ŠKOLNÍ MĚŘENÍ
A
EVALUACE VÝSLEDKŮ VZDĚLÁVÁNÍ VE ŠKOLE

Studijní materiál pro interní potřebu učitelů základních a středních škol

PhDr. Antonín Mezera

ÚVOD

Ve školní praxi se velmi často objevuje a řeší otázka **školní výkonnosti a školní úspěšnosti žáků**, stejně jako i otázka jejich interindividuálních rozdílů, které výrazným způsobem ovlivňují školní výsledky.

Školní výkonnost je do jisté míry užší pojem, který označuje podstatnou složku školní úspěšnosti, projevující se v úrovni víceméně objektivně měřitelných školních výkonů.

Školní úspěšnost žáka potom v převážné míře vyjadřuje úroveň sociálního hodnocení, které reflektuje skutečnost, jak konkrétní činnost žáka odpovídá požadavkům školy, zatímco **školní zdatnost** lze chápat jako komplex dispozic, které tuto úspěšnost podmiňují. Takto vysvětluje Hrabal (1989) tři základní pojmy, které nejsou v pedagogické praxi příliš rozlišovány.

Jakou roli a funkci vlastně sehrává známka v životě žáka a jeho rodiče? Školní známka, která ve své podstatě neodráží výkon školáka, ale spíše hodnocení tohoto výkonu učitelem (jedná se v jistém slova smyslu o subjektivní formu evaluace žáka učitelem, má v životě školáka své nesporné motivační momenty, které mají své psychologické (pro žáka) a sociální (pro celou rodinu) konsekvence. Aktu hodnocení a klasifikace známkou je již delší dobu zcela oprávněně přisuzována vysoká míra subjektivity (haló-efekt, figura-pozadí, generalizace, ”soukromé teorie osobnosti”, tendence k průměrnému hodnocení žáka učitelem, neochota měnit známku u žáka aj.). Od počátku 20.století se hledal nástroj, který by byl schopen poskytnout poněkud odlišnou diagnostickou informaci než je samotný prospěch a tímto nástrojem se stal **didaktický test,** který vznikal na obdobných metodologických základech jako *psychologický test*.

TEORIE A PEDAGOGICKÁ EVALUACE

Pedagogický výzkum v oblasti školního měření a evaluace výsledků vzdělávání je ve své podstatě na Západě dodnes výrazně psychologizován. Projevuje se to tím, že v něm nenacházíme tu izolovanost pedagogiky od psychologie, která byla velmi charakteristická pro československý výzkum (Průcha, 1992). Pro západní aplikovaný výzkum je naopak příznačné to, že hypotézy, metodologické postupy a teoretické explanace jsou většinou zakládány na psychologických teoriích a přístupech. Příkladem tu může být výzkumná oblast označovaná jako ”učení z textu”, v níž se propojují didaktické přístupy a metody s psychologickými teoriemi učen a teoriemi poznávacích procesů apod. Jiným překladem propojení pedagogiky s psychologií je právě nejen nepře formování interdisciplinární ”vědy o učení a vyučování” (instructional science), ale zejména oblast školního testování a měření školního výkonu žáka didaktickými testy. Jejich přínos je především spatřován v tom, že:

- jsou oproštěny od vlivu individuálních zvláštností učitelova hodnocení a měří relativně ”čistý” projev žákových dispozic, znalostí nebo dovedností,
- jsou ovlivněny odlišným způsobem reprodukce znalostí, než je ústní zkoušení,
- údaje jsou získávány ve standardní situaci pro všechny žáky a na vyšší či nižší úrovni statistického zpracování je standardizováno i hodnocení výsledků většiny didaktických testů.

Současně je třeba si uvědomit, že pojmy hodnocení a klasifikace nejsou zcela ekvivalentními pojmy, i když se v běžné školní praxi velmi často ztotožňují. **Hodnocení** je proces, kterým se hodnotí určitá činnost (akt, produkt) za použití verbálních i neverbálních prostředků a nikoliv pouze známkou. Jde vlastně o daleko širší pojem než je samotná klasifikace. **Klasifikace** (známkování) je pak výsledkem tohoto procesu hodnocení. V běžné školní praxi tedy figuruje proces hodnocení jako rozšířený soud učitele, kterým komentuje známku, vysvětluje její hodnotu, vede žáka k poznání toho, co již umí a co ještě ne. Tento poslední fakt je velmi důležitý z hlediska diagnostického, protože umožňuje učiteli i žákovi hledat efektivní cesty v procesu dalšího vyučování a učení (Dittrich, 1993).

Významnou charakteristikou didaktických testů (Schultests, Achievement tests) je *orientace na objektivní zjišťování úrovně zvládnutí obsahu učiva v jednotlivých vyučovacích předmětech.* V současné škole je to jeden z významných způsobů ověřování výsledků vzdělávání, který se od ostatních zkoušek liší především zvýšeným důrazem na objektivitu. Tím nabývá školní zkouška charakteru **měření** (measurement), založeného na užití řady statistických a metodologických procedur, které svým charakterem do jisté míry tvoří bariéru pro mnohé učitele zejména jazykových a společenskovědních oborů. Je nesporné, že kvantifikace je sice profilující charakteristikou školního testu, ale není však jeho cílovou vlastností. Gnoseologickou stejně jako metodologickou otázkou v mnoha směrech ovšem zůstává, zda školní měření didaktickými testy není současně testováním intelektových či kognitivních schopností. Rozdílnost v inteligenci a nerovnost ve vzdělávání byla sice před rokem 1989 záměrně obcházena, ale přesto řada teoretických prací(Alan, 1974) a empirických výzkumů poukázala na patrné rozdíly v mentálních schopnostech žáků jednotlivých typů škol. Např. Kolejková a Mazálková (In: Průcha, 1992) zjistily na vzorku populace 15letých žáků signifikantní rozdíly v intelektových schopnostech vázané na druh vzdělávání (učni vs. žáci gymnázia). V jiném šetření Švanda (1989, In: Průcha, 1992) prokázal značné rozdíly v IQ žáků v závislosti na druhu střední školy: Nejvyšší úroveň inteligence byla zjištěna u žáků gymnázia (IQ = 116 - 119), dále u žáků SOŠ (IQ = 88 - 94), nejnižší u žáků SOU (IQ = 88 - 94). Nepochybně tedy problém vztahu mezi intelektovými rozdíly a vzděláváním existuje i nás a při školním testování a evaluaci výsledků vzdělávání je nutné s tímto fenoménem počítat. Rozdíly v intelektových schopnostech ovšem velmi často nejsou ani tak odrazem intelektového deficitu žáků některých typů škol, jako spíše odrazem *rozdílných podmínek kognitivní socializace,* které lze dokumentovat signifikantně vyšším počtem rodičů s vysokoškolským vzděláním u žáků gymnázií ve srovnání s rodiči žáků SOŠ a SOU, kteří dosahují středoškolského nebo základního vzdělání. Je zjevné, že žáci SOŠ a SOU jsou ve srovnání se žáky gymnázií v nevýhodě, protože ve svém sociokulturním prostředí mají odlišné podmínky, které ovlivňují jejich motivaci. Obdobná nevýhoda je u těchto žáků spatřována při zacházení s formálními testy a formálními procedurami, protože ve svém sociokulturním prostředí se s nimi setkávají méně často než žáci navštěvující náročnější typy výběrových středních škol. Rozdíly mezi žáky jednotlivých typů škol, které jsou ve své podstatě výběrovými školami na různé úrovni náročnosti, je nesporně ovlivněn úrovní jejich kognitivních dovedností (cognitive skills), tj. schopností manipulovat se slovy a čísly, zpracovávat informace a vytvářet logické inference apod." Výběrové školy s různou mírou náročnosti jednotlivých studijních oborů působí v tomto smyslu selektivně tak, že tyto rozdíly ještě více akcentují, což se pak projevuje v rozdílných vzdělávacích a životních drahách jednotlivých žáků (Jencks, 1972). Například žáci gymnázií jsou nejen daleko více vybaveni, jak zacházet s formálním jazykem, s verbálně vyjádřenými logickými algoritmy a procedurami, ale v průběhu svého studia jsou v těchto kognitivních dovednostech podstatně

více ”trénováni” než žáci jiných středních odborných škol (např. SOŠ a SOU), u nichž tyto dovednosti nejsou hlavní součástí studijní orientace a dominantní náplní kurikulárního rámce vzdělávání, ale i školní komunikace (viz např. Jencksova teorie deficitu a Bernsteinova teorie jazykového deficitu).

Teorie a výzkum v oblasti *pedagogické evaluac*e (educational evaluation, assessment, measurement) a koncepce evaluace výsledků vzdělávání jsou na druhé straně ovlivňovány nejen těmito teoriemi, ale i svou roli v této oblasti aplikovaného výzkumu sehrává i Bloomova teorie *mastery learning.* Mastery learning je psychodidaktická teorie opírající se o předpoklad, že ovládnutí (mastery) nějaké souboru poznatků nebo dovedností je teoreticky možné u všech žáků, mají-li k tomu vhodné podmínky, a to *způsob vyučování* optimálně přizpůsobený subjektu a takové *množství času*, které žák k ovládnutí potřebuje (jinak také *teorie ”vhodných podmínek”* nebo *teorie ”učení směřující k ovládnutí”*). Bloomova psychodidaktická teorie je vyvozována z analýz učebních výkonů žáků v konkrétních vyučovacích předmětech a na tyto předměty je také aplikována. Bloomovy úvahy vycházejí ze skutečnosti, že žák současné školy je začleněn do vyučování zhruba po dobu 1200 hodin ročně. To znamená, že žák v průběhu svého vzdělávání absolvuje (včetně střední školy) přibližně 12 - 16 000 hodin. V průběhu této doby vznikají u žáků rozdílné výsledky v učebních výkonech. Na těchto diferencích v učebních výsledcích je pak založena selekce žáků určující přístup k vyššímu vzdělávání a ve svých důsledcích určující i značnou část jejich pozdější životní dráhy. Bloom se snaží svou teorií prokázat, že tyto diference mezi žáky nejsou podmíněny jejich rozdílnými intelektovými schopnostmi, ale mohou být odstraněny při takovém typu vyučování, jehož hlavním kritériem je to, že žáci nejsou omezováni v čase, v jakém potřebují dosáhnout osvojení určitého učiva. V tomto smyslu je nutné zásadně odlišit ”*individuální rozdíly v učení”* a ”*individuální rozdíly mezi žáky”*. Zatímco individuální rozdíly mezi žáky jsou ve své podstatě konstantní, objevují se individuální rozdíly v učení poměrně brzy (většinou ve 3.ročníku školy) a zvětšují se v průběhu školní docházky.

EVALUAČNÍ PEDAGOGICKÝ VÝZKUM A MĚŘENÍ VZDĚLÁVACÍCH VÝSLEDKŮ

Pedagogický evaluační výzkum zahrnuje řadu speciálních skupin:

- hodnocení vzdělávacích potřeb (needs assessment)
- hodnocení vzdělávací báze (base-line assessment)
- hodnocení vzdělávacích výsledků a výkonů (achievement evaluation, product evaluation)
- hodnocení efektivnosti výuky a škol (effectiveness evaluation)
- hodnocení učebních plánů, osnov a učebnic (curriculum evaluation, textbook evaluation)

  hodnocení pedagogických pracovníků a institucí (personel, institutional evaluation)

V oblasti evaluace výsledků vzdělávání jsou nejčastěji užívány:

- testy vědomostí (pokrývající učivo přesně vymezených vyučovacích předmětů) např. ve formě statisticko-normativních (norm-referenced) nebo kritériálních (criterion-referenced) testů či výkonových (achievement) testů nebo testů schopností (aptitude tests).
- dotazníky pro žáky (zjišťující ”příležitosti k učení”, tj. různé charakteristiky reálné výuky)
- dotazníky pro školy (zjišťující např. časové kvóty věnované na určité předměty, témata aj.)

Tyto testové a dotazníkové nástroje jsou konstruovány pro různé věkové skupiny žáků jednotně (viz národní srovnávací studie) nebo dochází k *testování na míru* (tailored testing), tj. na místo stejných úloh pro všechny testované žáky v určité věkové populaci bez rozdílu jsou testy a testové úlohy přizpůsobeny rozdílný schopnostním dispozicím žáků jednotlivých typů škol. Tradiční dotazníky, které většinou sledují nealternativní proměnné (pohlaví, příslušnost k etnické skupině, socioekonomický status, typ bydliště) a ”sociální pozadí” žáků a škol, nejsou dostatečně spolehlivým zdrojem informací o ”alternativních proměnných”, jako je charakter výuky (množství a kvalita výuky aj.). Proto jsou vyvíjeny dotazníky nového typu, které zjišťují tzv. příležitosti žáků k učení, zkušenosti žáků se zacházením s technickými prostředky a školními pomůckami apod.

