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

Petr Novotný[1]

[1] Přírodovědecká fakulta, Univerzita Karlova, Viničná 7, 128 00 Praha 2; petr.novotny@natur.cuni.cz

Předkládám svůj pohled na nedávno publikovaný článek autorského kolektivu pod vedením Jiřího Říčana. Stavbu článku, jeho obsah i odborné pojetí považuji za vadné, příslušnou argumentaci a jednotlivé námitky uvádím dále. Tyto námitky jsou řazeny v posloupnosti čtenáře textu, nikoli podle jejich vnímané závažnosti. Předem se autorům omlouvám, pokud některá námitka nebude věcně správná, srozumitelně vyargumentována či by byla pouhým přehlédnutím – každý máme své limity. Je-li to možné, odkazuji se na pasáže ve vysázeném textu publikovaném ve webovém archivu[1] číslem stránky ve stránkování celého čísla a číslem odstavce; 18:5 je tedy pátý odstavec (český abstrakt) na straně 18 (titulce článku).

Článek má zavádějící jméno. Článek skutečně porovnává dvě skupiny testů, ale žádným způsobem neprezentuje metody ani data, kterými by bylo možné sledovat vliv získaných výsledků na úspěšnost žáka. Neuvádí ani předchozí či výslednou roční klasifikaci sledovaných žáků, data pouze popisují, jakou známku dostali žáci v jednotlivých testech. Jedinou připomínkou myšlenkové větve „vliv na úspěšnost žáků“ je gradace do dvou rétorických otázek k problematice tvorby závěrečné klasifikace v Závěru.

Článek je slohově slabý. Text je psán z mého pohledu pseudo-odborným jazykem, který přispívá k zastření podstaty sdělení; přebytek cizích synonym vnímám jako zástěrku myšlenkové mělčiny. 18:5 „explorace diskrepancí“ je „zjišťování nesouladu“. Cizí slova vytvářejí dojem exaktních termínů, čímž se text vyhýbá chybějící argumentaci, zde například z jakých příčin autoři předpokládají existenci nesouladu mezi testy, které tvoří učitel či didaktik.

Text je rovněž negativně zatížen užíváním klišé, jež jsou svou vágností obtížně vyvratitelná či alespoň nesrozumitelná, text znepřehledňují, znejasňují a rozmývají. Například neumím určit, co znamená věta 28:3 „Přes prokázané diference (v mezích míry reliability a validity této případové studie) mezi … [testy] … musíme mít na paměti intervenující proměnné (interní validita šetření) bránící jednoznačné interpretaci výsledků (tak jak je to v neexaktních vědách pravidlem).“ Na straně 28:2 vyúsťuje text v „nejpodstatnější zjištění této studie“. Výsledné sdělení daného odstavce není ale díky užitému jazyku vůbec jasné – připomenu, že autoři navrhli testy s deseti otázkami a asi je chtějí hodnotit jednotným počtem bodů (viz str. 20 – „akceptují binární skórování“, dále str. 27), zároveň pracují s technistním návrhem 5 % = snížení stupně. Rozumím správně danému odstavci, že článek přichází s představou, že žák čtvrté třídy ZŠ, který neodpoví na dvě otázky z deseti, je hodnocen nedostatečně? U kratších testů paní učitelky by toto rozhraní bylo již kolem jedné nesplněné otázky – tento myšlenkový závěr popisované pasáže se mi nejeví jako zamýšlený, ale přesto mne k němu daný text vede. Nejasnost ve formulaci i údajně nejpodstatnějšího zjištění tak blokuje diskusi o zjištění samotném. Text odborné stati musí být srozumitelný, jednoznačný a logicky stavěný, což jsou zásady, které článek podle mého názoru nesplňuje.

