Reakce na „Kritika textu Říčan J. et al. – Komparace kvality tzv. teacher made testů s didaktickými testy a jejich vliv na úspěšnost žáků: případová studie. *Scientia in educatione*, *12*(2), 2021“

Jaroslav Říčan[1,*], Jiří Škoda[1], Viktorie Hermanová[1], Barbora Lanková[1]

[1] Pedagogická fakulta, Univerzita J. E. Purkyně, Hoření 13, 400 96 Ústí nad Labem; jaroslav.rican@ujep.cz

Článek má zavádějící jméno . . . neuvádí ani předchozí či výslednou roční klasifikaci sledovaných žáků, data pouze popisují, jakou známku dostali žáci v jednotlivých testech. . .

Zde došlo k nepochopení záměru autorů ze strany kritika. „Vlivem“ jsme neměli na mysli záležitosti kauzálního charakteru typické pro experimentální šetření (tj. kauzální výzkumný problém, následkové hypotézy). Školní úspěšností byla míněna úspěšnost v příslušných testech (testy vytvořené učitelkou a testy vytvořené autory). Jedná se o ústřední podstatu příspěvku – chtěli jsme poukázat na skutečnost v rozdílech úspěšnosti žáků (skórování) v závislosti na tom, zda se jednalo o testy vytvořené učitelkou nebo testy vytvořené autory. Jestliže bychom uváděli známky žáků (roční klasifikace), tak to nijak danou hodnotu nezvyšuje. Nevíme, jak s nimi daná pedagožka pracovala. Co by nám tedy informace o klasifikaci poskytly v návaznosti na porovnání dvou typů testů? Nešlo nám o výsledky žáků v daném předmětu, ale o výsledky žáků v návaznosti na daný typ testu. Poukazujeme tak na odlišnost v úspěšnosti žáků, která se odráží v konstruování daného testu, přičemž se jedná o stejnou tematiku/tematický celek. „Školní úspěšnost“ (academic success) je „definována jako úroveň výsledek žáků a studentů v závislosti na jejich učebních (učících) zkušenostech v jakékoliv disciplíně“ (Kanadli, 2016, s. 2062). V žádném případě nebylo zapotřebí uvádět výslednou roční klasifikaci – uchopované proměnné podléhají záměru–cíli šetření, nikoliv obráceně.

Kanadli, S. (2016). A meta-analysis on the effect of instructional designs based on the learning styles models on academic achievement, attitude and retention. *Educational Sciences: Theory & Practice*, *16*(6), 2057–2086.

. . . Proč je ve 4. třídě ZŠ tak automatický požadavek na větší zastoupení položek s vyšší kognitivní náročností napříč testem (odkaz na skripta není odborná odpověď) – trochu laicky se domnívám, že prostě v raném školním věku pracujeme pouze s nejnižšími cíli a vyšší mety přicházejí postupně – proto situaci, kdy test u malých dětí cílí převážně na spodní úrovně, *a priori* neodmítám a v textu marně hledám zdůvodnění, je-li tomu jinak.

Domníváme se, že analýzou a diskuzí každého výroku by neúměrně narostl obsah studie. V textu rovněž nikde explicitně neodmítáme testy/úkoly na nižší myšlenkové operace. Nicméně nemůžeme souhlasit se situací, kdy jsou testové položky za celé jedno pololetí dominantně tvořeny položkami vyžadujícími nižší myšlenkové operace. V kontextu např. porozumění čtenému, což vnímáme jako zcela esenciální kompetenci pro úspěšné „zvládání“ dalších předmětů (Piercy & Piercy, 2011), se hovoří o „fourth grade slump“ (Chall, 1996) jako přístupu, kdy přibližně v daném ročníku by mělo docházet ke specifickým intervencím (úkolovým situacím zaměřeným na „higher order thinking“). Nicméně takový přístup je podporován i v začátcích školní docházky, jelikož může být benefitní (Ford-Connors et al., 2015). Přístup, kdy do rozvoje formálních operací není možné pracovat se žáky na úkolech vyššího řádu (a tedy ani testovat), vychází z Piagetovské tradice (v mnohém ohledu v tomto kontextu překonané), což dokládají empirické studie i u žáků před vstupem do základního vzdělávání (Urban & Urban, 2019).

Chall, J. S. (1996). American reading achievement: Should we worry? *Research in the Teaching of English*, *30*(3), 303–310.

Ford-Connors, E., Robertson, D. A., Leighton, C. M., Paratore, J. R., Proctor, C. P., & Carney, M. (2015). Comprehension instruction within the context of the common core standards. In S. Parris, & K. Headley (Eds), *Comprehension instruction: Research-based practices* (pp. 105–122). Guilford Press.