Obdobně jsou vyvíjeny programy na měření *efektivnosti výuky* (effectiveness of teaching and learning) a *efektivnosti škol* (school effectiveness research), zaměřené zejména na: (1) přípravu a plánování výuky, (2) realizaci a řízení výuky ve třídách, (3) učební klima ve třídách, (4) stimulaci učení apod.

STANDARDIZOVANÉ A NESTANDARDIZOVANÉ TESTY

Dnešní praxe na školách vytváří určitý tlak na učitele, aby si sami konstruovali vlastní **nestandardizované didaktické testy** pro svou potřebu. Ty jim slouží především k objektivizaci hodnocení a k získání diagnostických podkladů pro jejich další práci při projektování výuky. Sem patří zejména didaktické testy vstupních vědomostí, které jsou učiteli užívány především na začátku školního roku ke zjišťování úrovně předchozích znalostí získaných v předchozích letech školní přípravy žáka. Testy ovšem zároveň slouží k získání diagnostických údajů pro žáky samé. Druhým obdobím velmi časté aplikace didaktických školních testů je většinou závěr studia určitého obsahového celku, který je zdrojem řady zpětnovazebních informací, z nichž profitují nejen žáci (při volbě další vzdělávací cesty), ale učitelé (při plánování výuky žáků nižších postupných ročníků).

Mezi základní vlastnosti didaktického školního testu nesporně patří:

- ***Objektivita***

  Vlastně každý examinátor nemůže bez porušení základních pravidel školního testování dojít při interpretaci výsledku didaktického testu k jiným výsledkům než jiný examinátor. Uvedené zásadě je vlastně podřízena celá konstrukce testu (formulace otázek, jasná pravidla zadání a hodnocení testu), přičemž tato neuzaujatost není vždy zaručena právě u klasifikace například písemné zkoušky několika učiteli.

- ***Validita (platnost)***

  Didaktický test totiž měří ve většině případů jen to, k čemu byl konstruován. Požadavky na osvojení např. vědomostí a dovedností v českém jazyce nebo matematice jsou ovšem daleko širší než je kterýkoli test schopen splnit, proto jej nelze například použít jako podklad pro klasifikaci žáka v závěru školního roku. V tomto smyslu je jakýkoli didaktický test zjevně nevalidní a ve své podstatě žádný test v tomto pojetí nedosahuje takto očekávanou absolutní validitu. Ve školní praxi se jedná vlastně pouze o určitý stupeň validity, který se ovšem velmi často mění v závislosti na pedagogickém kontextu, v němž byl didaktický test použit.

- *R**ealiabilita (spolehlivost)***

  Spolehlivý didaktický test splňuje ve většině případů poměrně náročné požadavky na přesnost a spolehlivost, které jsou u tohoto měřícího nástroje základní podmínkou jeho užití ve školní praxi. Čím je test reliabilnější, tím je výsledek méně zatížen náhodnými a situačně podmíněnými vlivy. Reliabilita je ovšem ovlivněna nejen testem samotným, ale i momentálním fyzickým a psychickým stavem žáka, stejně jako i skutečností, do jaké míry je žák seznámen s tímto typem do jisté míry modelových situací. V odborné literatuře se uvádí spodní hranice diagnostického využití testu u jednotlivce hodnota reliability 0,7, dobrý standardizovaný test by měl mít reliabilitu nad 0,85, nejlépe však blíží-li se hodnotě 0,95.

- ***Ekonomičnost***

  Efektivita školních testů je často na první pohled zřejmá, protože didaktický test je schopen učiteli v poměrně krátkém čase poskytnout daleko více informací, než individuální zkoušení jednotlivých žáků.

Účely využití výsledků školního testování jsou často velmi odlišné a mohou sloužit k:

- zařazení žáka do školského systému nebo do speciální či výběrové školy(vnější diferenciace)
- homogennímu seskupování žáků uvnitř škol (vnitřní diferenciace)
- pedagogické diagnostice (identifikaci talentovaných nebo naopak dificilních žáků, nápravě, poradenství a poradenské intervenci
- hodnocení žáka, zvýšení jeho motivace, je-li testu užíváno v průběhu vyučování (partie, tématu)
- hodnocení osnov a učebních programů

1. PLÁNOVÁNÍ A KONSTRUKCE TESTU

Ve školních podmínkách je didaktický test konstruován na základě určitého předem stanoveného cíle (účelu), případně na pokladě pracovní hypotézy, kterou chce učitel verifikovat (žáci se například špatně orientují v gramatických pravidlech nebo fyzikálních tabulkách a histogramech). Součástí procesu plánování testu je současně nejen ***analýza podmínek a populace pro kterou je test určen(věk, pohlaví, úroveň potřebných znalostí a dovedností, sociální zázemí aj.)***, ale i ***analýza konkrétního učiva, ze kterého je test navrhován.*** Při konstrukci didaktického testu by měl učitel vycházet jednak z taxonomie vzdělávacích cílů (Bloom, Tollingerová aj.), z didaktické analýzy učiva (např. Sup a Švec), ale i ze spektra učebních úloh, kterými se snaží ověřovat specifickou kognitivní oblast znalostí nebo dovedností žáků (Mareš aj.). **Postup při tvorbě didaktických testů** bychom mohli v krátkosti shrnout do několika následujících kroků:

- Analýza učiva a stanovení cíle didaktického testu
- Stanovení proporčnosti zastoupených položek (např. procentuálním vyjádřením učiva). Tvoří-li 20% z tematického celku určitá látka, pak by se 20% položek mělo týkat této látky.
- Tvorba testových položek, aniž by docházelo k doslovné formulaci položek převzatých z učebnic. Moderní didaktické testy by měly zahrnovat nejen faktografické znalosti žáka, ale i úlohy vyžadující jeho usuzování, aplikaci poznatků a dovedností, a měly by diagnostikovat úroveň tvořivého myšlení žáka. U standardizovaných didaktických testů se k těmto účelům většinou sestavuje tzv. ***obsahově operační tabulka***.

Například:

	PSANÍ		MATEMATIKA
1.	Znalost písma	1.	Znalost číselné soustavy a jednotlivých operací
2.	Porozumění gramatickým pravidlům	2.	Porozumění pojmům a procesům
3.	Aplikace (schopnost psát souvislý text)	3.	Schopnost řešit matematické úlohy

	PŘÍRODOPIS		VLASTIVĚDA
1.	Znalost terminologie a faktů	1.	Znalost společenských pojmů
2.	Porozumění přírodním zákonitostem	2.	Porozumění společenským jevům
3.	Aplikace znalostí v neznámé situaci		

Bloomova taxonomie vzdělávacích cílů: Kognitivní oblast (1956)

	Taxonomická klasifikace:	Příklady kognitivních činnosti:	Příklady učiva:
1.00	**Znalosti**		
1.10	Znalost základních pojmů	Pochopení, definování, identifikace, vysvětlení, znovuvybavení aj.	Pojmy, definice, názvy, významy, obsahy aj.
1.11	Znalost terminologie	Definování, vymezení, identifikace, pochopení, osvojení, aj.	Slovní zásoba, termíny, terminologie, významy, definice, názvy, aj.
1.12	Znalost specifických poznatků	Znovuvybavení, pochopení, identifikace, osvojení, aj.	Fakta a faktické informace, zdroje, jména, data, zprávy, osobnosti, místa, letopočty, příklady, jevy, aj.
1.13	Rozpoznání významu pojmu v kontextu učiva		
1.20	Definování pojmu vlastními slovy		
1.21	Znalost obvyklých poznatků	Znovuvybavení, identifikace, pochopení, osvojení, aj.	Formální a obvykle užívané symboly, způsoby a pravidla užívání, pracovní postupy, formy řešení, aj.
1.22	Znalost vývoje a souvislostí	Znovuvybavení, identifikace, pochopení, osvojení, aj.	Činnosti, procesy, vývojové trendy, souvislosti, vztahy, síly, vlivy, příčiny, aj.
1.23	Znalost klasifikací a kategorií	Znovuvybavení, identifikace, pochopení, osvojení, aj.	Oblasti, typy, třídy, rysy, skupiny, kategorie, klasifikace
1.24	Znalost kritérií	Znovuvybavení, identifikace, pochopení, osvojení, aj.	Kritéria, základní hlediska, části aj.
1.25	Znalost metodologie	Znovuvybavení, identifikace, pochopení, osvojení, aj.	Metody, techniky, přístupy, postupy, aj.
1.30	Znalost obecných a abstraktních poznatků a pojmů		
1.31	Znalost principů a generalizací	Znovuvybavení, identifikace, pochopení, osvojení, aj.	Principy, generalizace, zákony, implikace, základní tvrzení, zásady, aj.
1.32	Znalost teorie a struktury	Znovuvybavení, identifikace, pochopení, osvojení, aj.	Teorie, základy, vzájemné vztahy, struktury, organizace, formulace, aj.
1.40	Rozlišení forem správného a nesprávného užití pojmu		

1.50	Rozlišení mezi dvěma podobnými pojmy na podkladě jejich významu		
1.60	Vytvoření vlastní věty s použitím daného pojmu		
2.00	**Porozumění**		
2.10	Vysvětlení pojmu a principu	Vysvětlení a formulace vlastními slovy, transformace, a reformulace pojmu, ilustrace principu, aj.	Významy, příklady, definice, abstraktní pojmy, slova, fráze, aj.
2.20	Schopnost interpretace	Interpretování, reorganizace, reformulace, vymezení, pojmenování, vysvětlení, demonstrace, aj.	Vztahy a souvztažnost pojmů nebo jevů, aspektů, kvalifikace, závěry, metody, teorie, abstrakce, aj.
2.30	Předpověď výskytu daného jevu (extrapolace)	Predikce a odhad, vyvozování závěrů a předvídání vývoje, vymezení, vymezování vnitřních a vnějších vztahů, aj.	Souvislosti, implikace, závěry, faktory, významy, vlivy, pravděpodobnost výskytu, aj.
3.00	**Aplikace**	Aplikace principu nebo algoritmu řešení na novou situaci, zevšeobecňování, rozvoj, uplatňování, přenos, restrukturalizace, klasifikování	Principy, zákonitosti, závěry, vlivy, metody, teorie, abstrakce, situace, generalizace, procesy, jevy, postupy, aj.
4.00	**Analýza**		
4.10	Analýza základních poznatků	Vymezení, stanovení, upřesnění, identifikace, klasifikace, pochopení, kategorizace, dedukce	Části učiva, hypotézy, závěry, předpoklady, tvrzení, analýza argumenty, jednotlivostí, interpretace grafů a diagramů
4.20	Analýza vztahů	Analyzování, vymezení protikladů, srovnávání, vymezení a dedukce	Vztahy, souvislosti, témata, vlivy, příčiny, následky, argumenty, myšlenky, předpoklady aj.
4.30	Analýza organizačních principů	Analyzování, vymezení, stanovení, dedukce, formulace	Formy, vzory, typy a způsoby řešení, techniky, struktury, organizace, témata aj.
5.00	**Syntéza**		
5.10	Dovednost originálního řešení nebo kritického myšlení	Písemná formulace, sdělení, vytvoření, konstrukce, modifikace, demonstrace	Struktury, vzorce, písemné práce, výrobky, kompozice
5.20	Dovednost plánování a řešení situace souborem operací	Navrhování, plánování, úprava, modifikace, aj.	Plány, cíle, specifikace, schémata, operace, významy, způsoby řešení

5.30	Vyvozování závěrů o abstraktních vztazích	Vyvozování, usuzování, kombinování, syntetizace, dedukce, klasifikace, formulování, modifikování	Jevy, taxonomie, koncepce, teorie, schémata, vztahy, abstrakce, generalizace, hypotézy, dojmy, objevy a pokusy
6.00	**Hodnocení**		
6.10	Rozlišení mezi faktem a názorem	Usuzování, hodnocení, úsudek, validizace, zkoumání, šetření, rozhodování	Upřesňující principy, zákony, zákonitosti, postupy a procesy ověřování spolehlivosti,
6.20	Rozlišení mezi podstatnou a nepodstatnou informací	Usuzování, argumentace, srovnávání, standardizace, vymezování rozdílů a protikladů	Závěry, význam a smysl, efektivita, ekonomie, použitelnost, standardy, teorie, generalizace
6.30	Rozpoznání klamné argumentace v psaném textu		
6.40	Rozpoznání platnosti omezení prezentovaných dat		
6.50	Vyvozování platných závěrů z prezentovaných dat		
6.60	Rozpoznání předpokladu podmiňujícího předložené závěry		

Příklad analýzy kurikulárního rámce v oblasti přírodopisu a specifikace vzdělávacích cílů

