Ověřování a optimalizace didaktického testu

Jak již bylo řečeno každá testová úloha je pro kvalitní test důležitá, a proto budeme prověřovat základní vlastnosti každé úlohy. Těmito základními vlastnostmi rozumíme: obtížnost úlohy, citlivost úlohy a zřetel také klademe na tzv. nenormované odpovědi (viz níže). Po provedeném rozboru testových úloh, který nazýváme položkovou analýzou, vyloučíme úlohy, které mají nevhodné vlastnosti a tím vlastně náš test optimalizujeme pro další použití.

Obtížnost úlohy.

Prvním důležitým kritériem pro zjišťování vhodnosti testové úlohy je její obtížnost, tj. kolik studentů úlohu umí správně vyřešit. M. Chráska rozlišuje v [ ] hodnotu obtížnosti Q a index obtížnosti P. Hodnota obtížnosti udává procento studentů ve vzorku, kteří danou úlohu zodpověděli nesprávně anebo ji vynechali.

$$Q = 100\,\frac{n_n}{n}\ ,$$

kde Q je hodnota obtížnosti, n_n je počet studentů ve skupině, kteří odpověděli nesprávně a nebo neodpověděli a n je celkový počet studentů ve vzorku.

Index obtížnosti je procento žáků ve skupině, kteří danou úlohu zodpověděli správně.

$$P = 100\frac{n_s}{n},$$

kde P je index obtížnosti, n_s počet studentů, kteří odpověděli správně a n je celkový počet studentů ve vzorku. Zřejmě platí následující vztah mezi hodnotou obtížnosti a indexem obtížnosti testové úlohy:

$$Q = 100 - P$$

U NR – testů , což je náš případ se úlohy s $P < 20$ a $P > 80$ považují za extrémně obtížné, resp. extrémně snadné, a do testů se bez zvláštních důvodů nezařazují. Vykazuje-li některá z hodnot extrémní hodnotu P, má význam zkoumat, zda extrémní hodnota není způsobena technickými nedostatky, které mohly uniknout kompetentům úlohy. Je-li tomu tak musíme rozhodnout, zda daná úloha bude vyřazena nebo upravena při dalším použití.

Citlivost úlohy.

Dalším důležitým ukazatelem pro rozlišení (určení) kvality testové úlohy je ukazatel citlivosti úlohy. Je totiž legitimní požadovat u každé úlohy, aby každou z úloh testu řešilo správně více „lepších“ než „horších“ studentů. Za tzv. „lepší“ studenty se považují studenti, kteří v testu dosáhli v celém testu většího úspěchu. Citlivost úlohy tedy vyjadřuje, jak dalece daná úlohy zvýhodňuje studenty, mající lepší vědomosti, před studenty, kteří mají vědomosti horší. Jak se postupuje při stanovení koeficientu citlivosti? Nejdříve se vzorek studentů rozdělí na dvě části podle dosaženého hrubého skoré v testu. Horní polovinu označíme jako „lepší“ (L), spodní polovinu jako „horší“ (H). V případě, že máme dostatečně velký vzorek studentů zvýšíme účinnost tohoto kritéria vytvořením kontrastnějších skupin pro srovnávání. Skupinu (L) může tvořit např. pouze 33 % nejúspěšnějších studentů v testu, skupinu (H) 33 %

nejhorších studentů v testu. V zásadě však není příliš velkého rozdílu ve výsledcích, použije-li se jakýchkoliv kontrastních skupin s rozsahem mezi 20 – 33 % testovaných..

