Komparace kvality tzv. teacher made testů s didaktickými testy a jejich vliv na úspěšnost žáků: případová studie

Comparison of the quality of the teacher made tests with achievement tests and their influence on the success of pupils: case study

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

Příspěvek se zaobírá komparací didaktických testů a tzv. „teacher made“ testů. Hlavním cílem této případové studie je explorace diskrepancí ve skórování žáků na základě toho, je-li test stejného učiva tvořen pedagožkou nebo badatelem. Z toho důvodu autoři zjišťují: (1) kvalitu teacher made testů a didaktických testů vytvořených badateli ve vztahu k zákonitostem tvorby didaktických testů a (2) existenci spojitosti mezi výsledky žáků v rámci jejich procentuálního skórování v teacher made testech a didaktických testech. Za účelem naplnění cílů bylo sestaveno badateli 5 didaktických testů z témat, která žáci probírali ve čtvrtém ročníku 1. stupně ZŠ v předmětech vlastivěda a přírodověda, a byly komparovány s 5 teacher made testy vyhotovených pedagožkou. Studie poukazuje na nedostatečné kvalitativní aspekty teacher made testů (reliabilita, poměr subjektivně a objektivně skórovatelných úloh, poměr úloh vyžadujících nižší a vyšší kognitivní operace, způsob skórování). Přestože korelační analýza odhalila středně vysokou spojitost mezi procentuálním skóre žáků v teacher made testech a didaktických testech, očekávali bychom spíše spojitost blížící se absolutní hodnotě. Výsledky jsou diskutovány ve vztahu k současnému paradigmatu tvorby didaktických testů, dále je poukázáno na limity studie a zároveň jsou naznačeny potenciální cesty dalšího empirického směřování.

Klíčová slova:
první stupeň ZŠ, přírodověda, vlastivěda, teacher made test, didaktický test.

Zasláno 7/2020
Revidováno 10/2020
Přijato 2/2021

This article compares achievement tests and so-called “teacher-made” tests. The main objective of this case study is to explore and discrepate the results of pupils’ scorings based on whether is test of the same subject has been matter created by the educator or by the researcher. For this reason, the authors ascertain: (1) the quality of teacher-made tests and achievement tests created by the researchers in relation to the regularities of creating achievement tests and (2) the existence of a connection between pupils’ results within their percentage scoring in teacher-made tests and in achievement tests. In order to meet the objectives, the researchers compiled 5 achievement tests on topics that the pupils had been subjected to in the fourth grade of elementary school, during primary science classes. These were then compared to 5 teacher-made tests made by the educator. The study indicates insufficient qualitative aspects of the teacher-made tests (reliability; the ratio of subjectively and objectively scoreable tasks; the ratio of tasks requiring lower and higher cognitive operations; scoring method). Although the correlation analysis revealed a moderately high correlation between the percentage score of pupils in teacher-made tests and achievement tests, we would rather expect a continuity approaching absolute continuity. The results are discussed in relation to the current paradigm of creating achievement tests. The article also points out the limitations of the study and indicates potential ways of further empirical direction.

Key words:
primary education, primary science subject, teacher made test, achievement test.

Received 7/2020
Revised 10/2020
Accepted 2/2021

1 Úvod

Jednou z nejdůležitějších činností učitele je objektivní hodnocení (Junková, 2006), které funguje jako zpětná vazba především pro učitele, ale i pro žáky a jejich rodiče. Učitel z výsledků může zjistit, v jaké míře si žáci osvojili učivo a zda si ho zvládli zařadit mezi dříve získané vědomosti. Je to důležité zvláště v oblasti kognitivní, kdy učitel zjišťuje, jestli žák probíranou látku pochopil (Jeřábek & Bílek, 2010). Získané výsledky učiteli nadále naznačují, zdali si žáci učivo adekvátně osvojili, tedy dosáhli vytyčeného vzdělávacího cíle, a zároveň také, zda jim byl daný obsah vhodně předán, což je právě úkolem didaktického testu (Byčkovský, 1982; Chráska, 2007; Komenda & Mazuchová, 1995; Švamberk Šauerová, 2016). Didaktický test není náhradou zkoušky ústní či praktické, ale je jejím vhodným doplňkem (Schindler, 2006). Měl by být sestavován osobou, která nejenže bude dobrým pedagogem a odborníkem na danou oblast, ale zároveň bude znát základy statistiky (Chráska, 2016). Při jeho konstrukci a vyhodnocování by měl tvůrce vycházet z poznatků didaktické teorie a specifického vzdělávacího kontextu (Cizek, 2004).

Kvalita didaktického testu klade na učitele vysoké nároky. Ve školní praxi se setkáváme s názvem nestandardizovaný didaktický test, tedy takový test, který je sestaven dostatečně odborně, ale nejsou zde splněny veškeré náležitosti, a nelze proto ověřit všechny vlastnosti testu (Kalhoust & Obst, 2002; Skutil, 2011). Dále k nestandardizovanému didaktickému testu budeme referovat v terminologii Škody a Doulíka (2007) jako k tzv. teacher made testu. Každý učitel by měl být schopný vytvořit kvalitní didaktický test, dokázat ho vyhodnotit a analyzovat jeho kvalitu, přesto v České republice není situace ohledně tvorby testů uspokojivá. Chráska (1999) popisuje, že v názorech mnoha pedagogů přetrvává nedůvěra a skepse ohledně testů a měření v pedagogice. Spousta učitelů nemá dostatečné zkušenosti s tvorbou didaktických testů a neuvědomují si, jaké parametry musí kvalitní didaktický test splňovat. Pokud chceme, aby zjišťování výsledků bylo dostatečně spolehlivé a platné, musí být při konstrukci testu použity osvědčené postupy. V praxi se často setkáváme s případy, kdy sestavené testy vyžadují pouze namemorované poznatky a jejich následné hodnocení probíhá konzervativním způsobem známkování (Čapek, 2015). Tento přístup je minimálně problematický.

Cílem této studie je popsat parametry pěti vytvořených teacher made testů jedné pedagožky a pěti didaktických testů vytvořených badateli ke stejnému učivu. Zároveň chceme poukázat na sílu spojitosti v žákovském skórování v rámci procentuálního výsledku z teacher made testů a didaktických testů.

2 Historická kontinuita a současný stav testování

Historie testování ve školní praxi sahá podle Vrány (1948) do 16. století (jezuitské školy). Na našem území můžeme sledovat kořeny didaktického testování ve druhé polovině 19. století (Jeřábek & Bílek, 2010). O rozvoj metod objektivního hodnocení pro srovnání vzdělávacích výsledků se zasloužil Joseph Mayer Rice z USA, který roku 1894 navrhl a vyzkoušel metodu používání objektivních měřítek při nácviku pravopisu, po níž mohly být výsledky srovnány s dalšími školami (Fayol et al., 2012). Škoda, Doulík a Hajerová-Müllerová (2006) jako další důležitý mezník považují rok 1908, kdy C. W. Stone sestavil v USA první standardizovaný test, což způsobilo velký rozvoj tvorby didaktických testů. Příznivá situace byla pro rozvoj české (československé) pedagogické diagnostiky po první světové válce, a to hlavně díky Václavu Příhodovi, který didaktické testy považoval jako nástroj hodnocení a do školního zkoušení zahrnoval objektivitu, srovnatelnost zkoušek a výkonnost. Další významný pedagog té doby, Otokar Chlup, vystupoval proti didaktickým testům, které mu připadaly neúčelné. Byl názoru, že školství nepodporuje tvořivé myšlení žáků.

> Mechanismus této metody [myšleno bylo didaktické testování] nemůže správně hodnotit ani soustavu vědění a myšlení žákova, ani jednotlivých poznatků, jejichž důležitost odhadnouti jest koneckonců vždy věcí subjektivního posouzení. Měřením odvrací se žactvo od vážného myšlení a tvořivá činnost žákova, rozvíjející se pod vedením učitelovým, ubíjí se v bičovaném jezuitském závodění. (Chlup, 1931, s. 56)

Spory O. Chlupa a V. Příhody o didaktické testy se projevovaly ve dvou rovinách. V té první vedl spor ke zdokonalování didaktických testů a hledání nových způsobů měření a hodnocení žákovských výstupů, zatímco ve druhé vedl k nepoužívání didaktických testů. S rozvojem empirického bádání v kontextu kvality psychometrických vlastností testů, nástrojů a přístupů, akcentem na komparaci výsledků z mezinárodních srovnávacích testů (PIRLS, TIMSS) či problematikou přijímacího řízení na střední a vysoké školy je nepochybné, že didaktický test „hraje“ jednu z ústředních rolí v pedagogickém dění, a to v rovině teorie i praxe.