Práce s literaturou je špatná. Citační přístup považuji za povrchní, vybírající si dílčí větu či tvrzení, a pomíjející přitom argumentaci, která k danému tvrzení (snad) v originální práci ústí. Příkladem může být 27:5 „Šatánek a Hubalovská … tvrdili, že by každá úloha měla být ohodnocena podle náročnosti. Tím pádem by testové položky měly být hodnoceny jiným počtem bodů…“ Namísto představení argumentace citovaného přístupu, vztažení k tématu či protinázorům je použita pouze jeho zkratka, se kterou je pak nakládáno podle potřeby. Zásadní námitkou je citace na 22:2, kde je nedostatečná velikost vzorku pro kvantitativní metodiku užitou v textu obhajována článkem *Flybvjerg 2006* věnovanému kvalitativnímu (!!) výzkumu. Citovaná práce dle mého pročtení žádné sdělení na podporu tohoto tvrzení nepřináší. Konstatování, že z jednoho případu lze někdy získat více informací než z velkých dat, nemůže být chápáno jako omluva pro nedostatečnou velikost dat při kvantitativní analýze. Označení článku za „případovou studii“ rovněž nemá vliv na principiální požadavky užitých kvantitativních metod. Zaujalo mne také, že jsou citována dvě vydání Chráskových *Metod pedagogického výzkumu* – nejedná se nutně o chybu, ale smysl tohoto počinu mi uniká. Značná část teoretického úvodu je založena na studijních textech téhož autorského kolektivu, nikoli na primárních pramenech.

Text nedodržuje základní členění výzkumné stati. Teoretická kapitola *Kvalita didaktického testu* obsahuje na straně 20 čtyři sekce, které ve skutečnosti popisují použité metody (jak se dělaly testy, jak dlouhé byly a jak byly skórovány, výpočet reliability, zdroj obsahu pro testy). Podobně je nakládáno s limity studie – autoři si stále něco „uvědomují“, ale přesto dál pokračují v myšlenkových

[1] https://ojs.cuni.cz/scied/article/view/1837/1624

konstrukcích, které souběžně negují. U popisu *Nástroje, procedura a výzkumný vzorek* rozumím zařazení limitů, byť neobratnému. Limity jsou ale uváděny i v úvodu *Syntézy: komparace testů*. Rozmytí struktury článku je patrné i v absenci kapitoly *Výsledky*. Přestože je text koncipován jako výzkumná studie, jsou výsledky představovány formou volného členění metodické kapitoly *Analýza a zpracování dat*. Studie má přitom poměrně kompaktní metodiku, zpracování dat a snad i výsledky, ale jejich nesystematická prezentace vede společně s již zmíněným slohem opět k zamlžení textu. Podkapitola *Doporučený způsob skórování* je mi včetně přepočtů bodů a procent nesrozumitelná, nevím, kterou pozici autoři obhajují, zda se mají subjektivně skórované úlohy hodnotit binárně, a co vlastně sdělují. Úvahy mi připadají mimořádně technistní, ale jsou formulovány bez tvaru, se kterým by šlo polemizovat.

Didaktické testy nejsou testy tvořené didaktiky. Z textu na mne jako čtenáře čiší přesvědčení, že pokud test tvoří pracovník učitelské katedry (jak bych asi já definoval onoho didaktika), jde o didaktický test (a tento je lepší než test zkoumané učitelky). Přestože článek na 21:1 správně uvádí, že „didaktický test … se liší hlavně daným metodologickým přístupem“, nic takového sám nečiní. Použité testy nejsou součástí publikované práce, není ani jasné, které z uvedených témat (uvozeno písmeny) odpovídá kterému testu (uvozeno číslicemi), a těžko si o nich dělat úsudek. V textu je sice poskytnuta informace o subjektivně/objektivně skórovatelných položkách, resp. jejich poměrech, ale „didaktické“ testy nebyly před použitím ve studii validovány o nic více než testy oné paní učitelky. Zde zdůrazňuji významnou námitku – přeci pouze tato (nerealizovaná) validace, která předchází vlastnímu výzkumu, nás opravňuje nazývat dané testy didaktickými! „Sulcovitá“ stavba textu uhýbá každému mému pokusu o jednoznačný závěr a budu rád, pokud mne jiný pozorný čtenář opraví. Já ale text čtu tak, že testy obou typů byly přímo použity pro sběr dat, na didaktických testech byla poté spočítána reliabilita (22:3), a tím „naplněny“ metodické požadavky. Správný postup, tedy validace testů předcházející jejich použití, je v dané situaci asi složitý, ALE – proč nebyla určena reliabilita i pro testy paní učitelky? 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“. Jenomže to je úplně jiný příběh, než článek představuje.