Citlivost úlohy se dá exaktně posoudit pomocí výpočtu některého z koeficientů, kterých je celá řada. Nejznámější z nich je tzv. koeficient ULI (upper-lower-index) zavedený A.P. Johnsonem, dále se často používají tetrachorický koeficient citlivosti a bodově biseriální koeficient citlivosti. Všechny tyto koeficienty mohou nabývat hodnot od -1 do $+1$, přičemž platí, že čím vyšší hodnotu koeficient má, tím lépe úloha rozlišuje mezi studenty s lepšími vědomostmi a studenty s horšími vědomostmi. Např. v případě, že koeficient dosáhne hodnoty 0 (nebo blízké 0) znamená to, že úloha vůbec nerozlišuje mezi oběma skupinami studentů. Z tohoto úhlu pohledu se tato úloha jeví pro test jako nevhodná. Stejný názor zastáváme i u úloh , v nichž koeficient citlivosti dosahuje záporných hodnot.

Koeficient citlivosti označujeme písmenem *d* (z lat. discriminare = rozlišovat). Protože se v praxi ukázalo se, že nejčastěji používané ukazatele citlivosti mají zhruba stejnou účinnost, budeme se zabývat pouze nejjednodušším z nich – koeficientem ULI. Např. v [ ] je uvedena metoda jeho výpočtu.

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

kde d je koeficient citlivosti ULI, n_L je počet studentů z horní poloviny, kteří vyřešili úlohu správně, n_H je počet studentů z dolní poloviny, kteří vyřešili úlohu správně, N je celkový počet studentů. Pro vhodnost zařazení úloh do testu se požaduje, aby koeficient ULI u úloh s hodnotou obtížnosti $30 – 70$ bylo d alespoň $0,25$ a úloh s hodnotou obtížnost $20 – 30$ a $70 – 80$ alespoň 0,15.

V případě, že použijeme „ostřejších“ kontrastních skupin musíme vztah modifikovat.

$$d = \frac{n_L - n_H}{f \cdot N},$$

kde f … poměr četnosti kontrastní skupiny k četnosti testovaných (při 30 % kontrastní skupině je $f = 0,3$),

n_1', n_1'' … počet správných odpovědí v horní, resp.v dolní, kontrastní skupině,

Analýza nenormovaných odpovědí

Kromě zjišťování obtížnosti testových úloh a citlivosti testových úloh se v rámci analýzy vlastností úloh provádí také analýzy tzv. **nenormovaných odpovědí,** tj. rozbor nesprávných a vynechaných odpovědí.

Rozbor vynechaných odpovědí

Jestliže zjistíme, že některé odpovědi jsou vynechány, může to znamenat vedle neznalostí učiva také neporozumění formulaci úlohy, nedostatek času k vypracování odpovědi atd. V literatuře se uvádí, že je třeba věnovat zvýšenou pozornost úlohám, které vynechalo

a) 30 – 40 % žáků u otevřených úloh,
b) více než 20 % žáků u uzavřených úloh (v případě, že jsou penalizovány nesprávné odpovědi).

Rozbor nesprávných odpovědí

Rozbor nesprávných odpovědí je velmi jednoduchý u **úloh s výběrem odpovědi**. V tomto případě postačí překontrolovat, zda všechny nabídnuté distraktory jsou pro žáky stejně atraktivní. Zda jsou všechny dostatečně atraktivní poznáme podle toho jak si žáci jednotlivé distraktory vybírají. Ten distraktor, který si nikdo (nebo téměř nikdo) nevybírá, neplní svoji funkci a měl by být nahrazen jiným pokud možno atraktivnějším distraktorem. Při správně fungujících distraktorech by měl žák, který správnou odpověď nezná, měl více méně náhodně volit jednu z nabídek. Nalezení dostatečného počtu vhodných distraktorů jedním z nejobtížnějších problémů při konstrukci úloh s výběrem odpovědi.

U **otevřených úloh** je rozbor nesprávných odpovědí poněkud obtížnější. Doporučuje se veškeré chyby žáků rozdělit do dvou kategorií, na tzv. **základní a vedlejší chyby.**

Základní chyby – jsou způsobeny skutečnou neznalostí učiva, jeho nepochopením nebo nezvládnutím,

Vedlejší chyby – jsou způsobené různými náhodnými vlivy, např. přehlédnutí, numerická chyba ve výpočtu, nepřesnosti, špatnou čitelností textu atd.