3 Kvalita didaktického testu

Didaktický test má řadu funkcí. Popham (2017) popisuje tři hlavní účely pedagogického testování, kterými je srovnávání mezi testovanými, zlepšování výuky a hodnocení výuky. Hališka (1999) hovoří o funkci diagnostické, motivační a stimulační, klasifikační, kontrolní a prognostické. Nesmíme zapomenout na funkci formativní, která je v současné době zřejmě nejvíce akcentována (Laufková & Starý, 2016). Jak však může jakýkoliv test naplňovat tyto funkce, když panují pochybnosti o jeho kvalitě (psychometrické vlastnosti testu)? Termín „didaktický test“ je definován s nuancemi. Podle Chrásky (1999, s. 12) didaktický test „měří hlavně vědomosti a dovednosti žáků.“ Byčkovský (1982) nazval didaktický test nástrojem systematického zjišťování, při kterém se zjišťují výsledky výuky. Za nejbližší termín k pojmu didaktický test užívaný v zahraniční literatuře považujeme označení *achievement test*, který je definován jako nástroj určený k měření relativního úspěchu v dané oblasti prostřednictvím jednoho ze dvou typů, mezi které patří testy zjišťující úspěch a testy diagnostické (Hawkes et al., 1936; Ward et al., 1996).

Škoda a Doulík (2007, s. 11) popisují test jako „zkoušku, jejíž podmínky jsou pro všechny testované jedince shodné a jejíž výsledky mají číselný charakter.“ Z této závěrečné definice excerpujeme první ze základních vlastností didaktického testu: test by měl být maximálně (1) objektivní. Objektivitu testu lze dle Smékala, Švece a Zajace (1973) zajistit tím, že bude daný test sestaven z otázek s možností jednoznačné odpovědi, nadále je klíčové, aby hodnotitel mohl u jednotlivých odpovědí rozhodnout bez pochyb o správnosti odpovědi, a v posledním případě je nezbytné, aby výkon žáka byl posuzován a interpretován na základě systému norem. Klíčová je nadále jednoznačná formulace testových úloh a při jejich vyhodnocování by výsledky neměly být ovlivněny postoji a názory hodnotitele (Linn, 2008; Schindlera, 2006). Výše uvedeným kritériím odpovídají z větší části úlohy objektivně skórovatelné, k jejichž převaze při konstruování testu se přiklání i někteří odborníci (Pulpán, 1991; Škoda & Doulík, 2007). Při tvorbě didaktických testů se badatelé drželi co možná největší míry objektivity rozložení testových položek (92 % položek bylo objektivně skórovatelných). Někteří autoři opírají míru objektivity také o Bloomovu taxonomii a rozdělení jednotlivých úloh dle jejich cílení na využívání vyšších a nižších kognitivních operací (Anderson et al., 2000).

Na základě aktuálního paradigmatu tvorby didaktického testu uvádíme další vlastnosti, kterými jsou (Chráska, 1999): (2) reliabilita, přesnost a spolehlivost, kdy se výsledky testů musí co nejméně lišit od skutečných hodnot. Výsledky didaktického testu jsou tvořeny dvěma složkami, a to pevnou složkou, ve které se jedná o skutečné vědomosti, a náhodnou složkou, kdy na žáka působí vnější činitelé, jako je stav žáka, prostředí ve třídě, hluk z ulice apod. U správně vypracovaného didaktického testu by neměl být vliv náhodné složky vysoký, a pokud test poskytuje výsledky, na které mají vnější činitelé minimální vliv, mluvíme o testu s vysokou reliabilitou. Pokud chceme posuzovat míru reliability, slouží nám k tomu tzv. koeficient reliability, který v praxi nabývá hodnot od 0 (naprostá nespolehlivost) až po hodnoty, které se blíží k číslu 1 (naprostá spolehlivost). Existuje řada přístupů pro zjišťování reliability (split half, opakované měření, ...), mnohé z nich souvisí s charakterem testové položky. V případě binárně skórovatelných položek didaktického testu používáme Kuder-Richardsonův test, v případě, že testové položky jsou skórovatelné v intervalu $\langle 0; 1 \rangle$ a zároveň je prokázána jednofaktorová dimenzionalita, můžeme použít Cronbachovo alfa (Urbánek, 2002). Používat binární skórování (každá položka je hodnocena jedním bodem, nezávisle na obtížnosti dané položky) je doporučováno autory Škodou, Doulíkem a Hajerovou-Müllerovou (2006). Opakem je tzv. vážené skórování, kdy se jednotlivým položkám v testu přiřazuje různé bodové ohodnocení. Vážené skórování se ve školní praxi užívá častěji, své opodstatnění však má pouze tehdy, vyžadují-li některé úlohy na své řešení výrazně více času než ostatní úlohy (rovněž Pulpán, 1991). Podobně se vyjadřuje Chráska (2002), že „široké testové úlohy se navrhují poměrně snadno, ale jejich hlavní nevýhodou je nemožnost objektivního skórování“ (s. 183). V teacher made testech vyhotovených pedagožkou ani v didaktických testech vytvořených badateli se nevyskytovaly položky, jejichž vyplnění by zabralo žákům významně více času než vyplňování položek jiných (jako v případě položek otevřených se strukturou nevymezenou – Škoda et al., 2006; podobně tzv. široké testové úlohy – Chráska, 2002). V tomto případě proto akcentujeme binární skórování.

Na základě faktorové analýzy se v této studii nepotvrdila jednodimenzionalita umožňující užití Cronbachovo alfy, a tak bylo využito Split-half reliability, kdy se testové položky rozdělí na dvě přibližně stejné části. Tento způsob výpočtu reliability je u didaktických testů velmi rozšířený (Johnson & Penny, 2005; Marinova et al., 2005). Nevýhoda podstatného snížení hodnoty koeficientu korelace u Split-half reliability se koriguje použitím Spearman-Brownova vzorce (Johnson & Penny, 2005), který stanovuje reliabilitu pro celý nezkrácený test. (3) Obsahová validita (testová doména) je shoda mezi obsahem testu a obsahem výuky, tj. poměr počtu položek v didaktickém testu reprezentující subtémata okruhu, pro který je test určen. Prakticky to znamená, že pokud učitel věnoval např. subtématu č. 1 30 % výuky, mělo by být i 30 % testových položek konstruováno k tomuto obsahu. Tento aspekt byl při tvorbě didaktických testů problematizován dvěma faktory: (i) při konstrukci jsme výhradně vycházeli ze zaznamenaných informací žáků (výpisky v sešitu, odkazy na domácí úkoly v pracovním sešitu) a nebylo tak reflektováno realizované kurikulum (reálná výuka) a (ii) některé testy žáci absolvovali po několika hodinách. Tuto záležitost blíže řešíme v empirické části tohoto textu. Disman (2005) uvádí, že (4) konstruktová validita referuje k tomu, jaké pedagogické nebo psychologické konstrukty jsou vlastně diagnostikovány (vyjádření míry vztahu mezi nástrojem – testem a teoretickým konstruktem – v našem případě myšlenkové úrovně dle klasifikace Blooma). V kontextu didaktického testování se jedná o to, zdali se diagnostikují vyšší nebo nižší kognitivní operace. Jestliže učitel věnoval subtématu č. 1 30 % výuky, mělo by být 20 % konstruovaných položek na vyšší a 10 % na nižší úrovni kognitivních operací (poměr 2 : 1). Na tento způsob konstrukce testových položek v didaktickém testu poukazují například autoři Škoda, Doulík a Hajerová-Müllerová (2006) nebo Junková (2006), která dále uvádí možné využívání specifikační tabulky či techniky seznamu výukových cílů.

Za další vlastnost didaktického testu můžeme považovat senzibilitu (citlivost). Test musí být pro žáka rovněž přiměřený, jelikož příliš obtížné, či příliš snadné úlohy, by ho objektivně nehodnotily (hodnota

obtížnosti úloh). Závěrem můžeme konstatovat, že didaktický test se od jiných diagnostických způsobů (hlavně v komparaci s teacher made testy) liší hlavně daným metodologickým postupem s cílem zajištění co možná nejvíce optimálních vlastností s akcentem na reliabilitu, validitu a objektivitu (Marinova at al., 2005).

