Fórum:Testy/Reliabilita

Jsou výsledky testu konzistentní? Reliabilita neboli spolehlivost měří, nakolik jsou výsledky zopakovatelné, nakolik *výstřely padají blízko sebe*. Je tak jednou z fundamentálních vlastností testu. Neexistuje však pouze jediný správný způsob jejího odhadování [1]. Použijeme-li kterýkoliv z přístupů odhadu reliability, měli bychom rozumět principům a teoriím, na kterých je založen, jakož i omezením, která s sebou předpoklady daného přístupu přinášejí [2]. Jaké jsou tedy možnosti odhadu reliability?

Odhad metodou test-retest

Přímo z definice reliability se nabízí myšlenka zadat test studentům dvakrát po sobě a měřit závislost mezi dvěma výsledky pomocí korelačního koeficientu. Tento odhad reliability je často používán např. pro odhad měřicích přístrojů (tlaku, váhy apod.). V případě didaktického testování je ale jeho využití velmi omezené. Studenti si totiž otázky pamatují a mají tendenci odpovídat konzistentně, což by odhad reliability nadhodnotilo. Na rozdíl od opakovaného střílení na terč nebo měření tlaku proto nelze dvě administrace stejného znalostního testu považovat za nezávislé. V případě delšího časového intervalu se zase studenti správná řešení mohou naučit (nebo je zapomenout). Korelace mezi dvěma časově vzdálenějšími výsledky pak spíš vypovídají o **stabilitě znalosti v čase**.

Odhad pomocí paralelních forem testu

Abychom vyloučili vliv paměti při použití metody test-retest, lze vytvořit dvě tzv. paralelní formy testu. Stručně řečeno to jsou dva testy, které měří danou znalost stejným způsobem, mají stejné průměry celkových skóre, stejné směrodatné odchylky a stejné korelace s jinými testy, a to v jakékoliv populaci. Paralelní formy lze vytvořit např. změnou číselných údajů v zadání jednotlivých příkladů apod. Odhad reliability je pak opět založen na korelaci mezi dvěma celkovými skóre. Tento způsob odhadu reliability lze v didaktickém testování nadmíru doporučit. Jak lze tušit, omezením je obtížnost vytvoření dvou verzí testu, které by byly skutečně paralelními formami. Odhad pak může do jisté míry odrážet, nakolik jsou dvě verze testu skutečně ekvivalentní. Dalším omezením může být neochota respondentů k dvojímu testování. Výsledky v druhém testu také mohou být ovlivněny zvýšenou únavou studentů.

Odhad metodou split-half

V mnoha případech není možné testovat studenty vícekrát, a je tedy zapotřebí odhadovat reliabilitu z jediné administrace testu. Jedním z přístupů je rozdělení testu na dvě poloviny a zkoumání korelace mezi celkovými skóre v těchto dvou polovinách. Pokud test rovnou konstruujeme jako dvojice paralelních položek, můžeme tento přístup vnímat jako odhad reliability paralelních forem, které však mají poloviční délku. Reliabilita je závislá na počtu položek (testy s více položkami mívají reliabilitu vyšší), je proto nutná úprava odhadu na skutečnou délku testu pomocí tzv. *Spearmanovy-Brownovy formule* [3], [4]. Výslednému odhadu se říká **split-half reliabilita**.

Pokud test nekonstruujeme přímo jako dvojice odpovídajících si položek, může být oříškem, které z možných rozpůlení zvolit. Nabízí se rozdělení na první a druhou polovinu, na sudé a liché položky, nebo lze položky rozdělit náhodně. Každé rozpůlení dá ale jiný odhad spolehlivosti. Jedním z řešení je rozdělit test postupně na všechny možné poloviny a získané odhady zprůměrovat, odhadu se pak říká **průměrný split-half koeficient**. Číselně blízké a přitom výpočetně o dost jednodušší je rozdělit test na nejmenší možné části (tedy na jednotlivé položky) a odhadovat reliabilitu pomocí korelací mezi nimi. To je základem myšlenky Cronbachova alfa.