Jestliže převažují vedlejší chyby nad hlavními, může to znamenat, že v úloze závisí úspěch více na jiných (náhodných) okolnostech než na stupni zvládnutí učiva. Takovou úlohu je třeba jako nevyhovující z didaktického testu vyloučit. V dobré testové úloze by počet hlavních chyb měl být vždy větší než počet chyb vedlejších.

Vyhodnocování testových úloh a analýza nenormovaných odpovědí

Příklad z přijímaček na FaME

Úprava vytvořeného didaktického testu

Nevhodná testová úloha se vyznačuje následujícími vlastnostmi:

- Úloha je příliš obtížná nebo snadná (hodnota obtížnosti $Q > 80$ nebo $Q < 20$).
- Úloha málo rozlišuje mezi žáky s dobrými a špatnými vědomostmi (např. koeficient citlivosti d je u středně obtížných úloh menší než 0,25.
- V testovací úloze je příliš vynechaných odpovědí (u otevřených úloh např. více než 30 – 40 %, u uzavřených úloh více než 20 %).
- Počet vedlejších chyb převažuje nad počtem hlavních chyb (u úloh otevřených).
- Žáci nevybírají ze všech nabídnutých distraktorů v úloze (u úloh s výběrem odpovědi).

Vytvoření definitivní podoby didaktického testu

- Nevhodné nebo podezřelé úlohy je nutné z testu vyřadit a nahradit vhodnějšími.
- Navrhujeme více úloh, aby bylo z čeho vybrat.
- Provedeme korekci úlohy (kdy, jak?).
- V případě, že test obsahuje různé typy úloh doporučuje se úlohy stejného typu soustředit do jedné části testu.
- Úlohy řadíme do testu podle vzrůstající obtížnosti.

Vytvoření ekvivalentních forem

Kdy a jak?

Standardizované testy

Kvalitně zpracovaný nestandardizovaný test může být pro učitele cenným zdrojem informací o průběhu a výsledcích výuky.
Nestandardizované testy – dosažené výsledky v testu nelze srovnávat s ostatními žáky, nemáme k dispozici klasifikační nebo jiný standard, na jehož základě bychom mohli objektivně hodnotit a klasifikovat. Většinou nevíme, jak dalece jsou výsledky dotčeny náhodou, jak dalece jsou spolehlivé a přesné, tj. reliabilní.

Určení reliability didaktického testu

Existují různé postupy pro určení tzv. koeficientu reliability

Reliabilita testu pomocí Kuderova – Richardsonava vzorce

Tento model výpočtu je vhodný pro didaktické testy úrovně, které jsou složeny z obsahově homogenních úloh. Výpočet koeficientu reliability se provádí pomocí Kuderova Richardsonova vzorce

$$r_{rk} = \frac{k}{k-1}\left(1 - \frac{\sum pq}{s^2}\right), \text{ kde}$$

k je počet úloh v testu,
p je podíl žáků ve vzorku, kteří řešili určitou úlohu v testu správně,
$q = 1 - p$ a
s je směrodatná odchylka pro celkové výsledky žáků v testu.

Reliabilita testu metodou půlení

Podmínky:

- sudý počet úloh,
- úlohy jsou řazeny dle vzrůstající obtížnosti.

Test se rozdělí na dvě poloviny tím způsobem, že jednu polovinu tvoří úlohy s lichým pořadovým číslem, druhou polovinu úlohy se sudým pořadovým číslem. Výsledky dosažené v testu se vzájemně korelují.

Výpočet se provádí dle Spearmanova – Brownova vzorce

$$r_{sb} = \frac{2 \cdot r_p}{1 + r_p}, \text{ kde}$$

r_{sb} je koeficient reliability a
r_p je koeficient korelace mezi výsledky žáků v obou polovinách didaktického testu.