4 Cíle, výzkumné problémy, předpoklady a hypotéza

Cílem této studie je popsat parametry pěti vytvořených teacher made testů jedné pedagožky a pěti didaktických testů vytvořených badateli ke stejnému učivu. Zároveň chceme poukázat na sílu spojitosti v žákovském skórování v rámci procentuálního výsledku z teacher made testů a didaktických testů. Jelikož se studie zabývá jevem, který je běžný v pedagogické praxi, tedy konstrukcí testů, a zaměřuje se na práci jednoho učitele, můžeme hovořit o deskriptivní (popisující) případové studii (Mareš, 2015). Johanson (2003) uvádí, že jedním z možných způsobů usuzování v rámci případové studie je prostřednictvím verifikace hypotéz testovat teorii. Z toho důvodu jsme na základě vytyčených cílů projektovali jak deskriptivní, tak i relační výzkumný problém (VP).

VP_1 (deskriptivní): Jaké jsou vlastnosti teacher made testů v komparaci s aktuálním paradigmatem tvorby didaktických testů?

V souvislosti s výše uvedeným deskriptivním výzkumným problémem jsme vytvořili následující předpoklady (P):

P_1: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému počtu testových položek.

P_2: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému poměru objektivně a subjektivně skórovatelných testových položek.

P_3: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému poměru testových položek na vyšší a nižší kognitivní operace.

P_4: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému způsobu skórování (úlohy jsou hodnoceny jiným počtem bodů).

VP_2 (relační): Jaká je spojitost mezi výsledky žáků v rámci jejich procentuálního skórování v teacher made testech a didaktických testech?

V souvislosti s výše uvedeným relačním výzkumným problémem jsme vytvořili následující hypotézu (H):

H (relační): Spojitost mezi výsledky žáků v rámci jejich procentuálního skórování v teacher made testech a didaktických testech je pozitivní.

5 Nástroje, procedura a výzkumný vzorek

Za účelem naplnění vytyčených cílů sestavili badatelé (autoři této studie) celkem pět didaktických testů z témat, která žáci probírali ve čtvrtém ročníku 1. stupně základní školy (Sedláčková, 1993). Autoři při tom výhradně čerpali ze zaznamenané práce v hodině (tj. výpisků žáků v sešitech, včetně odkazů na domácí cvičení). Uvědomujeme si, že vztahové kritérium „zaznamenaná práce v hodině“ pro konstrukci didaktických testů je nekompletním přístupem, jelikož nezahrnuje kritérium „skutečná práce v hodině“. Jedná se o nedostatek, který si autoři tohoto příspěvku uvědomují, avšak dlouhodobé záměrné pozorování bylo pedagožkou odmítnuto. Testy byly konstruovány pro dva vyučovací předměty, a to vlastivědu a přírodovědu. Vzdělávací obsah těchto předmětů lze snadno strukturovat, vykazuje vysokou míru integrace[1] poznatků různých vědních disciplín, lze ho posuzovat s ohledem na mezinárodně platné vymezení přírodovědné gramotnosti. K danému vzdělávacímu obsahu se snadno vytvářejí úlohy vyžadující různé úrovně osvojení poznatků a při konstrukci a hodnocení těchto úloh se lze snadno inspirovat pravidelnými mezinárodními srovnávacími studiemi (např. TIMSS).

V rámci prezentované výzkumné studie se jednalo o následující témata: a) Má vlast (vlastivěda); b) Živá a neživá příroda a společné znaky rostlin (přírodověda); c) Orientace v krajině (vlastivěda); d) Rozmnožování rostlin (přírodověda); e) Stavba rostliny, kořen, stonek (přírodověda).

Každý z didaktických testů obsahoval vždy deset položek, což je podle Chrásky (1999) minimální počet pro zajištění akceptovatelné reliability testu, pokud bereme v potaz chybu měření. V testu se vždy nacházely rozmanité typy testových položek (blíže k typologii Kalhous & Obst, 2002). Žáci v dané škole nebyli informování o způsobu vyhodnocení testů, aby nevznikaly nejasnosti při započítání do běžné klasifikace. Uvědomujeme si, že způsob skórování by měl být žákům zpřístupněn, jelikož

[1] Posuzování testů společně pro vlastivědu a přírodovědu vnímáme za legitimní (viz vzdělávací oblast Člověk a jeho svět – RVP, 2017, s. 42).

> . . . pokud máme informace o formě testu (zkoušky), přizpůsobíme tomu svou strategii učení, vyčleníme si čas na učení, nebo zpracujeme zpětnou vazbu efektivněji, než když tyto informace o podobě testu (zkoušky) nemáme. (Dutke et al., 2010, s. 195)

Ve výzkumu (zadávání didaktických testů autory tohoto příspěvku žákům) jsme však zachovali konzistenci s přístupem pedagožky (žákům nezpřístupňuje bodové vyhodnocování testových položek). Žákům bylo řečeno, že se jedná o pracovní list, po jehož vyplnění obdrží zpětnou vazbu. Žáci vždy nejprve vypracovávali teacher made test se svou pedagožkou, a poté vyhotovili didaktický test. Didaktické testy byly vyplněny v těchto datech: 11. 10. 2019 (a) dva dny po testu pedagožky; 8. 11. 2019 (b) den po testu pedagožky; 15. 11. 2019 (c) dva dny po testu pedagožky; 18. 11. 2019 (d) čtyři dny po testu pedagožky; 2. 12. 2019 (e) čtyři dny po testu pedagožky. Časový odstup mezi vyplňováním obou typů testů je samozřejmě intervenující proměnnou. Odstup byl způsoben dvěma faktory: (1) bylo zapotřebí přijmout disponibilní možnosti pedagožky pro zadávání didaktických testů (kdy a v jaké hodině proběhne zadávání tak, aby nenarušilo „chod“ výuky) a (2) vypracování didaktického testu bezprostředně po vyhotovení testu zadaného pedagožkou by nemuselo být k žákům přiměřené (únava, kognitivní vyčerpání). Zadávání teacher made testů ani didaktických testů nebylo realizováno první ani poslední vyučovací hodinu daného dne. Výzkum byl realizován po dobu 5 měsíců (9/2019–1/2020) v jedné 4. třídě středočeského kraje. Přestože se jednalo pouze o dostupný výběr, tedy jednoho pedagoga a jednu 4. třídu ZŠ, jak uvádí Flybvjerg (2006), i na základě jednoho individuálního případu (případové studii) lze dojít k určitému rozvoji vědeckého poznání.

6 Analýza a zpracování dat

6.1 Didaktické testy

Abychom zajistili co nejvyšší platnost výzkumných zjištění, bylo zapotřebí nejprve učinit analýzu vlastností vyhotovených didaktických testů, jakožto vztahových kritérií při komparaci s teacher made testy. Reliabilita byla nejprve určena na základě relativní velikosti chyby měření, která se zvyšuje snižujícím se počtem úloh. Z toho důvodu všech pět didaktických testů obsahovalo deset testových položek. Dále za účelem stanovení spolehlivosti didaktických testů byla použita metoda Split-half reliability a korigována použitím Spearman-Brownova vzorce. Hodnota reliability pro jednotlivé didaktické testy činila: 0,75; 0,705; 0,667; 0,928; 0,904. Obecně přijatelné hodnoty reliability jsou stanoveny v intervalu 0,70–0,60 (Sekaran, 1992) a 0,95 (Cronbach & Meehl, 1955; Tavakol & Dennick, 2011). Z hlediska spolehlivosti se tak všechny didaktické testy jeví jako dostatečně spolehlivé. Poměr objektivně vs. subjektivně skórovatelných úloh činil 10 vs. 0 (did. test č. 1) a 9 vs. 1 (did. testy 2–5). Pro účely určení charakteru testových položek na základě jejich kognitivní náročnosti jsme podnikli doplňující šetření, ve kterém jsme využili tři expertní[2] posudky (stejný přístup jsme podnikli u teacher made testů – viz navazující oddíl).