Obsahová část testu

1.1. Vědy o Zemi

1.1.1. Charakteristické rysy Země, geografická sféra

1.1.1.1. Složení a stavba zemského tělesa

1.1.1.2. Tvary zemského povrchu

1.1.1.3. Hydrosféra

1.1.1.4. Atmosféra

1.1.1.5. Horniny, nerosty, půda

1.1.2 . Zemětvorné procesy

1.1.3. Země ve vesmíru

1.2. Vědy o živé přírodě

1.3. Vědy o neživé přírodě

1.4. Přírodní vědy, technika a matematika

1.5. Historie přírodních věd a techniky

1.6. Přírodní vědy, životní prostředí a vodních zdrojů

1.7. Povaha přírodních věd

1.8. Přírodní vědy a jiné vědní disciplíny

Kognitivně operační část testu

1.0. Znalosti

2.1. Porozumění

2.1.1. Jednoduché operace

2.1.2. Složitější operace

2.1.3. Tematická informace

2.2. Teoretické uvažování, rozbor a řešení problémů

2.2.1. Zobecňování a odvozování vědeckých principů

2.2.2. Používání vědeckých principů při kvantitativním řešení problémů

2.2.3. Používání vědeckých principů při vysvětlování

2.2.4. Vytváření, interpretace a aplikace modelů

2.2.5. Rozhodování

2.3. Používání nástrojů a provádění rutinních a vědeckých postupů

2.3.1. Používání přístrojů, laboratorního vybavení a počítačů

2.3.2. Získávání dat

2.3.3. Provádění výzkumu

2.4.4. Interpretace získaných dat

2.4.5. Formulace závěrů ze získaných dat

2.5. Komunikace

2.5.1. Získávání a zpracovávání informací

2.5.2. Sdílení informací

Perspektivy

3.1. Postoje k přírodním vědám, matematice a technice

3.1.1. Kladný vztah k přírodním vědám, matematice a technice

3.1.2. Skeptický postoj k využití přírodních věd a techniky

3.2. Volba povolání souvisejícího s přírodními vědami, matematikou a technikou

3.2.1. Podněty pro volbu povolání souvisejícího s přírodními vědami, matematikou a technikou

3.2.2. Zdůrazňování významu přírodních věd, matematiky a techniky i pro "netechnická" povolání

3.3. Zlepšení zapojení slabých skupin žáků do výuky přírodovědných předmětů a matematiky

3.4. Zvyšování zájmu o přírodní vědy, matematiku a techniku

3.5. Bezpečnost při provádění přírodovědných experimentů

3.6. Rozvoj vědeckého způsobu myšlení

Testové položky, vytvářené v rámci tohoto konceptuálního rámce zařazujeme do položkové databanky, označujeme například dvěma číselnými kódy.

Příklad:

"Uveďte důvod, proč máme žízeň a musíme pít, když je horko?"

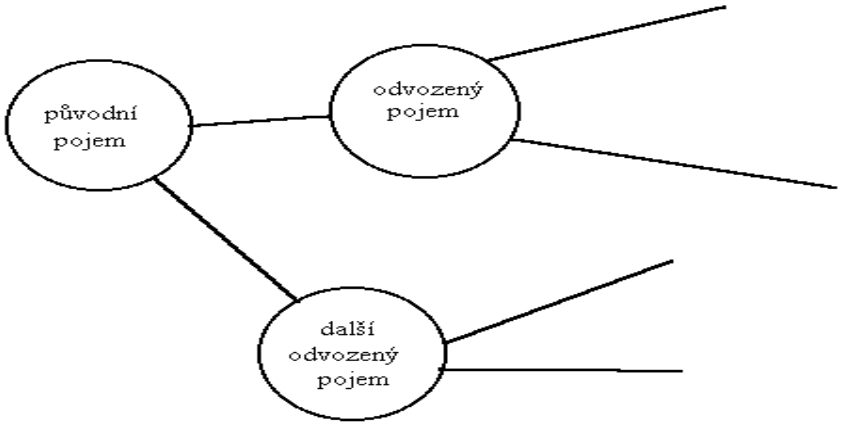
Kód O = 1.2.5, 1.2.2.2 Kód K = 2.1.2.

Má-li školní test co nejrovnoměrněji pokrývat celou oblast sledovaného učiva (typické je to zejména pro testy CR kritériálního typu) je výhodnější využít k rozboru daného učiva (tématu) orientační tabulky:

TESTOVÁ ŠABLONA didaktického testu z českého jazyka a literatury

Test Blueprint	Gramatika (50%)	Větná skladba (20%)	Literatura (30%)
Hodnocení			
Syntéza	5		
Analýza	2		2
Porozumění	3	3	2
Znalosti	15	7	8
Celkem	25	10	12

K rozboru daného učiva (tématu, partie) můžeme využít i orientačního grafu. Vyjdeme z učebnice nebo materiálů, podle kterých byla daná partie učiva probírána a zaznamenáváme jednotlivé nově probírané pojmy a pojmy, které na ně zákonitě navazují. Každý další pojem, který je s předcházejícím spojen nebo z něho odvozen, spojíme s výchozím šipkou. Jednotlivým oblastem učiva potom přidělujeme větší či menší váhu počtem přidělených položek.



Na základě předchozí analýzy provádíme ***volbu typu úloh a způsobu skórování***, popřípadě stanovujeme přibližný ***časový limit.***

2. TYPY DIDAKTICKÝCH TESTŮ

Z hlediska užití můžeme hovořit o konstrukci a interpretaci tzv. učitelských (tj. učitelem konstruovaných - ”teacher-made”) testů, které mají charakter standardizovaných na normy orientovaných didaktických testů (NR - norm-referenced test). Pro potřeby školního měření (edukometrie) se časem vyvinul test odlišné konstrukce - kritériální (CR - criterion-referenced) test. Poněkud odlišný charakter má *písemná zkouška testového typu*, která je označována jako essay test (E test).

Klasifikace didaktických testů (Byčkovský, 1980, podle Niemierka):			
Klasifikační hlediska	TESTY		
Měřená charakteristika výkonu	rychlosti úrovně		
Dokonalost přípravy a vybavení testu	standardizované kvazistandardizované	nestandardizované	
Povaha činnosti testovaného	kognitivní	psychomotorické	
Míra specifičnosti učení	výsledky výuky	studijní předpoklady	
Interpretace výkonu	rozlišující (relativního výkonu)	ověřující (absolutního výkonu)	
Časové zařazení do výuky	vstupní	průběžné	výstupní
Rozsah obsahového zaměření	monotematicképolytematické		
Míra objektivity skórování	objektivně	Kvaziobjektivně	subjektivně

2.1. DIDAKTICKÉ TESTY RELATIVNÍHO VÝKONU (NORM-REFERENCED TESTS)

Výkon žáka je srovnáván s výkonem populace žáků stejného věku nebo stejného ročníku, přičemž populaci při standardizaci testu zastupuje reprezentativní výběrový vzorek (obvykle 400 - 500 žáků). Přestože NR testy dosáhly vysokého stupně objektivity a citlivě diferencují mezi žáky, byly a jsou předmětem trvalé kritiky. Některé sporné rysy NR testů spočívají v:

- přeceňování predikční hodnoty testů při přijímacích zkouškách
- poškozování kognitivního stylu při opakovaném a častém testování., neboť formulace některých položek může vést žáky k představě, že existuje jen jediná správná odpověď
- destruktivním vlivu trvale slabých výsledků na sebehodnocení a motivaci některých žáků
- nebezpečí nadměrné koncentrace didaktického úsilí učitele pouze na práci s testem a v podceňování ostatních metod evaluace výsledků vzdělávání (např.autentické hodnocení, praktické úlohy apod.)

2.2. DIDAKTICKÉ TESTY ABSOLUTNÍHO VÝKONU (CRITERION-REFERENCED)

Kritériální testy informují uživatele zejména o *stupni zvládnutí učiva* žákem. Jsou konstruovány tak, že *kritériem úspěchu je obsah výuky, učivo, tj. předem stanovený stupeň jeho zvládnutí.* Typický kritériální test vyžaduje u vybraných poznatků transformovaných na testové položky jejich *úplné zvládnutí*. Výsledky v tomto typu testu jsou vyjadřovány alternativně - žák učivo zvládl / zná - nezvládl / nezná, přičemž kvantifikace výsledků tu není tak rozhodující jako u NR testu. Řeší se takové otázky jako např. ”Které pravopisné jevy jsou tak základní, že je musí ovládnout každý žák...?” Jsou složité algoritmy u základních početních úkonů v době kalkulaček ještě základním učivem?” V principu tyto otázky řeší osnovy či kmenové učivo, ovšem čím jsou osnovy obecnější a čím více poskytují učiteli více volnosti, tím více na něj přesunují rozhodování. Tvorbě kritériálního testu proto předchází vyřešení naznačených otázek. Konstrukce NR testu má podobné základní kroky jako je tomu u NR testu, ale chybí tvorba norem a kvantifikace získaných výsledků zde nemá tak dominantní postavení. Každý testovaný jev musí být ovšem pokryt větším počtem položek. Sekvence odpovědi na citlivou otázku by měla být: chybně - chybně - správně - správně a nikoliv správně - chybně - chybně - správně (viz znaménkový test). CR testy jsou toho daleko více časově náročné než NR testy ve smyslu ”mastery learning” - to je učení až do úplného zvládnutí základního učiva (podle Blooma). Jestliže totiž sledujeme poměrně široký rozsah školního kurikula, vede to k tvorbě obsahově a časově náročných kritériálních testů, které zatěžují nejen žáka v průběhu testování, ale na druhé straně i učitele ve fázi vyhodnocování a analýzy výsledků. CR testy jsou dominantně uplatňovány zejména v matematice, přírodních vědách a v gramatice jazyků, v oblasti čtení, psaní a počtů, tj. v předmětech, kde je hierarchicky uspořádané učivo a kde další postup výuky závisí na zvládnutí předchozího kroku. Vhodnější a přiměřenější jsou naopak NR testy (případně essay testy) ve vyučovacích předmětech, kde existuje volnější spojení mezi základními idejemi (např. v dějepisu nebo dalších společenskovědních oborech).

2.3. DIDAKTICKÉ ”ESSAY” TESTY

Essay testy jsou písemné zkoušky testového typu s menším počtem širších otázek, na které žáci odpovídají vlastními formulacemi. Poskytují pochopitelně větší prostor pro uplatnění žákovy individuality a umožňují aktivizovat a hodnotit náročnější myšlenkové operace (aplikace, srovnávání, analýzu), formulovat a rozvíjet formulační a stylistické dovednosti. Aplikace obecných pravidel a postupů v E testech *zmírňuje slabiny běžných písemných zkoušek*, zvláště nedostatečnou objektivitu, která vzniká nepromyšlenými a těžko objektivně hodnotitelnými otázkami a obtížemi při hodnocení značně odlišných odpovědí na tutéž otázku. Riziko subjektivního hodnocení lze však snížit oddělenou klasifikací jednotlivých otázek, anonymitou testu při hodnocení atd. Uplatnění E testu je dominantní v předmětech s vyšší mírou kognitivních než pamětních dispozic (jazyky).

	Norm-referenced tests Rozlišovací testy	Criterion-referenced tests Kritériální / ověřovací testy
Hlavní funkce	1. Měření individuálních rozdílů ve školním výkonu žáků	1. Vymezení učiva, které si žák osvojil a zvládl 2. Stanovení, co žák umí a neumí
Základní charakteristiky (Niemierko, 1990)	Nástroj pedagogického měření je: 1. určen ke zjišťování rozdílů ve školních výsledcích žáků 2. sestaven z úloh co nejlépe rozlišujících tyto školní výsledky 3. schopen poskytnout možnost komparace výsledků každého žáka s výsledky žáků dané populace (s populační normou)	Nástroj pedagogického měření je: 1. určen ke zjišťování/ověřování výsledků učení/vzdělávání žáků 2. sestaven z úloh reprezentujících vymezenou oblast školního učiva 3. schopen poskytnout možnost interpretace výsledků každého testovaného žáka s ohledem k sledované oblasti školního učiva
Interpretace výkonu	Srovnání výkonu žáka s výkonem ostatních spolužáků	Vyjádření výkonu žáka v jasně vymezeném oboru školního učiva
Rozsah učiva	Rozsáhlý	Poměrně málo rozsáhlý
Plán testu	Specifikační tabulka (blueprint)	Jasně vymezený a pokud možno uspořádaný obor učiva (domain)
Výběr testových úloh	Velmi snadné a velmi obtížné testové položky se vyřazují	Úlohy, které vyčerpávajícím způsobem reprezentují obor testovaného učiva
Testové normy	Populační	U ”mastery testů” - výkonové normyU ”domain testů” - se výkonové normy neužívají
Posuzování reliability a chyby měření	Nejčastěji je sledována vnitřní konzistence (KR-20) Paralelní formy testu	U ”mastery testu” je reliabilita většinou stanovena rozhodnutím o zvládnutí (masters) a nezvládnutí (non-masters) učiva. U ”domain testu” odhad reliability ”domain skóru”

4. KONSTRUKCE TESTU

Vlastní konstrukce testu začíná návrhy testových položek, zkoumáním jejich obsahové validity a sestavením prototypu experimentální verze testu.