Piercy, T. D., & Piercy, W. (2011). *Disciplinary literacy: Redefining deep understanding and leadership for 21st-century demands*. Lead & Learn Press.

Urban, K., & Urban, M. (2018). Influence of fluid intelligence on accuracy of metacognive monitoring in preschool children fades with calibration feedback. *Studia Psychologica*, *60*(2), 123–136.

Článek je slohově slabý. Text je psán z mého pohledu pseudo-odborným jazykem . . . přebytek cizích synonym vnímám jako zástěrku myšlenkové mělčiny. 18:5 *„explorace diskrepancí“* je „zjišťování nesouladu“. . . Text je rovněž negativně zatížen užíváním klišé, jež jsou svou vágností. . .

Z našeho pohledu subjektivní pohled autora kritiky. Nebylo záměrem autorského kolektivu „maskovat“ slovy autora „myšlenkovou mělčinu“. Rozumíme tomu, že autoři textu mají formulovat své myšlenky tak, aby byly čtivé, pochopitelné, jasné. Každý autor má svůj „jazyk“, a je tedy možné, že se mu nepodaří vždy své myšlenky adekvátně a adresně formulovat.

Práce s literaturou je špatná. . . Příkladem může být 27:5 „Šatánek a Hubalovská. . . “

Šatálek a Hubalovská (1972) hovoří v příspěvku o objektivitě v rámci hodnocení. Poukazují i na využívání analýzy chyb v rámci hodnocení jednotlivých položek. Každá položka je bodována jiným počtem bodů, dle obtížnosti. Nevnímáme danou větu za „vytrženou z kontextu“, jak uvádí kritik, tedy jakousi zkratku, vnímáme tuto myšlenku za podstatu příspěvku! Proč by měla být tedy považována za využitou dle potřeb autorů?! Zkratka by to za nás byla v případě, že ji vytrhneme z diskuse a opomeneme/neuvedeme uváděné protiargumenty v původním zdroji, ale v tomto případě o dané skutečnosti pojednává celý příspěvek.

. . . v textu obhajována článkem Flybvjerg 2006 věnovanému kvalitativnímu (!!) výzkumu. . .

Proč by daný zdroj nemohl být využit? Ano, jedná se o kvalitativní výzkum, ale není dáno, že případová studie nemůže být založena na kvantitativním šetření. Autoři nikterak neomlouvají velikost vzorku, od začátku pracují s tím, že menší výzkumný vzorek i tak poukazuje na odlišnosti v rámci edukační praxe a že i z výzkumného pohledu tady tato možnost existuje.

Sedláček (2014) konstatuje, že „případová studie je skutečnou výzkumnou strategií, a nikoli jednotlivou technikou, neboť badatel kromě **více informačních zdrojů** využívá veškeré dostupné metody sběru dat. . . Apriorně vyloučeny však nebývají ani metody uplatňované tradičně v kvantitativních šetřeních. Vhodnost použití je posuzována vždy s ohledem na výzkumnou otázku a charakteristiku případu“ (s. 99). Podobně Walterová a Starý (2006) upozorňují na „možnost výhradně kvantitativních případových studií, využívajících baterií testů a souboru **deskriptivních proměnných. . .** upozorňují však na smíšený typ případových studií, frekventovaně využívaných“ (s. 85).

Walterová, E., & Starý, K. (2006). Potenciál změny v realitě školy: Strategie případové studie. *Orbis scholae*, *1*(1), 77–97.

Sedláček, M. (2014). Případová studie. In R. Švaříček, K. Šeďová et al. (Eds.), *Kvalitativní výzkum v pedagogických vědách* (s. 96–112). Portál.

Dále k předpokladu ohledně existence nesouladu mezi testy: V kontextu uvedené výhrady autora ohledně vyhýbání se argumentaci (což platí i pro některé další texty autora).

V jaký moment je argumentace dostatečná a kdy nikoliv? Má empirická stať obsahovat rozbor „každého“ zdroje, o který se opírá? Zcela souhlasíme s filosofií nutné argumentace, avšak v určitém bodu by samotný podrobný (detailní) rozbor odváděl pozornost od hlavní linie textu.

V textu 22:3 *„. . . jsme využili tři expertní posudky. . . “*, v tab. 1 tamtéž je pět expertů.

Ano, to je chyba, mělo být uvedeno „pět expertních posudků“.

„Pseudovalidace“ didaktických testů je nejasná. Byla použita EFA pro určení dimenzionality – se kterou rotací, se kterým korelačním koeficientem, na binarizovaných výsledcích? Když byla zjištěna multidimenzionalita testů, jak potom byla aplikována split-half metoda?

Ano, souhlasíme, není v textu rozebrán detailní popis.