Tab. 1: Expertní posouzení poměru testových položek. Poměr položek vyšší : nižší kognitivní náročnosti pro jednotlivé didaktické testy

	Did. test č. 1	Did. test č. 2	Did test č. 3	Did. test č. 4	Did. test č. 5
Expert 1	1 : 9	4 : 6	3 : 7	5 : 5	4 : 6
Expert 2	2 : 8	5 : 5	3 : 7	5 : 5	5 : 5
Expert 3	2 : 8	5 : 5	2 : 8	5 : 5	2 : 8
Expert 4	1 : 9	5 : 5	3 : 7	4 : 6	2 : 8
Expert 5	2 : 8	4 : 6	2 : 8	5 : 5	3 : 7
Me	2 : 8	5 : 5	3 : 7	5 : 5	3 : 7

Zdroj: autoři

Výše uvedené expertní posouzení nekoresponduje s dostatečným poměrem úloh na vyšší a nižší kognitivní operace (2 : 1 ve prospěch úloh na vyšší kognitivní operace). Autoři si uvědomují, že tento fakt bude mít dopad na celkovou platnost zjištěných nálezů, avšak: a) teacher made testy byly zadávány po výuce (podle našeho názoru) nízkého počtu realizovaných hodin (test č. 1: 3 vyučovací hodiny; testy č. 2 a 5: 4 vyučovací hodiny; testy č. 3 a 4: 2 vyučovací hodiny), což problematizovalo konstrukci položek v didaktickém testu (a to z hlediska konstruktové i obsahové validity) a b) při konstrukci didaktických testů jsme vycházeli výhradně ze zaznamenané práce žáky v hodině a z tohoto důvodu jsme čelili dilematu

[2]Experta jsme definovali jako osobu s vzdělaností úrovní ISCED 8 připravující budoucí učitele prvního stupně v oblasti obecné didaktiky nebo didaktiky přírodovědných předmětů (příp. kombinací těchto předmětů).

vytvoření teoreticky žádoucí úlohy vyšší kognitivní náročnosti, avšak bez písemného dokladu o tom, jestli je vůbec možné od žáků očekávat její vypracování.

Dále jsme posuzovali testové položky z hlediska jejich obtížnosti (p) a citlivosti (d). Vhodné úlohy nabývají indexu obtížnosti $p = \langle 20; 80 \rangle$, za podezřelé považujeme ty s hodnotami $p < 20$ (velmi obtížné) a $p > 80$ (velmi snadné) a zakázané jsou ty úlohy, ve kterých se p blíží k 0 (Škoda et al., 2006). Chráska (1999) rovněž označuje úlohy s indexem obtížnosti s $p > 80$ za podezřelé, avšak ty s $p < 20$ za zakázané. Pro vyhodnocení indexu obtížnosti jsme adoptovali přístup Chrásky (1999). Index obtížnosti určuje tu procentuální část celkového počtu žáků, kteří mají úlohu vyřešenou správně. Počítáme ho podle vzorce

$$p = 100 \cdot \frac{n_s}{n},$$

kde p je index obtížnosti, n_s jsou žáci, kteří odpovídali správně a n představuje celkový počet testovaných žáků.

Hodnota citlivosti pro vhodné úlohy má být $d > 0{,}25$ pro $p = \langle 30; 70 \rangle$ a $d > 0{,}15$ pro $p = \langle 20; 30 \rangle$ a $\langle 70; 80 \rangle$. Podezřelé hodnoty proto nabývají hodnot $d = \langle 0; 0{,}15 \rangle$ až $\langle 0; 0{,}25 \rangle$ a zakázané úlohy jsou ty se zápornou hodnotou ($d < 0$).

Koeficient citlivosti ULI počítáme podle vzorce

$$d = \frac{n_L - n_H}{0{,}5n},$$

kde d je koeficient citlivosti ULI, n_L je počet žáků s vyšším skórem (lepší polovina z celkového počtu testovaných žáků), kteří mají úlohu řešenou správně, n_H je počet žáků s horším skórem (horší polovina z celkového počtu testovaných žáků), již mají úlohu řešenou správně a $0{,}5n$ představuje polovinu všech testovaných žáků.

Tab. 2: Přehled vhodných, podezřelých a zakázaných úloh na základě určení indexu obtížnostia citlivosti

	Test č. 1	Test č. 2	Test č. 3	Test č. 4	Test č. 5	Celkem
Vhodné úlohy (p)	2	5	4	7	2	20
Podezřelé úlohy (p)	6	5	6	3	8	28
Zakázané úlohy (p)	2	0	0	0	0	2
Vhodné úlohy (d)	4	5	4	9	4	26
Podezřelé úlohy (d)	5	5	6	1	6	23
Zakázané úlohy (d)	1	0	0	0	0	1

Zdroj: autoři

Autoři si uvědomují relativně vysoký počet podezřelých úloh (velmi snadných; $p > 80$) objevujících se ve vytvořených didaktických testech. Autorům však není známa publikace, která by definovala jejich přijatelný počet (poměr).

6.2 Teacher made testy

Abychom odpověděli na vyřčené předpoklady, v tabulce 3 uvádíme analýzu teacher made testů ve vztahu k (1) počtu testových položek (P_1), (2) poměru objektivně a subjektivně skórovatelných položek (P_2), (3) poměru testových položek na vyšší a nižší kognitivní operace (P_3) a (4) způsobu skórování (P_4).

6.3 Syntéza: komparace testů

Abychom odpověděli na druhý výzkumný problém korelačního charakteru: „Jaká je spojitost mezi výsledky žáků v rámci jejich procentuálního skórování v teacher made testech a didaktických testech?“, využili jsme korelační analýzu.

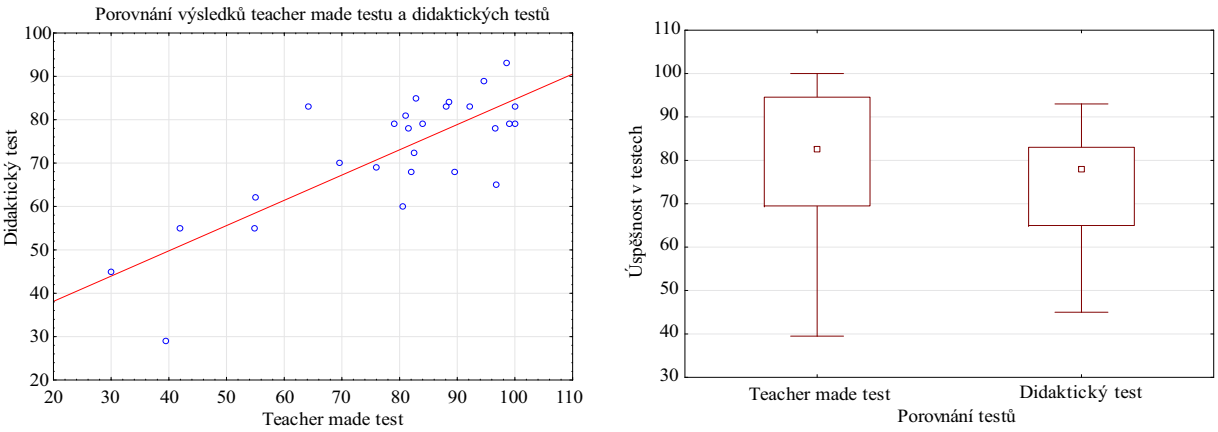
Z toho důvodu, že každý z teacher made testů byl tvořen jiným počtem položek (test č. 1: 8 úloh, test č. 2: 6 úloh, test č. 3: 7 úloh, test č. 4: 4 úlohy, test č. 5: 6 úloh), nemůžeme uvažovat o rovnoměrném rozestupu (typické pro metrická data). Rovněž si uvědomujeme, že takto vyvozovat z výsledků žáků průměrné procentuální skóre je a priory z matematického hlediska nesmyslné (průměrujeme výsledky testů, přičemž v každém z nich existovaly jiné rozestupy v důsledku rozmanitého počtu úloh a jejich skórování). Na druhou stranu tato studie odráží problematiku praxe a věříme, že výsledky přináší důležitý vhled a podnět do otázky konstrukce testů pedagogy. Také si uvědomujeme, že počty respondentů pro statistické zpracování dat induktivním přístupem jsou hraniční (1 pedagog, 27 žáků). Normalitu dat jsme tedy z důvodu rozdílných rozestupů v kontextu skórování a počtu úloh v případě teacher made testů neověřovali a rovnou jsme přistoupili k užití neparametrického Spearmanova koeficientu pořadové korelace.