Obecné zásady:

- *Délka položky musí být přiměřená věku.*
- *Zadání musí být korektní a nesmí obsahovat "chytáky".*
- *Položky musí na sobě nezávislé. Jedna položka by neměla navazovat na druhou.*
- *Položky by měly být nápadité a srozumitelné.*
- *Položky by měly být náležité přitažlivé grafickou úpravou.*

4.1. TVORBA TESTOVÝCH POLOŽEK

4.1.1. UZAVŘENÉ POLOŽKY

4.1.1.1. Klasické uzavřené položky s nabízenou odpovědí (multiple-choice)

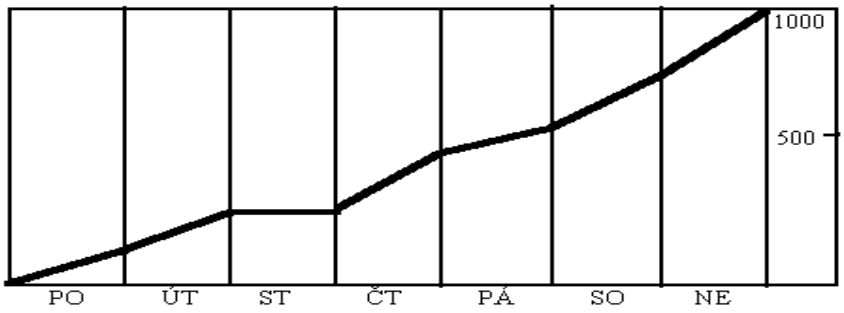
Tento typ testových úloh je také označován jako položky mnohonásobné volby nebo položky s mnohonásobným výběrem.. Každá taková položka obsahuje *jádro (kmen)*, kde je definován problém a několik *alternativ,* které představují možné řešení a několik distraktorů. Jádro může mít podobu otázky (položky rozhodovací) nebo má podobu neúplného tvrzení (položky doplňovací). Pravděpodobnost, že žák odpověď uhádne, přestože ji nezná, je u položek tohoto typu poměrně vysoká (p = 1/ n nabízených řešení). V klasickém testu se 4 nebo 5 variantami odpovědi je pravděpodobnost uhádnutí správné odpovědi 25% - 20%. Z tohoto důvodu je nutné u testových položek tohoto typu počítat tzv. *korekci pro náhodné hádání* (viz dále).

Příklad:

Česká republika sousedí s:	a) Maďarskem	distraktor
	b) Itálií	distraktor
	c) Ukrajinou	distraktor
	d) Německem	

4.1.1.2 Položky situační a interpretační

Testové položky tohoto typu většinou nabízejí větší počet distraktorů, aniž by byly taxativně vyjmenovány. Například: ”Jachtař na moři si pečlivě zaznamenával, kolik mil urazil každý den. Poznámky si dělal celý týden a výsledky zanesl do grafu. Který den bylo bezvětří?” Žák v tomto příkladu vlastně vybírá ze 7 možných odpovědí, které však nejsou taxativně vyjmenovány.



Také další položku lze zařadit mezi položky s mnohonásobným výběrem, kde žák vybírá z deseti čísel 0,1,2, až 9 (p = 10%).

Příklad:

”Doplňte na prázdné místo takovou číslici, aby číslo, které dostanete, bylo dělitelné 9!”

674__351

Mezi přednosti a nedostatky testových položek s mnohonásobným výběrem patří zejména:

- Objektivita skórování i vyšších kognitivních funkcí jako je porozumění a schopnost aplikace
- Nebezpečí náhodného uhádnutí správného řešení
- Neověřování aktivní znalosti, ale pouze tzv. znovupoznání.(to je patrné při zjišťování znalostí vzorců, termínů nebo schémat, které žák mezi nabízenými variantami pozná)
- Špatná volba tzv. nepravděpodobných řešení (distraktorů) nebo neúmyslné napovězení správné volby, kterou je žák schopen dedukovat vyvodit z textu nebo kontextu zadání.

Obecné zásady při tvorbě položek s mnohonásobnou volbou:

- *Tvořte alternativní odpovědi, aby byly gramaticky konzistentní s jádrem a obdobné svou formou.*
- *Správná odpověď nesmí být kratší nebo delší než ostatní.*
- *Otázku v negativní formě raději nepoužívejte.*
- *Nepoužívejte testových položek s dvojím záporem a složitých formulací, které jsou matoucí.*
- *Vždy měňte pozici správné odpovědi mezi distraktory.*
- *U testových položek, vyžadujících výpočet, volte raději položku s volnou odpovědí.*

4.1.1.3. Položky přiřazovací a uspořádací (matching items)

Testové položky tohoto typu jsou svou povahou uzavřené, poskytují více správných odpovědí nebo distraktorů, ale na druhé straně jsou poměrně obtížně skórovatelné, protože mezi úplně špatnou a zcela správnou odpovědí je řada částečně správných alternativ.

Příklad:

Pokus se přiřadit jednotlivé vynálezy k jejich autorům

Vynález	**vynálezce**
1. žárovka	A. B.Bell
2. telefon	B. J.Watt
3. parní stroj	C. S.Morse
4. telegraf	D. Gutenberg
5. knihtisk	E. bratři Veverkové

Příklad uspořádací položky:

”Seřaďte následující autory chronologicky (od současníků k nejstarším)!
Hájek z Libočan - Ovidius - Rolland - Hugo - Updike - Třebízský - Shakespeare

Obecné zásady:

- *Prvky v obou skupinách (návěští a doplňky) musí být homogenní.*
- *Určete přesně na jakém principu má být přiřazování provedeno.*
- *Do žádné skupiny (návěští a doplňků) nezařazujte více než 10 prvků.*
- *Doplňky řaďte systematicky (například abecedně)*

4.1.1.4. Položky s alternativní odpovědí ANO - NE (FALSE - TRUE)

Jednodušší vzdělávací cíle (znalosti a porozumění) jsou většinou zachycovány také testovými položkami s nucenou volbou typu: ANO-NE nebo FAKT - NÁZOR

Příklad:

Bitva u Kresčaku se odehrála na území dnešního Nizozemí.	ANO	- NE
Václav Havel je nejoblíbenějším prezidentem	FAKT	- NÁZOR
Tomáš Garrigue Masaryk byl prvním prezidentem ČSR.	FAKT	- NÁZOR

Položky s nucenou volbou typu ANO - NE nejen nesporně kladou důraz spíše na memorování a znovupoznání učiva, ale současně výrazným způsobem omezují možnost interpretace pro poměrně vysoké riziko nespolehlivosti, protože klient má při řešení možnost 50% úspěšnosti, přestože odpověď vůbec nezná. Při formulaci jednotlivých položek je vhodné vyloučit složitější souvětí, větná spojení s užitím negace nebo negace negace a testové položky převzaté přímo u učebních textů.

4.2. OTEVŘENÉ POLOŽKY

4.2.1. Doplňovací položky

Doplňovací položky jsou určitým přechodem mezi uzavřenými úlohami a typickými položkami s otevřenou odpovědí. Testová položka tohoto typu většinou obsahuje tvrzení, kde má žák doplnit slovo, frázi nebo termín.

Příklad:

Teplota vzduchu byla dnes ráno 17 _____, rychlost větru 7 _____ a tlak 760 _____.

Položky tohoto typu volíme všude tam, kde sledujeme znalost, nikoli znovupoznání, a do jisté míry jsou schopny zjišťovat aktivní ovládání faktických informací. Opakem jsou testové položky s nabízenou odpovědí, kde sledujeme spíše znovupoznávání a nikoli znalosti. Mezi chyby, kterých se dopouštíme při tvorbě těchto položek, patří zejména *nejednoznačné zadání.*

Příklad:

Doplňte následující větu:

Na_____ a v _____byli_____koncem II.světové války.

Obecné zásady
Vynechané slovo musí mít podstatnou úlohu ve větě.

- *V položce by mělo být doplněno jen jedno slovo, doplnění více slov je pro žáka matoucí svým zadáním.*
- *Vyhýbejte se nápovědám, plynoucím z gramatické konstrukce tvrzení.*
- *Položky v testu organizujte přehledně, např. do sloupců, nikoliv do souvislého textu.*
- *Předem si připravte klíč, obsahující všechny akceptovatelné odpovědi.*

4.2.2. Položky s krátkou a stručnou odpovědí

Tyto testové položky se snadno konstruují a jsou poměrně snadno a objektivně skórovatelné. Používáme je při zjišťování znalostí faktů, vzorců a specifických informací.

Příklad:

”Plyn nejhojněji zastoupený v atmosféře je ..... (klíč: dusík - N_2 - bezbarvý - ho málo)

4.2.3. Položky s otevřenou (omezenou či rozšířenou) odpovědí (Essay items)

Kmen položky je záměrně vždy široce formulován a umožňuje značně individuální přístup žáka. Omezení je dáno jedině časem, ale může být pochopitelně na různé pravopisné a jazykové úrovni. Essay položky používáme všude tam, kde chceme zjistit, jak žák ”umí myslet”. Naproti tomu, kde se jedná o zjišťování konkrétních znalostí či dovedností, otázky se samostatnou odpovědí zbytečně komplikují zkoušení. Pozitivem je, že vedou žáky k co nejvyšší efektivitě v organizaci vlastních myšlenek a jejich vyjadřování. Projevují vysokou volnost projevu vlastní individuality žáka a učiteli umožňují nahlédnout do procesu myšlení a uvažování žáků. Měří většinou vyšší kognitivní funkce, schopnost samostatného vyjádření a argumentace (zobecňování, abstrakce, analýza a hodnocení). Např. žáci musí uvést 2 hypotézy, které podporují předchozí závěry. Testovat tímto typem položek má význam, jestliže jsou žáci k tomuto způsobu práce systematicky vedeni. Žáci musí být současně jasnými a pádnými argumenty informováni o způsobu hodnocení. Většina položek tohoto typu je tzv. subjektivně skórovatelná a proto vyžaduje většinou hodnocení nejméně dvou examinátorů.

Příklad:

”Srovnej úvahu a pojednání jako slohový útvar.”

Obecná pravidla

- *Test by měl obsahovat jen několik položek tohoto typu.*
- *V žádném případě nemůžeme použít otázek z učebnice, které by umožňovaly žákům prostou reprodukci.*
- *Požadavek (zadání položky) musí být jasně formulován. Formulace otázky musí být v těsném vztahu k měřenému konstruktu.*
- *V zadání testové položky vždy specifikujte, co považujete za podstatné a co za okrajové, pokud právě toto rozlišení není vlastním předmětem dotazu.*
- *Poskytněte respondentům dostatek času a sdělte jim předem časový limit.*

Při skórování uplatněte jeden z následujících principů:

- *Hodnocení vychází ze vzorové odpovědi se vzorovým hodnocením jednotlivých etap řešení. Počet bodů musí odpovídat složitosti kroků nebo operací.*
- *Skórujte nejprve tutéž položku u všech žáků a vzájemně srovnávejte jejich odpovědi. Odpovědi rozdělte na tolik skupin, kolika stupňovou škálu jste volili.*
- *Opravujte testy vždy anonymně a pravopisné a jazykové chyby skórujte jen, je-li to účelné (např.v jazykovém testu)*

5. PILOTÁŽNÍ OVĚŘENÍ A ÚPRAVA TESTU

Při ověřování a úpravách nově vytvořeného didaktického testu je většinou vhodné dodržovat následující obecně platné zásady:

- Návrh standardizovaného testu ověřujeme 2 - 3 pilotními průzkumy, s jejichž pomocí provádíme položkovou analýzu a analýzu kódování jednotlivých testových položek. Klademe si otázky typu: ”Jaká je obtížnost testových položek a jaká je jejich rozlišovací schopnost? Odpovídá délka testu a postup administrace věkové úrovni žáků?
- Po nutných úpravách provedeme konečnou redakci a stanovíme minimální požadavky (např. u CR testu), časový limit a další podmínky pro zadání testu.
- Před administrací zajistíme kvalitní výcvik všech examinátorů a test administrujeme na reprezentativním stratifikovaném vzorku žáků.
- Počet připravených položek by měl být cca 4x větší než je zamýšlený konečný počet položek v testu. Je to nutné proto, že budeme vždy připravovat variantu A a B a že z připravených otázek nebudeme moci některé z různých důvodů použít.
- Navrhovanou verzi testu je vhodné dát k oponentuře několika kompetentním osobám.