Na straně 28:2 vyúsťuje text v *„nejpodstatnější zjištění této studie“* . . . zároveň pracují s technistním návrhem 5 % = snížení stupně. Rozumím správně dotčenému odstavci, že článek přichází s představou, že žák čtvrté třídy ZŠ, který neodpoví na dvě otázky z deseti, je hodnocen nedostatečně?

Článek nepřichází s výše uvedenou „představou“ – článek pouze poukazuje na fakt, jak by vypadalo hodnocení žáka v situaci, kdy by se přijal navrhovaný způsob hodnocení Sedláčkovou (1993).

Didaktické testy nejsou testy tvořené didaktiky.

Nikde v textu neuvádíme, že „pokud test tvoří pracovník učitelské katedry. . . “ jedná se o didaktický test. Expert v kontextu posouzení konstruktové validity byl pregnantně definován, což je typické i v dalších

studiích (např. Neuenhaus, 2011; vymezení pojmu „expert“ v příslušném kontextu, účelu studie). Dále: „Nic takového sám nečiní“ – co tím má kritik konkrétně na mysli? Že tvorba vytvořených testů autorským kolektivem nesleduje zákonitosti tvorby didaktických testů? S tím nesouhlasíme – viz kapitola č. 3 *Kvalita didaktického testu* (primárně objektivita, reliabilita, obsahová validita, konstruktová validita). Snažili jsme se maximálně držet daného protokolu s tím, že zejména v kontextu expertního hodnocení nedošlo k naplnění konstruktové validity v intencích doporučení (poměr úloh na vyšší × nižší myšlenkové operace).

> Neuenhaus, N. (2011). *Metakognition und Leistung: Eine Längsschnittuntersuchung in den Bereichen Lesen und Englisch bei Schülerinnen und Schülern der fünften und sechsten Jahrgangsstufe* [Doctoral dissertation, Universität Otto-Friedrich, Bamberg, Germany].

Kritik dodává: **přeci pouze tato (nerealizovaná) validace, která předchází vlastnímu výzkumu, nás opravňuje nazývat dané testy didaktickými.**

Rozumíme tomu dobře, že didaktický test je možné označit „didaktickým testem“ pouze v momentě, kdy byl použit v předchozím šetření (je k dispozici zdroj)? Doplňujeme, že v příspěvku nepíšeme, že testy autorů sledující protokol pro tvorbu didaktických testů jsou jednoznačně lepší/vhodnější ve všech sledovaných aspektech, ale zkoumáme, jestli se liší a jaký vliv v případě jejich užití (vůči testům používaných pedagožkou) mají na skór žáků.

A znovu – kde vzniká oprávnění nazývat první skupinu jako didaktické testy – není-li možné je předem validovat, nechť studie přizná, že porovnává „testy učitelky a ty, co jsme sami vytvořili podle zápisků ze sešitů dětí“.

Proč by autoři studie nemohli dané testy označit za „didaktické testy“? Zormanová (2017) přímo uvádí, že od běžné zkoušky se didaktický test liší tím, že je navrhován, hodnocen, ověřován a interpretován podle určitých pravidel. Pravidel – protokolu pro tvorbu didaktických testů – jsme se snažili maximálně držet.

> Zormanová, L. (2017). *Didaktika dospělých.* Grada.

Použité testy nejsou součástí publikované práce.

Podle našeho názoru se v tomto bodu nemůže jednat o objektivní výhradu. V řadě článků (včetně J_{imp}) nejsou mnohdy použité nástroje k dispozici (důvodem je i fakt ohledně omezeného množství znaků při publikování studie).

Kritika ve vztahu k velikosti vzorku:

Autoři explicitně uvádějí: „Záměrem této práce je vzbudit pozornost.“ Jsme si vědomi toho, že není možná generalizace, a na tuto skutečnost v textu upozorňujeme: „Design celého výzkumného směřování (dostupný výběr, jedna třída, nízký počet žáků, pět teacher made testů) neumožňuje generalizaci výsledků.“

Kritika „Podkapitola Doporučený způsob skórování“

Kritik píše: „Doporučený způsob skórování je mi včetně přepočtů bodů a procent nesrozumitelný, nevím, kterou pozici autoři obhajují, zda se mají subjektivně skórované úlohy hodnotit binárně, a co vlastně sdělují.“ Kapitola se nejmenuje Doporučený způsob skórování, ale **Doporučovaný** způsob skórování, a představuje tedy syntézu názorů odlišných autorů. Vlastní kritika je tedy dána chybným čtením nadpisu ze strany kritika. V oddíle jsou připomenuty podstatné informace z teoretické části (podle kterých autorů je v jakém případě doporučováno hodnotit úlohy rozdílným počtem bodů). Dále k části „způsob skórování, která je mi včetně přepočtů bodů a procent nesrozumitelná, nevím, kterou pozici autoři obhajují“ – zde autoři nic neobhajují. Přepočet bodů/procenta souvisejí právě s paradoxní situací, kdy v některých testech vytvořených pedagožkou žák řeší větší % úloh, avšak vzhledem k rozdílnému bodování úloh získává (paradoxně) horší známku.