Tab. 3: Základní deskripce parametrů teacher made testů

	Test č. 1	Test č. 2	Test č. 3	Test č. 4	Test č. 5
Počet úloh	8	6	7	4	6
Poměr objektivně vs. subjektivně skórovatelných úloh	1 vs. 7	2 vs. 4	0 vs. 7	0 vs. 4	0 vs. 6
Poměr úloh na vyšší : nižší kognitivní operace	0 : 8 × 0 : 8× 0 : 8 × 0 : 8× 0 : 8 **Me : 0 : 8**	0 : 6 × 1 : 5× 0 : 6 × 3 : 3× 1 : 5 **Me : 1 : 5**	1 : 6 × 0 : 7× 0 : 7 × 1 : 6× 1 : 6 **Me : 1 : 6**	1 : 3 × 0 : 4× 0 : 4 × 1 : 3× 0 : 4 **Me : 0 : 4**	1 : 5 × 1 : 5× 0 : 6 × 0 : 6× 1 : 5 **Me : 1 : 5**
Způsob skórování	1 bod: 7 úloh 7 bodů: 1 úloha	1 bod: 5 úloh 2,5 bodů: 1 úloha	1 bod: 3 úlohy 2 body: 4 úlohy	1 bod: 3 úlohy 2 body: 1 úloha	1 bod: 6 úloh
Procentuální převod (%) a klasifikace	$100 - 90 = 1$ $89 - 79 = 1 - 2$ $78 - 68 = 2$ $67 - 57 = 2 - 3$ $56 - 46 = 3$ $45 - 54 = 3 - 4$ $34 - 24 = 4$ $23 - 13 = 4 - 5$ $12 - 0 = 5$	$100 - 93 = 1$ $92 - 85 = 1 - 2$ $84 - 70 = 2$ $69 - 62 = 2 - 3$ $61 - 54 = 3$ $53 - 45 = 3 - 4$ $44 - 37 = 4$ $36 - 29 = 4 - 5$ $28 - 0 = 5$	$100 - 96 = 1$ $95 - 85 = 1 - 2$ $84 - 65 = 2$ $64 - 54 = 2 - 3$ $53 - 45 = 3$ $44 - 33 = 3 - 4$ $32 - 24 = 4$ $23 - 15 = 4 - 5$ $14 - 0 = 5$	$100 - 90 = 1$ $89 - 71 = 1 - 2$ $70 - 61 = 2$ $60 - 51 = 2 - 3$ $50 - 41 = 3$ $40 - 31 = 3 - 4$ $30 - 21 = 4$ $20 - 11 = 4 - 5$ $10 - 0 = 5$	$100 - 93 = 1$ $92 - 68 = 1 - 2$ $67 - 60 = 2$ $59 - 51 = 2 - 3$ $50 - 40 = 3$ $39 - 28 = 3 - 4$ $27 - 16 = 4$ $15 - 8 = 4 - 5$ $7 - 0 = 5$

Zdroj: autoři

Hodnota Spearmanova korelačního koeficientu byla $r = 0{,}662$ a koeficient determinace $r^2 = 43{,}8\ \%$. Ačkoliv je možné nulovou hypotézu o nulovém korelačním koeficientu zamítnout na jednoprocentní hladině významnosti ($p = 0{,}000\,2$), jedná se stále pouze o středně silnou asociaci (Hendl, 2012; Chráska, 2016). Vzhledem k charakteru testů bychom očekávali spíše vysokou až absolutní korelaci (přímou úměrnost). Ke stejnému závěru bychom došli také na základě Kendalova tau ($\tau = 0{,}501$). Srovnání jednotlivých testů je dobře patrné z korelačního grafu (zároveň data také vizualizujeme prostřednictvím kvartilového grafu, přestože pro tyto účely nebývá primárně využíván).



Obr. 1: Porovnání teacher made testu a didaktického testu na základě bodového a kvartilového grafu

Z bodového grafu je patrné, že závislost nebude silná, jelikož vykreslené hodnoty se značně odchylují od přímky (lineární interpolace), a to bez ohledu na to, jakou analytickou křivku využijeme. Z kvartilového grafu je dobře patrné rozložení obou měření. V případě didaktického testu dochází k posunu na vertikále směrem dolů a „ořezání“ krajních hodnot (velmi dobrého a velmi špatného výsledku žáka).

Tab. 4: Srovnání typů testů. V tabulce uvádíme počet získaných procent každého z testů. V závorkách je u teacher made testů uvedena klasifikace pedagožky. První sloupec: teacher made test, druhý sloupec: didaktický test. Symbol „×“ vyjadřuje neúčast žáka při vyplňování testu. Závěrečný sloupec prezentuje průměrné procentuální skóre žáků z obou typů testů (TMT = teacher made test; DT = didaktický test)

Žák	Test č. 1		Test č. 2		Test č. 3		Test č. 4		Test č. 5		∅ skór	
											TMT	DT
Dívka 1	82,86 (1)	76	86,67 (1–2)	76,7	55 (2–3)	80	90 (1)	95	100 (1)	96,7	82,91	85
Dívka 2	64,29 (2–3)	60,7	86,67 (1–2)	80	82 (2)	60	80 (1–2)	65	100 (1)	96,7	82,59	72,4
Chlapec 1	92,86 (1)	70	100 (1)	100	100 (1)	87,5	90 (1)	× ×	100 (1)	96,7	94,57	89
Dívka 3	96,43 (1)	64	86,67 (1–2)	80	100 (1)	100	100 (1)	65	100 (1)	80	96,62	78
Chlapec 2	64,23 (3)	64	66,67 (2–3)	80	45 (3)	45	80 (1–2)	68,6	91,67 (1–2)	90	69,51	70
Chlapec 3	71,43 (2)	62	46,67 (3–4)	60	45 (3)	30	30 (4)	61,4	16,66 (4)	63,3	41,95	55
Dívka 4	85,71 (1–2)	76	86,67 (1–2)	70	× ×	85	80 (1–2)	88,6	100 (1)	93,3	88,10	83
Chlapec 4	64,29 (2–3)	76	86,67 (1–2)	90	100 (1)	80	100 (1)	85	91,67 (1–2)	90	88,53	84
Dívka 5	× ×	32	40 (4)	50	41 (3–4)	70	40 (3–4)	45,7	0 (5)	25	30	45
Chlapec 5	64,29 (2–3)	64	66,67 (2–3)	86,7	95 (1–2)	70	100 (1)	71,4	× ×	100	81,49	78
Dívka 6	85,71 (1–2)	66	93,33 (1)	93,3	100 (1)	65	90 (1)	92,1	91,67 (1–2)	100	92,14	83
Dívka 7	100 (1)	72	× ×	70	100 (1)	80	70 (2)	22,9	58,33 (2–3)	96,7	82	68
Chlapec 6	71,43 (2)	60	80 (2)	76,7	91 (1–2)	45	80 (1–2)	30	× ×	86,7	80,61	60
Dívka 8	71,43 (2)	64	86,67 (1–2)	96,7	100 (1)	50	90 (1)	40	100 (1)	90	89,62	68
Chlapec 7	× ×	× ×	× ×	80	100 (1)	90	100 (1)	80	× ×	66,7	79	79
Dívka 9	78,57 (1–2)	60	73,33 (2)	60	64 (2–3)	80	80 (1–2)	70	66,66 (2)	70	100	79
Dívka 10	50 (3)	60	86,67 (1–2)	90	100 % (1)	70	× ×	28,6	66,66 (2)	96,7	76	69
Chlapec 8	92,86 (1)	68	100 (1)	100	100 (1)	100	100 (1)	98,6	100 (1)	100	98,57	93
Chlapec 9	100 (1)	80	93,33 (1)	83,3	100 (1)	75	100 (1)	67,1	100 (1)	90	99	79
Chlapec 10	100 (1)	86	100 (1)	100	100 (1)	85	100 (1)	48,6	100 (1)	96,7	100	83
Dívka 11	92,86 (1)	62	93,33 (1)	90	73 (2)	72,5	70 (2)	87,1	83,33 (2)	93,3	81	81
Dívka 12	64,29 (2–3)	46	73,33 (2)	53,3	27 (4)	47,5	60 (2)	× ×	50 (3)	71,7	54,92	55
Chlapec 11	85,71 (1–2)	90	80 (2)	76,7	55 (2–3)	77,5	50 (3)	98,6	50 (3)	73,3	64,14	83
Chlapec 12	50 (3)	54	66,67 (2–3)	76,7	73 (2)	35	70 (2)	64,3	16,66 (4)	78,3	55	62
Dívka 13	× ×	62	86,67 (1–2)	60	100 (1)	65	100 (1)	50	100 (1)	86,7	96,67	65
Chlapec 13	× ×	44	26,67 (5)	3,3	73 (2)	70	0 (5)	× ×	58,33 (2–3)	0	39,50	29
Chlapec 14	78,57 (2)	76	73,33 (2)	76,7	× ×	80	× ×	60	100 (1)	100	83,97	79

Zdroj: autoři

7 Interpretace dat a diskuze

7.1 Počet testových položek

První předpoklad P_1: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému počtu testových položek referuje k reliabilitě užitých nástrojů. Námi vytvořených pět didaktických testů vždy obsahovalo deset testových položek, což koresponduje s doporučením aktuální literatury (Škoda et al., 2006). Na druhou stranu pět teacher made testů bylo tvořeno 8, 6, 7, 4 a 6 testovými položkami s průměrnou reliabilitou. Podle odborné literatury se nejedná o dostatečnou hranici vzhledem k tomu, že relativní velikost chyby měření se zvyšuje se snižujícím se počtem úloh. Ve výsledku to znamená, že zatímco v případě didaktických testů ovlivňují nahodilé jevy (únava, vnější biologické podmínky atp.) výsledné hodnocení z 10 %, v případě teacher made testů se jedná průměrně o 17 %.

