholog při psychologické diagnostice: „Přesáhl respondent určitý kritický skór? Liší se skóre z prvního a druhého měření? Který z respondentů získal vyšší hodnocení? Je tento intraindividuální rozdíl klinický významný?“ atd.

Je nicméně evidentní, že jde pouze o ilustrační příklady; například všude tam, kde se setkáváme s efektem regrese k průměru, záleží na výběrové populaci[21]. Zároveň konkrétní provedení statistického testu je závislé na „diagnostické otázce“, kterou si v tu kterou chvíli psycholog klade, a je vhodné jej občas do určité míry upravovat.

Určitá podobnost jednotlivých postupů by mohla čtenáře vést k názoru, že to je vše vlastně jedno, a že stačí použít běžné intervaly spolehlivosti. V tom případě bychom rádi apelovali alespoň na dodržení tří věcí: zaprvé nezanedbávat regresi k průměru při opakovaném testování, zvláště u méně přesných testů – je vhodné si zjistit, jakým způsobem byly u metody konstruovány intervaly spolehlivosti, a případně alespoň předpokládat „průměrnější“ retestové skóre, než jaké bylo naměřeno poprvé. Za druhé použít při srovnávání dvou měření interval spolehlivosti dvakrát, pokud nepracujeme s chybou rozdílu (intervaly spolehlivosti obou měření se nesmí překrývat). A za třetí, nepřikládat „klinickou významnost“ rozdílu testů, které se sice liší (např. při použití

[21] Protože testovaná (klinická) populace se často liší od normální (standardizační) populace, dosahuje i jiných průměrů a testové výsledky proto regredují k jiné hodnotě „lokálního průměrného skóre“. Protože však příslušnost ke klinické skupině předem neznáme, může vést apriorní přesvědčení o „klinickosti“ klienta ke zkresleným závěrům, a proto se zdá být nejvhodnějším postupem regrese k průměru, určeného standardizačním výběrovým souborem.

nepřekrývajících se intervalů spolehlivosti), ale jejich vzájemná korelace je zanedbatelná (oba měří něco jiného).

I přes zdánlivou obtížnost jsou všechny výše uvedené postupy velmi jednoduché – v dnešní době není problém mít veškeré vzorce nachystané např. v Excelu a při diagnostice je jen použít. Druhou možností je pak využít funkcí některé on-line diagnostické služby, např. v současnosti vyvíjené Diagnostické kalkulačky (http://kalkulacka.testforum.cz), nebo zahraniční služby PsychoCalc (https://begavett.shinyapps.io/PsychoCalc/), ani jedna z nich však v době psaní tohoto textu nefunguje ve všech ohledech zcela dokonale.

Závěrem chceme říci, že použití adekvátních statistických postupů pro vyčíslení chyby měření nejenže zkvalitní provedené diagnostické závěry, ale je také nezbytné pro zajištění etického a férového vyšetření. Nejnovější verze Standardů pro pedagogické a psychologické testování (AERA, 2014) doslova uvádí: *„Pro každý celkový skór, subskór či kombinaci skórů, které jsou interpretovány, musí být reportován také relevantní odhad reliability či přesnosti měření,“* (standard 2.3), a dále pak *„Pokud interpretace testu zdůrazňuje rozdíl mezi dvěma pozorovanými skóry jedince či dva průměry skupin, reliabilita či chyba měření včetně standardních chyb tohoto rozdílu musí být poskytnuta,“* (standard 2.4).[22] Tyto požadavky jsou primárně kladeny na distributora testu. Pokud však u nás tyto informace v běžně používaných testech chybí, leží zodpovědnost za dodržení příslušných standardů i na psychologovi – uživateli testu.

„Intuitivní“ odhady velikosti chyb obecně přeceňují skutečný význam pozorovaných rozdílů. Obvykle se nám rozdíly zdají významnější, než jsou ve skutečnosti, čemuž přispívají i automatizované diagnostické zprávy různých testů záměrně zdůrazňujících intrapsychický profil různými grafy a křivkami (bez uvedení intervalů spolehlivosti). Bez adekvátního zhodnocení chyby měření tento jev může vést ke zbytečnému „patologizování“ klientů (ať už směrem k podprůměru, nadprůměru či „nevyrovnanosti“ skórů), zbytečnému škatulkování a ve výsledku tak k jejich přímému poškození.

[22] Standard 2.3: „For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.“ Standard 2.4: „When a test score interpretation emphasizes differences between two observed scores of an individual or two averages of group, reliability/precision data, including standard errors, should be provided for such differences.“

Cígler, H., & Šmíra, M.: Error of measurement and the estimation of true score: Selected methods of Classical test theory

One of the elementary skills involved in the interpretation of the psychological results is handling the error of measurement. Unfortunately, many Czech psychological tests do not include all the necessary information about the error of measurement (e.g. confidence intervals and standard errors of measurement for different purposes). Even if such information is available, we might need to consider other circumstances of the assessment, and adjust the method of estimation and its application properly – it is not always possible to rely on the test developer in such cases. Since there are not many applications for such computations easily available for the test users, they should be capable of doing many of the elementary computations by hand. This paper briefly summarizes common techniques for the interpretation of the error of measurement using confidence intervals in the framework of Classical Test Theory. The theory is supported by detailed examples that should be helpful for applying these procedures in practice.