Kritika: Zpracování dat je chybné

Skutečnost, že není v textu zmíněna práce s odlehlými hodnotami, neznamená chybu. Doplňme, že „neparametrické postupy (například použitý Spearmanův korelační koeficient) typicky převádějí původní kvantitativní hodnoty proměnných na pořadí (‚rank‘), a tím se od vlivu odlehlých hodnot oprošťují“ (Dušek et al., 2019). Stejné problematice se věnují také další autoři (např. Caruso & Cliff, 1997).

Caruso, J. C, & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's Rho. *Educational and Psychological Measurement*, *57*(4), 637–654.

Dušek, L., Pavlík, T., Jarkovský, J., & Koptíková, J. (2019). Analýza dat v neurologii LXXIV. – Neparametrický Spearmanův koeficient korelace. *Česká a slovenská neurologie a neurochirurgie*, *82*(2), 236–239.

Kritika: Zpracování dat je chybné – obrázek č. 1

S kritikem je možné souhlasit ve výtce týkající se způsobu vizualizace (nepoměr osy x a y u bodového grafu), což ovšem neznačí explicitní chybu. Ano, jsme si vědomi toho, že procenta jsou pouze do 100 a v grafu osa pokračuje chybně do 110. Tato skutečnost byla dána snahou autorů o lepší vizualizaci dat (maximum tak není „nalepené" na horní části grafu). Jsme přesvědčeni, že čtenář si je schopný s tímto poradit (umí s grafem naložit).

Kritika: Mechanismus výpočtu sumárního skóre k porovnání obou skupin testů

K této části se autoři doznávají přímo v textu článku, netřeba tedy opětovně reagovat. Hodnota korelačního koeficientu a jeho sledovaný význam jsou v textu také zmíněny.

Kritika – velikost vzorku

Autoři sami v textu udávají: „Dalším limitem této studie je malý výzkumný vzorek (1 pedagog, 5 teacher made testů a didaktických testů, 27 dětí jedné 4. třídy ZŠ) a jeho dostupný výběr znemožňující generalizaci výsledků."

Autoři se nesnaží o to, aby článek suploval nebo plně naplňoval paradigma vědy (o kterém si kritik obrázek udělal), a proto přímo v textu píší: „Design celého výzkumného směřování (dostupný výběr, jedna třída, nízký počet žáků, pět teacher made testů) neumožňuje generalizaci výsledků (externí validita) a existuje řada intervenujících proměnných potenciálně zkreslujících jednoznačnou interpretaci výsledků (interní validita), včetně uvažování nad vyhotovenými pěti didaktickými testy badateli jako potenciálního vztahového kritéria kvality vůči teacher made testům pedagožky. Záměrem této práce je vzbudit pozornost odborné (akademiků, praktikujících učitelů) i laické (municipalita, rodiče) veřejnosti k této problematice, jelikož se podle našeho mínění jedná o zcela zásadní téma, a zároveň nabídnout badatelům na tomto poli potenciální způsob při realizaci podobné studie."

Úvahy mi připadají mimořádně technistní, ale jsou formulovány bez tvaru, se kterým by šlo polemizovat.

Bohužel, na takto obecně posuzované konstatování nemůžeme reagovat.

Takováto velikost vzorku neumožňuje kvantitativní přístup ani zobecnění, jaké je stavěno ve vzletných formulacích výzkumných problémů, předpokladů a hypotéz, které jsou svou povahou „útočné" 21:5–8 „teacher made testy neodpovídají současnému paradigmatu. . . ".

Nikde autoři nehovoří o generalizaci výsledků, jelikož v souvislosti s výběrem vzorku nešlo o náhodný výběr; dále – proč „útočné"? Formulovali jsme předpoklad, to je celé.

Závěrem

Řadu komentářů od kritika shledáváme za ryze subjektivní. Chápeme i vyjádření autora kritiky „Předem se autorům omlouvám, pokud některá námitka nebude věcně správná, srozumitelně vyargumentována či by byla pouhým přehlédnutím – každý máme své limity." Osobně však zastáváme stanovisko, že má-li se jednat o věcnou kritiku či diskuzi, měla by být kritika objektivní, ideálně opřena o zdroje. Text má jistě objektivní nedostatky (není rozebrána EFA, reliabilita testů pedagožky aj.), avšak řadu dalších nedostatků ve studii explicitně pojmenováváme a přiznáváme.