5. 1. VYVÁŽENOST DVOU TESTOVÝCH FOREM

Kromě požadavků na objektivitu, jednoznačnou skórovatelnost a stabilitu v čase musí obě varianty testu splňovat i požadavek dostatečné reliability. Otázky tedy musí být tedy nejen přiměřené, ale i různě obtížné. Proto je vhodné do položkové databanky zařazovat skupiny položek, které jsou lehké (startovací), středně a velmi těžké. Z těchto připravených položek lze vytvořit z**kušební (pracovní) variantu testu** s písemnou instrukcí pro respondenty. Žáci se v tuto chvíli stávají jakýmisi spolutvůrci testu, protože jej komentují, ptají se na položky, které jsou jim nejasné atd. Tato fáze je určitou zpětnovazební částí práce na testu, přičemž pracovní varianta testu by měla být zadána alespoň ve dvou paralelních třídách (optimum je cca 100 žáků vzhledem k dodržení statistických norem pro tvorbu testů a dotazníků). Souběžně s tvorbou testu je nutné vytvořit **systém hodnocení (bodování** nebo také **kódování)**.

5.2. KÓDOVÁNÍ TESTOVÝCH POLOŽEK

Příklad kódování testových položek s volnou odpovědí:

30 správná odpověď včetně správného řešení	10 minimální odpověď
31 správná odpověď	70 nesprávná odpověď
39 jiná správná odpověď	91 přeškrtnuté řešení
20 částečně správná odpověď	99 prázdné řešení

Zastoupení jednotlivých typů odpovědí

kód	30	31	39	20	10	70	90	99
%	15	45	8	8	2	18	3	1

6. MĚŘENÍ V EDUKOMETRII A PSYCHOMETRII

Ke kvantitativnímu vyjádření výsledků jsou v oblasti školního testování užívány nejčastěji:

1. rating (posuzování). Ten je do jisté míry svéráznou kvantitativní operací, při které dochází na straně examinátora nejen k intuitivnímu zobecňování z řady pozorovaných jevů chování, ale i k srovnávání řady posuzovaných žáků. Výsledkem je udání individuální míry určité charakteristiky žáka, která je vymezována na ratingové škále (např. v hodnotách Lickertovy bodové škály 1-2-3-4-5-6-7),
2. bodový systém (skórování), který je nejčastější metodou užívanou v edukometrii a psychometrii. Žák v určitém testu, který se skládá řady testových položek (items), získá určitý počet bodů, jejich součet tvoří "hrubý skór" (HS - raw score). Ten již umožňuje a určitých podmínek užití *intervalové stupnice*, na které můžeme provádět všechny matematické operace. Zda ovšem skutečně měříme na intervalové stupnici, je často problematické (stupnice vymezené jednotkami HS např. v inteligenčním testu může, ale nemusí být intervalová - viz plošná normalizace asymetrické distribuce).

6.1. STANDARDNÍ SKÓRY

Do kategorie standardních skórů patří již zmíněný *hrubý skór.* Může to být např. čas potřebný k vyřešení série úloh, známka získaná na ratingové škále, počet bodů v didaktickém testu nebo v dotazníku. Význam hrubého skóru je například podstatný u CR testu, jestliže například stanovíme, že u žáka budeme hodnotit osvojení určité dovednosti, když ze 6 testových úloh úspěšně vyřešil 5 úkolových situací. V případě NR testu má však hrubý skór pro nás ryze informativní hodnotu. Žák například ve stanoveném časovém limitu vyřešil 20 úloh z 35. Co to pro nás znamená? Přitom délka didaktického testu a stanoveného časovému limitu pro nás nemá ve své podstatě žádný zásadní význam, protože to je věc konvence a praktických zřetelů, které testováním sledujeme. To, co odlišuje výkon žáka od výkonu ostatních spolužáků, je míra jeho *(směrodatné)* odchylky od průměrného výkonu ostatních. Standardní skóry mají totiž právě tu výhodu, že umožňují *intervalové srovnávání*, tj. srovnávání mezi žáky, a umožňují současně (v případě NR testu) *interindividuální srovnávání*. Stejně zajímavé je zjištěni, do jaké míry jsou získané výsledky administrovaného testu rozptýleny od středu distribuce, od průměru. K tomu bychom mohli použít *rozpětí (mezi nejvyšší a nejnižší získanou hodnotou)* nebo *průměrné odchylky od průměru.* Nejběžněji se však užívá *variance (rozptylu).* Variance je odvozena z odchylek od průměru. Vzorec variance je potom druhou mocninou směrodatné odchylky.

$$SD^2 = \frac{suma(X - M)^2}{N}$$

Vyjadřování v jednotkách směrodatné odchylky (sigma) se také nazývá ”z-skórování”. Standardní distribuce je potom charakterizována hodnotami M = 0, VAR = 1 a SD = 1. Je-li distribuce získaných výsledků normální, pak můžeme dokonce určit, jaké procento případů spadá do jednotlivých intervalů, vymezených z - skóry. Například mezi z-skóry 2 a 3 spadá 2,14%, mezi z-skóry -1 a 0 spadá 34,13% a tato procenta jsou *nezávislá na velikosti směrodatné odchylky.* Za zapamatování stojí, že hranice -2 SD (z-skór -2) odděluje zhruba 2% z celkového počtu případů, hranice -1 SD odděluje zhruba ”spodních” 16%.

$$z = \frac{X - M}{SD}$$

odchylka od průměru	- 2 SD	- 1 SD	0	+ 1 SD	+ 2 SD
odpovídající percentil	2	16	50	84	98

Metoda užití z - skóru nám umožňuje porovnání výsledků žáka v často velmi rozdílných didaktických testech.

Příklad srovnání výkonů žáka ve dvou didaktických zkouškách

zkouška	maximální počet bodů	Karel získal ....
matematika	100	46
český jazyk	100	56

	AM	SD
matematika	40	6
český jazyk	50	15

Z výkonu žáka v obou didaktických zkouškách je zřejmé, že Karel dosáhl stejných výsledků v obou didaktických zkouškách.

V některých případech je ovšem interpretace s pomocí z-skóru méně častější právě pro záporné hodnoty, které občas získáme (např. z = - 0,32). V takovém případě máme možnost transformace z - skóru na jiné míry (např. AM = 100, SD = 50 nebo AM = 500, SD = 100).

6.2. RŮZNÉ FORMÁTY STANDARDNÍCH SKÓRŮ

Základním standardním skórem je z-skór. Záporné hodnoty z-skóru nás však vedou k různým transformacím, při nichž z-skór násobíme zvolenou konstantou a k výsledku přičítáme další konstantu. Výsledkem je rovněž standardní skór ovšem v širším slova smyslu. Do této kategorie standardních skórů patří např.:

1. Deviační IQ= z . SD + 100 (Terrman užívá např. SD=16, naopak Wechsler SD=15)

2. T škála (skór) = z . 10 + 50 (M = 50, SD = 10)

3. Steny = jsou užívány u dotazníkových metod a osobnostních testů.

1	2	3	4	5	6	7	8	9	10
2,5%	4,5%	9%	15%	19%	19%	15%	9%	4,5%	2,5%

4. Staniny

jsou vlastně devítistupňovou stupnicí, kde střed tvoří 5.stanin s 20%, což je málo, nebo 4-6 stanin s 54%, což je zase příliš mnoho.

5. Percentil

je užíván především pro svou snadnou srozumitelnost při sdělování výsledků, ale má také podstatnou nevýhodu, protože vlastně není Lineární transformací standardních skórů a nemůžeme s ním zacházet jako s intervalovou škálou (nelze např. vypočítat průměr skupiny na základě percentilů). Percentily jsou vlastně pořadová data, která přeceňují rozdíly v pásmu průměru a podceňují rozdíly na obou koncích normální distribuce.

Příklad:

Tabulka distribučního rozložení absolutních a relativních četností

Položka	Absolutní četnost správných odpovědí HS	Relativní četnost správných odpovědí %	Percentilová hodnota
6^{min}	1	2,8	3
7	2	5.6	9
8	3	8,3	17
9	5	13,9	31
10	7	19,4	50
11	6	16,7	67
12	4	11,1	78
13	3	8,3	86
14	3	8,3	94
15	1	2,8	97
16^{max}	1	2,8	100

Tabulka převodu percentilových norem na známky

PR	známka
100 - 90	1
89 - 80	2
79 - 65	3
64 - 51	4
50 – 00	5

Tabulka A převodu percentilových norem standardizovaných testů na známky

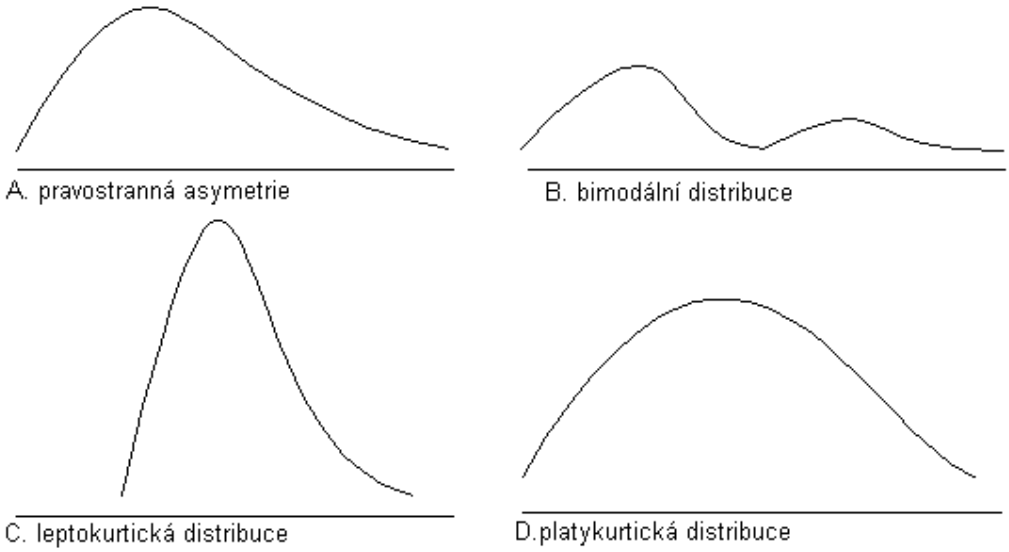
PR	známka
100 - 95	A
94 - 70	B
69 - 30	C
29 - 5	D
4 - 0	E

Tabulka B převodu percentilových norem standardizovaných testů na známky

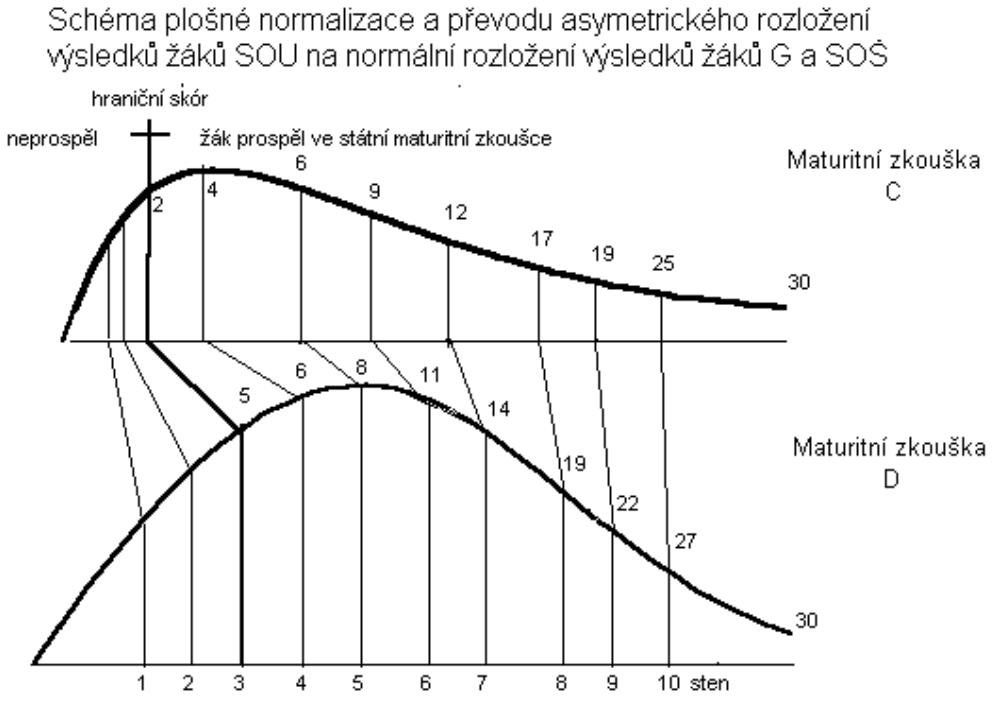
Kategorie	Varianta I.	Varianta II.	Varianta III.
vyniká	95 a více	90 a více	80 a více
vyhověl	95 - 5	90 - 10	80 - 20
nevyhověl	do 5	do 10	do 20

6.3. PLOŠNÁ NORMALIZACE

Při sběru dat se občas setkáme se skutečností, že hrubé skóry standardizovaného testu nemají normální distribuční rozložení, ale jsou rozloženy *asymetricky* (více vlevo nebo vpravo od průměru) nebo *bimodálně* (distribuční křivka má dva vrcholy), *leptokurticky* (distribuční křivka je příliš špičatá) nebo *platykurticky* (distribuční křivka je naopak plošší než normálně).



