Univerzita Karlova v Praze, Pedagogická fakulta

Katedra [illegible]

Využívání počítačů k testování

Disertační práce

Zuzana Bažantová

Školitel: Doc. Ing. Petr Byčkovský, CSc.

Praha, Duben 2007

Děkuji všem, kteří mi při mém studiu i přípravě disertační práce věnovali svůj čas a podělili se se mnou o své zkušenosti.

Zejména pak mému školiteli Doc. Ing. Petr Byčkovskému, CSc. za jeho vedení, podporu, rady i poskytnuté informace.

Prohlašuji, že jsem tuto disertační práci vypracovala samostatně s použitím literatury, která je uvedena v seznamu.

Duben 2007

Zuzana Bažantová

OBSAH

ÚVOD

V posledních třiceti letech se ve vzdělávání stále častěji využívá informačních a komunikačních technologií, tj. multimediálních technologií a zejména počítačů. Využívání počítačů ve výuce je dnes rozšířeno nejen v USA a v jiných anglofonních zemích, ale i téměř ve všech evropských státech včetně České republiky. Počítačem podporovaná výuka se neustále dále vyvíjí a její projekty, průběh a produkty jsou v mnoha zemích podrobovány výzkumům. O jejích výhodách a nevýhodách se vedou četné diskuze. Názory na využívání počítačů k výuce se často liší. Jinak je tomu u použití počítačů při testování, kde počítače mohou usnadnit řadu různých aktivit, ať již jde o zadávání testů či jejich vytváření. Neocenitelná je jejich pomoc zejména při vytváření bank testových úloh, analýze testových výsledků a analýze testových úloh. Předkládaná práce se zaměřuje především na tuto problematiku.

Disertační práce se zabývá využíváním počítačů k testování. Skládá se ze dvou částí: teoretické a výzkumné. Je zaměřená na řešení dvou problémů: přehled charakteristiky současného stavu využívání počítačů při testování (1) a porovnání výhod a nevýhod využívání metod klasické teorie testu (KTT) a teorie odpovědi na položku (IRT) při tvorbě testu a analýze jeho výsledků (2).

Teoretická část (problém 1) má být dílčím příspěvkem k této problematice. Popisujeme zde historický vývoj využívání počítačů ve vzdělávání s důrazem na jejich uplatnění při testování, tj. při hodnocení výsledků vzdělávání. Jednu rozsáhlou kapitolu (kap. 4) věnujeme tvorbě, administraci a analýze výsledků testu, protože nároky kladené na tvorbu kvalitních testů zadávaných formou tužka-papír musí splňovat i testy počítačové. Jsou zde uvedeny i základní myšlenky IRT. Na tuto kapitolu navazuje výzkumná část.

Cílem výzkumné části (problém 2) je seznámit se současným stavem využívání počítačů při testování. Je nepochybné, že využívání počítačů zefektivňuje a usnadňuje zadávání, analýzu skórů a vytváření testových variant podle předem stanovených požadavků. Stejně tak se zdá, že využití IRT k tvorbě testu a analýze jeho výsledků má před KTT své výhody. Významné je zejména využití počítačů k položkové analýze pomocí IRT, protože v IRT jsou charakteristiky položek (obtížnost a citlivost) měřeny nezávisle na schopnostech testovaných na stejné škále, což umožňuje vybrat či odebrat úlohy odpovídající určité úrovni schopnosti testovaných, a tím vytvářet zcela paralelní varianty testu. Na druhé straně zvládnutí teorie IRT a relativně značné požadavky na dodržení určitých podmínek znesnadňují její využití v plném rozsahu.
Druhý problém je přitom řešen jak z hlediska teoretického, tak z praktického, a to při aplikaci na výsledky čtyř variant testu Obecné studijní předpoklady (OSP), které byly použity v přijímacím řízení (v řádném termínu) na Pedagogickou fakultu UK v roce 2006. Celkově bylo otestováno 1696 uchazečů o studium.

Ve výzkumné části uvádíme také využití speciálních software pro provádění klasické analýzy testových výsledků (ITEMAN a LERTAP 5) a pro jedno-, dvou- a tří-parametrovou IRT analýzu (BILOG-MG). Pokud je mi známo, porovnání položkové analýzy prováděné těmito programy u nás dosud nebylo uvedeno.

V práci popisujeme také adaptivní testování, kterým se ve světě, ale u nás zatím jen ojediněle, zabývá mnoho expertů a které použití počítače vyžaduje.

I TEORETICKÁ ČÁST

1 Historický vývoj využívání počítačů ve vzdělávání

V této kapitole uvádím stručný přehled o vývoji využívání počítačů ve vzdělávání. Počátky využívání počítačů ke vzdělávání sahají do 60. let minulého století. Tehdy došlo k ohromnému pokroku v počítačových technologiích a jejich dostupnosti. Počítačů se využívalo jen na velkých univerzitách ke čtení a psaní textů. Šlo o velmi rozměrné sálové počítače, jen zřídka o středně velké. Za historický počátek rozsáhlejšího využívání počítačů ve vzdělávání jsou považovány dva velké americkou vládou financované projekty PLATO (Molnar 1997) a TICCIT (Alessi; Trollip 2001). PLATO (*P*rogrammed *L*ogic for *A*utomatic *T*eaching *O*perations) byl vytvořen v letech 1959-60 na University of Illinois v USA. Tento systém umožňoval počítačem podporovanou výuku (computer-based instruction, CBI), při které využíval integraci textu a grafiky, a tak seznámil učitele s jedním z prvních programovaných prostředí pro vyučování na počítači. PLATO v podstatě položil základy dnes běžně známým aplikacím jako on-line fórum, nástěnka, on-line testování, e-mail, chat aj. (Černochová 2006). Základ systému TICCIT (*T*ime-shared, *I*nteractive, *C*omputer-*C*ontrolled *I*nformation *T*elevision) vyvinuli v roce 1972 inženýři společnosti Mitre. Systém umožňoval výuku na počítači, kterou si mohl řídit sám student. Byl dotvořen počítačovými odborníky z Univerzity Texas v Austinu a Brigham Young Univerzity, na nichž byl systém používán. Tento systém přispěl později k vývoji tzv. *component design theory* [1] (Merrill 1983).

Tyto první vzdělávací počítačové systémy byly propracovány a měly rysy podobné těm dostupným dnes na webu. Jejich nevýhody oproti těm dnešním byly především finančního rázu. Komunikace a jiné náklady byly vysoké. To spolu s nástupem mikropočítačů (menších a levnějších počítačů s uživatelsky příjemnějším rozhraním) v polovině 70. let vedlo k jejich zániku a jejich pozici převzaly stolní počítače. Do té doby se díky vysokým nákladům, jejich rozměrnosti a komplikovanosti s počítači pracovalo zejména na vysokých školách. Vývoj mikropočítačů však

[1] Component Display Theory popsal v roce 1983 M.D. Merrill. Component Display Theory (CDT) klasifikuje učeni na jeho obsah (fakta, postupy, principy) a výkonové procesy (zapamatování, používání, zevšeobecňování). Specifikuje čtyři primární prezentační formy: pravidla, příklady, vyvolávání z paměti, procvičování. Sekundární prezentační formy zahrnují nezbytné předpoklady, cíle, pomoci, mnemotechniky a feedback. Podle této teorie je vyučování efektivnější do té míry, že obsahuje všechny nutné primární a sekundární formy. Teorie předpokládá, že pro daný předmět a učícího se jedince existuje jedinečná kombinace prezentačních forem, které ústí v jeho nejefektivnější učící zkušenost. Významným aspektem CDT je kontrola řízená studentem. Jde o myšlenku, že studenti si mohou vybrat své vlastní výukové strategie ve vztahu k obsahu a prezentaci učení. V tomto smyslu poskytuje vyučování podle CDT vysoký stupeň individualizace, protože studenti mohou přizpůsobit učení tak, aby vyhovovalo jejich vlastním potřebám a stylům. CDT specifikuje, jak navrhnout vyučování pro jakoukoli kognitivní oblast. CDT poskytuje základ pro návrh vyučovací hodiny v TICCIT počítačem podporovaném výukovém systému.
1994 představil Merrill novou verzi CDT nazvanou Component Design Theory. Tato nová verze má globálnější zaměření než originální teorie s důrazem na struktury kurzu (místo jednotlivé hodiny) a vyučovací transakce než prezentační formy. Navíc poradce převzal kontrolu nad učením studenta místa studenta samotného. Vývoj nové CDT je těsně spojen s prací na expertních systémech a autorských nástrojích pro návrh vyučování.

přispěl k rychlému rozšíření počítačů nejen do vysokých škol (1975 mělo v USA přístup k počítačům 55% škol a 23% škol používalo počítače primárně k výuce, i když zpočátku jich nebylo na školách mnoho, velmi často byl v učebně pouze jeden počítač, a to pro učitele; Molnar 1997). Na konci 70. let byly v USA osobní počítače všude – v kancelářích, školních učebnách, domácnostech, v laboratořích a knihovnách. Počítače se staly nutností pro mnoho škol a univerzit. Mnoho univerzit dokonce vyžadovalo po nově nastupujících studentech, aby doma vlastnili počítač (Molnar 1997).

Mikropočítače nejen umožnily využívat grafiku a zvuk, ale dovolovaly i jejich vzájemné propojení do sítí, takže mezi institucemi mohly být vyměňovány informace. Prvním, pro americké základní a střední školy široce dostupným mikropočítačem byl Apple II, vyvinutý v roce 1978. V roce 1981 přišla firma IBM s osobním počítačem využívaným především v průmyslu a obchodu. Ten však nebyl příliš vhodný pro zavedení do škol, protože postrádal vzdělávací programy, náklady na jeho pořízení a provoz byly vyšší a integrace textu, grafiky a barev byla horší než v případě počítače Apple II. V roce 1984 uvedla firma Macintosh na trh nový počítač Apple, který měl snazší ovládání (myší vedle klávesnice), umožňoval lepší integraci textu a grafiky, lepší zvuk a kreslení na obrazovce. Avšak zejména nedostatek učebních programů (courseware), nedostatek barev a jeho cena bránily v uplatnění ve vzdělávání. Ačkoli Macintosh usnadnil ovládání počítače pomocí myši vedle klávesnice, IBM počítače si rychle získaly značné zastoupení na trhu s laptopy a stolními počítači, zejména poté, co firma Microsoft přišla s vylepšením operačního systému Windows, který se podobal systému od firmy Macintosh. Používání osobních počítačů neustále narůstalo, když společnosti vyvinuly a zveřejnily systémy, které umožňovaly počítačům vzájemnou komunikaci. Začalo se používat propojení několika lokálně blízkých počítačů do sítí, tzv. local area networks (LAN). Postupem času došlo k rozšíření rozlehlých sítí pro dálkový přenos dat, do tzv. wide area networks (WAN), jejichž vyvrcholením se stal v 80. letech internet (celosvětový soubor vzájemně propojených LAN a WAN sítí). Do té doby využívaly v USA internet pouze vláda a akademické kruhy, a to jen k výměně textových materiálů (Alessi, Trollip 2001). Začátkem 90. let byl jako součást internetu vyvinut World Wide Web (www).

Současně s dalším zdokonalováním osobních počítačů, s poklesem jejich ceny, možnostmi jejich propojování do sítí (u nás až zhruba od roku 1997), s rozvojem internetu a webu (u nás se od roku 1999 začaly organizace a instituce hromadně připojovat k internetu a využívat jeho službu WWW ke vzdělávání) a s vývojem značného množství výukových multimediálních programů (i když ne vždy kvalitních) dochází v dalším období (v 90. letech 20. století) k výraznějšímu rozšíření počítačem podporované výuky na školách, kdy byl obsah distribuován na disketách nebo CD-ROM (Zounek 2006). Internet dnes používají k různým vzdělávacím i jiným účelům (chatování, psaní e-mailů, telefonování přes Skype, nakupování, seznamování, k výzkumu, výměně textových, grafických či audiovizuálních informací, ke studiu apod.) stovky miliónů lidí na celém světě.

Významným podnětem pro využívání počítačů ve vzdělávání byl počátkem 60. let vývoj jazyka BASIC (Beginner´s All-purpose Symbolic Instruction Code), který nahradil do té doby používaný programovací jazyk FORTRAN (Molnar 1997). BASIC se pro svou jednoduchost rychle uplatnil při tvorbě výukových počítačem podporovaných materiálů (nejprve pro výuku matematiky). V polovině 60. let přichází S. Papert z MIT (Massachusetts Institute of Technology) s myšlenkou vyvinout prostředí, v němž by i děti mohly samy řídit a ovlivňovat chování počítače. 1967 položil spolu s W. Feurzeigem, B. Berankem, M. Newmanem, M. Minskym a J. McCarthym základy první verze jazyka LOGO (kde se objekt „želva" učí novým příkazům) vycházejícího mimo jiné z umělé inteligence a vývojové psychologie. Tento program našel v různých verzích uplatnění ve výuce řady předmětů ve školách na celém světě včetně ČR. U nás se v polovině 80. let pracovalo s želvičkou Zofkou, v první polovině 90. let s programem ComeniusLOGO a dnes se používá program Imagine (Černochová 2006).