7.2 Poměr objektivně a subjektivně skórovatelných položek

Analýzu druhého předpokladu P_2: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému poměru objektivně a subjektivně skórovatelných testových položek prezentujeme prostřednictvím tabulky.

Tab. 5: Procentuální výsledek počtu objektivně skórovatelných úloh v jednotlivých testech

Testy	Test č. 1	Test č. 2	Test č. 3	Test č. 4	Test č. 5	Průměr
Teacher made test	87,5 %	66,67 %	100 %	100 %	0 %	70,8 %
Didaktický test	100 %	90 %	90 %	90 %	90 %	92 %

Zdroj: autoři

Zjištěné průměrné hodnoty je možné interpretovat tak, že v případě teacher made testů vstupuje subjektivita do výsledného hodnocení z 29,2 %, zatímco v případě didaktických testů se jedná pouze o 8 %. V žádném případě nechceme polemizovat nad tím, zda subjektivita do vyučovacího procesu patří (z našeho úhlu pohledu jednoznačně patří). Jsme však přesvědčeni, že subjektivita má své místo v případě hodnocení úsilí, schopnosti kooperace, verbálního projevu apod. V případě testování (testujeme-li znalosti) by měl být její vliv minimalizován. Vyhodnocování subjektivně skórovatelných úloh může podléhat percepčním chybám učitele (kvalita písma žáka, vnímání žáka učitelem, gender aj.). Percepční chyby jsou mentální zkratky (heuristiky – zrychlený úsudek o druhém člověku) z důvodu nedostatku času či zájmu. V praxi se jednotlivé chyby doplňují a navazují na sebe, čímž je umocněn jejich efekt. Mohou si však také odporovat. Holeček (2014) zdůrazňuje významnost nezaujatosti učitele při posuzování žáků: „Vnímání učitele by mělo být objektivní a nezkreslené…“ (s. 36), přestože je zřejmé, že každý člověk, tedy i učitel, je neustále ovlivňován mnoha vjemy, které následně zpracovává a utváří si z nich názor a postoj k ostatním. Pokud si však vyučující dostatečně neuvědomuje své vlastní percepční omezení, a nesnaží se je napravovat, jeho chování může přinášet dalekosáhlé nežádoucí dopady na školní výkony žáka, na jeho školní úspěšnost, budoucí studium i celý jeho život. „Každý člověk, tedy i každý učitel, má předsudky. Podmínkou pro odstranění předsudků je jejich rozpoznání. Snaha uvědomit si vlastní předsudky je akt zodpovědnosti.“ (Lazarová & Pol, 2002, s. 5) Rozborem percepčních chyb bychom se již dostali mimo ústřední linii tohoto příspěvku, nicméně považujeme za nutné výběrově uvést příklady, ve kterých se subjektivita posuzovatele promítla do hodnocení (blíže Man et al., 2000). Například v kontextu neměnnosti očekávání učitele vůči žákům (tzv. perseverační tendence) bylo již v roce 1928 prokázáno, že u zdatnějších žáků měli učitelé tendenci v náhodně vybraných diktátech přehlížet chyby častěji než u méně zdatných žáků. Další příklad uvádí, že když byly dvěma skupinám učitelů předloženy dvě stejné písemné práce s informací, že práce byla napsána jazykově nadaným žákem s výborným rodinným zázemím nebo průměrným žákem z průměrného rodinného zázemí, učitelé hodnotili stejnou práci fiktivně lepšího žáka s dobrým zázemím značně lépe než průměrného žáka. Za určitých okolností proto může nepřiměřené očekávání nabýt podoby sebenaplňující předpovědi (pygmalion efekt, golem efekt). Jiné studie prokazují vliv vzhledu žáka na výkon (Fitzpatrick et al., 2016), poruchy učení, pozornosti, chování (učitelé předpovídali vyšší úspěch studentům, kteří četli pod úrovní požadované čtenářské gramotnosti a zároveň nebyli označení jako žáci s poruchou učení, pozornosti či chování (Tournaki, 2003), nebo genderu (Hofer, 2015). Subjektivita by měla být co nejvíce eliminována právě konstrukcí objektivně skórovatelných položek.

7.3 Poměr testových položek na nižší a vyšší kognitivní operace

Třetí předpoklad P_3: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému poměru testových položek na vyšší a nižší kognitivní operace souvisí s tzv. konstruktovou validitou, tj. zda (a jak dobře) určený nástroj zjišťuje specifické pedagogické nebo psychologické konstrukty (v našem případě úroveň kognitivních operací dle Bloomovy taxonomie). Pro určení typu položek jsme využili expertní validizace (též mínění soudců – Disman, 2005). Každý z pěti expertů pro každou z testových položek v případě didaktických i teacher made testů dichotomicky určil, zda se jedná o položku vyžadující vyšší nebo nižší kognitivní operace. Jednotlivé testy byly tvořeny následujícími položkami (poměr nižší : vyšší kognitivní náročnost):

- Didaktické testy (*Me* expertního posouzení): 2 : 8, 5 : 5, 3 : 7, 5 : 5, 3 : 7. Celkem: 18 : 32 (36 % : 64 %).
- Teacher made testy (*Me* expertního posouzení): 0 : 8, 1 : 5, 1 : 6, 0 : 4, 1 : 5. Celkem: 3 : 24 (11 % : 89 %).

Na základě expertního posouzení můžeme konstatovat, že v teacher made testech nebyl reflektován poměr otázek na vyšší a nižší kognitivní operace, jenž je doporučován odbornou literaturou. Na druhou stranu nebyl tento poměr splněn ani v kontextu didaktických testů. Jak však již bylo jednou uvedeno, konstrukce didaktických testů byla limitována nízkým počtem realizovaných vyučovacích hodin (2–4), ze kterých pedagožka vytvářela teacher made testy, a při jejich tvorbě jsme vycházeli výhradně ze zaznamenané práce žáky v hodině (nebylo nám umožněno přímé pozorování). Z tohoto důvodu jsme čelili dilematu vytvoření teoreticky žádoucí úlohy vyšší kognitivní náročnosti, avšak bez písemného dokladu o tom, jestli je vůbec možné od žáků očekávat její vypracování.

7.4 Doporučovaný způsob skórování

Závěrečný předpoklad (P_4: Teacher made testy neodpovídají současnému paradigmatu tvorby didaktických testů ve vztahu k doporučovanému způsobu skórování – úlohy jsou hodnoceny jiným počtem bodů) odkazuje k bodování jednotlivých testových položek. Šatánek a Hubalovská téměř před půl stoletím (1972) tvrdili, že by každá úloha měla být ohodnocena podle náročnosti. Tím pádem by testové položky měly být hodnoceny jiným počtem bodů jako ve většině námi analyzovaných teacher made testech. Jejich výzkum se věnuje třem žákům, kdy každý z žáků zvládl stejný počet úloh, ale každá úloha byla jiné náročnosti, proto nevidí důvod, proč by žáci měli být ohodnoceni stejně. Jak však bylo výše uvedeno, v případě, kdy test neobsahuje široké úlohy vyžadující delší čas na jejich vyplnění, doporučuje se binární skórování (Chráska, 1999; Škoda et al., 2006). V teacher made testech nebyl tento přístup respektován. Tento fakt může vyústit v paradoxní situaci:

- Chlapec 4, který v teacher made testu č. 1 vyřešil 3 úlohy z 8 (vyřešil 37,5 % testu), získal 9 bodů (jelikož první úloha byla hodnocena maximem 7 bodů) ze 14 možných a (64,3 %) byl hodnocen známkou 2–3. Chlapec 13 ve stejném testu vyřešil 4 úlohy za 1 bod a z první úlohy za 7 bodů získal 3 body (tuto úlohu řešil na 42,9 %). Z hlediska testové domény tento žák vyřešil téměř 4,5 úlohy (56,3 % testu), získal však 7 bodů ze 14 (50 % bodového maxima) a byl ve výsledku hodnocen hůře (známka 3) než chlapec 4.
- Nejednotnost v bodovém ohodnocení jednotlivých testových položek a jejich různý počet v jednotlivých teacher made testech způsobuje druhou paradoxní situaci – zisk 60 % bodů z celkového množství znamenalo v případě teacher made testu č. 1 a č. 3 hodnocení 2–3, testu č. 2 hodnocení 3, testu č. 4 a 5 hodnocení 2.