Standardní skóry získané pomocí standardní odchylky (z-skóry) nebudou zjevně u těchto výběrových souborů plnit svou funkci. Musíme proto hledat jiný způsob standardizace. Uvedené příklady můžeme například chápat jako výsledky několika písemných maturitních testů, z nichž některé jsou obtížnější pro žáky SOŠ a SOU (úroveň C) a jiné naopak snadnější pro žáky gymnázií (úroveň D). Ve smyslu axiomu normality mají ovšem duševní vlastnosti (a mezi ně řadíme pochopitelně i školní výsledky žáků) v populaci - nezávisle na tom, jaká je distribuce testů, které je měří - vždy normální distribuci. Nyní tedy musíme najít takovou *transformaci hrubých skórů*, která zajistí *normální distribuci transformovaných standardních skórů*. Nejsnadnější a nejjednodušší transformací tohoto typu je ***McCallova plošná transformace***. Vyjdeme například z následující asymetrické distribuce výsledků maturitní zkoušky, kterou se ve smyslu axiómu normality snažíme převést na normální, tedy *normalizovat*.



Následujícím způsobem budeme normalizovat asymetrickou distribuci výsledků, kterou zároveň převedeme z HS na stenové hodnoty:

Výsledky v maturitní zkoušce (HS)	Frekvence hrubých skórů	Kumulativní frekvence (HS)	Steny	Hranice stenů v %	Hranice stenů u 189 žáků	Hranice stenů v HS
1	1	1	1			
2	0	1	2			
3	2	3	2	2,28	4,3	2,7
4	5	8	3			
5	9	17	4	6,68	12,6	4,1
6	10	27	4	15,87	30,0	5,7
7	14	41	5	30,85	58,5	7,6
8	16	57	5			
9	13	70	5			
10	15	85	5	69,15	130,7	14,1
11	12	97	6			
12	12	109	6			
12	10	119	6			
14	6	125	6			
15	5	132	7			
16	5	142	7			
17	6	148	7			
18	6	153	7	84,13	159,0	18,5
19	5	158	8			
20	3	164	8			
21	4	169	8			
22	2	172	8	93,32	176,4	22,6
23	4	176	9			
24	2	178	9			
25	1	182	9			
26	0	184	9	97,72	184,7	26,7
27	2	185	10			
28	1	187	10			
29	1	187	10			
30	1	189	10			

V uvedené tabulce jsme zjistili percentilové hranice stenů (5.sloupec) a tyto percentilové hranice přepočítáme na příslušný počet např. maturujících žáků (N = 189). Tato operace je potom provedena v 6.sloupci. Spolu s normalizací jsme tedy provedli i transformaci na standardní skóry zvoleného formátu., tj. na steny (někdy se také říká na "plošné steny" nebo "normalizované steny"). Normalizací jsme ovšem ztratili původní intervaly a museli jsme je nahradit jinými.

V případě výrazné asymetrie výsledků ovšem vždy stojí za zamyšlení, zda by nebylo vhodné prodloužit či zkrátit časový limit testu nebo upravit testové úlohy. Nepravidelnost na horním (pravém) konci distribuce hrubých skórů může být zaviněna nejen různou mírou obtížnosti testu, ale i např. nepřiměřenou volbou časového limitu nebo nepromyšlenou bodovou bonifikací za rychlost. Silně platykurtická (plochá) nebo dokonce bimodální (dvouvrcholová) distribuce může svědčit také o tom, že náš standardizační vzorek se skládá ze dvou skupin, které se svými výkony v daném testu značně liší. Asymetrický test je ve většině případů citlivější v pásmu vysokých nebo naopak nízkých skórů a následnou normalizací tato specifika ztratíme ze zřetele. V každém případě je důležité stanovit *standardní chybu měření* a *standardní chybu odhadu* odděleně pro různá pásma výkonu v testu. ***Standardní chyba měření*** je vlastně směrodatná odchylka aritmetického průměru, přičemž vysoká reliabilita testu je většinou provázena nízkou standardní chybou měření.

$$SE = SD \cdot \sqrt{1 - R}$$

kde SD je *směrodatná odchylka* a R je *reliabilita*

6.4. STANDARDIZAČNÍ VZOREK

Standardizační vzorek ve své podstatě představuje populaci zkoumaných osob, pro kterou didaktický test připravujeme, a v tomto smyslu by měl být reprezentativní výběrovou skupinou pro danou populaci (např. patnáctiletých nebo maturantů). Populace je většinou chápána jako *soubor jedinců vymezený územními, národnostními, biologickými a sociálními ukazateli.* Poněkud jiným typem výběrového souboru je ***stratifikovaný vzorek,*** ve kterém musí být zastoupeny všechny vrstvy a skupiny žáků dané populace přiměřeným procentem.

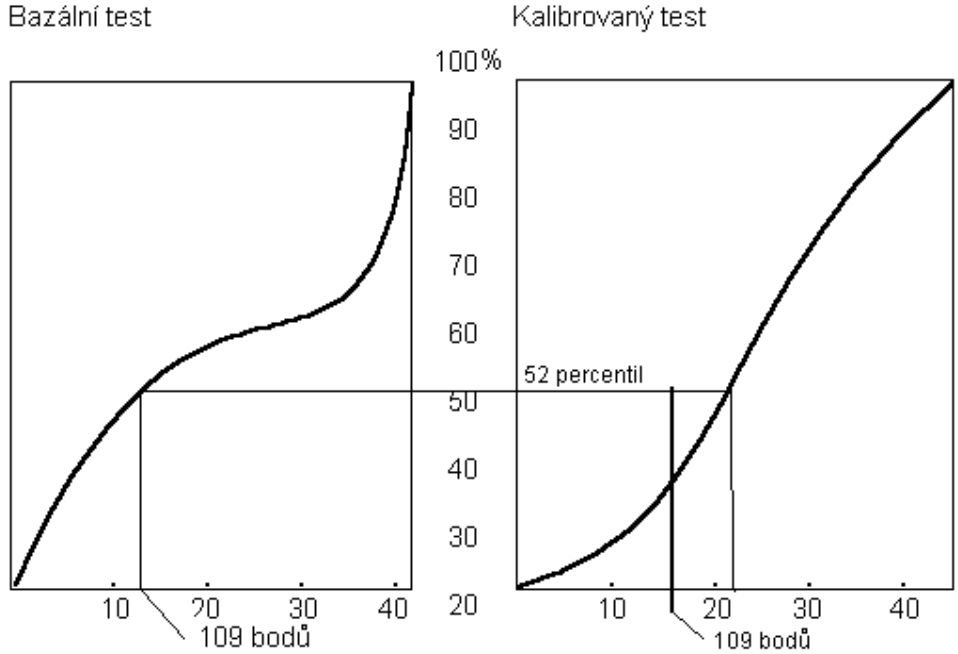
Mezi základní stratifikační kritéria při standardizaci testu nesporně patří zejména:

- √ Pohlaví - Normy založené na jednom pohlaví bývají totiž u většiny výkonových testů zkreslené. Proto v rámci stratifikovaného vzorku volíme dívky a chlapce v takovém poměru, v jakém se například v populaci patnáctiletých nebo maturantů skutečně vyskytují
- √ Vzdělání - Ve standardizačním vzorku by měly být např. zastoupeni vedle studentů gymnázií a středních škol i učni SOU, a to opět proporcionálně k četnosti těchto vrstev v populaci.
- √ Sociální status - V našich podmínkách se např. zdá účelné omezit se na vzdělání rodičů nebo typ státní vs. soukromé školy, která může být ukazatelem sociálního statutu maturantů.
- √ Lokalita - Rozdíly mezi městskými a venkovskými školami jsou v poslední době stále menší, ale nelze vyloučit, že mají svůj signifikantní vliv na řadu proměnných, které mohou ovlivňovat výkonnost žáků ve standardizovaném testu. Správná stratifikace pole lokality s sebou ovšem nese značné obtíže (nelze např. zdaleka říci, že by každá pražská škola byla reprezentativní pro Prahu).

6.5. PROBLÉM NEREPREZENTATIVNÍHO VZORKU

V některých případů nejsme z řady technických, organizačních nebo finančních důvodů schopni standardizovat určitý test na standardizačním vzorku a nemůžeme získat data od osob na všech úrovních. V takovém případě máme dvě možnosti:

- √ Zakalkulování netestované části populace do norem. Netestovanou část populace prostě umístíme na spodní okraj frekvenční distribuce a provedeme plošnou normalizaci.
- √ Kalibrací. Při kalibraci nového testu můžeme zvolit následující postup. Máme například standardizovaný "bazální didaktický test" (např. pro žáky z gymnázií) a podle něj chceme kalibrovat nový test. Oběma testy potom otestujeme přiměřený vzorek např. 200 žáků. Pro každý HS nového testu zjistíme na našem vzorku odpovídající percentil - např. HS 22 nového testu bude odpovídat 52 PR. Stejný percentil má u bazálního testu hodnotu HS 12, které odpovídá v původní reprezentativní standardizaci standardní skór 109. Proto přiřadíme hrubému skóru 17, který je průměrem hrubých skórů 12 a 22, v novém testu standardní skór 109 a tímto způsobem sestavíme celou tabulku norem pro nový kalibrovaný didaktický test.



Požadavek standardní testové situace se ovšem netýká jen standardního prostředí a stejného chování examinátora, ale zahrnuje i nutnost standardizace v takovém čase, v jaké bude test používán (např. na počátku nebo v závěru školního roku).

7. POLOŽKOVÁ ANALÝZA (item analysis)

Při konstrukci testu většinou postupujeme tak, že sestavíme předběžnou verzi, zjistíme distribuci hrubých skórů a validitu, případně vnitřní konsistenci. Dávají-li tyto globální údaje naději, že test bude plnit funkci, pro kterou je určen, zvláště jestliže jsme zjistili aspoň statisticky významnou validitu, má smysl zkušební verzi didaktického testu a jeho výsledky podrobit alespoň zkrácené **položkové analýze,** tj. zkoumáme vlastnosti jednotlivých položek a jejich přínos k žádoucí funkci testu. *Položková analýza* nám umožní odstranit "mrtvé" položky a zdokonalení testu, aby měřil to, co má skutečně měřit.

V průběhu položkové analýzy většinou ověřujeme:

- Obtížnost testových položek (mj. i efektivitu distraktorů)
- Korelaci mezi položkami a testem
- Korelaci mezi položkami a kritériem
- Časovou ekonomii testových položek

Obtížností položky rozumíme *pravděpodobnost jejího nesprávného řešení v dané populaci.* Odhadujeme tedy poměr počtu nesprávných řešení k celkovému počtu osob ve zkoumaném vzorku. Pravděpodobnost správného řešení značíme p a pravděpodobnost nesprávného řešení q, takže p + q = 1. Ačkoliv obtížnost je q, je zvykem soustřeďovat se spíše na p, tedy vlastně na "snadnost" položky. Tím vznikají drobní terminologické nedůslednosti, které však příliš nevadí. Jestliže tedy konstruujeme test pro účely depistáže slabých žáků, volíme testové položky s nízkou obtížností p = 0,5 - 0,8. Naopak, snažíme-li se vyhledávat talentované žáky v matematice, volíme testové položky s vysokou obtížností p = 0,2 - 0,5. Má-li být test všestranný, je vhodné volit testové položky s průměrnou obtížností 0,5, které mají *největší varianci:*

variance položky $VAR_i^* = p_i \cdot q_i$

- vzorec pro binární řešení (1,0)

kde, p_i = pravděpodobnost správného řešeníq_i = pravděpodobnost nesprávného řešeníObtížnost testových položek úzce souvisí s distribucí hrubých skórů testu, tj. s **rozlišovací schopností (účinností - r_{it})**. Při převaze snadných položek bude distribuce negativně zešikmena a budou převládat vysoké skóry a naopak. Při rovnováze snadných a obtížných položek se bude distribuce sice blížit normálnímu rozložení, avšak budou chybět položky střední obtížnosti, distribuce bude leptokurtická ("špičatá"), což je nevýhodné. Variance v tomto případě bude maximální, budou-li mít všechny položky p = 0,5. Obdobně záleží variance na korelaci mezi položkami, čím je korelace vyšší, tím větší je variance testu.