Reliabilita jako vnitřní konzistence položek – Cronbachovo alfa

Nejčastěji používaným odhadem reliability je **Cronbachovo alfa**, jehož vzorec lze najít v příloze. Je oblíbený z několika důvodů: jeho výpočet je jednoznačný, srozumitelný a je implementován do všech statistických programů. Cronbachovo alfa měří závislosti mezi jednotlivými položkami, je tudíž mírou vnitřní konzistence testu. Za jeho používáním coby odhadu spolehlivosti stojí představa, že všechny položky testu měří jedinou vlastnost, síla jejich závislosti je vysoká a jediné rozdíly jsou tudíž způsobené chybou měření.

Cronbachovo alfa je pojmenováno po Lee Cronbachovi, který jej proslavil v roce 1951 [5]. Hodnotu rovnou jedné dostaneme, pokud jsou položky svázány lineárně (v takovém případě by ovšem stačila položka jediná). Malá hodnota naopak vypovídá o nízké vnitřní konzistenci položek, nebo nízké spolehlivosti [6].

- **Bacha na Cronbacha!**

Psychometři upozorňují na mnohá omezení a dezinterpretace Cronbachova alfa [7]. Jedním z omezení je již naznačená skutečnost, že je pouze dolní mezí reliability a silně ji podhodnocuje v případě, kdy položky nepopisují stejnou oblast vědomostí. Alternativní koeficienty vhodné pro případ složitější struktury znalostí vyžadovaných testem jsou popsány v článku [8].

Cronbachovo alfa také nefunguje správně, pokud jsou vedle náhodných chyb v měření obsaženy další chyby, např. efekt posluchárny, ve které byl test skládán (skládají-li studenti test v různých posluchárnách, lze si představit, že v některých posluchárnách studenty ruší hluk z ulice), efekt zkoušejícího (zkouší-li více examinátorů) apod. S

takovými situacemi se vypořádává tzv. teorie zobecnitelnosti [9], [10], která používá složitějších modelů analýzy rozptylu.

Cronbachovo alfa vychází z modelu normálně rozdělených položkových skóre. Je tedy na místě otázka, nakolik správné je jeho použití v případě položek typu ano/ne. Některá zobecnění byla navržena v článcích [11] a [12].

Nakonec upozornění, které platí pro všechny typy odhadů reliability, avšak pro Cronachovo alfa obzvlášť [13]: Již v definici reliability je obsažena důležitá vlastnost, a sice že je závislá na homogenitě testovaných jedinců. Budeme-li testovat skupinu studentů s podobnými znalostmi (např. jednu studijní skupinu), odhad reliability bude nižší než v případě, kdy budeme testovat skupinu méně homogenní – např. studenty různých ročníků. Odhadujeme-li reliabilitu testu, je proto nutné předem stanovit, pro jakou skupinu je test určen (jeden obor, celý ročník, celou školu) a odhady je pak nutné počítat z odpovídajících dat.

Demonstrujme význam Cronbachova alfa na následujícím příkladu:

Představme si, že chceme zkoušet sčítání čísel od jedné do deseti. Snadno sestavíme test, ve kterém bude větší množství (řekněme padesát) doplňovacích otázek typu „3 + 4 = ...“. Ten, kdo sčítat umí, odpoví správně na všechny otázky, nebo nanejvýš udělá jen ojedinělé nahodilé chyby. Naopak ten, kdo sčítat vůbec neumí, se jen ojediněle strefí do správného řešení. Takto sestavený test můžeme označit za vnitřně konzistentní – testuje jediný koncept (sčítání v daném oboru čísel). Cronbachovo alfa se bude blížit jedné.