7.5 Souvislost výsledků žáků v rámci užití dvou nástrojů měření

V závěru předkládané případové studie jsme testovali nulovou hypotézu tvrdící, že: „Neexistuje spojitost mezi výsledky žáků v rámci jejich procentuálního skórování v teacher made testech a didaktických testech.“ Na jednoprocentní hladině významnosti jsme odmítli nulovou hypotézu a přijali hypotézu alternativní: spojitost mezi výsledky žáků v rámci jejich procentuálního skórování v teacher made testech a didaktických testech je pozitivní. Potvrzení této hypotézy se samozřejmě dalo předpokládat, naše pozornost však směřovala k hodnotě Spearmanova korelačního koeficientu ($r = 0{,}662$) vyjadřující pouze středně silnou spojitost (Hendl, 2012; Chráska, 2016).

Vzhledem k tomu, že žáci pětkrát po sobě vyplňovali vždy dva testy na stejný obsah vzdělávání, očekávali bychom spojitost blížící se až k absolutní korelaci. Tento fakt je rovněž patrný z celkového průměrného procentuálního skóre, které prezentujeme v tabulce 6.

Tab. 6: Syntéza – celkový průměrný procentuální skór žáků

Typ testu	Teacher made test	Didaktický test
Celkový průměrný výsledek žáků	78,91	72

Zdroj: autoři

Podle našeho mínění tato skutečnost vyvolává otázky ohledně činitelů, které tuto situaci zapříčinily (spolehlivost testů? orientace teacher made testů primárně na položky na nižší kognitivní operace a s tím související nižší schopnost žáků řešit položky na vyšší kognitivní operace v didaktických testech), a zároveň otevírá pole pro budoucí výzkumy v této věci.

Z didaktického testu měli žáci většinou horší výsledky než z testu pedagožky (viz tab. 4). Sedláčková (1993) uvádí možnost klasifikace na základě procenta správných odpovědí. Podle autorky v případě běžné klasifikace představuje každá ztráta 10 % jeden klasifikační stupeň, avšak ve velmi přísně nastaveném měřítku je to již 5 %. Uvažujeme-li v intencích výsledků této studie a při nastavení běžného klasifikačního měřítka Sedláčkové (1993), 12 z 27 žáků by obdrželo jiné hodnocení v závislosti na tom, zda by byl jako evaluační nástroj použit teacher made test, nebo didaktický test. V osmi případech se průměrné výsledky z obou typů testů lišily o 5–9 % a pouze v sedmi případech se výsledky lišily méně než o 4 % (což by podle autorky nemělo vliv na výsledné hodnocení). To je podle našeho názoru nejpodstatnější zjištění této studie. I zde se otevírá možnost dalšího šetření, jelikož potenciálně přínosné by mohlo být interview se sedmi žáky, kteří v této studii (poněkud paradoxně) získali vyšší průměrné skóre z didaktických testů než z teacher made testů pedagožky.

8 Limity studie

Přes prokázané diference (v intencích míry reliability a validity této případové studie) mezi teacher made testy a didaktickými testy musíme mít na paměti intervenující proměnné (interní validita šetření) bránící jednoznačné interpretaci výsledků (tak, jak je to v neexaktních vědách pravidlem). Zde uvádíme činitele, které mohly zkreslit výsledky: efekt měření odkazuje k situaci, kdy si žáci z prvního testu osvojí určité informace nebo dovednosti, což následně zkreslí výsledky druhého testu. Prodleva mezi testem zadaným pedagožkou a námi zadaným testem činila 1–4 dny. Žáci mohli po testu zadaném pedagožkou diskutovat jeho obsah a správné odpovědi na otázky nebo si své odpovědi ověřit ze zápisků v sešitu (na druhou stranu by pak zřejmě neobdrželi průměrně horší hodnocení, jak se také stalo). Ne všechny testy byly vyplňovány stejným počtem žáků (všech 10 testů vyplnilo 15 žáků, 9 žáků vyplnilo 9 testů, 2 žáci vyplnili 8 testů a 1 žák vyplnil 6 testů; ve 13 případech se jednalo o absenci při vyplňování teacher made testů, ve 4 případech při vyplňování didaktických testů). Za značný limit této studie shledáváme konstruktovou validitu didaktických testů (nebyl splněn doporučovaný poměr položek na vyšší a nižší kognitivní operace). Příčiny této situace byly diskutovány výše. 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. Rovněž způsob konstrukce teacher made testů (nejednotné skórování úloh) znemožňuje ověření normality a případné použití rigoróznějších parametrických testů. Zároveň si uvědomujeme, že došlo k porovnávání známek napříč dvěma předměty (vlastivěda, přírodověda), což může vyvolávat otázku ve vztahu k parametrům testů pedagožky v jiných vyučovacích předmětech.

9 Závěr

Trčková ve své práci (2013) uvádí, že 97 % učitelů používá didaktické testy. Můžeme si pokládat otázku, co si přesně učitel představuje pod pojmem *didaktický test*. Autorka dále zmiňuje, že pouhých 7 % učitelů si dle jejího výzkumu tvoří didaktický test sami, ostatní čerpají z učebnic nebo jiné literatury. Nabízí se otázka o kvalitě „testů“ (spíše úloh nebo otázek?) uvedených v učebnicích (ve smyslu jejich reliability, validity a objektivity). Zároveň je důležité podotknout, že je stále nedostatečné množství zdrojů, které by poukazovaly na jednoznačná a sjednocující pravidla pro psaní jednotlivých položek v textu, stále se tak jedná o určitý proces, ve kterém se pedagogové řídí spíše svým osobním pocitem (Millman & Greene, 1993; Haladyna et al., 2002). Přestože využívání teacher made testů lze považovat za jeden z efektivních nástrojů pro zajišťování zpětné vazby od žáků, je stále nutné myslet na skutečnost, že ne vždy disponují dostatečnou mírou reliablity a validity a zároveň že ne každý pedagog má dostatečné dovednosti k jejich správné evaluaci (srov. Mpofu, 2011; Hartell & Strimel, 2019). Učitelé dávají nejčastěji do popředí jednoduché úlohy na pamětní operace (Germ & Harms, 2008). Janík a Stuchlíková (2010) poukazují na fakt, že takto koncipované úlohy snižují zájem žáků například o přírodovědné obory. Suchoradský (2008)

považuje testy jako jednoduchou a zejména pak spravedlivou formu zkoušky. Oceňuje objektivní a rychlý způsob, jak prověřit schopnosti žáka, ale zdůrazňuje, že by se nemělo jednat o jediný způsob hodnocení. K tomuto stanovisku se připojujeme i my. Je potřeba hledat rozmanité formy a způsoby hodnocení, didaktické testy představují pouze jednu z využitelných možností.

Nebylo cílem zde didaktický test představovat jako „ideální či optimální" způsob evaluace žákovské činnosti. Jsme si vědomi, že vše má svá pozitiva a negativa, což samozřejmě platí i pro didaktický test. Za silné stránky se považují stejné podmínky pro všechny žáky při testu, reliabilita, validita, objektivita nebo způsob vyhodnocení (praktičnost). Za slabé stránky se naopak shledává časová náročnost při tvorbě, absence mluveného projevu žáka (nemožnost si tak ověřit určité zvládnuté kompetence mluveného projevu) nebo ztráta možnosti interakce s učitelem, která může navést žáka správným směrem při zodpovězení otázek (Hališka, 1999). Záměrem této studie v žádném případě nebylo ani poukazovat na potenciální nedostatky v oblasti kompetencí pedagogů při tvorbě didaktických testů, avšak z našeho úhlu pohledu si tato oblast zaslouží zvláštní pozornost. Předkládaná deskriptivní případová studie měla za cíl poukázat na kvalitu vyhotovených testů jedné pedagožky ve dvou vyučovacích předmětech. 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). Závěrem si dovolíme položit dvě otázky: Vyplývají známky na vysvědčeních žáků ze spolehlivých, platných a objektivních nástrojů měření a je vůbec možné je poměřovat? Je možné uvažovat o budoucí hypotetické možnosti vyvážení autonomie učitelů při tvorbě testových materiálů s organizovaným (centrálním) modelem testování vycházejícího z více rigorózního přístupu při tvorbě testových materiálů?