Standardizace

Hodnota každé zkoušky bude jistě vyšší, když zkouška bude standardizována. Standardizace je chápána ve dvou směrech:

- rámcový obsah zkoušky se nemění, zůstane každý rok stejný, jestliže se nezmění osnovy příslušného předmětu ,
- rozumíme ji shromáždění a zpracování testových výsledků do testových standardů, umožňujících vyjádřit výkon testovaného ve vztahu k výkonu populace, pro kterou je test určen.

Smyslem standardizace je vytvoření testového standardu (testové normy), který umožní zařadit studenta podle dosaženého počtu bodů do určitého žebříčku (stupnice, škály). Jednodušší metody standardizace, ke kterým směřujeme jsou na zjišťování procent studentů, kteří v reprezentativním vzorku dosáhli určitého výsledku. Popíšeme jednoduchou metodu, ve které se standardizace provádí pomocí percentilů a kterou jsme použili i v přijímacích testech. Metoda se nazývá Percentilová škála je uvedena např. v [ ]. U této metody se ke každému dosaženému počtu bodů (hrubému skóre) přiřadí tzv. percentilové pořadí, které udává kolik procent studentů ve vzorku dosáhlo horšího výkonu. To umožňuje posoudit, jaké je relativní postavení studenta ve skupině.

Percentilové pořadí pro určitý výsledek v testu lze vypočítat podle vzorce

$$PR = 100 \cdot \frac{n_k - \frac{n_i}{2}}{n} \text{ , kde}$$

PR je percentilové pořadí studenta pro daný výsledek v testu,

n_k je kumulativní četnost u daného výsledku,

n_i je četnost daného výsledku a

n je počet testovaných studentů.

Kumulativní četnost je četnost určitého výsledku v tabulce Tab. 2 a četnost všech slabších výsledků dohromady. Vypočítáme pecentilovou normu pro test Jednoduché stroje. Test byl zadán vzorku 339 žáků. Výsledky jsou uvedeny v tabulce.

Konstrukce percentilové normy pro test Jednoduché stroje

Počet respondentů 339

Počet bodů	Četnost	Kumulativní četnost	Percentilové pořadí
0	8	8	1
1	18	26	5
2	33	59	13
3	44	103	24
4	54	157	38
5	57	214	55
6	50	264	71
7	31	295	82
8	20	315	90
9	16	331	95
10	8	339	99

Používání didaktických testů ve školní praxi

Z výsledků didaktického testu by měl učitel získat co nejvíce informací k hodnocení žáků, ale také k optimálnímu řízení svého dalšího pedagogického procesu. Co tedy lze didaktickým testem zjistit a jak budeme interpretovat získané výsledky.

Diagnostický rozbor výsledků testu

Diagnostický rozbor výsledků by měl následovat prakticky po každém použití didaktického testu. Při tomto rozboru si učitel všímá zejména chyb, kterých se žáci dopustili, a hledá jejich pravděpodobné příčiny. Forma rozboru výsledků závisí na druhu použitého testu, zejména na druhu použitých testových úloh. Pokud byly v testu použity uzavřené testové úlohy, může mít rozbor podobu, kterou uvádí následující příklad.

Diagnostický rozbor výsledků u úloh s výběrem odpovědi

Žáci (10) sedmého ročníku základní školy vypracovali didaktický test Mechanické vlastnosti kapalin a plynů. Tento test obsahoval celkem 10 uzavřených úloh s jednou správnou odpovědí ze 4 možných odpovědí a byl skórován tak, že za každou správnou odpověď získávali žáci vždy 1 bod. Jeden z vhodných způsobů analýzy je uveden v Tab. 3.