I v současné době je stále vyučování na počítači (instructional computing) ve svém vývoji. Došlo k ohromnému pokroku, ale stále ještě není všude jasná představa o tom, jak počítače ve vzdělávání

využít. Sice se již nenaráží na problém neslučitelnosti software a hardware, ale o kvalitních vzdělávacích software, určených zejména pro web, by se mohly vést diskuze. Ve vyspělých zemích včetně České republiky je vybavena počítačovými učebnami a připojena k internetu velká část základních a středních škol (u nás je dnes vybavena a k internetu připojena zhruba polovina škol). Ani tiskárny, scanner, digitální fotoaparáty, někdy i digitální videokamery apod. nejsou dnes v mnohých školách ničím neobvyklým. Samozřejmě však existují rozdíly mezi jednotlivými školami u nás i v zahraničí například ve vybavení škol, způsobu využívání počítačů ve výuce, při učení žáků či řízení školy a její komunikaci s rodiči (Černochová 2006). Jedním z významnějších ukazatelů k porovnání stupně integrace ICT do škol jsou finanční částky věnované na rozvoj infrastruktury škol, školení učitelů a vývoj softwarových aplikací. Stále více rodin také již vlastní počítač. Podle zprávy OECD z roku 2003 měly v Dánsku, Německu, Švédsku a Švýcarsku zhruba 2/3 domácností počítač. V Anglii má přes 20% školáků ve věku 12-19 let dokonce vlastní počítač ve svém pokoji (Livingstone; Bober 2004). Počítač však v domácnostech většinou není primárně používán ke vzdělávání. Podle studie PISA (Program for International Student Assessment) z roku 2003 používají studenti počítač doma především k posílání e-mailů a chatování (55%), k vyhledávání informací na internetu (55%) a k hraní počítačových her (53%), k práci s textovým editorem (48%) a až potom ke studiu výukových materiálů (30%), k programování (24%) a k práci s výukovým software (18%).

2 Vývoj využívání počítačů k testování

Dosavadní vývoj využívání počítačů k testování rozdělují Bunderson, Inouye a Olsen (1989) do čtyř etap (generací):

1. **generace - běžné počítačové testování** (Computerized testing = CT)
2. **generace - adaptivní počítačové testování** (Computerized adaptive testing = CAT)
3. **generace - kontinuální testování** (Continuous measurement = CM)
4. **generace - sofistikované testování** (Intelligent measurement = IM)

Všechny čtyři generace umožňují počítačem kontrolované zadávání testových úloh, rychlé skórování a zpracování výsledků. Použitý software obvykle podporuje nové typy zobrazování úloh a odpovědí, velkou kapacitu ukládání informací a také síťovou komunikaci. První generace nevyžaduje ještě rychlý matematický procesor (floating-point processor) pro průběžné výpočty, které je nutný pro adaptivní algoritmus ve druhé generaci. Ve třetí generaci je testování začleněno do výuky a přestává být samostatnou součástí výuky. Umělá inteligence umožňuje sofistikovanější skórování a interpretaci výsledků ve čtvrté generaci.

Vědecké základy testování se v průběhu čtyř generací liší. První generace bývá často charakterizována používáním klasické teorie testu nebo absencí základní psychometrické teorie. Odborníci často implementují testy takovým způsobem, že si buď neuvědomují nebo jsou lhostejní vůči validitě a reliabilitě testu a považují papírově či individuálně zadávané verze testu za shodné s počítačovými (Bunderson; Inouye; Olsen 1989). Teoretická východiska druhé a vyšších generací jsou širší. Druhá generace utvořila adaptivní algoritmus založený na teorii dopovědi na položku (IRT) a je proto limitována situacemi, v kterých může být prokázán předpoklad unidimenzionality základní škály (předpoklad IRT). Třetí generace potřebuje k dosažení svého plného potenciálu rozšíření IRT či vznik nových psychometrických teorií, které umožní něco jiného než kalibrování úloh. Nutné je také testování profilů učícího se subjektu. Profily představují rozdílné učební dovednosti, styly a záliby a jejich používání je spjato s rozvojem kontinuálního testování (CM). Ke kontinuálnímu testování je zapotřebí vytvořit principy zobrazení a techniky implementace. Uživatelé musejí projít důkladným a rozsáhlým tréninkem. Čtvrtá generace představuje nové vědecké základy zahrnující modely znalostního inženýrství do výukových a testovacích aplikací. Tyto aplikace integrují znalost skórování komplexních úloh, profesionální znalost interpretovat profily studentů a znalost expertů schopných použít data z CM a žákovských profilů, a tím poskytnout preskriptivní doporučení žákům i učitelům.

Co se týče procesů administrace testu, jednotlivé generace se od sebe příliš neliší. Hlavní rozdíly mezi systémy lze nalézt v tom jak rozsáhlou podporu poskytují při vysvětlení výsledků. Dosud mají nezastupitelnou roli při interpretaci výsledků poradci nebo učitelé v roli poradců. Některá vysvětlení a doporučení systém poskytuje i při kontinuálním testování.

První dvě generace se obvykle zabývají statickým testováním, zatímco zbylé dvě kladou důraz na testování dynamické. Tato skutečnost je v těsném vztahu s jejich vzdělávacími cíli. První dvě generace primárně plní vzdělávací funkce, protože psychometrické testy zadávané dosud na počítači jsou obvykle variacemi běžných testů používaných ke vzdělávacím účelům. Třetí a čtvrtá generace se zaměřují na individuální vzdělávací účely. Testovací (měřící) škály se v jednotlivých generacích liší v jejich složitosti. V první generaci bylo možné užít různé typy škál, neformální a ordinální, nominální či intervalové. Druhá generace vyžaduje unidimenzionální škálu se shodnou velikostí intervalů pro testy založené na IRT. Třetí generace požaduje, abychom se zabývali multidimenzionalitou neodmyslitelnou při učení jakékoli komplexní oblasti znalostí. K tomu je nutné vytvořit kombinované škály, aby bylo možné poskytnout žákovi zprávy o všech jeho pokrocích. Ke skórování a interpretaci výsledků v procesu učení klade čtvrtá generace na škálování stejné požadavky jako třetí generace testování. Sofistikované testování můžeme také použít ke zlepšení skórování testů z první a druhé generace, v těchto případech se však vrátíme k jednodušším škálám. IRT poskytla hlavní vědecký pokrok ve vývoji příslušných škál pro vzdělávání. Umožňuje nám získat hodnoty jak pro osoby, tak pro úlohy na stejné unidimenzionální škále. IRT škálování je využíváno v aplikacích druhé generace a také v některých neadaptivních IRT testech již v první generaci. Kalibrace úloh komplexnějších než multiple choice úloh je již jedním ze znaků, které definují třetí generace.

Stupeň kontroly se vztahuje k vyhrazeným podmínkám, za kterých testování může být standardizováno. První dvě generace připouští značnou kontrolu nad zobrazením a sledem vizuálních a audio podnětných materiálů, nad typem odpovědi a časováním odpovědí. Třetí generace představuje dodatečnou kontrolu nad „vzděláváním" a neklade takový důraz na rozlišování mezi „vzděláváním" a testováním. Čtvrtá generace by mohla být použita při zavádění jiného stupně kontroly - kontroly výuky žákem samotným.

Dále budou detailněji popsány jednotlivé generace.

2. 1 První generace – běžné počítačové testování

První generace, **běžné počítačové testování** (Computerized testing = CT) zahrnuje tvorbu nových neadaptivních počítačem generovaných testů, které jsou podobné konvenčně zadávaným testům formy tužka - papír. Neadaptivní jsou takové testy, kde počet úloh, jejich pořadí, obsah a časování zadávání úloh nezávisí na předchozích odpovědích testovaného. Tvorba testů vyžaduje existenci rozsáhlých bank úloh. Teoretickým základem tvorby testů a jejich variant je klasická teorie testů (viz např. LORD, 1980). Tato generace podporuje administraci testů typu tužka-papír v digitální (počítačové) formě. Zahrnuje i testování on-line a rychlé vyhodnocení testů. Ačkoli přináší administrace testu na počítači některé nové problémy, přináší i řadu pokroků při zobrazování úloh na monitoru počítače, při získávání a kódování odpovědí, skórování, zpracování a vyhodnocování výsledků.

Běžné počítačové testování umožňuje *větší standardizaci*. Časování zobrazování úloh na PC umožňuje precizní kontrolu nad tím, co testovaný vidí a slyší. Administrace testu na počítači umožňuje novými způsoby, které nejsou u manuálně administrovaných testů možné, standardizovat podmínky, pokyny a postupy administrace. Větší standardizace s sebou přináší nutně větší obtížnost upravit testovací podmínky pro místní potřeby. Testovací systém může být naprogramován, aby byl odolný k úpravám testovacích podmínek při zadávání nebo může ponechat testujícím flexibilitu v úpravě testovacích podmínek určitými předepsanými způsoby jako je rozdělení testu do dvou časově rozdílných částí.

Počítačové testování také poskytuje *větší bezpečnost při testování*. Neexistují žádné papírové kopie nebo klíče správných odpovědí, které by někdo mohl odcizit či okopírovat nebo jinak zneužít. Testy administrované na počítači mohou zahrnovat několik úrovní bezpečnostních opatření (heslo, kódování aj.), aby bylo cizím osobám zabráněno v přístupu k testovým materiálům, bankám úloh či klíčům správných odpovědí. Pořadí testových úloh může být také náhodně změněno, je-li to požadováno, aby student neměl potřebu sledovat monitor jiného testovaného.
Podoba testu v tištěné podobě má své silné stránky, ale i určitá omezení. Může snadno zobrazovat text a obrázky, s většími náklady i fotografie. V rámci tištěných testů nebylo možné provádět časování (timing), změny pořadí zobrazení úloh v testu, animaci či pohyb. Tištěnou stránku nahrazuje v počítačovém formátu zobrazení na monitoru. Nabízí *větší variabilitu zobrazení* testu, jehož kvalita záleží na technických parametrech počítače. Počítač umožňuje použití *nových typů úloh* založených např. na animacích, komplexní grafice, zvuku. Počítačové verze běžných standardizovaných testů nabízejí významnou *úsporu času při administraci i vyplňování testu.* Vyplnění správných odpovědí na úlohy testu v počítači vyžaduje méně času ve srovnání se skenonovatelnými papírovými záznamovými archy. Počítače zjednodušují respondentům zadávání opovědí na některé typy úloh jako je např. přiřazování slov textu či označení obrázku (myší). V budoucnu bude zřejmě možné vyplňovat test i ústně a komunikovat s počítačem pouze pomocí mikrofonu.

Větší rychlost některých počítačem zadávaných testů proti podobným manuálně zadávaným testům je umožněna pokrokem v získávání a kódování odpovědí, ale i zvýšenou kontrolou nad prezentací a rychlostí zobrazování informací na monitoru. Při prezentaci pouze jedné úlohy na monitoru, počítač automaticky spojuje odpovědi s číslem úlohy a ukazuje výběry odpovědí pro bezprostřední zkontrolování. Vhodné znaky pro změnu odpovědí mohou nahradit časově náročnější mazání gumou na vytištěných záznamových arších. Počítače mohou také zadávat jiné typy úloh než jsou multiple-choice úlohy. Tvořené odpovědi zahrnující čísla či vzorce zapsané na klávesnici jsou celkem snadno interpretovatelné jednoznačně. Stručné odpovědi jako klíčová slova vyžadují sofistikovanější techniky, ale je možné je velmi přesně kódovat (s jistou tolerancí v pravopisu) dokonce bez technik na zpracování lidského jazyka. Úlohy tak mohou být kódovány jako správné, částečně správné či chybné. Počítače mohou zjednodušit jiné formáty odpovědí jako např. označování slov v textu nebo částí obrázků či kreseb. Tyto formy odpovědi jsou stále běžnější pro populace testovaných, jejichž počítačová gramotnost rok od roku roste. A umožňuji větší standardizaci a automatizaci procesu vyhodnocování testů a stávají se důležitou součástí portfolia typů úloh počítačového testování.

Počítačově zadávané testy umožňují nejen rychlejší zpracování, ale také eliminují některé tradiční typy chyb jako např. špatné přiřazení odpovědi k otázce. Zobrazení pouze jediné otázky v daný moment namísto celého testu, jak je obvyklé u testů papír-tužka, umožňuje studentovi soustředit se pouze na jeden problém. Pomáhá tedy získat přesnější výsledek u studentů, kteří mají potíže s koncentrací nebo čtením. Chyby spojené s vyhodnocováním testů jsou díky počítači téměř odstraněny. Jde zejména o chybné identifikace správných odpovědí ať už selháním lidského faktoru nebo chybou vzniklou při skenování záznamových archů. Pokud by přesto došlo k chybě, např. chybnému zadání klíče správných odpovědí do počítače, je velmi snadné tuto chybu dodatečně odstranit a přepočítat výsledky v testu. Na druhé straně je množství vizuální informace dostupné žákovi v jednom okamžiku na monitoru obvykle limitováno zobrazovacími možnostmi monitoru. To by mohl být problém při práci s delšími pasážemi textu, které není možné zobrazit na jednu obrazovku, a tím jsou testovány do jisté míry i paměťové schopnosti studenta, pokud text neumožňuje scrolování. Měření času potřebného na jednotlivé úlohy nabízí dodatečné informace, které je možné analyzovat, např. jak rychle žák čte, jak rychle analyzuje komplexní úlohy či grafické a zvukové informace.