Poděkování

Tento příspěvek vznikl za podpory grantové agentury UJEP-SGS-2020-43-008-2.

Literatura

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, abridged edition.* Pearson PLC.

Byčkovský, P. (1982). *Základy měření výsledku výuky.* Tvorba didaktického testu, ČVUT VÚIS.

Cizek, G. J. (2004). Achievement tests. *Encyclopedia of Applied Psychology*, *1*, 41–42.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Čapek, R. (2015). *Moderní didaktika: lexikon výukových a hodnoticích metod.* Grada.

Disman, M. (2005). *Jak se vyrábí sociologická znalost.* Karolinum.

Dutke, S., Barenberg, J., & Leopold, C. (2010). Learning from text: Knowing the test format enhanced metacognitive monitoring. *Metacognition and Learning*, *5*(2), 195–206. https://doi.org/10.1007/s11409-010-9057-1

Fayol, M., Alamargot, D., & Berninger, V. W. (2012). *Translation of thought to written text while composing: advancing theory, knowledge, research methods, tools, and applications.* Psychology Press/Taylor & Francis Group.

Fitzpatrick, C., Côté-Lussier, C., & Blair, C. (2016). Dressed and groomed for success in elementary school: Student appearance and academic adjustment. *Elementary school journal*, *117*(1), 30–45. https://doi.org/10.1086/687753

Flybvjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative inquiry*, *12*(2), 219–245. https://doi.org/10.1177/1077800405284363

Germ, M., & Harms, U. (2008). What do biology tests look like in German grammar schools? A descriptive study about task formats and teachers intentions for surveying different cognitive dimensions. In M. Hamman, M. Reiss, C. Boulter, & S. D. Tunnincliffe (Eds.), *Biology in Context: Learning and Teaching for the twenty-first century* (pp. 248–258). Institute of Education.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hališka, J. (1999). *Jak testy sestavit a pracovat s nimi.* Středisko služeb školám.

Hartell, E., & Strimel, G. J. (2019). What is it called and how does it work: examining content validity and item design of teacher-made tests. *International Journal of Technology and Design Education*, *29*(4), 781–802. https://doi.org/10.1007/s10798-018-9463-2

Hawkes, H. E., Lindquist, E. F., & Mann, C. R. (1936). *Construction and use of achievement examinations.* Houghton Millin Company.

Hendl, J. (2012). *Přehled statistických metod.* Portál.

Hofer, S. I. (2015). Studying gender bias in physics grading: The role of teaching experience and country. *International Journal of Science Education*, *37*, 2879–2905. https://doi.org/10.1080/09500693.2015.1114190

Holeček, V. (2014). *Psychologie v učitelské praxi.* Grada.

Chlup, O. (1931). *O školu měšťanskou.* Nové školy. Knihovny nových škol.

Chráska, M. (1999). *Didaktické testy: příručka pro učitele a studenty učitelství.* Paido.

Chráska, M. (2002). *Didaktické testy ve školní praxi. Brno.*

Chráska, M. (2007). *Metody pedagogického výzkumu.* Grada.

Chráska, M (2016). *Metody pedagogického výzkumu.* Grada.

Janík, T., & Stuchlíková, I. (2010). Oborové didaktiky na vzestupu: přehled aktuálních vývojových tendencí. *Scientia in educatione*, *1*, 5–32.

Jeřábek, O., & Bílek, M. (2010). *Teorie a praxe tvorby didaktických testů* [online] [cit. 10. 5. 2020]. UPOL. Dostupné z http://zvyp.upol.cz/publikace/bilek_jerabek.pdf.

Johanson, R. (2003). Case study methodology. *Acta Linguistica Hungarica – ACTA LINGUIST HUNG, 32*, 22–24.

Johnson, R., & Penny, J. (2005). Split-Half Reliability. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 649–654). Elsveir.

Junková, J. (2006). *Didaktické testování* [online] [cit. 17. 5. 2020]. Dostupné z https://is.muni.cz/el/1441/podzim2009/ZS1BK_PDD/didakticke_testovani.pdf.

Kalhous, Z., & Obst, O. (2002). *Školní didaktika.* Portál.

Komenda, S., & Mazuchová, J. (1995). *Tvorba a testování testu.* UP.

Laufková, V., & Starý, K. (2016). *Formativní hodnocení ve výuce.* Portál.

Lazarová, B., & Pol, M. (2002). *Multikulturalita a rovné příležitosti v české škole.* Institut pedagogicko-psychologického poradenství ČR.

Linn, R. L. (2008). *Measurement and assessment in teaching.* Pearson Education India.

Man, F., Mareš, J., & Stuchlíková, I. (2000). Paradoxní účinky učitelových motivačních postupů. *Pedagogika*, *50*(3), 224–235.

Mareš, J. (2015). Tvorba případových studií pro výzkumné účely. *Pedagogika*, *65*(2), 113–142.

Marinova, V., Tsvetkov, D., & Hristov, L. (2005). On the reliability of didactic tests. *Pedagogical Almanac*, *13*(1), 242–248.

Millman, J., & Greene, J. (1993). The specification and development of tests of achiievement and ability. In R. L. Lin (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 335–366). Macmillan Publishing Co, Inc; American Council on Education.

Mpofu, B. (2011). Formative evaluation versus summative evaluation. Longman.

Popham, W. J. (2017). *The ABCs of educational testing: demystifying the tools that shape our schools.* Corwin.

Pulpán, Z. (1991). *Základy sestavování a klasického vyhodnocování didaktických testů.* Kotva.

Rámcový vzdělávací program pro základní vzdělávání. (2017). MŠMT. [on-line] [cit 5. 3. 2020]. Dostupné z: https://www.msmt.cz/file/41216/

Sedláčková, J. (1993). *Diagnostické metody ve vyučování matematice.* PřF UPOL.

Sekaran, U. (1992). *Research methods for business: A skill building approach.* 2nd ed. Wiley.

Schindler, R. (2006). *Rukověť autora testových úloh.* Centrum pro zjišťování výsledků vzdělávání.

Skutil, M. (2011). *Základy pedagogicko-psychologického výzkumu pro studenty učitelství.* Portál.

Smékal, V., Švec, V., & Zajac, J. (1973). *Didaktické testy a jejich vyhodnocování.* Středisko pro výzkum učebních metod a prostředků.

Suchoradský, O. (2008). *Testy a jejich užití při hodnocení žáků.* Metodický portál [on-line] [cit. 11. 3. 2020]. Dostupné z https://clanky.rvp.cz/clanek/c/Z/2666/TESTY-A-JEJICH-UZITI-PRI-HODNOCENI-ZAKU.html.

Šatánek, A., & Hubálovská, H. (1972). Příspěvek k hodnocení výkonů v testových zkouškách. *Pedagogika*, *2*, 185–189.

Škoda, J., & Doulík, P. (2007). *Tvorba a hodnocení didaktických testů: cvičebnice pro studenty učitelství a účastníky kurzu DPS.* PF UJEP.

Škoda, J., Doulík, P., & Hajerová-Müllerová, L. (2006). *Zásady správné tvorby, použití a hodnocení didaktických testů v přípravě budoucích učitelů* [on-line] [cit. 2. 2. 2020]. Dostupné z http://cvicebnice.ujep.cz/cvicebnice/FRVS1973F5d/

Švamberk Šauerová, M. (2016). *Hyperaktivita nebo hypoaktivita – výchovný problém.* Wolters Kluwer.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tournaki, N. (2003). Effect of student characteristics on teachers' predictions of student success. *Journal of Educational Research*, *96*(5), 310–319. https://doi.org/10.1080/00220670309597643

Trčková, V. (2013). *Didaktické testování jako profesní kompetence učitele matematiky na základní škole* (Diplomová práce). [online] [cit. 9. 3. 2020]. Dostupné z https://theses.cz/id/03mwq9/.

Urbánek, T. (2002). *Základy psychometriky.* Masarykova univerzita.

Vrána, S. (1948). *Zkoušení a známkování.* Komenium, učitelské nakladatelství.

Ward, A., Stoker, H. W., & Murray-Ward, M. (1996). *Educational measurement: Theories and applications* (Vol. 2). University Press of America.