Výsledky žáků v testu Mechanické vlastnosti kapalin a plynů

Jméno žáka	Počet bodů	1 *D*	2 *C*	3 *B*	4 *A*	5 *A*	6 *D*	7 *B*	8 *B*	9 *C*	10 *A*
1 ……	10	/	/	/	/	/	/	/	/	/	/
2 ……	10	/	/	/	/	/	/	/	/	/	/
3 ……	9	/	/	*A*	/	/	/	/	/	/	/
4 ……	8	/	*D*	/	/	*B*	/	/	/	/	/
5 ……	7	/	/	*C*	/	/	*A*	/	/	/	*B*
6 ……	6	/	/	*A*	*D*	*C*	/	*C*	/	/	/
7 ……	5	/	*A*	*A*	*D*	/	/	*D*	*A*	/	/
8 ……	5	/	/	*A*	/	/	*B*	/	*D*	*B*	*C*
9 ……	5	/	*A*	*A*	*D*	/	/	*D*	*A*	/	/
10 …….	3	/	*B*	*B*	/	*C*	*C*	*C*	*C*	*A*	/
Celkem	68	10	6	3	7	7	7	6	6	8	8
P (%)	100	100	60	30	70	70	70	60	60	80	80

Tab. 3

Posouzení celkových výsledků třídy

Dosažené výsledky třídy, případně školy, se obyčejně posuzují podle průměrného počtu dosažených bodů. Aritmetický průměr počítáme u výsledků didaktických testů nejvýhodněji podle vzorce

$$\bar{x} = \frac{1}{n}\sum n_i x_i \ , \quad \text{kde}$$

$\bar{x}$ je aritmetický průměr výsledků žáků v testu,

n je celkový počet testovaných žáků,

x_i jednotlivé dosažené počty bodů,

n_i počty žáků, kteří dosáhli výsledku x_i.

Dosažené výsledky testování je výhodné znázornit graficky, protože z grafického znázornění je možno získat také informace o rozložení výsledků ve třídě. Nejčastěji se používá tzv. **histogram četnosti.** Histogram četnosti je v podstatě sloupcový diagram, u

něhož se na vodorovnou osu nanáší dosažené výsledky testování (body) a na svislou osu počty žáků (četnosti).

Výpočet průměrného počtu bodů na žáka v testu

Počet bodů x_i	Četnost n_i	$x_i \cdot n_i$
10	2	20
9	1	9
8	1	8
7	1	7
6	1	6
5	3	15
4	0	0
3	1	3
Celkem	10	68

Tab. 4

Výpočet průměru třídy

$$\bar{x} = \frac{68}{10} = 6,8 .$$

Grafické znázornění výsledků testování

Výhodné je znázornění výsledků testu z Tab. 3 pomocí histogramu četnosti.

Histogram četnosti pro školní třídu

Klasifikace výsledků testu

Mnoho nejasností je mezi učiteli v otázce převodu bodového hodnocení na klasifikační stupně. Je to vcelku pochopitelné, protože ani v teorii nebyl tento problém dosud spolehlivě a jednoznačně rozřešen. Uveďme některé používané přístupy

a) Intuitivní přístup ke klasifikaci

Někteří učitelé přistupují k převodu bodových výsledků na klasifikační stupně zcela subjektivně a sami víceméně určují, kolik bodů je potřeba na dosažení určité známky. Pokud se jedná o učitele s velkou pedagogickou a odbornou zkušeností, většinou je jejich hodnocení odpovídající.

Někteří odborníci doporučují jako optimální řešení této otázky tzv. normativní přiřazování klasifikačních stupňů na základě posudku skupiny odborníků. Z jednotlivých posudků můžeme určit průměr.

b) Klasifikace na základě procenta správných odpovědí

Někdy se při převodu bodových výsledků na klasifikační stupně vychází z **procenta správných odpovědí,** kterého žák v testu dosáhl.V Tab. 5 jsou návrhy klasifikací podle procenta správných odpovědí.