V textu 22:3 „…jsme využili tři expertní posudky…“, v tabulce 1 tamtéž je pět expertů spolu s tajemným „Me“.

„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?

Není artikulován obsah pojmů subjektivně/objektivně skórovatelné otázky, tedy není jasné, zda nedochází (chybně) k záměně za dvojici termínů otevřené/uzavřené otázky – nicméně text neobsahuje dostatek informací, aby se to dalo posoudit.

Podobně postrádám diskusi, 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.

Zpracování dat je chybné. Článek neobsahuje informaci o přístupu k odlehlým nebo neúplným hodnotám, zato čtenáři nabízí (23:2) vzorec na výpočet procent z celku (!). Vím, že určité detaily mohou vždy proklouznout pozorností redakčního procesu, ale když najdeme u obrázku č. 1 osy procentuálního výkonu žáků na škále od 20(30) % do 110 % (společně s chybějícími jednotkami či vizuálním nepoměrem osy x a y u bodového grafu, přestože obě osy mají stejný rozsah), nevím, jak s tím mám jako čtenář naložit.

Podívejme se na mechanismus výpočtu sumárního skóre k porovnání obou skupin testů. Předpokládám, že snad v průběhu recenzního řízení někdo poukázal na fakt, že postup průměrování procent různých testů je vadný (a bylo by namístě skóry k sobě některou metodou standardizovat), což si autoři v publikovaném textu „uvědomují“ (23:10), nicméně ignorují, neboť „studie odráží problematiku praxe“. Bohužel zde se text nezastavuje a nespokojeně komentuje výslednou korelaci „… očekávali [bychom] spíše vysokou až absolutní korelaci“, později 27:8 navazuje „potvrzení této hypotézy se samozřejmě dalo předpokládat, naše pozornost ale směřovala k hodnotě… [korelačního koeficientu]“. Připomínám, že obě skupiny se liší poměrem subjektivně skórovatelných úloh, a tak by stálo možná za diskusi vliv testu versus vliv hodnotitele. Nejpodstatnější je ale uvědomit si, že testové skóre měří latentní proměnou (znalost nějakého výukového tématu, zde spíše několika faktorů bez bližší specifikace), a toto se děje s poměrně značnou chybou; i u mnohem delších a metodicky sofistikovanějších testů třeba kolem jedné směrodatné odchylky; a extrémní hodnoty jsou měřeny z principu ještě více nepřesně. Netroufám si hodnotit, co vypovídá zjištěná korelace, neznáme odhad chyby měření a upřímně taková otázka přesahuje můj obzor statistické simulace – očekávat plnou korelaci (velkou chybou zatížených měření různých témat různými nástroji zprůměrovaných přes sebe) je ale nemožné. Závěrem pak připomenu problémy s výpočtem reliability

uvedené již dříve a velikost vzorku, kterou vidíme na straně 25 – dvacet sedm žáků v pěti testech vedlo ke dvěma proměnným – průměr z jedné skupiny testů, průměr ze druhé skupiny testů – obě v kvantitě $n = 27$. 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…“. Docházím tak k závěru, že ani po trojím prostudování článku nevím, jak kvalitní testy používá ona učitelka, ale o tom, jak se článek blíží k paradigmatu vědy, jsem si obrázek udělal dostatečný.

Jakkoli je nepříjemné psát i číst takovouto kritiku, musím závěrem shrnout – vysoká míra formálních chyb, nízká odborná a metodologická úroveň článku zahrnující zjevně chybné postupy – to vše nás staví před věčnou otázku, jaké charakteristiky má mít kvalitní didaktický výzkum. Tato definice nevzniká dílem jednotlivce, nýbrž dynamickým balancováním komunity: co se píše, co se tiskne, co se čte a co diskutuje – předmětný článek dosáhl všech těchto met. *Zaslouženě?*