Využívání počítačů eliminovalo chyby při výpočtu skóru v testu, umožnilo snadné získání dílčích skórů v testu, které mají také svou vypovídající hodnotu. Čas potřebný pro zpracování dat a přípravu výsledků klesl z dnů až týdnů na řádově minuty. To výrazně zvýšilo praktický přínos testu

pro hodnocení procesu výuky. Elektronická verze výsledků zjednodušuje jejich přenos a archivaci. Výsledky v testu lze tak s téměř nulovou chybovostí a nízkými náklady přenést do centrálního počítače (úložiště), kde jsou k dispozici pro položkovou analýzu a jiné analýzy či čistě archivní účely. Počáteční přenos přes telekomunikační sítě či posláním magnetického disku nebo pásky byl posléze nahrazen efektivnějším digitálním přenosem v komunikačních sítích.

První generace počítačového testování je úzce spojena s pokrokem v oblasti přípravy testů. Začaly vznikat rozsáhlé databáze testových položek (tzv. banky úloh). Testy tak bylo možno výrazně individualizovat na základě aktuálních potřeb či cílů. I přípravu testů zadávaných a vyhodnocovaných v klasické papírové podobě bylo možné zefektivnit použitím bank testových položek, textů či grafiky. Pro přípravu takových testů je možné využít jak specializovaný software, tak i běžný počítačový software a laserovou tiskárnu. Pokročilejší metody přípravy testů využívají algoritmy pro automatické generování testů z banky úloh na základě požadavků zadavatele testu. Výsledkem může být sada testů ekvivalentních z hlediska obsahu, formátu a možností skórování.

V počátcích počítačového testování byla obecně velmi nízká úroveň standardizace jak v oblasti hardware, tak v oblasti software. Teprve časem se začala prosazovat standardizovaná, intuitivní a uživatelsky přívětivá rozhraní. Počítačové testování umožňuje snadné vyhodnocování rozdílů mezi skupinami studentů v závislosti na typu školy, pohlaví, sociálnímu zázemí atd.

2. 2 Druhá generace – adaptivní počítačové testování

Druhou generací je **adaptivní počítačové testování** (Computerized adaptive testing = CAT), při kterém jsou testovanému zadávány úlohy na základě jeho předchozí odpovědi (viz např. Weiss, 1988). Obtížnost úloh se přizpůsobuje úrovni studentových znalostí. Nejdříve počítač předběžně odhadne úroveň znalostí studenta a podle toho vybírá z banky úloh počáteční úlohy odpovídající úrovni studenta, které mu zadá. Pokud student odpoví správně, počítač vybere následující úlohu s vyšší obtížností, pokud chybně, úlohu s nižší obtížností. Počítač ukončí testování v okamžiku, kdy odhadne výkon studenta s předem stanovenou velikostí chyby.

Hlavní výhoda adaptivního testování oproti neadaptivnímu spočívá v kratších testech, tj. testech s menším počtem úloh při jinak stejné přesnosti odhadu výkonu testovaného. Adaptivní testování však vyžaduje ještě rozsáhlejší banku úloh než je tomu u neadaptivního testování, ve které navíc musí být úlohy tříděny na základě parametrů získaných aplikacemi testových software založených na teorii odpovědi na položku (item response theory; viz např. LORD 1980, Baker 1985, Hambleton; Swaminathan; Rogers 1991). Kromě tohoto základního pojetí adaptivního testování uvádějí někteří autoři i alternativní přístupy (např. Alessi; Trollip 2001).

Základní rozdíl CAT od první generace je ve způsobu zadávání testových otázek. Adaptivní testování probíhá tak, že každá další otázka je generována pomocí výpočtu založeného na předchozích odpovědích testovaného. To významně urychluje proces testování, protože pro změření znalostí respondenta se stejnou přesností je zapotřebí méně otázek. Větší nároky na výpočetní algoritmy nejsou díky vývoji počítačů překážkou rozvoje adaptivních testů. Existují tři případy adaptivních testů: přizpůsobení otázek na základě IRT parametrů zejména parametru obtížnosti, přizpůsobení doby určené k odpovědi na základě rychlosti odpovědi na předchozí otázky a nakonec přizpůsobení obsahu otázky na základě předchozích odpovědí. Ve všech těchto případech lze přizpůsobovat délku testu.

Přizpůsobení (adaptace) rychlosti prezentace otázek

Počítačem zadávaný test přizpůsobený rychlosti zobrazení otázky byl vyvinut ve WICAT Educational Institute (vzdělávacím institutu) v roce 1983 a sloužil k posouzení rychlosti vnímání při učení (perceptual speed). Test zkoumal spolupráci obou hemisfér mozku. Šlo o tzv. Word/Shape

Matching test. Respondent se měl rozhodnout, zda nabízená obrazová a slovní odpověď (např. slovo kruh a obrázek čtverce) tvoří dvojici či nikoli a stisknout příslušné tlačítko. Pokud se testovaný nebyl schopen rozhodnout ve stanoveném čase, systém přešel k další otázce a patřičně prodloužil časový interval pro odpověď. Naopak po správné odpovědi následovalo zkrácení intervalu. Výsledkem byly dvě veličiny: procento správných odpovědí a odhad časového intervalu potřebného k odpovědi (rychlost odpovědi). Předpokládalo se, že rychlost (speed score) by mohla dobře predikovat úspěšnost v jiných typech rychlostních úloh (speeded tasks) vyžadujících současně zpracování verbálních i obrazových podnětů; že poměr rychlosti a chyb by byl užitečným ukazatelem o kognitivním stylu; a že nízké skóry by byly jedním z ukazatelů potenciální neschopnosti učení (learning disability). Test byl zadán skupinám žáků základních škol, kteří neměli při práci s testem žádné obtíže.

Přizpůsobení (adaptace) obsahu otázek

Simulační úlohy jsou vždy adaptivní, neboť další informace poskytnutá uživateli závisí na jeho odpovědích. Některé simulace také zohledňují u jednotlivých úloh reakční dobu. Simulační úlohy patří obvykle k tomu nejlepšímu, nejpropracovanějšímu, co může počítačové testování nabídnout, a to jak z hlediska komplexnosti úloh, tak i atraktivity pro respondenty. Příkladem může být simulované stanovení diagnózy pacienta s využíváním audiovizuálních vstupů předkládaných na základě rozhodnutí testovaného (zde: medika) nebo oprava nefunkční věci (automobilu, počítačové sítě apod.). Studenti dělají řadu rozhodnutí v simulovaném prostředí, skórování se provádí vyhodnocením jejich výsledků, strategií, ale i komplikovanosti cesty k výsledku. V USA byly zkoumány možnosti využívání simulace jako součást procesu certifikace lékařů (prováděl The National Board of Medical Examiners). Jiným příkladem testů s adaptivním obsahem vedle simulací může být počítačem zadávaný test pro odhad části studijního profilu testovaného (1983, byl součástí počáteční WICAT Learner Profile Battery). Tento test byl založen na porovnání bipolárních dimenzí (bipolar dimensions): logického a analytického myšlení oproti citům (feeling) a mezilidským vztahům (interpersonal preference) a intuitivního, holistického zpracování oproti kontrolovaným následným procesům. Šlo o úlohy párových srovnání výroků typu: „Rád bych věci rozložil na části, abych věděl, jak fungují", „Rád kreslím imaginární věci", „Mám rád na stole pořádek". Podobně jako jsou ve sportu párovány vítězové s vítězi a poražení s poraženými, dokud není sestaven konečný žebříček, byly dílčí preference vzájemně porovnávány, dokud nevznikl uspořádaný seznam.

Přizpůsobení (adaptace)na základě položkových parametrů

V tradičních testech forma tužka-papír či počítačových je většina otázek pro testovaného příliš lehká nebo naopak příliš obtížná, testovaný pravděpodobně zodpoví všechny lehké otázky správně a nezvládne ty obtížné. Příliš lehké či příliš obtížné otázky poskytují jen málo informací o měřené úrovni schopností respondenta. Přestože kořeny myšlenek a metod adaptivního testování sahají až do prací Bineta (1909), Birnbauma (1968) a Lorda (1970), mohly být do praxe uvedeny až s rozvojem a rozšířením počítačů. Počítač umí rychle vypočítat odhady úrovně schopností a odhady chyby a zjistit, zda bylo dosaženo kritérium pro ukončení testu. Pomocí CAT lze úroveň schopnosti každého testovaného měřit se stejnou přesností. Oproti tomu u tradičních testů (ať formy tužka-papír či počítačových) lze měřit přesněji skóry uprostřed měřící škály než na jejich okrajích. CAT vyžaduje rozsáhlou banku kalibrovaných úloh, tj. úloh, kterým jsou přiřazeny parametry odpovídající teoretické křivce odpovědi na úlohu (theoretical item response curve). Každá teoretická křivka je funkcí pravděpodobnosti správné odpovědi ve vztahu k úrovni schopnosti testovaného. Počítačový adaptivní test dále vyžaduje, aby odpovědi na úlohy byly lokálně nezávislé, tj. nebyly ovlivněny odpověďmi na jiné úlohy.

Počítačové adaptivní testování je založeno na psychometrické teorii zvané teorie odpovědi na položku (IRT) vyvinuté a vyložené např. Birnbaumem (1968), Hambletonem (1989), Hambletonem a Swaminathanem (1984), Lordem (1952), Raschem (1960) a dalšími. IRT požaduje, aby se

testovaní lišili ve svých schopnostech na unidimenzionální škále od nízké po vysokou schopnost (ability). Pro každého testovaného je pravděpodobnost správné odpovědi závislá na aktuálním odhadu jeho schopnosti a vlastnostech křivky odpovědi v konkrétní úloze. Křivky odpovědí jsou obvykle specifikovány až třemi parametry: parametrem obtížnosti, citlivosti a pseudonáhodným parametrem hádání.

Kroky při administraci adaptivního testu

Administrace adaptivního testu se dělí na čtyři kroky:

1. Vytvoření předběžného (preliminary) odhadu schopnosti testovaného.
2. Výběr a zadání úlohy, která poskytne maximální informaci na úrovni schopnosti daného testovaného. Tuto informační hodnotu úlohy lze vypočítat on-line nebo ji uložit do provizorní informační matice (precomputed information matrix). Odpoví-li testovaný správně, je mu předložena obtížnější úloha, pokud chybně, jednodušší úloha. Ze všech dostupných úloh je však vždy vybrána úloha, která poskytne maximální informaci pro aktuálně odhadnutou úroveň schopnosti testovaného.
3. Odhad úrovně schopnosti je po vyřešení každé zadané úlohy přepočítán pomocí jedné z několika existujících metod.
4. Testovací proces pokračuje, dokud není splněno testovací kritérium (např. pevně stanovený počet testových úloh, kdy směrodatná chyba měření je menší nebo rovna požadované hodnotě; či kdy informační funkce testu dosáhne nebo překročí požadovanou hodnotu).

Kalibrováni bank úloh

CAT testy vyžadují pečlivý vývoj a kalibraci relativně velké banky úloh (min. 100 úloh). Tyto úlohy jsou zadávány velkému počtu testovaných z cílové populace a pro každého testovaného jsou získány vektory odpovědí (response vectors). S 500-1000 vektory je schopen speciální počítačový software odhadnout parametry vybrané křivky odpovědi na položku. Jsou-li jednou úlohy kalibrovány, mohou být přidány do pracovní (operational) banky úloh a použity v CAT systémech. Počítačových programů na kalibraci úloh dnes existuje několik, nejrozšířenějšími jsou PARSCALE 3.2 (Muraki, Bock, 2003), LOGIST (Wingersky a spol., 1982) a BILOG 3 (Mislevy a Bock, 1990) či BILOG-MG (Zimowski, Muraki, Mislevy, Bock, 1996, 2003); dále se používá např. BICAL (Wright a spol., 1979), ASCAL (Assessment Systems Corporation, 1988), MICROSCALE (Mediax Interactive Technologies, 1986), MULTILOG 6 (Thissen, 1991), TESTFACT (Wilson, Wood, Gibbon, 1991), RUMM (Sheridan, Andrich, Luo, 1996) či MIRTE (Carlson, 1987).

Pohled do historie počítačových adaptivních testovacích systémů

S objevem mikropočítačů a se snížením ceny multiprocesorů bylo možné realizovat počítačové adaptivní testování. Americká armáda financovala několik rozsáhlých projektů. První vojenský prototyp CAT systému byl vyvinut pro Apple III počítače ve výzkumném centru Naval Personnel Research a Development Center (1984). Tento prototyp byl určen k první velkoplošné počítačové adaptivní administraci subtestů z testu Armed Services Vocational Aptitude Battery (ASVAB). Po úspěšném výzkumu validity a reliability tohoto počítačového testu ve srovnání s jeho papírovou verzí, uzavřelo Ministerstvo obrany (Department of Defence) smlouvu se třemi nezávislými společnostmi, které měly navrhnout a vyvinout CAT systémy (WICAT systémy). Nakonec se však armáda rozhodla tento proces iniciovaný těmito kontrakty nedokončit.

CAT systémy zahrnují standardní hardware (např. počítač, síťové prvky, rozhraní, periferie jako monitor, klávesnici, myš, tiskárnu aj.) a speciální software. Software slouží pro vývoj testů v centrálním výzkumném centru i pro konkrétní testovací centra a následně pro testování jak v testovacích centrech, tak na přenosných počítačích. CAT systémy vyvinuté v 80. letech profesionálními vzdělávacími a psychologickými testovacími organizacemi byly mnohem jednodušší než původní vojenské. Současné systémy zpravidla fungují na osobních počítačích.