Keywords: Classical Test Theory, CTT, standard error of measurement, SEM, interpretation of the test results

Podpořeno z projektů OPVK:
SOVA-21 – Internacionalizace, inovace, praxe: sociálně-vědní vzdělávání pro 21. století, CZ.1.07/2.2.00/28.0225
INZA – Inovací bakalářských studijních programů k lepší zaměstnatelnosti, CZ.1.07/2.2.00/28.0238

Přílohy

Příloha 1: Odvození rovnic 15 a 16

Testujeme hypotézu o shodě pravého skóre testu X (τ_x) s pravým skóre testu Y (τ_y), převedeného na škálu testu X (τ'_y):

$$E(\tau_x - \tau'_y) = E(\tau_x) - E(\tau'_y) = 0$$

přičemž

$$E(\tau'_y) = M + \frac{\sigma_{x'}}{\sigma_{y'}}\left[E(\tau_y) - M\right]$$

kde $\sigma_{x'}$ a $\sigma_{y'}$ jsou směrodatné odchylky pravých skórů testů X a Y. Protože platí

$$\frac{\sigma_{x'}}{\sigma_{y'}} = \sqrt{\frac{r_{xx'}}{r_{yy'}}}$$

a zároveň můžeme odhadnout $E(\tau_y)$ podle vzorce 7, po dosazení

$$E(\tau_x - \tau'_y) = r_{xx'}X + (1 - r_{xx'})M - \left\{M + \sqrt{\frac{r_{xx'}}{r_{yy'}}}\left[r_{yy'}Y + (1 - r_{yy'})M - M\right]\right\}$$

a tedy

$$E(\tau_x - \tau'_y) = r_{xx'}(X - M) - \sqrt{r_{xx'}r_{yy'}}(Y - M)$$

Standardní chyba rozdílu má potom hodnotu

$$\sigma_{\Delta\tau_x\tau'_y} = \sqrt{\left(\sigma_{xy}\sqrt{r_{xx'}}\sqrt{1 - r_{xx'}}\right)^2 + \left(\sqrt{\frac{r_{xx'}}{r_{yy'}}}\sigma_{xy}\sqrt{r_{yy'}}\sqrt{1 - r_{yy'}}\right)^2}$$
$$= \sigma_{xy}\sqrt{r_{xx'}}\sqrt{2 - r_{xx'} - r_{yy'}}$$

Protože testová statistika má podobu podílu předchozích dvou rovnic, tedy

$$\frac{E(\tau_x - \tau'_y)}{\sigma_{\Delta\tau_x\tau'_y}}$$

lze oba předchozí vzorce podělit $\sqrt{r_{xx'}}$. Výsledná úprava je pak nezávislá na tom, na škálu kterého pravého skóru byly oba testy převedeny, výsledky jsou zcela shodné.

Výsledné úpravy jsou prezentovány v rovnicích 15 a 16, první z nich uvádí odhad rozdílu pravých skórů převedených na stejnou škálu, druhá pak chybu tohoto odhadu.

Zdroje

AERA, APA & NCME (2014). *Standards for Educational and Psychological Testing.* Washington: American Educational Research Association.

Charter, R. A. (2009). Differences Scores: Regression-Based Reliable Difference and the Regression-Based Confidence Interval. *Journal of clinical psychology, 65*(4), 456-460. doi: 10.1002/jclp.20554

Charter, R. A., & Feldt, L. S. (2000). The Relationship between Two Methods of Evaluating an Examinee's Difference Scores. *Journal of Psychoeducational Assessment, 18*(2), 125-142. doi: 10.1177/073428290001800203.

Dudek, F. J. (1979). The Continuing Misinterpretation of the Standard Error of Measurement. *Psychological Bulletin 86*(2), 335-337. doi: 10.1037/0033-2909.86.2.335

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics : an introduction*. Los Angeles: Sage.

Harvill, L. M. (1991). Standard Error of Measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41. doi: 10.1111/j.1745-3992.1991.tb00195.x

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* London: Addison-Wesley Publishing.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Payne, R. W. and Gwynne Jones, H. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology 13*(2), 115–121. doi: 10.1002/1097-4679(195704)13:2<115::AID-JCLP2270130203>3.0.CO;2-1

Revelle, W. (2015). Chapter 7: Classical Test Theory and the Measurement of Reliability. In W. Revelle, *An introduction to psychometric theory with applications in R*, pp. 205–239. http://www.personality-project.org/r/book/Chapter7.pdf.