Klasifikace podle procenta správných odpovědí

Procento správně vyřešených úloh v testu			Klasifikační stupeň
Klasifikace běžná	Klasifikace přísná	Klasifikace velmi přísná	
91 – 100	96 – 100	95 – 100	1
81 – 90	88 – 95	90 – 94	2
71 – 80	82 – 87	85 – 89	3
61 – 70	70 – 81	80 – 84	4
0 – 60	0 – 69	0 – 79	5

Tab. 5

Klasifikace podle procenta správných odpovědí
(Ostravský)

Procento správně vyřešených úloh v testu	Klasifikační stupeň (2004/2005)
85 – 100	A (11)
76 – 84	B (19)
68 – 75	C (31)
59 – 67	D (32)
50 – 58	E (35)
0 – 49	F (24)

c) Klasifikace na základě normálního rozdělení

Bodové výsledky žáků v testu můžeme klasifikovat na základě **normálního rozdělení četnosti.** Předpokládá se, že výkony dosažené v testu odpovídají tzv. Gaussově křivce (velké soubory).

Posouzení objektivity klasifikace pomocí didaktického testu

Zdrojem důležitých a zajímavých informací se může stát výpočet korelace korelace mezi výsledky žáků v testu a mezi klasifikací v daném předmětu. Jestliže např. máme k dispozici dobrý a ověřený didaktický test (validní a reliabilní), který zkouší učivo, které bylo předmětem klasifikace, můžeme výsledků testování využít k posouzení míry objektivity provedené klasifikace. Na začátku školního roku můžeme např. zadat žákům test, který zkouší úroveň vědomostí a dovedností, které si žáci měli osvojit v předcházejícím roce. Pokud půjde o skutečně validní test, můžeme na základě srovnání výsledků testování s klasifikací na konci minulého roku posoudit, jak dalece klasifikace odpovídá skutečným vědomostem žáků, tzn. jak dalece byla provedena objektivně.

Výsledky žáků v testu je možné porovnat s výsledky klasifikace ne základě výpočtu Spearmanova koeficientu pořadové korelace. Tento koeficient se vypočítá podle vzorce

$$r_s = 1 - \frac{6 \cdot \sum d^2}{n\left(n^2 - 1\right)} \quad , \text{ kde}$$

r_s je Spearmanův koeficient pořadové korelace,

n je počet sledovaných žáků,

d je rozdíl mezi pořadím žáka ve třídě podle klasifikace a pořadí tohoto žáka podle výsledku v testu.

Spearmanův koeficient pořadové korelace vypovídá o těsnosti vztahu mezi klasifikací a výsledky testu. Teoreticky může tento koeficient nabývat hodnot od -1 přes 0 do + 1.

2.12.Plánování didaktického testu

Aby konstruktér testu navrhl vyvážený test je nutné, aby při jeho plánování postupoval systematicky. Etapa plánování je velmi důležitá a v přípravě dobrého didaktického testu se nedá vynechat. V této etapě si musí konstruktér k jakému účelu bude test sloužit a jaký bude jeho pravděpodobný rámcový obsah. Na základě těchto znalostí konstruktér testu zpracuje testové specifikace. V nich se uvádějí technické údaje o testu jako je počet úloh, druh testových úloh, čas vymezený k testování apod.. Hlavní částí testové specifikace je samozřejmě vymezení (upřesnění) obsahu, jež slouží jako základ návrhu testových položek (úloh). Konkrétně se touto problematikou budeme zabývat v kapitole .

2.13. Praktičnost testu

Praktičností testu se rozumí především snadnost zadávání, skórování a interpretace výsledků. Pochopitelně, že by bylo vhodné při větším počtu respondentů provádět skórovaní testu strojově. Výsledky jsou okamžitě k dispozici v takové formě, jak je požadované. Dále při strojovém skórování je chyba způsobená fyzickou osobou ať již vědomě či nevědomě prakticky vyloučena. Nezanedbatelnou výhodou strojového zpracování je možnost získat celou řadu dalších zajímavých výsledků. Je zřejmé, že při takto pojatém skórování má zkouška vysokou míru objektivity. Využívání strojového skórování u přijímacích testů výrazně zvyšuje objektivitu hodnocení. Podrobně se si vše ukážeme v kapitole .