Educational Testing Service a College Board vyvinuly CAT testovací systém pro IBM počítače pro testování základních dovedností v angličtině a matematice na úrovni střední školy (Abernathy, 1986; Ward aj., 1986). Poté následovaly další systémy např. od Assessment Systems Corporation (MicroCAT), Psychological Corporation či The Waterford Testing Center (více Bunderson, Inouye, Olsen, 1989).

Výhody CAT

Protože počítačové adaptivní testy jsou také zadávány počítačem, všechny výhody oproti „papírovému" testování zmíněné u běžných počítačových testů v první generaci platí i pro adaptivní testování. Ve stručnosti jsou to následující tyto výhody:

- zvýšená kontrola nad zobrazováním úloh na monitoru počítače;
- zlepšení bezpečnosti testu;
- větší možnosti zobrazení testu;
- zlepšení v získávání, kódování a vyhodnocování odpovědí;
- snížení výskytu chyb měření;
- automatizaci individuálně zadávaných testů;
- získání záznamů o testování v centrálním počítači;
- schopnost sestavovat testy a vytvářet úlohy na počítači.

IRT také poskytuje mnoho výhod při procesu skórování, protože každá úloha má kalibrovanou pozici na škále schopnosti (ability scale). Výhody CAT (Wainer 1983, 1984; Ward 1986) jsou zejména tyto:

- *zvýšení přesnosti měření.* Oproti počítačem zadávaným testům či testům zadávaným formou tužka-papír zachovává počítačový adaptivní test vysokou přesnost měření nejen uprostřed měřící škály (kolem průměrného testového skóru), nýbrž na všech úrovních schopnosti (ability levels), tedy i na okrajích škály. Všichni testovaní jsou tedy měřeni se stejnou přesností;
- *ekvivalentní odhady schopnosti za kratší testovací čas.* Počítačový adaptivní test oproti běžnému papírovému či počítačovému testu vyžaduje administraci mnohem méně úloh, protože testovaným jsou předkládány pouze úlohy odpovídající zhruba jejich úrovni schopnosti. Olsen, Maynes a Ho (1986) uvádějí, že u CAT je potřeba k získání ekvivalentní úrovně přesnosti pouze 30% až 50% testových úloh ve srovnání s papírovým testováním;
- schopnost měřit *latentní rysy odpovědi* (response latencies) jak u úloh, tak jejich částí;
- *další zlepšení bezpečnosti testu.* Jelikož každý testovaný dostává při CAT test jemu „šitý na míru", je tím velmi omezeno vyzrazení otázek a také opisování. K ochraně bank úloh a testových výsledků bývá použito šifrování. Neexistují žádné papírové kopie testů, záznamových archů či klíčů správných odpovědí.

Výzkumné problémy s počítačovým adaptivním testováním

Wainer a Kiely (1987) identifikují některé výzkumné problémy v CAT systémech.

- *Vlivy souvisejících otázek/ kontextové vlivy* (context efects). V běžném testování dostane každý testovaný ty samé úlohy, a to ve stejném pořadí, takže kontextové vlivy jsou pro všechny stejné. U CAT testů jsou ale možné různé kontextové vlivy u jednotlivých testovaných, protože každý obdrží individuálně sestavený test. Jedním možným řešením tohoto problému je vyrovnat párování a posloupnost úloh u všech úloh v bance, výsledné kalibrace parametrů by měly tedy vyrovnat kontextové vlivy. Jedním z kontextových vlivů je křížení informací (cross-information), kdy k zodpovězení úlohy je zapotřebí odpovědi na předcházející úlohu. Tento vliv by šel sice velmi pečlivou kontrolou všech úloh odstranit, ale zpravidla není možné posoudit všechny možné dvojice úlohy z rozsáhlé banky, tato technická kontrola samotná nestačí. Je možné také využít poloautomatizovanou kontrolu počítačem, kdy počítač vyhledá identické pojmy či synonyma v úlohách a příslušných odpovědích v bance úloh, které následně analyzuje expert na přípravu úloh;

- *Nevyvážený kontext* (unbalanced context). Tento problém nastává, pokud jsou úlohy v testu vybírány spíše z jedné tematické oblasti či okruhu dovedností. Je nutné vybírat otázky napříč tematickými okruhy a požadovanými dovednostmi. Možným řešením je doplnit k počítačovému adaptivnímu testu ještě dodatečná specifická kritéria pro výběr tematické oblasti;
- *Nedostatek robustnosti* (lack of robustness). CAT umožňuje výrazně kratší počítačové testy než běžné testování, proto má každá nesprávně fungující otázka mnohem vážnější dopad na celý test. Jedním řešením je vyžadovat mnohem přísnější kritérium pro ukončení adaptivního testu (delší fixní délku testu, menší standardní chybu odhadu nebo vyšší hodnoty informace testu). Výzkum ukázal, že CAT může zredukovat délku testu o 35%, v některých aplikacích dokonce o 50% či 75% při zachování stejné přesnosti. Dílčím řešením je tedy testovanému zadat více úloh než je dle teorie nutné pro získání požadované přesnosti, a tím zvýšit robustnost testu;
- *Stupňování obtížnosti úloh* (item difficulty ordering). V běžných testech bývají na začátek zpravidla zařazovány snadné úlohy pro motivaci, v adaptivních testech bez počátečního odhadu schopnosti testovaného se obvykle zadává jako první průměrně těžká úloha. To ale není optimální ani pro testované s nízkými schopnostmi ani pro ty s vysokými. Možným řešením je začít adaptivní test 5-6 otázkami, které pokryjí celé spektrum obtížností úloh v bance ke stanovení počátečního odhadu schopnosti jedince. Jiným řešením je začít adaptivní test úlohou nižší obtížnosti (např. 30%-35% místo 50%). V průměru se adaptivní test poněkud prodlouží, ale bylo by zajištěno, že většina testovaných má zkušenost s přiměřeně snadnými úlohami.

2. 3 Třetí generace – kontinuální testování

Třetí generace, **kontinuální testování** (Continuous measurement = CM) sleduje pomocí kalibrovaných škál založených na item response theory (IRT) dynamické změny v úrovni učení studenta a umožňuje tak vytvořit profil jeho výkonů v delším časovém období. Úlohami měřícími studentův výkon mohou být běžné testové úlohy a skupiny úloh (item clusters) či úlohy vyžadující rozsáhlejší odpověď nebo úroveň projektů samostatně zpracovávaných studentem. Kontinuální měření nabízí studentům a učitelům kontinuální monitorování pokroku v učení, a tak poskytuje informace, které jsou východiskem pro účinnější učení studenta a individuální přístup učitele při vyučování. Testování je integrováno do výuky, stává se jeho součástí (např. Wainer 2000).

Kontinuální testování předpokládá definici kurikula ve dvou úrovních: a) získávání zkušeností, které mají studentům pomoci růst směrem k určitému ukončení vzdělávání; b) množina standardů, požadavků, jejichž splnění umožňuje dosažení jedné ze tří úrovní: počáteční, středně pokročilé a výstupní úrovně.

Vlastnosti kontinuálního testování

- Požadavky na hardware počítače jsou stejné jako u vzdělávacího počítačového systému (computer-aided education, CAE) s dostatečnou rychlostí a kapacitou pro výpočty CAT. CAE systém je obvykle umístěn v učebně, kde probíhá procvičování a testování;
- Testování je prováděno kontinuálně, je zahrnuto do kurikula. Procvičovací moduly (exercise modules) jsou kalibrovány obdobně jako jsou kalibrovány úlohy ve druhé generaci;
- Kontinuální testování je nenásilně včleněno do výuky jako součást jejich obvyklých učebních aktivit, není zařazeno odděleně;
- Kontinuální testování se od prvních dvou generací odlišuje důrazem na dynamické testování primárně pro individuální účely, ne pro účely vzdělávacích institucí;
- Data získaná kontinuálním testováním (zvládnuté oblasti znalostí, odbornost) by měla být přístupná jak studentům, tak učitelům. Individuální pokroky studentů jsou zachyceny v tzv. mastery mapě (mastery map);

- Škálování je v CM komplexnější než u CAT. Místo jednorozměrných škál v CAT generaci používá CM mnohočetné (multiple) a často vícerozměrné škály (multidimensional scales), které je ale možno shrnout do jednoduchého složeného skóru (single composite score) nebo cílové funkce ke sledování pokroku každého studenta na individualizované mastery mapě;
- Referenční úlohy (reference tasks) jsou kalibrovány v CM generaci na intervalové škále. Tyto úlohy mohou simulovat požadavky na výkony z reálného života či ze zaměstnání. Referenční úlohy jsou zpravidla komplexnější než jednotlivé úlohy a vyžadují mnohočetné odpovědi. Když se studenti setkají s těmito různými úlohami, kontinuální odhad změn v jejich výkonu může být zpřesněn adaptivním odhadem aktuálně dosažené úrovně.

Příklady dílčích CM systémů

Jedním z částečných CM systémů byl již zmíněný systém TICCIT (Timeshared Interactive Copmuter-Controlled Information Television) vyvinutý inženýry ze společnosti Mitre v letech 1971 a 1972. Výukový software k němu vytvořila skupina počítačových vědců s vědci z oblasti vzdělávání z univerzit Texas a Brighamn Young. TICCIT byl jedním ze dvou hlavních systémů financovaných National Science Foundation na počátku 70. let, tím druhým byl systém PLATO vyvinutý na univerzitě v Illinois. Jiným příkladem je test ASVAB (Armed Services Vocational Aptitude Battery) z počátku 80. let, používaný americkou armádou, který byl a je stále nabízen s funkčnostmi CM (Wainer 2000).

Kalibrace výukových cvičení zadávaných počítačem

Výzkum ve vzdělávacím institutu WICAT v roce 1980 poskytl příklad některých výhod kalibrace cvičení na porozumění čtenému textu začleněných do kurikula, konkrétně WICAT Elementary Reading Curriculum. Tento systém byl stejně jako TICCIT založen na výuce na počítači, kterou si mohl řídit sám student. Byl adaptivní povahy – studenti dostávali cvičení na porozumění textu různých obtížností podle dosažení určité úrovně. Úlohy byly kalibrovány za použití počítačového programu BILOG. Autoři testu si mysleli, že tato cvičení mohou být kalibrována tak snadno jako jednotlivé úlohy a že parametry obtížnosti získané kalibrací budou poskytovat výbornou stupnici pro cvičení na porozumění čtenému textu (více Bunderson, Inouye, Olsen, 1989).

Rozdíly v užitečnosti mezi CM a běžnými školními výkonovými testy (School Achievement Tests)

Užitečností se zde rozumí praktičnost a jednoduché užívání ve skutečném výukovém prostředí. V případě standardizovaných testů jsou předem stanovené „testovací dny", či v jednotlivých předmětech testovací hodiny rušivým zásahem do školního týdne. Pro většinu studentů jsou tyto dny poměrně traumatickým zážitkem. Oproti tomu počítačové testování se od testování formou tužka-papír velmi liší. Na začátku roku představí učitelé žákům počítačovou učebnu a stanoví základní pravidla. Během roku docházejí studenti do této učebny, pustí se do studia (např. porozumění čtenému textu) na svých pracovištích. Sami studenti si určují, kdy si půjdou své schopnosti otestovat, a volí si také své tempo učení. Testování je tedy součástí výuky.

Mastery Assessment Systems jako kontinuální testování

Pojetí mastery assessment systémů bylo vyvinuto během let 1986 a 1987 výzkumníky z Educational Testing Service (ETS). Poprvé je popsali Forehand a Bunderson (1987a, 1987b). Tyto systémy mají dva základní rysy. Zaprvé mastery systém hraje svou roli v plánování kurikula, ale sám není jeho součástí. A zadruhé termín „mastery" se nevztahuje jen na minimální schopnosti. Vývojáři mastery testovacích systémů vynakládají úsilí do zmapování definovaných oblastí znalostí a jejich zapsání do cílů, vybírají subset měřitelných milníků (milestones) na cestě učení z většího souboru mastery assessment systémů a zasazují je do vlastního kurikula. Svou práci pak předají testovací organizaci. „Mastery" (suverénní ovládnutí) značí dosažení osobních učebních cílů, které sahají nad rámec minimálních kompetencí. Je ho dosaženo hlavně vytrvalostí a důsledností. Hodnocení vyšších úrovní mastery musí obsahovat jedinečné tvoření, např. řešení komplexního problému, ústní

prezentaci, písemné analýzy, portfolia. „Assessment" zahrnuje použití standardizovaného testování kompetencí a rady (guidance) pro posouzení úrovní předcházejících dosažení mastery. Udílené rady bývají založeny na disciplinovaném subjektivním či inteligentním počítačovém skórování.