Pokud bychom nyní v testu vyměnili polovinu úloh za příklady typu „12 : 3 = ...“, situace se změní. Dáme-li takto změněný test žákům prvních či druhých tříd základní školy, budeme testovat dva koncepty: sčítání a dělení. Lze si představit, že část žáků bude umět dobře sčítat, ale zcela pohoří v dělení. Test již nebude tak konzistentní, jako v předešlém případě; nemůžeme už také říci, že kterékoliv dvě otázky z testu testují totéž. Cronbachovo alfa se sníží.

Mluvíme-li o vnitřní konzistenci testu, měli bychom si uvědomit, že nezávisí jen na samotných otázkách, ale také na cílové skupině. Pokud bychom totiž dali onen upravený test s jednoduchými početními úlohami gymnaziálním studentům, pravděpodobně by se jevil opět jako vnitřně konzistentní a Cronbachovo alfa by se blížilo jedné: z pohledu takovéto pokročilejší skupiny testovaných totiž zkoušíme opět jediný koncept – základní početní úkony. Zda je konkrétní úloha věnovaná sčítání nebo dělení, bude v tomto případě lhostejné.

Z uvedených příkladů vyplývá, proč by Cronbachovo alfa konkrétního testu nemělo být ani příliš nízké, ani příliš vysoké. Je-li test nekonzistentní, budou se nám špatně interpretovat jeho bodové výsledky. Představme si, že náš test s úlohami na sčítání a dělení dáme žákům druhých tříd. Podle dosaženého počtu bodů asi poměrně snadno rozpoznáme skupinu těch, kteří umí dobře sčítat i dělit, a skupinu žáků, kteří sčítat ani dělit neumí vůbec. Mezi nimi budou žáci, kteří sčítají i dělí, ovšem s mnoha chybami, ale také ti, kteří výborně sčítají, neumí však vůbec dělit. Z výsledku takového testu nepoznáme, zda konkrétní žák obstál v obou činnostech srovnatelně, nebo byl v jedné výborný a v druhé propadá; pravděpodobně by bylo vhodné namísto jednoho testu použít dva samostatné, každý zaměřený na jinou dovednost.

Pokud se naopak Cronbachovo alfa blíží jedné, znamená to, že mnoho studentů z dané skupiny odpovědělo buď na všechny otázky správně, nebo na všechny otázky špatně. Jinými slovy, odpověděl-li student správně na několik prvních otázek, odpovídal správně i na všechny ostatní a obráceně. V uvedeném testu sestaveném pouze z příkladů na sčítání by asi bylo zbytečné dávat žákům padesát otázek – pokud bychom test zkrátili, dostali bychom pravděpodobně zcela srovnatelné výsledky. Test s velmi vysokým Cronbachovým alfa navíc nemusí dostatečně jemně rozlišovat mezi různými úrovněmi znalostí.

Shoda v rozhodnutí o úspěchu/neúspěchu

Dosud jsme se reliabilitou zabývali v kontextu testů relativního výkonu, které mají za úkol rozlišit mezi jednotlivými studenty. Uvědomme si nyní, že odhad reliability založený na korelaci mezi výsledky dvou testů (např. dvou paralelních forem) nebere v potaz obtížnost těchto testů, a může tak nadhodnotit spolehlivost z perspektivy testu absolutního výkonu. Představme si, že testujeme dvěma testy pět studentů a obdržíme tyto výsledky:

Tab. 8.2 Tabulka ukázkových výsledků pěti studentů ve dvou testech

	počet bodů v prvním testu	počet bodů v druhém testu
1. student	18	48

2. student	45	75
3. student	33	63
4. student	48	78
5. student	51	81

Pearsonův korelační koeficient je roven 1, mezi dvěma výsledky je přímo lineární závislost (druhý je vždy právě o 30 bodů vyšší než první). Reliabilita (coby korelace paralelních forem) je tedy vysoká. Budeme-li však rozhodovat o složení zkoušky na základě dosažení 50 bodů, první test úspěšně absolvuje 1 student z 5, zatímco v druhém to budou 4 studenti z 5. Pro popsání shody v rozhodnutí o úspěchu/neúspěchu proto nelze použít dosud zmíněné nástroje. Místo toho se odhadují tzv. **indexy shody**. Jedním z nich je například **Cohenovo kappa**. To se počítá z procentuální shody mezi testy: v našem případě by testy rozhodly o (ne)úspěchu shodně jen u prvního a posledního studenta, tedy ve 40 % případů. Cohenovo kappa poměřuje procento shody s pravděpodobností náhodné shody, vychází tak v našem případě ještě o něco menší, a to 0,12 (viz také příloha). Hodnoty kappa vyšší než 0,7 ukazují na velmi dobrou shodu, hodnoty mezi 0,4 a 0,7 na dostatečnou shodu, menší hodnoty, stejně jako v našem případě, na nedostatečnou shodu. Jak plyne z uvedeného příkladu, dva testy mohou mít pro posouzení relativního výkonu vysokou míru ekvivalence, a přesto mohou mít jeho dvě verze nízkou míru shody co do hodnocení absolutního výkonu. Pro odhadování spolehlivosti je proto nutné rozlišit, za jakým účelem je test zadáván a použít pak správné míry spolehlivosti.

Shoda posuzovatelů, zkoušejících nebo komisí

Pokud hodnotí výkon studentů různí posuzovatelé (např. při ústním zkoušení, při hodnocení esejí apod.), je nutné zajistit jejich srovnatelnost. Jak tedy odhadnout, nakolik dva zkoušející hodnotí stejně? Prvním krokem může být spočítat průměrná hodnocení, která jednotliví zkoušející udělili. Pokud např. jeden zkoušející dává v průměru známku 2,4 a druhý 3,6, každý student bude vědět, kterého zkoušejícího si vybrat (bude-li mít tu možnost). Průměrná hodnocení ale mohou být ovlivněna tím, jaké studenty ten který zkoušející hodnotil. Pokud například první zkoušející hodnotil pouze v předtermínu (dá se tedy očekávat, že šlo spíše o snaživější a lépe připravené studenty) a druhý naopak hodnotil jen studenty, kteří měli problémy s obdržením zápočtu, mohlo být hodnocení prvního zkoušejícího ve skutečnosti přísnější než hodnocení druhého. Proto jediným možným způsobem posouzení shody posuzovatelů je zajistit (alespoň na části zkoušených) nezávislé hodnocení od obou zkoušejících. Shodu mezi dvěma výsledky pak už můžeme odhadovat pomocí výše popsaných metod, např. pomocí koeficientu kappa (jde-li o hodnocení úspěchu/neúspěchu), nebo dalších indexů shody.