Tabulka položkové analýzy (pro r_{it})

položka	H	S	D	(H+S+D)	P	(H-D)	r_{it}

Legenda:

H počet žáků v horní (nejúspěšnější) třetině podle celkového

S počet žáků střední třetiny

D počet žáků dolní třetiny podle HS (nejslabší)

P index obtížnosti otázky, tj.% správných odpovědí u celého testovaného souboru

r_{it} rozlišovací hodnota položky vzniká tak, že vydělíme údaje (H-D) maximálním možným rozdílem mezi H a D (tj.1/3 celkového souboru žáků)

Vypočteného indexu r_{it} se užívá k hodnocení diagnostické hodnoty položek a pro úpravy definitivní verze testu. Doporučuje se do ní nezařazovat položky s p v intervalu 0 - 0,3, které jsou velmi těžké a v intervalu p = 0,7-0,8, které jsou naopak velmi lehké, a kterých můžeme použít jako ”startovacích” pretestových položek v úvodu didaktického Z diagnostického hlediska jsou nejproduktivnější testové položky s p = 0,3 - 0,7.

$$\text{Obtížnost položek } OP = \frac{\text{suma správných odpovědí}}{N} \times 100$$

Snadné položky p = 0,7 - 0,9, obtížné položky p = 0,1 - 0,3

Obtížnost položek je tedy odvozována z míry úspěšnosti žáků při jejich řešení, přičemž stupeň úspěšnosti vyjadřujeme většinou v procentech (100%= 1.0, 50%= 0.5 je optimální, 0%= 0,0).

U rychlostních (speed) testů hraje obtížnost položek poněkud odlišnou roli. Ideální obtížnost 0,5 zde nemá ve své podstatě význam. V případě aplikace rychlostního testu užíváme většinou takových položek, které bez velkých nesnází vyřeší kterýkoli žák, pokud se k nim v daném časovém limitu dostane a pokud je v časovém stressu neřeší příliš rychle a nepozorně. Analýza obtížnosti je zde stejně důležitá, čekáme však od ní spíše odpověď, jakým způsobem žáci řešili jednotlivé položky, což přispívá k poznání pojmové validity testu.

Odhad obtížnosti položek testů schopností a dovedností, kde odpověď záleží v zaškrtnutí správné alternativy, může být zkreslen *náhodným hádáním.* Kdyby při výběru ze dvou možností všichni žáci náhodně hádali, zjistili bychom přibližně p = 0,5, i když by obtížnost testové položky byla třeba 0,1. Můžeme předpokládat, že všichni žáci, kteří na danou neznají správnou odpověď, náhodně hádají (např. u položky s nabízenou odpovědí a-b-c-d). Je-li tomu tak, máme k dispozici *korekci pro náhodné hádání* pod vzorce:

$$\text{opravené } \underline{p} = \text{zjištěné } \underline{p} - \frac{\text{zjištěné q}}{k - 1}$$

kde, k = počet alternativ, ze kterých žák vybírá

Například chceme ověřovat diskriminační schopnost škálových položek, které by byly schopny diferencovat u experimentální skupiny budoucích obchodníků jejich dispozice stát se ”úspěšnými obchodníky”. Administrací 5 škálových položek u experimentální a kontrolní skupiny jsme získali následující výsledky.

Položka	”Úspěšní” obchodníci	”Špatní” obchodníci
1. Rád se setkávám s lidmi	100 %	100 %
2. Hodně jsem cestoval	5 %	5 %
3. Mám zkušenosti v různých odvětvích	85 %	80%
4. Trávím volný čas ve společnosti	40 %	55 % otočíme klíč
5. Plánuji a organizuji si práci	90 %	60 %

Ze získaných výsledků je zřejmé, že škálové položky č. 1, 2 a 3 nediferencují a je vhodné je buď přeformulovat nebo z testu vyloučit. Naopak u položky č.4 dosahují ”špatní” obchodníci vyšších výsledků než ”dobří”. Klíč škálového hodnocení je proto vhodné změnit. Z uvedených škálových položek dobře diferencuje pouze položka č.5, kterou je vhodné v testu ponechat.

Při sledování diskriminační schopnosti jednotlivých položek si všímáme zejména distribučního rozložení výsledků zejména mezi 27% nejlepšími a 27% nejhoršími výsledky (diskriminační koeficient je většinou obsažen ve všech stávajících programech na položkovou analýzu)

$$\text{Distribuční koeficient} = \frac{DP - ŠP}{N / 2} \times 100$$

$$\text{Koeficient citlivosti} = \frac{\text{suma S odp - suma Š odp.}}{N}$$

Příklad položkové analýzy 4 testových položek

Položka	Skupina nejlepších 10 žáků	Skupina nejhorších 10 žáků	Obtížnost	Distribuční koeficient
1	10	10	1,0	0,0
2	7	5	0,6	0,2
3	5	1	0,3	0,4
4	3	5	0,4	- 0,2

Příklad analýzy funkčního zařazení distraktorů u položky 1

	N	A správné řešení	B	C	D
dobří žáci	10	5	4	0	1
špatní žáci	10	3	2	0	5

Z výsledků je patrné, že distraktor B je nutné přeformulovat, protože ”úspěšní” respondenti volí tuto variantu daleko častěji než ”neúspěšní”. Distraktor C prakticky vůbec neměří a je vhodné ho vyloučit nebo změnit. Jedině distraktor D je formulován správně. Při konstrukci distraktorů se doporučuje volit distraktory s obtížností 0,3 - 0,7.

KORELACE MEZI POLOŽKAMI A TESTEM

Předpokládejme, že sledujeme vztah mezi vícestupňovou proměnnou /testem a dichotomickou proměnnou / testovou položkou nabývající hodnoty 1 a 0. Tento vztah jsme schopni vyčíslit s pomocí *Guilfordova biseriálního korelačního koeficientu:*

$$r_{bis} = \frac{M_p - M_q}{SD_t} \times \frac{p \cdot q}{y}$$

kde je **p, q** pravděpodobnost správného, resp.nesprávného řešení

M_p průměr vícestupňové proměnní pro skupinu těch, co vyřešili danou položku správně, Resp. získali bod pro měřenou vlastnost

M_q průměr pro skupinu těch, kteří nezískali bod pro měřenou vlastnost

y hodnota normální distribuce v bodě na ose x, který je dán rozdělením plochy normální distribuce v poměru p : q

Sd_t standardní odchylka vícestupňové proměnné pro celý vzorek

Korelace mezi položkou a testem je vedle korelace položky s kritériem *nejdůležitějším psychometrickým parametrem položky.* Je-li tato korelace nulová, je nutné položku vyloučit, protože zjevně neměří. Test takovými položkami by potom nebyl vnitřně konsistentní. Případ nulové korelace položky s testem schopností je velmi řídký, opakem jsou testy osobnosti, kde tento případ velmi častý. V řadě případů je vhodné použít spíše faktorové analýzy položek. Korelace položek s testem je současně vhodné doplnit o korelaci položek se subtesty, do nichž byly tyto položky zařazeny.

KORELACE MEZI POLOŽKOU A KRITÉRIEM: EMPIRICKÉ ŠKÁLY

Obdobným způsobem můžeme korelovat testové položky s kritériem, které chceme testem predikovat. Dnes jsou tyto čistě empirické škály považovány za zastaralé, právě pro svou pragmatičnost, protože znamenají zanedbávání pojmové validity testu.

ČASOVÁ EKONOMIE POLOŽKY

Princip položkové analýzy nám vlastně umožňuje docílit maximální validity a reliability testu při minimálním počtu položek. Počet položek však není sám o sobě rozhodující pro ekonomičnost testu. Položky se totiž mohou často lišit co do časové náročnosti. Jedna testová položka může být na rozdíl od druhé například 5x časově náročnější. Časová ekonomie testových položek se většinou bohužel nezkoumá nebo se zkoumá jen velmi povrchně. Položky nemají působit jako ”chytáky”, to žáky zbytečně znervózňuje. Naopak je dobré, když žák, který nemůže položku správně vyřešit, protože na ni nestačí, najde pokud možno snadno chybné řešení, které ho uspokojí. Žák potom přejde k další položce a neztrácí zbytečně čas, který je zejména u časově omezených testů rozhodující. Pečlivý rozbor chybných řešení na druhé straně nahrazuje alespoň zčásti společný nedostatek testů schopností, kde nesledujeme většinou proces řešení, ale jeho výsledek.

POLOŽKOVÁ ANALÝZA U OVĚŘOVACÍHO TESTU

U kritériálního testu vlastně měříme, zda žáci zvládli nějaký tématický celek učební látky a na základě získaných výsledků potom konstatujeme, že jej žáci zvládli nebo nezvládli. Měříme vlastně rozdíl mezi vstupem a výstupem.

Položka Vstupní test B Výstupní test A	1 B	1 A	2 B	2 A	3 B	3 A	4 B	4 A	5 B	5 A
Václav Novák	-	+	+	+	-	-	+	-	-	+
Věra Syrová	-	+	+	+	-	+	-	+	+	+
Hana Pěkná	-	+	+	+	-	+	-	+	+	+

ANALÝZA FREKVENCE

Kromě položkové analýzy slouží k určení diagnostické hodnoty didaktického testu i **analýza frekvence HS (hrubý skór)**. Globální charakteristiky zjistíme pomocí **tabulky frekvencí HS**.

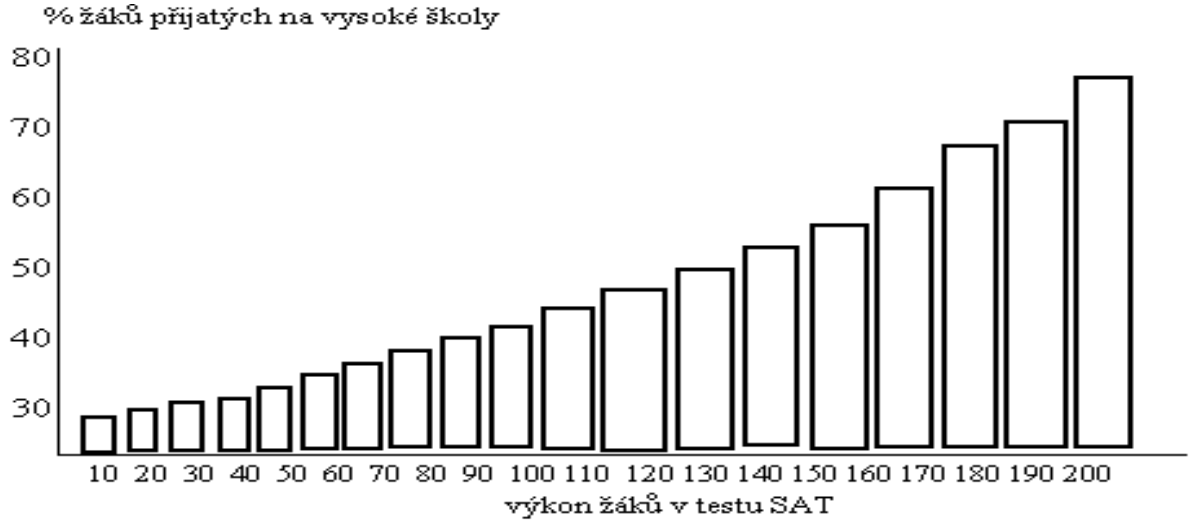
HS	*f*	*F*	*f/2*	*F- f/2*	*F-f/2 : N . 100*	*PR*
min.						
3	2	2	1	1	3,33	3
4	0	-	-	-	-	-
5	4	6	2	4	13,33	13
6	9	15	6,5	10,5	35.00	35
.	.	.	.	.	.	.
10	5	.	.	.	.	.
max.	1	30	0,5	29,5	98,33	98

f frekvenceF kumulovaná frekvence $F = f_1 + f_2 + ... + f_n$

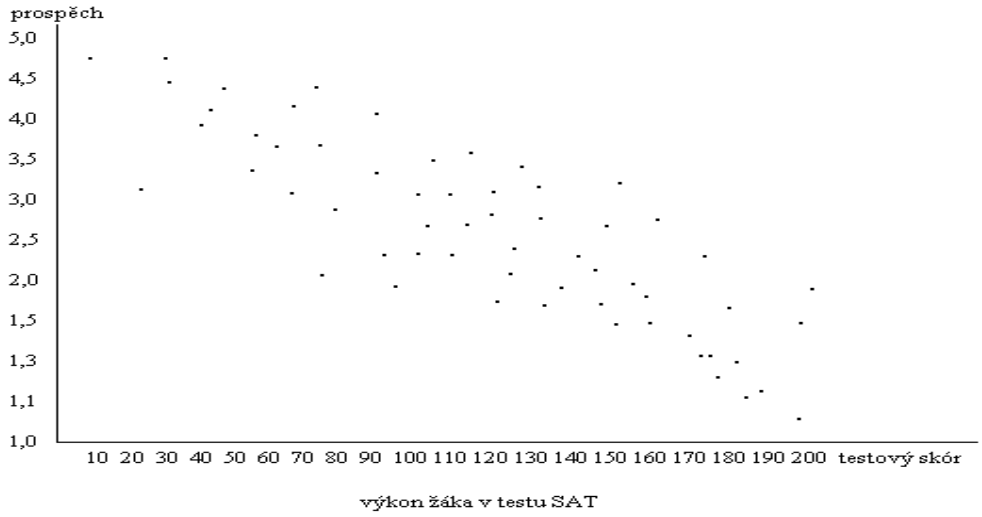
8. VALIDITA

Pojem validity je spojován s jeho českým ekvivalentem *platnost* a rozumí se jí *schopnost testu diagnostikovat, měřit nebo predikovat určitou psychologickou či pedagogickou veličinu*, která je předmětem našeho zájmu. V tomto smyslu rozlišujeme především *empirickou a pojmovou* validitu testu.