Části mastery systému

Mastery systém vyžaduje CAE systém, který však není tak složitý jako v případě systémů TICCIT či WICAT. Hlavními nehardwarovými komponenty jsou:

- *mastery mapa*, kterou používají jak studenti, tak učitelé a která jim slouží jako plán učebního procesu;
- *referenční úlohy*, které jsou zpravidla komplexnější než jednotlivé úlohy, jde o testlet vyžadující rozsáhlejší odpověď. Referenční úlohy se vztahují k reálnému životu;
- *kalibrace úloh a referenčních úloh* – referenční úlohy mohou být umístěny na škálách, úlohy seskupeny do trsů úloh, testletů mohou být také umístěny na těchto škálách. Škálové hodnoty jsou statistickými hodnotami obtížnosti založenými na IRT;
- *skórovací systém* orientovaný na výuku pro každou referenční úlohu – úlohy jsou obvykle skórovány dichotomně (správně, nesprávně). Je též možné skórovat referenční úlohy kvalitněji spojením výkonu s výukovou strategií. Studenti jsou zařazeny do tří kategorií: na ty, jejichž výkony dosáhly požadované úrovně; na ty, kteří potřebují a jsou připraveni na procvičování; a na ty, kteří nejsou připraveni na procvičování dané referenční úlohy. Skórovací algoritmus je založen na přesnosti správných odpovědí a nepřítomnosti částečných odpovědí, které naznačují mylné představy. Algoritmus je kombinací expertního posouzení a systematického pozorování výkonu velmi schopných studentů ve srovnání s těmi méně schopnými;
- *profesionálně vyvinutý program pro učitele* poskytuje tzv. sledovací systém (tracking systém), takže učitelé mohou učinit profesionální rozhodnutí o tom, jak vést třídu či jednotlivé žáky k pokrokům v učení.

Jak mastery systémy mohou sloužit individuálním studentům a učitelům?

Mastery systémy nabízejí řadu způsobů, kterými může testování sloužit individuálním studentům a učitelům. Tradiční testování často klade důraz na vzdělávací účely jako připuštění, certifikace, umístění do zaměstnání, klasifikace. A také slouží jedincům pomocí poradenství, speciálního uznání, umístění v učebním programu, diagnózy učebních problémů a sebehodnocení vlastních znalostí a růstu. Mastery systémy znásobují tyto příležitosti tak, že se soustředí na testování růstu v dovednostech (skills) a znalostech (knowledge). Cílem mastery systému je postup studentů směrem k mastery (suverénnímu ovládnutí). Poskytuje služby jak pomoci a způsoby, jak jich použít.

Proměny nového profilu studenta

S kontinuálním testováním se objevují nové profily studentů. Bunderson, Inouye a Olsen (1989) uvádějí příklad z WICAT reading curriculum. Při vyhodnocování učebních profilů studentů zde byly identifikovány některé charakteristické vzorce chování. Jeden extrémní přístup volili studenti, kteří se neobávali nesprávné odpovědi. Uvědomovali si, že mohou cvičení opakovat mnohokrát, aniž by byly jejich neúspěšné pokusy někde zaznamenávány. Tito studenti vždy rychle zkoušeli odpovědět na otázky, i když často nesprávně, ale posléze se naučili, co bylo požadováno a poté odpovídali pečlivě a úspěšně. Jiný extrém byl pozorován u studentů, kteří se vyhýbali jakékoli možnosti nesprávné odpovědi, jako by se obávali, že to negativně ovlivní jejich hodnocení. Tito studenti raději vynechávali otázky, pokud si nebyli jisti svou odpovědí.
Výzkumníci identifikovali různé strategie jako úspěch-neúspěch, únik-úspěch, neúspěch-vyhýbání se, únik-vyhýbání se. Diskuze s učiteli potvrdily, že tyto vzorce chování byly charakteristické i pro výkony studentů při běžné výuce ve třídě. Nejvážnější extrém (únik-vyhýbání se) byl pozorován u studentů velmi nejistých a málo sebevědomých nebo naopak u těch, kteří se jednoduše vyhýbali práci. Schopnost počítačového systému definovat a měřit tyto nové pozorovatelné učební strategie

studentů může pomoci zvolit vhodný individuální přístup. Je-li systém vhodně koncipován, mohou si studenti sami vybrat, jaký přístup jim nejvíce vyhovuje. Systém TICCIT nabízel studentům volbu mezi grafickou či slovní formou, mohli si vybrat, zda preferují definice a pravidla či praktická cvičení a ukázkové příklady. Upřednostňovaný přístup bylo možno zvolit i pro nápovědu.

2. 4 Čtvrtá generace – sofistikované testování

Čtvrtá generace, **sofistikované testování** (Intelligent measurement = IM) je dalším rozšířením kontinuálního testování, které umožňuje automatickou a podrobnou interpretaci profilu studenta, a tak učitelům i studentům poskytuje závěry o charakteristikách studentova profilu a rady prostřednictvím adaptivních expertních systémů. Výsledky testování tvořící individuální profil mohou být uloženy jako databáze poznatků v symbolické podobě (knowledge base) a kdykoli vyvolány.

Sofistikované testování je definováno jako integrace znalostního inženýrství (knowledge-based computing) do jakéhokoli dílčího procesu testování ve vzdělávání. Užívá se v tomto kontextu spíše výraz znalostní inženýrství než známější termín umělá inteligence k zdůraznění, že znalosti a zkušenosti odborníků v testování mohou být uloženy v paměti počítače v symbolické formě zvané báze znalostí (knowledge base). Sofistikované testování umožňuje uchovávat znalosti, kopírovat je ve formě počítačového systému, který může interagovat s uživatelem jako odborný konzultant či poradce, a šířit odborné znalosti do mnoha míst. Příkladem použití programu expertní podpory jsou procesy vývoje testu (podpora analýzy úloh, specifikace testu, programů pro tvorbu úloh a testu), procesy administrace testu (individuálně zadávaných testů, skórování tvořených odpovědí, inteligentní tutoring) a analýza a výzkumné účely (podpora statistické analýzy, kalibrace experimentálních úloh, sběru dat pro studijní účely ve školním prostředí).

Čtvrtá generace předpokládá existenci počítače vybaveného znalostními expertními systémy (knowledge-based computing features, expert systems, Wainer 2000) ať už v rámci hardware nebo software. Hlavním rozdílem mezi čtvrtou a dřívějšími generacemi co se týče automatizace procesu zadávání testu spočívá ve schopnosti sofistikované interpretace jak statických měření, tak měření během dynamického vzdělávacího procesu.

Tři potenciální přínosy IM v administraci testu

Sofistikované testování může použít počítačovou inteligenci k
- skórování komplexních tvořených odpovědí zahrnutých v úlohách a referenčních úlohách
- vytváření interpretace založené na individuálních profilech skórů
- poskytnutí poradenství studentům i učitelům.

Inteligentní tutoriální systémy

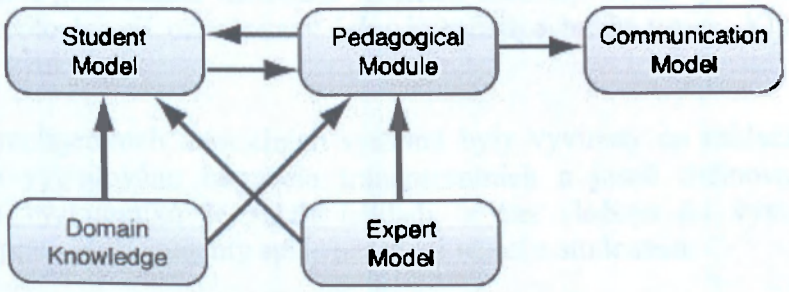
Inteligentní *tutoriální systémy* (intelligent tutoring systems, ITS) nabízejí značnou flexibilitu v prezentaci materiálu a větší možnosti odpovídat na individuální potřeby studentů. Tyto systémy dosáhly své „inteligence" díky pedagogickým rozhodnutím o tom, jak učit, tak jako informacím o studentech. ITS se ukázaly být vysoce efektivní ve zvyšování studentova pokroku v učení a motivace (Beck, Stern, Haugsjaa, 2004). Některé inteligentní tutoriální systémy lze považovat za prototypy systémů kontinuálního testování se znaky čtvrté generace (např. CAE kurz matematické logiky, LISP tutor; více Bunderson, Inouye a Olsen, 1989).

Součásti inteligentních tutoriálních systémů

Woolf (1992) identifikoval čtyři hlavní součásti: model studenta (student model), pedagogický modul (Pedagogical Module), modul znalostí (Domain Knowledge) a komunikační model

(Communication Model). Beck, Stern a Haugsjaa (2004) identifikovali ještě jednu součást, expertní model (Expert Model), který Woolf zahrnul do modulu znalostí.

Obr. 2-1 zobrazuje interakce mezi jednotlivými součástmi ITS.



Obr. 2-1 Interakce jednotlivých součástí ITS

Model studenta
Model studenta ukládá informace specifické pro každého individuálního studenta. Takovýto model sleduje, jak dobré výkony student podává. Může však také zaznamenávat neporozumění. Účelem modelu studenta je poskytnout údaje pedagogickému modulu systému, všechny nashromážděné informace by měly být přístupny tutorovi.

Pedagogický modul
Tato součást systému poskytuje model učebního procesu. Např. informace o tom, kdy přezkoušet, kdy začít novou látku a jakou látku zvolit, jsou kontrolovány tímto modulem. Vstupními daty jsou údaje z modelu studenta, a tak pedagogická rozhodnutí odrážejí různé potřeby každého studenta.

Modul znalostí
Modul znalostí obsahuje informace, kterým tutor vyučuje, a je tak nejdůležitější součástí systému. Zpravidla vyžaduje významné znalostní inženýrství k reprezentaci znalostní oblasti, tak aby jiné části systému ji mohly hodnotit.

Komunikační model
Interakce se studentem zahrnující dialog a okno obrazovky počítače jsou kontrolovány tímto modulem. Jak by měl být učební materiál co nejefektivněji prezentován studentovi?

Expertní model
Expertní model je podobný modulu znalostí v tom, že musí obsahovat informace, kterým jsou studenti vyučováni. Jde však o více než pouhou reprezentaci dat, jde o model, jak někdo školený v jisté oblasti reprezentuje znalost. Použitím expertního modulu může tutor srovnávat řešení studenta s řešením experta označením míst, kde má student obtíže.

Typy inteligentních tutorů
Existuje několik způsobů, jak dělit inteligentní tutoring systémy. Zpravidla se dělí na dvě dimenze: abstrakci učebního prostředí a znalostní typ výuky (Beck, Stern a Haugsjaa, 2004).

Mnoho systémů se pokouší poskytovat výuku simulací realistického pracovního prostředí, v kterém se student může učit úkoly. Realisticky simulované učební prostředí může redukovat jak náklady, tak rizika tréninku. Příkladem ITS založeném na simulaci je Advanced Cardiac Life Support (ACLS) Tutor (Eliot, Woolf, 1995), v kterém studenti přebírají roli vedoucího týmu při poskytnutí první pomoci pacientům postihnutým infarktem. Systém nejen že monitoruje studentovy činnosti, ale spouští realistickou simulaci pacientova stavu a udržuje prostředí, které odpovídá reálné životní situaci. Cílem tedy není jen testovat studentovy znalosti o postupech při první pomoci, ale také mu umožnit vyzkoušet si tyto postupy v praktičtějším prostředím než kterým je tradiční třída.

ITS se soustředí na učení jednoho typu znalostí. Nejbčžnější typ ITS učí procedurální poznatky (podle Blooma, 1956). Cílem je, aby se studenti naučili, jak provést konkrétní úkol. Takový systém je nazýván kognitivní tutor (cognitive tutor; např. SHERLOCK). Jiné ITS se soustředí na učení studentů konceptům a „mentálním modelům" (mental models). Tento typ modelů vyžaduje větší oblast znalostí, a proto bývají označovány jako knowledge based tutors. Kladou větší důraz na komunikaci během výuky.

Všechny moduly inteligentních tutoriálních systémů byly vyvinuty na základě výzkumů, ale jen několika z nich je vybudováno na zcela transparentních a jasně definovaných principech a modelech. Snahou výzkumníků je snížit náklady a čas vložené do vývoje těchto systémů a podporovat spolupráci mezi studenty spíše než mezi učitel a studentem.

Shrnutí vývoje využívání počítačů k testování

Cílem této kapitoly bylo popsat čtyři generace počítačového testování, jak je dělí Bunderson, Inouye a Olsen. Rozvoj počítačového testování byl umožněn pokrokem v informačních technologiích. Klíčovými technologiemi, které umožnily rozvoj počítačového testování jsou:

- pokles nákladů při současném nárůstu výkonu a úložné kapacity
- hardware a software pro síťovou komunikaci mezi pracovními stanicemi a stanicí, která uchovává výsledky
- vysokokapacitní paměťová média jako CD-ROM a DVD, která umožňují levnou distribuci kurikula a testových materiálů včetně videa, zvuku, grafiky a textu
- rozvoj WAN sítí, které umožňují centralizovanou správu testů a testových výsledků z více škol či regionů
- vývoj psychometrických postupů při kalibraci úloh a odhadu schopnosti jedinců (IRT)
- rozvoj znalostního inženýrství a expertních systémů pro interaktivní vyhledávání ve znalostních bázích.

Tyto technologie umožnily do jisté míry simulovat schopnosti, které byly dříve výlučně „lidské", napr. rozhodování, komunikace, porozumění. První generace zvýšila rychlost a přesnost administrace testů. Druhá generace poskytla novou teorii - adaptivní testování a zefektivnila administraci počítačových testů. Kalibrace úloh umožnila adaptivní výběr úloh během administrace testu. Ve třetí generaci se testování stává součástí procesu výuky a přestává být rušivým elementem ve výuce. Třetí generace nabízí studentům a učitelům průběžné informace o pokroku v dané oblasti v mastery mapě. Přesnější zpětná vazba je zajištěna pomocí referenčních úloh. Třetí generace by nedosáhla svých cílů bez nových nástrojů znalostního inženýrství označovaných jako čtvrtá generace.

Nyní se ukazuje, že současný trend ve využívání počítačů v testování se projevuje především v různých aplikacích v rámci distančního vzdělávání.