Odkazy

1. AERA,, APA a NCME. *Standardy pro pedagogické a psychologické testování.* 1. vydání. Praha : Testcentrum, 2001. 320 s. ISBN ISBN 80-86471-07-1..
2. SCHUWIRTH, Lambert WT a Cees PM VAN DER VLEUTEN. General overview of the theories used in assessment: AMEE Guide No.57. *Med Teach* [online]. 2011, roč. -, vol. 3, no. 10, s. 783-797, dostupné také z <https://informahealthcare.com/doi/abs/10.3109/0142159X.2011.611022>. sv. ISBN 978-1-903934-97-5. ISSN 0142-159X. PMID: 21942477 (https://www.ncbi.nlm.nih.gov/pubmed/21942477).DOI: 10.3109/0142159X.2011.611022 (http://dx.doi.org/10.3109%2F0142159X.2011.611022).
3. SPEARMAN, Charles. Correlation calculated from faulty data. *British Journal of Psychology.* 1910, roč. 7, vol. 3, no. 3, s. 271–295, ISSN 2044-8295. DOI: 10.1111/j.2044-8295.1910.tb00206.x (http://dx.doi.org/10.1111%2Fj.2044-8295.1910.tb00206.x).
4. BROWN, William. Some experimental results in the correlation of mental abilities. *British Journal of Psychology.* 1910, roč. 7, vol. 3, no. 3, s. 296–322, ISSN 2044-8295. DOI: 10.1111/j.2044-8295.1910.tb00207.x (http://dx.doi.org/10.1111%2Fj.2044-8295.1910.tb00207.x).
5. CRONBACH, Lee J. Coefficient alpha and the internal structure of tests. *Psychometrika* [online]. 1951, roč. -, vol. 16, no. 3, s. 297–334, dostupné také z <http://psych.colorado.edu/~carey/Courses/PSYC5112/Readings/alpha_Cronbach.pdf>. ISSN (print) 0033-3123 (online) 1860-0980.
6. TAVAKOL, Mohsen a Reg DENNICK. Making sense of Cronbach's alpha. *International Journal of Medical Education* [online]. 2011, roč. -, vol. -, no. 2, s. 53-55, dostupné také z <http://www.ijme.net/archive/2/cronbachs-alpha/>. ISSN 2042-6372. DOI: 10.5116/ijme.4dfb.8dfd (http://dx.doi.org/10.5116%2Fijme.4dfb.8dfd).
7. SIJTSMA, Klaas. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* [online]. 2009, roč. -, vol. 74, no. 1, s. 107-120, dostupné také z <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792363/>. ISSN (print) 0033-3123 (online) 1860-0980. PMID: 20037639 (https://www.ncbi.nlm.nih.gov/pubmed/20037639).DOI: 10.1007/s11336-008-9101-0 (http://dx.doi.org/10.1007%2Fs11336-008-9101-0).
8. ZINBARG, Richard E, William REVELLE a Iftah YOVEL, et al. Cronbach's alpha, Revelle's beta, and McDonald's omega: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika* [online]. 2005, roč. -, vol. 70, no. 1, s. 123-133, dostupné také z <http://personality-project.org/revelle/publications/zinbarg.revelle.pmet.05.pdf>. ISSN (print) 0033-3123 (online) 1860-0980. DOI: 10.1007/s11336-003-0974-7 (http://dx.doi.org/10.1007%2Fs11336-003-0974-7).

9. CRONBACH, Lee J, Nageswari RAJARATNAM a Goldine C GLESER. Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology.* 1963, roč. 17, vol. 16, no. 2, s. 137-163, ISSN 2044-8317. DOI: 10.1111/j.2044-8317.1963.tb00206.x (http://dx.doi.org/10.1111%2Fj.2044-8317.1963.tb00206.x).
10. SHAVELSON, Richard J a Noreen M WEBB. *Generalizability Theory: A Primer : [object Object].* 1. vydání. Newbury Park, Calif : Sage Publications, 1991. 152 s. ISBN 978-0803937451.
11. MARTINKOVÁ, Patrícia a Karel ZVÁRA. Reliability in the Rasch model. *Kybernetika* [online]. 2007, roč. -, vol. 43, no. 3, s. 315-326, dostupné také z <http://dml.cz/bitstream/handle/10338.dmlcz/135776/Kybernetika_43-2007-3_4.pdf>. ISSN 0023-5954.
12. MARTINKOVÁ, Patrícia a Karel ZVÁRA. Reliability of Composite Dichotomous Measurements. *European Journal for Biomedical Informatics* [online]. 2010, roč. -, vol. 6, no. 2, s. 14-23, dostupné také z <http://www.ejbi.org/en/ejbi/article/77-en-reliability-of-composite-dichotomous-measurements.html>. ISSN 1801-5603.
13. ZVÁRA, Karel. Měření reliability aneb Bacha na Cronbacha. *Informační bulletin České statistické společnost* [online]. 2002, roč. -, vol. 13, no. 2, s. 13-20, dostupné také z <http://www.statspol.cz/bulletiny/ib-02-2.pdf>. ISSN 1210-8022.