Snažíme se například vyjádřit vztah mezi výsledky testu školní způsobilosti (SAT) a úspěšným přijetím žáka střední školy do některého oboru vysoké školy (pozn. predikční validita testů SAT se např. pohybuje kolem hodnoty r = 0,4.). Získáme-li informace o vysokém vzájemném vztahu obou těchto veličin, můžeme do jisté míry hovořit o dobré *predikční* validitě takto konstruovaného testu, která nás může poměrně kvalitně informovat o skutečnosti, co můžeme očekávat na základě znalosti výsledků žáka v test. Na základě získaných dat jsme po několika letech například schopni sestavit *expektační graf*, který nás informuje o počtu přijatých žáků na vysokou školu, kteří v dříve administrovaném testu SAT dosáhli určitých výsledků.



Obdobným způsobem bychom mohli žákům I.ročníku střední školy administrovat test SAT, který má predikovat jejich úspěšnost ve studiu na daném typu odborné střední školy. Získané výsledky v testu a prospěchem žáků můžeme korelovat a získáme následující graf korelace mezi testem a školním prospěchem:



Validita testu je v tomto smyslu korelace mezi testem a kritériem. Kritériem je v obou příkladech jedna proměnná prospěl - neprospěl. Korelace je vlastně vztah mezi dvěma proměnnými, jejichž hodnoty jsou uspořádány ve dvojicích a kde míru vztahu mezi těmito veličinami můžeme vyjádřit *Pearsonovým korelačním koeficientem* (r), který může mít hodnotu 1,00 až -1,00. (Např.dobře sestavená baterie školních schopností může korelovat například s prospěchem v hlavních předmětech až -0,75 a čím výše, tím lépe škola využívá schopností žáků. Záporná korelace je dána tím, že známky vlastně vyjadřují neprospěch).

$$r = \frac{V_x \cdot V_y}{Co_{xy}}$$

V tomto smyslu je ovšem nutné varovat před usuzováním na efektivnost testu z matematicko- statistické významnosti korelace mezi testem a kritériem. Statistická významnost validity může být v určitém stádiu vývoje testu důležitá pro rozhodnutí, zda má vůbec smysl na testu dále pracovat, pro posouzení jeho praktické užitečnosti je však naprosto bez významu. Představuje totiž vždy příliš ”měkkou normu”. Zatímco při 30 žácích je statisticky vysoce významný teprve korelační koeficient 0,46, při 50 žácích je to už 0,36, při 70 žácích 0,30, při 100 žácích 0,25 a při 300 žácích 0,15. Termín ”významnost” v této souvislosti zjevně mate a daleko přesnější by bylo hovořit o ”průkaznosti”. Předchozí vztah dvou proměnných má zjevně podobu lineární (přímkové) korelace. Tento vztah ovšem může nabývat i podoby nelineární (křivkové) korelace.

Další pojem, se kterým se musíme seznámit, je ***kritérium***. To je proměnná, kterou se snažíme některým typem testu diagnostikovat nebo predikovat. Takovým kritériem může být např. úspěšnost žáka při studiu na střední škole nebo jeho přijetí na vysokou školu. Kritériem je zpravidla něco společensky důležitého, takže validita je zároveň praktickou užitečností testu.

Stejně tak bychom se ovšem mohli pokusit predikovat úspěšnost žáka v maturitních zkouškách, kterou bychom ověřovali didaktickým testem administrovaným v druhé nebo

třetím ročníku střední školy. Tento typ didaktického testu by do jisté míry mohl pomoci predikovat úspěšnost žáka v maturitních testech a současně by mohl pomoci k poznání *pojmové validity* maturitních testů. Jestliže tedy vztáhneme test k více jak jednomu kritériu, získáme různé údaje o několika typech validity daného testu. Validita v tomto smyslu není vlastnost testu, nýbrž jeho vztah ke kritériu, a kolik kritérií, tolik validit testu. Prokazatelnost validity většiny testů je totiž závislá na kvalitě zvoleného kritéria. Jednotlivá kritéria sice můžeme sloučit do jednoho tzv. *syntetického kritéria*, ale ve většině případů je daleko vhodnější zjišťovat validitu pro každé dílčí kritérium zvlášť.

Zjišťování prediktivní validity klasickou metodou (tj.testovat a čekat na kritérium, až nastane) je nesporně velmi zdlouhavé. Ve výzkumu proto můžeme sledovat nejen *prediktivní validitu* (tj. např. úspěšnost žáka v maturitní zkoušce, ale i *postdiktivní validitu*, tj. např. u úspěšných a neúspěšných dospělých studentů vysokých škol můžeme zjišťovat úroveň jejich ”maturitních” znalostí po několika letech od jejich odchodu ze střední školy. V takovém případě můžeme současně hovořit o sledování tzv. *souběžné validity.* Administrujeme například maturitní zkoušky úspěšným studentům vysokých škol, jejichž úspěšnost máme možnost souběžně měřit. V rámci validizačního projektu tohoto typu si ovšem musíme být vědomi řady neznámých, které nemáme možnost kontrolovat (např. motivace ke spolupráci, vliv zkušenosti při studiu na vysoké škole, spontánní úmrtnost neúspěšných studentů, kteří z vysoké školy odešli aj.).

9. RELIABILITA

Při konstrukci každého didaktického či psychologického testu se stanovuje jeho **reliabilita**, která je výrazem spolehlivosti a přesnosti testu, se kterou měří to, co má měřit. Na rozdíl od validity je reliabilita "vnitřní záležitostí" testu. Nejde zde o vztah k něčemu mimo test, ale o přesnost měření a reliabilitu testu vlastně sledujeme nezávisle na jeho validitě. Zjistíme-li potom, že test je vysoce reliabilní, plyne z toho že může být také více či méně validní ve vztahu k různým kritériím. Nedává to však žádnou záruku pojmové, ani empirické validity. Negativně můžeme reliabilitu vlastně definovat jako *nezávislost měření na náhodě.*

Předpokládejme, že například měříme výsledky žáka ZŠ v české jazyce, u něhož jsme didaktickým testem naměřili 190 bodů. Tuto hodnotu můžeme teoreticky rozdělit na skutečné výsledky žáka, kterých dosáhl v testu a na *chybu měření*, ke které v daném případě došlo vinou námi konstruovaného testu. Obdobně předpokládejme, že skutečná znalost žáka má hodnotu například 185 bodů, chyba měření má pak hodnotu 5 bodů:

$$190 = 185 + 5$$

obecně

$$\mathbf{t = T + e}$$

t = naměřený skór T= pravý skór e = chyba (chybový skór)

√ Zjištěný skór můžeme vždy hypoteticky rozdělit na *pravý a chybový skór*, přičemž pravý skór koreluje s chybovým skórem vždy nulově ($r_{Te} = 0$) a chybové skóry různých testů korelují navzájem nulově ($r_{e_1e_2} = 0$). Průměr chybových skórů je vždy nulový ($M_e = 0$). V předchozím příkladu jsme teoreticky vymezili pravý a chybový skór, přičemž oba tyto typy mají svou charakterististickou distribuci. Výsledný skór (v našem případě 190) je potom nejen součtem pravého a chybového skóru, ale všechny 3 typy skórů mají svou vlastní distribuční křivku, na nichž nás zajímá zejména jejich variance a vztahy mezi těmito variancemi.

$$\mathbf{SDt^2 = SD_T^2 + SD_e^2}$$

√ Varianci testu tedy můžeme - podobně jako testový skór - rozložit na pravou a chybovou komponentu.

√ Čím větší je potom podíl pravé variance na celkové varianci testu, tím je test realiabilnější. Tento podíl potom označujeme jako *koeficient reliability.*

$$\mathbf{r_{tt} = SD_T^2 : SD_t^2}$$

Kdyby tedy test neměl žádnou chybovou varianci, bylo by $r_{tt} = 1$ (rovnost čitatele a jmenovatele). V běžných podmínkách se ovšem hodnota r_{tt} pohybuje mezi hodnotou 0 a 1.Graficky bychom mohli pravou a chybovou varianci znázornit také takto:

Metody sledování reliability založené na opakované administraci testu:

Test - retest

Jedním ze specifických typů reliability je tzv. *dependabilita testu*, kterou můžeme vymezit jako *míru shody výsledků při opakovaném měření*, tedy jako reliabilitu testu vzhledem k času. Konkrétně potom hovoříme o korelačním koeficientu (koeficientu dependability) mezi dvěma administracemi testu nebo o testové a retestové korelaci (r_{dep}). U didaktických testů je tento koeficient relativním ukazatelem spolehlivosti testu, neboť retestováním času dochází k nácviku některých dovedností, které vlastně chceme měřit. Žák se již nezaměřuje na úlohy, které při testu vyřešil, ale při retestu se koncentruje na nezvládnuté úlohy, což do jisté míry mění spolehlivost testových výsledků při retestování.

$$r_{dep} = \Sigma (t_1 - M_1)(t_2 - M_2) / N . SD_{t1} . SD_{t2}$$

V případě testové reliability platí pravidlo, že čím delší je test, tím vyšší je většinou i testová reliabilita.

Reliabilita	Interpretace
0,95 a vyšší	výborný test
0,90 - 0,94	velmi dobrý test
0,85 - 0,89	dobrý test
0,80 - 0,84	použitelný test
0,75 - 0,79	použitelný test s velkou dávkou opatrnosti
0,70 - 0,74	mezní stabilita testu
0,65 - 0,69	test je zatížen velkou chybou
0,60 - 0,64	použitelný test pro získání výsledků třídy, nikoli pro hodnocení jednotlivých žáků
0,59 - nižší	je třeba se prohlédnout po jiném nástroji

DOPORUČENÁ LITERATURA:

Byčkovský,P.: Základy měření výsledků výuky: Tvorba didaktického testu. Praha, ČVUT, 1982 (skriptum)

Dittrich,P.: Pedagogicko-psychologická diagnostika. H+H, Jinočany 1993

Hniličková,J., Josífko,M., Tuček,A.: Didaktické testy a jejich statistické zpracování. Praha, SPN 1972

Hrabal,V.: Pedagogicko-psychologická diagnostika žáka. Praha, SPN 1989, s.40

Hrabal,V., Lustigová,Z., Valentová,L.: Testy a testování ve škole. Středisko vědeckých informací pedagogické fakulty University Karlovy (Informační bulletin, Supplementum 76), Praha 1992

Jencks,Ch.: Inequality. Harmondsworth, Penguin 1972

Kalous,J.: Užití statistiky v pedagogickém výzkumu. In: Skalková,J. a kol.: Úvod do metodologie a metod pedagogického výzkumu. SPN, Praha 1985

Kerlinger,F.N.: Základy výzkumu chování. Úvod do pedagogického výzkumu. ČSAV, Praha 1972

Komenda,S., Klementa,J.: Analýza náhodného v pedagogickém experimentu a praxi. Praha 1981

Mehrens,W.A., Lehmann,I.J.: Standardized Tests in Education. Holt, Rinehart and Winston, Inc., New York 1969

Mehrens,W.A., Lehmann,I.J.: Measurement and Evaluation in Education and Psychology. Holt, Rinehart and Winston, Inc., New York 1973

Průcha,J.: Pedagogické teorie a výzkumy na západě. Univerzita Karlova, Praha 1992

Průcha,J. Pedagogická evaluace. Brno 1996

Příhoda,V.: Psychologie a hygiena zkoušky. Dědictví Komenského, Praha 1924

Říčan, P.: Úvod do psychometrie. Psychodiagnostické a didaktické testy, Bratislava 1977