3 Využívání počítačů ve výuce

Stručný přehled možností využívání počítačů ve výuce je předmětem této kapitoly. Jsou zde uvedeny studie i staršího data o počítačem podporované výuce, protože vystihují účinnost počítačů ve výuce a ukazuje se v nich metodologie, která je vhodná pro syntézu výsledků experimentálních výzkumů – metaanalýza (viz např. Kulik a Kulik 1991, Kulik 1994).

Dodnes se provádějí nejrůznější výzkumy (InfoDev 2005) a stále se vedou četné diskuze (např. http://otec.uoregon.edu/arguments_against.htm) o tom, zda-li mohou počítače nebo jiné médium zlepšit vzdělávání (výkon žáků) či změnit metody, jak jsou žáci vzděláváni, zda má použití počítačů při výuce skutečně pozitivní vliv na učení, zda jsou počítače přínosem pro výuku. Výzkumná

zjištění jsou mnohdy velmi rozdílná. Jde také o otázky účelnosti a smysluplnosti nových technologií ve vztahu k obsahu kurikul, k učení žáků, k metodám výuky, k výsledkům vzdělávání aj. Nicméně ani v zahraniční literatuře nenajdeme dostatek uspokojujících odpovědí na tyto otázky (např. *Key Data on ICT in Schools in Europe* 2004).

V mezinárodní encyklopedii *Handbook on Information Technologies for Education and Training* (Adelsberger, et. al. 2002) najdeme sice popis nejnovějších prostředků informačních a komunikačních technologií (ICT), které mohou být nebo již jsou ve světě využívány ve vzdělávání (např. telelearning, web-based learning and teaching, tutorial software, virtual universities apod.), ale také realistický pohled na uplatňování ICT ve školním vzdělávání, který již takové nadšení nesdílí (Průcha 2006). Obdobné nálezy nacházíme i v mnoha jiných zahraničních zdrojích, např. v časopise *British Journal of Educational Technology*, který se soustavně věnuje problematice ICT ve vzdělávání. V roce 2006 zde byl zveřejněn článek odborníků z University of Pensylvania, který kriticky vyhodnocuje efekty počítačem podporovaného učení (computer-based learning, Wijekumar aj. 2006). Říká se v něm, že současné metaanalýzy dokládají, že počítačové technologie nemají zdaleka takový efekt na zlepšení učebních výsledků, jaký se očekával, naopak spíše minimální a někdy dokonce i negativní.

I americké studie z 90. let 20. století ukazují efektivitu a omezení použití počítačů ve výuce (Coley 1997). Výzkum největší světové organizace v oblasti testování, Educational Testing Service (ETS), z 90. let 20. století týkající se efektivity informačních technologií ukázal, že počátečni používání počítačů, např. použití výukového software k procvičování sčítání a odčítání, může být účinné. Ale pedagogičtěji komplexnější používání počítačů, např. používání internetu v malých skupinách, často nepřináší očekávané výsledky. Zdá se, že čím je výuka komplexnější a sofistikovanější, tím obtížněji ji lze hodnotit. Avšak v některých případech se ukazuje tento komplexní přístup jako slibný pro budoucí vyučování a učení.

V jiných zahraničních zdrojích však najdeme optimističtější nálezy. Na počátku 90. let 20. století shrnul J. Kulik z University of Michigan výsledky stovek výzkumných studií rozdílných rozsahů a různých vzdělávacích oblastí (základní, střední, vyšší vzdělávání a vzdělávání dospělých) o výhodách a nevýhodách využívání počítačů ve výuce. K výzkumu použil metaanalýzu a dospěl k závěrům, že učení se pomocí počítačů je ve většině případů účinné, zlepšuje výsledky studentů, i když v rozdílném rozsahu podle typu výuky (Kulik a Kulik 1991, Kulik 1994).[2] Během 10 let (do 1994) nashromáždil a prostudoval výzkumná zjištění z více než 500 samostatných studií (100 metaanalytických zpráv) o počítačem podporované výuce (computer-based instruction) a jejích účincích. Kromě toho shromažďoval a kriticky hodnotil dostupnou literaturu. Z velké části byly počítačové programy, které Kulik zkoumal, staršího data (vyvinuty před rokem 1990) a měly tendenci zdůrazňovat procvičování.

Pojem „metaanalýza" poprvé použil Gene Glass v roce 1976 při klasické syntéze prací o účinkách psychoterapie. Glass použil metaanalýzu k statistické analýze velkého souboru výsledků z individuálních studií za účelem jejich integrace. K měření používal tzv. index „*effect size*" (ES, *velikost účinku*). Index udává počet jednotek standardní odchylky, které oddělují výsledné skóry experimentálních (používajících novou metodu výuky – zde počítač) a kontrolních skupin (používajících běžné metody výuky). Velikost účinku se počítá jako podíl rozdílu průměrných skórů experimentální a kontrolní skupiny ($\mu_E - \mu_K$) a standardní odchylky měření (σ):

$$ES = \frac{\mu_E - \mu_K}{\sigma}.$$

[2] Metaanalýza je statistická analýza výsledků většího počtu srovnatelných studií za účelem jejich integrace a shrnutí výsledků. Cílem je porovnat výzkumy týkající se stejného tématu, vysvětlit možnou nekonzistentnost jejich závěrů a nakonec dojít k hodnověrnějšímu, syntetizujícímu závěru.

Např. v počítačově zadávaném testu SAT dosáhla skupina testovaných průměrného skóru 550 bodů (experimentální skupina), v běžně zadávaném testu SAT formou tužka-papír získala skupina testovaných průměrný skór 500 bodů. Standardní odchylka v SAT testu je 100. Tedy effect size je roven 0,5 ((550-500)/100).

Všechny zkoumané metaanalýzy ukázaly, že počítačem podporovaná výuka (i když ne všechny její typy) v průměru zlepšuje výsledky školní výuky, i když v rozdílném rozsahu.
Kulik dospěl po důkladném studiu metaanalýz k několika závěrům. Jsou-li studenti vyučováni za podpory počítače,

- obvykle se naučí více za kratší dobu (i když vliv počítače byl různě velký, vždy byl pozitivní). V průměru šlo o 34% redukci času při vysokoškolské výuce (17 studií) a o 24% redukci času při výuce dospělých (15 studií). Minimální průměrná velikost účinku byla 0,22 v 18 studiích ze společenských věd na základních a středních školách, maximální hodnoty 0,57 se dosáhlo v 18 studiích při výuce dospělých. Vážená průměrná velikost účinku 12 metaanalýz byla 0,35, což znamená, že průměrným účinkem počítačem podporované výuky bylo zlepšení výkonů studentů (jejich skórů) o 0,35 standardních odchylek nebo-li zlepšení z 50. na 64. percentil (Kulik a Kulik 1991);
- chodí do vyučovacích hodin raději. Průměrný účinek počítačové výuky v 22 studiích zlepšil postoj k vyučování o 0,28 standardních odchylek;
- mají pozitivnější postoje k počítačům (protože jim pomáhají při výuce). Průměrná velikost účinku v 19 studiích vzhledem k postoji k počítačům byla 0,34.

Nicméně se neprokázalo, že by počítače měly pozitivní účinky ve všech zkoumaných oblastech. Průměrný účinek počítačem podporované výuky ve 34 studiích, které prověřovaly studentův postoj k učební látce, se například pohyboval blízko nuly.
Všechny studie, provedeny na základních a středních školách, byly kvantitativního rázu. Výsledky ze tříd, kde probíhala výuka s pomocí počítače, byly porovnávány s výsledky ze tříd s běžnou výukou. Studie lze rozdělit podle způsobu používání počítače do několika skupin.

Dříve byly rozlišovány čtyři způsoby použití počítačů ve výuce: *procvičování, tutoriál, dialog a management* (např. Atkinson 1969). Pozdější taxonomie některé kategorie zrušily a přidaly nové. Taylor (1980) například rozlišoval tři způsoby použití počítačů ve školách: počítač jako *tutor, nástroj (tool) a konzultující (tutee).* Jako *tutor* počítač prezentuje učební materiál, hodnotí studentovy odpovědi, určuje, co se dále bude prezentovat a zaznamenává studentův pokrok. Většina použití počítačů ve výuce z dřívějších taxonomií spadají právě do této kategorie. Počítač slouží také jako *nástroj* (*tool*), studenti ho používají k statistickým analýzám, výpočtům nebo zpracování textu. Jako *konzultující* (*tutee*) počítač přijímá příkazy od studentů v programovacím jazyce, kterému rozumí.
Slavin (1989) zastával jiný způsob pohledu na inovace ve výuce. Inovace definoval různými stupni přesnosti. Prvním stupněm definuje nejasné inovace, které předkládají pouze zmatené, nepřesné modely pro praxi výuky a nemají jasný pojmový základ. Druhý stupeň představují zřetelněji specifikované inovace, které mají obvykle snadno popsatelný pojmový základ, ale v praxi jsou realizovány různými způsoby (podle Slavina např. kooperativní učení, mastery learning či individualizovaná výuka). Až třetí stupeň obsahuje precizně definované přístupy zahrnující specifické výukové materiály, dobře rozvinuté procvičovací postupy pro učitele a detailní normativní manuály (např. Stanfordův CCC-program).

Počítačem podporovaná výuka spadá podle Kulika do prvního stupně, protože se vztahuje na různé postupy s různými pojmovými základy. Nicméně některé dobře definované kategorie použití počítače lze zahrnout do druhého stupně. Jednou z nich je počítačový tutoring (computer tutoring). Většina programů počítačového tutoringu vychází ze Skinnerovy programované výuky (např. Pelikán 1998). Třetí stupeň inovací zahrnuje běžné výukové materiály, procvičovací postupy atd.

Z dnešního pohledu bychom počítačem podporovanou výuku asi zahrnuli spíše do druhého, výjimečně i do třetího stupně.
Kulik setřídil 97 metaanalytických studií podle hlavních typů použití počítačů. Nejprve stanovil ES u všech studií. Průměrná velikost účinku celé skupiny 97 studií je 0,32. To naznačuje, že průměrný student, který obdržel počítačovou výuku, podal výkon na 63. percentilu, zatímco průměrný student, který prošel běžnou výukou, se nacházel na 50. percentilu. Velikosti účinku (ES) lze také interpretovat v podobě měsíců na ekvivalentní škále. ES = 0,32 odpovídá přibližně 3 měsícům. Standardní odchylka rozložení všech velikostí účinku je 0, 39. To znamená, že přibližně dvě třetiny všech studií měly velikost účinku mezi - 0,1 a 0,7 a že 95% všech výsledků spadalo mezi -0,4 a 1,1.

Kulik stanovil na základě studií šest typů použití počítačů ve výuce:
- *tutor* (*tutoring*), který zahrnuje procvičování a tutoriál z dřívějších taxonomií a odpovídá kategorii Taylora počítač jako tutor;
- *řízení učení žáka* (*managing*), kdy počítač hodnotí studenty buď on-line nebo off-line, vede je k vhodným výukovým prostředkům a zaznamenává jejich pokrok v učení;
- *simulace*, kdy počítač generuje data, která vyhovují specifikacím studenta a prezentuje je numericky či graficky, aby ukázal vztahy v modelech;
- *obohacování výuky* (*enrichment*), při kterém počítač poskytuje studentovi relativně nestrukturovaná cvičení různých typů (hry, simulace, tutoring aj.), aby obohatil jeho zkušenost z výuky a motivoval ho;
- *programování* (*programming*), kdy studenti píší krátké programy v jazycích BASIC a ALGOL pro řešení matematických problémů. Očekává se, že tato zkušenost s programováním bude mít pozitivní účinky na studentovy schopnosti řešit problémy a na pojmové porozumění matematice;
- *programování v logech* (logo programming), kdy studenti dávají počítači instrukce ve značkách (tzv. logech) a sledují výsledky na obrazovce počítače. Očekává se, že to studentům pomůže především při řešení problémů, plánování a předvídání následků.

Velikosti účinku se u jednotlivých typů liší. Nejvyšší průměrnou ES mělo logo (0,58) a tutor (0,38), nejmenší programování (0,09) a simulace (0,10). Jediným typem aplikací počítače ve výuce s jen pozitivními výsledky na základních a středních školách byl počítačový tutoring. Na druhou stranu výsledky ze základního a středního školství jsou nevýrazné u jiných aplikací počítače ve výuce: řízení výuky, simulace, obohacování výuky (enrichment) a programování. Efektivita použití počítačů byla srovnatelná s výukou pomocí tištěných materiálů. Ani programování obvykle nemělo pozitivní efekt na učení studentů v matematice. Použití počítačových simulací v přírodovědných předmětech vedlo také k malému účinku na učení se přírodním vědám na základních a středních školách.

V roce 1994 zveřejnila Software Publishers Association (Sivin-Kachala; Bialo 1994) výzkumnou studii jedné nezávislé konzultantské firmy, která analyzovala 176 studií z let 1990 až 1995 o efektivitě nových technologií ve školách. Studie ukázala, že výkony studentů včetně těch se speciálními potřebami (od předškolního po vyšší vzdělávání) se ve všech hlavních předmětech zlepšily. Důsledně se zlepšovaly i postoje studentů k učení a jejich vlastní pojetí, když byly ve výuce používány počítače. Závěrem studie bylo, že použití nové technologie jako učebního nástroje může dělat měřitelný rozdíl ve výkonu studentů, v jejich postojích a interakcích s učiteli a ostatními studenty.

Jiné výzkumy ukázaly, že když školy začaly používat počítače ve výuce, došlo ke zlepšení v docházce studentů a v počtu propadajících. Dále se zjistilo, že se na studenty kladou větší nároky, že jsou více zaměstnaní prací a nezávislejší. Studenti byli povzbuzeni k experimentování a zkoumání nových hranic znalostí použitím počítače, získali více zodpovědnosti za své úkoly a odváděli kvalitnější práci. Studie dokazují, že počítačem podporovaná výuka může výuku

individualizovat a dát studentům okamžitou zpětnou vazbu či dokonce vysvětlit správnou odpověď. Počítač je nesmírně trpělivý a objektivní, což je pro studenty motivující.

Alessi; Trollip (2001) zastávají názor, že správné použití počítače ve výuce je tehdy, kdy se tím zjednodušuje či zefektivňuje proces učení. Mezi tyto situace patří takové, kdy náklady na vyučování za použití jiných metod jsou příliš vysoké (např. vojenský výcvik), kdy je v centru zájmu bezpečnost (např. chemické laboratoře), kdy je zbytečně náročné vyučovat jinými metodami (např. grafické znázornění), kdy je potřeba rozsáhlého individuálního procvičování žáků (např. procvičování slovní zásoby a gramatiky v cizích jazycích), kdy žákům chybí motivace (např. při učení starověké světové historie) apod. Žádná z těchto situací však nezaručuje, že počítač bude jako nástroj ve výuce účinný.

Roy Pea (ředitel SRI's Center for Technology in Learning in Menlo Park v Kalifornii) se dívá na způsob použití počítačů ze sociálního kontextu, který je podle něj rozhodující k porozumění toho, jak technologie mohou ovlivnit vyučování a učení. To znamená, že by mělo být věnováno více pozornosti výukovým strategiím používaných jak v počítačovém software, tak i „kolem něj" v prostředí třídy. Ani sebelepší software nemusí mít pozitivní účinky na výuku, pokud je použit nesprávně.

3. 1 Přínos využívání počítačů v klasicky koncipované výuce

Počítače lze využít v jedné, ve více či ve všech čtyřech hlavních fázích výuky: při prezentaci učiva, řízení činnosti studenta, procvičování i hodnocení studenta včetně plánování výuky.

Při **prezentaci učiva** může počítačový výukový program nahradit učitele v mnoha směrech. Může prezentovat nejen textové a obrazové instrukce příp. doprovázené zvukem, ale také zprostředkovat dynamické zobrazení pokusů ve zvětšeném nebo zmenšeném měřítku a navíc umožňuje, aby se samostatně učící student kdykoliv mohl k předchozím informacím vrátit. Informace jsou prezentovány především na příkladech. Hlavní roli v této fázi plní počítač či učitel. Ne vždy musí výuka začít prezentací informací.

Řízení činnosti studenta je více interaktivní povahy než první fáze a zahrnuje jak činnost studenta, tak použití počítače. Řízení studenta je zpravidla založeno na řešení úloh (aplikaci pravidel a principů, procvičování procedurálních dovedností) a problémů. Učitel či interaktivní médium pozoruje žáka a v okamžiku, kdy se dopouští chyb, mu poskytuje rozsáhlejší pomoc - opravuje jeho chyby a dává mu pokyny, jak se zlepšit. Ve třídě často učitel pokládá žákovi otázky, které musí zodpovědět. Když odpoví chybně, buď ho učitel opraví a řekne mu správnou odpověď nebo mu pokládá jiné otázky, aby žáka dovedl ke správné odpovědi. Když se někdo učí z knihy, někdy jsou v ní zahrnuty otázky nebo navrženy aktivity, ale oproti učení se ve třídě, když žák neodpoví na otázky správně, nikdo ho neopraví (někdy jsou uvedeny správné výsledky, ale většinou ne vysvětlení). Řízení činnosti studenta je ve vyučování důležité, protože nikdo se nenaučí všechno sám. Žáci dělají chyby a často si ani neuvědomují, že je dělají. Proto je třeba je na ně upozornit. Pokud žáci sami mají přijít na jisté pravidlo či vyřešit problém, je účinnější, když je jejich aktivita usměrňována.

Procvičování je třetí fází ve výuce. Vedle působení učitele nebo počítače zde hlavní roli hraje student. Vyžaduje se od něj, aby uplatňoval předtím získané dovednosti či prováděl činnosti rychle nebo plynule, někdy ve ztížených podmínkách, a přitom se nedopouštěl žádných nebo jen málo chyb. Někdy se vyžaduje, aby uplatňování dovedností přešlo dokonce ve vytvoření návyků. Ačkoli učitel nebo interaktivní médium sleduje činnost studenta a provádí opravy chyb, jež doplňuje velmi krátkým komentářem, těžištěm této fáze je procvičování žáka, v centru pozornosti tedy stojí žák. Plynulost a rychlost spolu sice souvisí, ale jde o trochu rozdílné aspekty dobře naučené informace. Plynule ovládat nějakou dovednost neznamená jen ji provádět rychle, ale provádět ji bez přemýšlení, automaticky (např. při učení se cizím jazykům, čtení, psaní, řízení auta, počítání).

Na druhou stranu některé informace nevyžadují plynulost (psaní kritiky, provádění chemického experimentu).

První tři fáze výuky zakončuje čtvrtá fáze a tou je **hodnocení studenta**. Nemůžeme předpokládat, že výuka byla stejnou měrou úspěšná pro všechny studenty, proto výsledek učení musí být zjištěn a hodnocen pomocí prostředků, mezi které patří také didaktické testy. Výsledky testů poskytují zpětnou vazbu, tj. informace nejen o úrovni učení se, ale i o úrovni vyučování, a tak mohou být východiskem pro budoucí změny v projektování a realizaci výuky. Počítač může sloužit ke generování, zadávání a vyhodnocování testů, a to jak pro jednotlivé studenty, tak pro skupiny studentů.

3. 2 Oblasti výuky vhodné pro využívání ICT

1 Počítač lze využít jako inteligentní nástroj v procesu poznávání

- ve výuce lze využít počítač k řešení problémových úloh (k strukturování a analýze pro zjednodušení), tj. k rozvíjení myšlení a dovedností žáků, k modelování situací a objektů (v přírodních vědách, matematice, medicíně), k simulaci (např. fyzikálních pokusů);
- počítač může žákům sloužit k získávání zkušeností v uměle vytvořených prostředích jako by byla reálná (virtuální realita), žáci zkoumají chování objektů (např. cestují lidským tělem a pozorují, jak fungují orgány v těle, apod.);
- počítač usnadňuje složité výpočetní úlohy, jejichž řešení by trvalo hodně dlouho (např. při stavbě lodí, letadel; Devlin 2005).

Zařazení nejnovějších poznatků různých vědních a technických oborů získaných pomocí výpočetní techniky do školního kurikula je zatím stále budoucností. Výzvou jsou symbolické systémy (manipulace se symboly, ikonami) s využitím vizuálního a funkčního programování.

2 Počítač jako podpora výuky (computer based education, CBE či computer assisted instruction, CAI)

- východiskem pro CAI byly v 70. letech umělá inteligence, kognitivní vědy a moderní technologie, což vedlo ke vzniku vzdělávacího prostředí SOPHIE (SOPHisticated Instructional Environment) a tutorů pro výuku algebry, geometrie;
- tutor – Euklidova geometrie. Výzkum ukázal, že výuka pomocí PC tutora byla dvakrát až třikrát efektivnější než běžná výuka (Molnar 1997);
- v současné době se stále více využívá internetu, sofistikovaných systémů typu LMS (Learning Management System), které učitelům nabízejí nástroje pro vývoj, distribuci vzdělávacích obsahů, komunikaci, monitorování a hodnocení práce žáků a žákům nástroje na podporu učení a vzájemnou spolupráci. Systémy tohoto typu se již implementují na dnešních školách a stávají se součástí jejich vzdělávacího prostředí (Mudrák 2005).

Pojmem e-learning je obecně označován proces učení podporovaný elektronickými technologiemi (které kromě ICT zahrnují i audiovizuální prostředky jako televizi, film, zvukové nosiče, ...). Obvykle se však e-learning chápe v užším slova smyslu jako učení podporované ICT (Zounek 2006), zejména pak internetem (WBT, *Web Based Training*) případně CD-ROM (bývá označován jako CBT, *Computer Based Training* či CAL, *Computer Assisted Learning*). e-learning je velmi rozšířený ve sféře distančního a doplňkového vzdělávání dospělých.

Zkušenosti s e-learningem vedly k přehodnocení očekávání, která se původně s e-learningem spojovala a obvykle bývá kombinován s dalšími formami výuky (tradiční výukou, workshopy, telekonferencemi). Tento přístup je označován jako „blended learning" (kombinované smíšené vzdělávání, Černochová 2006, Zounek 2006 aj.). Ve světě dnes již existuje řada tzv.virtuálních škol. Studium čistě formou e-learningu ve virtuální škole není vhodné pro každého. Nutným

předpokladem kromě vybavení a dovedností potřebných k používání ICT je silná motivace a disciplína. Virtuální škola může být vhodná např. pro dlouhodobě nemocné žáky či žáky, kteří z jiných důvodů nemohou pravidelně navštěvovat školu, může být vhodná i pro nadané žáky nebo pro žáky v učení velmi samostatné.

3 Počítač jako nástroj ke komunikaci

Počítač slouží ke zpracovávání informací a jejich sdílení, k spolupráci, provádění běžných činností (k nimž člověk dosud potřeboval tužku, papír, pravítko).
ICT podporuje efektivní komunikaci učitele a žáka, žáků mezi sebou i rodičů. Tato komunikace nemusí být omezená na pouhou distribuci informací (např. e-mailem či pomocí informační nástěnky). Je žádoucí povzbuzovat vlastní aktivitu žáků – aktivní komunikace s učiteli, experty na nejrůznější oblasti, žáky ostatních škol (např. v rámci výukových projektů). Setkání žáků s odborníky není mnohdy z časových a organizačních důvodů možné, s pomocí ICT (videokonference, chat, diskusní fóra) lze však žákům zprostředkovat setkání s odborníky, se kterými by se nebylo reálné setkat. Uspořádání takovéhoto virtuálního setkání je obvykle mnohem efektivnější než osobní návštěva spojená se zdlouhavou přepravou.

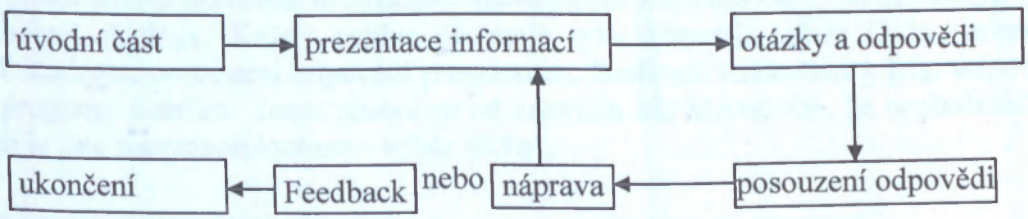
4 Počítač pomáhá řešit problém propojování informačních zdrojů

Počítač (internet) nám umožňuje vyhledávat a ukládat na web nejrůznější materiál (publikace, zprávy, aj.), propojení odborných záznamů a zdrojů a zpřístupnění výsledků lidského poznávání (např. wikipedia http://www.wikipedia.org/, http://cs.wikipedia.org/ či četné odborné encyklopedie http://planetmath.org/encyclopedia/, http://hyperphysics.phy-astr.gsu.edu/).

3. 3 Metody využívání počítačů k výuce a učení

Alessi, Trollip (2001) uvádějí řadu metod, jak pomocí **interaktivních multimedií** (Interactive Multimedia IMM) usnadnit proces učení. Jde o tutoriály[3] (tutorials), multimediální vzdělávací programy (hypermedia), procvičování, simulace (simultations), hry (games), nástroje (tools), prostředí pro otevřené učení (open-ended learning environments), testy (tests) a učení ve webovém rozhraní (web-based learning).

Tutoriály (dříve se jim říkalo *teachware*) jsou vzdělávací programy, které se zpravidla uplatňují v prvních dvou fázích vyučovacího procesu. Přebírají úlohu učitele při prezentaci informací a řízení činnosti žáka při počátečním osvojování. Obr. 1 znázorňuje strukturu typického tutoriálu (Alessi, Trollip 2001).



Obr. 3-1 Struktura tutoriálního programu

Tutoriál začíná úvodní částí, v které je student informován o účelu a povaze programu. Po tomto úvodu začne cyklus. Zobrazí a rozvine se informace. Student musí odpovědět na otázku. Program jeho odpověď posoudí, aby vyhodnotil porozumění či dovednost a poskytne mu feedback ke zlepšení stávajícího a dalších výkonů. Na konci každé iterace, udělá program postupně rozhodnutí, aby určil, jaká informace by měla být předložena v další iteraci. Cyklus pokračuje, dokud není výukový proces ukončen buď studentem nebo samotným programem. Ukončením na

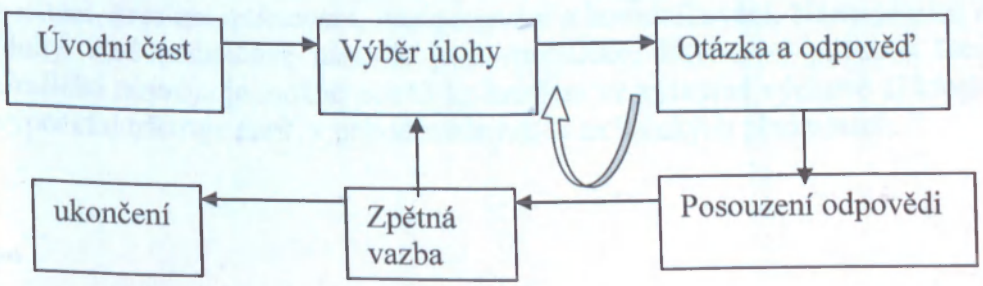
[3] Tutoriály odpovídají dřívější programované výuce.

konci cyklu se míní shrnutí a závěrečné poznámky. Ačkoli se ne všechny tutoriály zabývají všemi těmito aktivitami, nejefektivnější z nich zahrnují tyto či podobné komponenty.

Multimediální programy (hypermedia) jsou jinou metodologií prezentace či získávání informací, ale jsou vytvořeny pro otevřenější učení nebo učení se objevováním. Jsou méně strukturovány než tutoriály, a tak umožňují žákům vybrat si své cesty v procházení učebním materiálem (informacemi). Hypermédia pracují s hypertextem, což je text s odkazy na další informace, odkazy zprostředkovávají vazby mezi jednotlivými částmi většinou rozsáhlejšího textu, a kombinují ho se zvukem, obrazem či videem. Ačkoli myšlenka hypermedií se objevila již v 70. letech minulého století, až v 80. letech 20. století mohla být díky rozvoji informačních technologií uvedena do praxe. Dnes je hypertext a hypermédia běžně používán na webu, CD-ROMech a jiných digitálních médiích. Hypermédia představují knihy nové generace v elektronické podobě, zahrnují text, fotografie, video a audio nahrávky. Hypermédia na CD-ROMech a na webu obsahují informace klasických učebnic, encyklopedií a literárních prací a přidávají k nim audio, video, animaci a další formy zobrazování informací.

Zajímavý výukový multimediální portál lze nalézt na http://www.agutie.com/ *(„Geometry from the Land of the Incas"). Portál nabízí neobvyklý pohled na geometrii zábavnou až hravou formou. Je určen žákům i studentům a získal celou řadu ocenění ze strany odborných společností sdružujících učitele matematiky.*

Nacvičování (drills) – neustálé opakování na úlohách (např. násobilka, slovíčka), dokud látka není zcela osvojena) a většina výukových her pomáhají žákům zejména ve třetí fázi procvičit látku do trvalého a plynulého osvojení. Drill opakuje látku, dokud se ji žák zcela nenaučí. Počítačové drily bývají často kritizovány. Ačkoli někteří částečně drily brání (např. Decoo 1994), mnoho jiných teoretiků působících ve vzdělávání tvrdí, že drily nevyužívají potenciál počítačů (např. Streibel 1986, Jonassen 1988) a mohou být snadno uskutečněny i učebnicemi nebo pomocnými kartičkami s informacemi (flashcards). Stejně jako Alessi a Trollipovi (2001) se mi toto tvrzení zdá být nespravedlivé, neboť alespoň pro část žáků je aktivita spojená s počítačem zábavnější než procvičování pomocí papírových kartiček. Často je také drilům vyčítáno, že jich existuje příliš mnoho a jen málo jich je kvalitních; že neučí ničemu novému, ale jsou pouze procvičením pro ty studenty, kteří již částečně učební látku ovládli. To je sice pravda, ale je to povahou drilů, které nejsou rozšířením naučeného ve smyslu poskytování nových informací. Ve světě interaktivních multimédií předchází drilům často vhodný tutoriál nebo simulace. I když mnoho komerčně dostupných multimediálních drilů je méně kvalitních, jsou drily užitečné a zřejmě i nezbytné pro efektivní učení a neměly by být vynechávány.
Obr. 3-2 znázorňuje obecnou strukturu drilu (Alessi, Trollip 2001). Většina drilů, stejně jako tutoriály, mají docela pravidelnou strukturu skládající se z úvodní části následovanou cyklem, který se mnohokrát opakuje. Každý cyklus obsahuje tyto činnosti: výběr úlohy, zobrazení úlohy, odpověď studenta, posouzení odpovědi programem, feedback studentovi k jeho odpovědi. Po řadě úloh je program ukončen. Tento proces se od tutoriálu liší hlavně tím, že nepředkládá informace. V drilech je toto nahrazeno krokem - výběr úlohy.



Obr. 3-2 Obecná struktura drilů (↻ označuje opakující se cyklus)

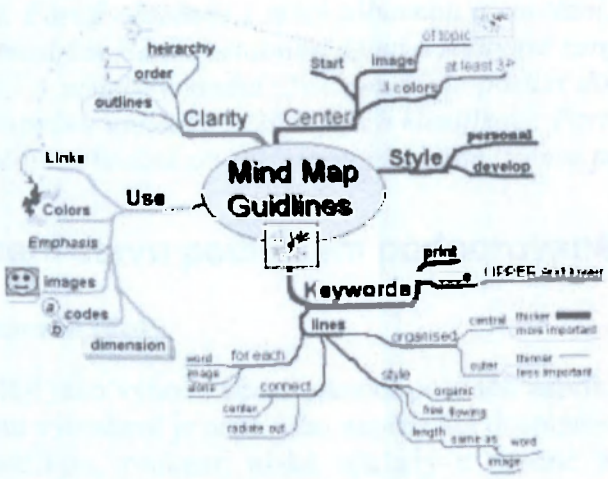
Nacvičování je často kombinováno z motivačních důvodů s hrami. Některé **výukové hry** se neopakují, takže neprocvičují látku do suverénního ovládnutí tak jako drily. Hry jsou významným vzdělávacím nástrojem, pokud jsou vhodně použity, protože mají motivační vliv na děti a v některých oblastech i na dospělé. Hry mohou být také dobré pro integraci učení napříč několika předmětovými oblastmi. Mohou přímo přispět k učebním cílům jako je soutěživost, spolupráce a týmová práce. Nicméně vytvoření úspěšné vzdělávací hry není snadné. Vzdělávací hry musejí splňovat tři základní požadavky: musejí mít užitečné učební cíle, být zábavné a herní cíle (vítězství) musí upevňovat cíle učební. Pokud jsou tyto požadavky splněny, nenahradí je žádné náročnější multimédium. Tvorba vzdělávacích her vyžaduje mnoho času a úsilí.

Simulace jsou komplikovanějším multimédiem, mohou být použity k prezentaci informací, k řízení činnosti žáka, k procvičování učiva, ale i hodnocení žákových znalostí. Jejich popularita roste. Alessi a Trollip dělí simulace do dvou skupin podle toho, zda jejich hlavním vzdělávacím cílem je učit něčemu nebo učit, jak něco dělat. Každou skupinu dále dělí do dvou subkategorií. Do první skupiny patří fyzické (physical) a iterativní simulace, do druhé procedurální a situační simulace. Simulace plně využívají potenciálu počítačů. Simulace jsou oproti tutoriálům a drilům zdokonalené v oblasti motivace, lepším přenosem učení, větší efektivitou a flexibilitou. Mají výhody praktičnosti, bezpečnosti a kontrolou nad reálnými zkušenostmi, poskytují dobrého předchůdce těmto zkušenostem. Mohou studentům zprostředkovávat nevšední zážitky. Na druhé straně jsou simulace nejnáročnější multimediální metodologií co do plánování a vývoje. Tvůrci potřebují více porozumět obsahu a studentům, musí se zabývat mnohými komplexními faktory a zvládnout náročnější programování k realizaci simulačního modelu a zasadit ji do efektivního učebního programu. Simulace lze kombinovat s hrami k rozvíjení učení pomocí objevování.
Některé instituce (NASA, CERN, MFF UK) nabízejí i anonymním uživatelům nabízejí možnost provádět reálné experimenty s využitím vzdálené laboratoře případné vzdálená měření. Do jejich průběhu lze zasahovat pomocí www rozhraní. Virtuální laboratoře, ať už provozované na vzdáleném serveru nebo lokálně na pracovní stanici ve třídě, mají široké možnosti užití ve výuce přírodních věd, zejména fyziky, chemie a biologie. Virtuální či vzdálené laboratoře (VLE Virtual Learning environment) přinášejí do výuky nové a užitečné prvky, na druhé straně není žádoucí z výuky eliminovat reálné snadno proveditelné experimenty (Bílek; Turčáni 2006).

http://www.xperiment.se/TaiwanUniv/ntnujava/index.html
Rozcestník koncipovaný jako virtuální laboratoř, vytvořený katedrou fyziky taiwanské Normal University. Obsahuje přes 50 simulací rozdělených dle oblastí fyziky, např. simulace zastavení auta při zadané počáteční rychlosti, tření povrchu atd.

Nástroji se rozumí počítačový software, který žáci používají společně s jiným médiem nebo aktivitami pro dosažení výukového cíle, tedy k učení, organizování a porozumění novým dovednostem nebo znalostem. V této souvislosti se staly populárními tzv. mindtools (Jonassen 2000), což je software, který zlepšuje kritické myšlení, posiluje kognitivní funkce, pomáhá při reorganizaci znalostí, je obecně aplikovatelný v různých situacích a je kontrolován studentem. Navíc mindtools[4] zpravidla pomáhají rozvíjet spolupráci, aktivní a konstruktivistické učení a jsou používány v autentických učebních kontextech. Počítačovými nástroji jsou např. nástroje na psaní, počítání, kreslení, plánování, komponování a komunikování. Nástroje nám mají pomoci řešit různé úlohy. Jiné počítačové nástroje jsou specifické. Jsou svou povahou otevřenější a flexibilnější. Grafické nástroje je možné použít ke kreslení ve výtvarné výchově či kreslení grafů v matematice, výpočetní nástroje např. v přírodovědných či technických předmětech.

[4] Více o mindtools na http://www.mindmapoptions.com/



Obr. 3-3 mind map z wikipedia.org

Prostředí pro otevřené učení (POU) jsou prostředí ke zkoumání (Hannafin, Land, Oliver 1999), která umožňují studentům určit si cíle a usilovat o jejich dosažení použitím metod, které považují za vhodné a žádoucí. Tato prostředí zdůrazňují řešení smysluplných problémů, experimentování, interpretování, analyzování celku spíše než jeho částí, nahlížení na problémy z různé perspektivy, učení se z chyb, testování a korigování znalostí a obvykle spolupráci s ostatními studenty. Student si sám může zvolit, jak a co se bude učit. Dobrá POU zahrnují motivační scénáře, přirozené a snadné ovládání, nástrojový software pro manipulaci a komunikaci a prostředky jako jsou databáze, multimediální knihovny a encyklopedie. Ačkoli mohou být použita k získávání a vytváření znalostí, jsou často prostředím, v kterých si žák procvičuje aplikování nových znalostí. Jde o progresivní, současně však i poměrně náročnou techniku učení, která klade vysoké nároky na učitele i žáky.

http://www.modern-education.net/aktivity/skoly/krajinazaskolou.asp
Jde o projekt vytvářený soukromou firmou Centrum moderního vzdělávání (CZ), s.r.o.
Projekt je zaměřený na rozvoj klíčových kompetencí žáků a studentů. Krajina za školou je projektová výuka s mezipředmětovými vazbami využívající moderní informační a komunikační technologie. Celý projekt je postaven na myšlence porovnání dvou fotografií, které byly pořízeny ze stejného místa, ovšem v různém čase (jedna fotografie stará 50 či více let a druhá z aktuální doby). Žáci tak získají plastický obraz změn, které nastaly v krajině, městě, kultuře. Tyto změny reprezentují historický vývoj, který budou studenti v rámci projektu sami objevovat, vnímat a popisovat. Autoři předpokládají realizaci projektu na celoevropské úrovni a vznik široké databáze konkrétních a detailně zdokumentovaných proměn krajiny, klimatu, architektury a kultury obecně. Databáze bude k dispozici všem školám, které budou moci data využít v rámci výuky.

Testy téměř vždy zastupují poslední fázi vyučovacího procesu, hodnocení toho, co mělo být naučeno. Zjišťují, co se žák naučil a v jakém rozsahu. Výjimkou jsou procvičovací testy či kvízy, které se běžně používají v procvičovací fázi. O testech je podrobně pohovořeno v kap. 4.

Učení ve webovém rozhraní se často kombinuje s ostatními metodologiemi, většinou s multimediálními programy. Lze jej použít ve všech fázích vyučovacího procesu.

http://skolazaskolou.cz
Jde o portál vytvářený soukromou firmou Centrum moderního vzdělávání (CZ), s.r.o.
http://www.modern-education.net/index.asp
Nabízí e-learningové kurzy doplňující předepsaný obsah výuky v českých školách (druhý stupeň ZŠ a střední školy). Klade si za cíl pomoci s domácí výukou, zejména domácí přípravou a procvičováním učiva před testováním ve škole. Portál nabízí také možnosti

on-line komunikace žáků s vrstevníky, kteří řeší stejný problém nebo si prohlubují znalosti stejné studijní látky. Portál obsahuje i sekci zábavnou a soutěžní, kde mohou projevit svoji soutěživost ve vědomostním klání s ostatními dětmi o zajímavé ceny.
Portál byl rozšířen i o učitelský modul. Ten umožňuje posílat domácí úkoly e-mailem, tisk testů a podporuje i správu výsledků žáků a jejich klasifikaci. Portál umožňuje také podporu výuky přímo ve vyučovací hodině po zakoupení příslušné licence pro školu.

3. 4 Zhodnocení stavu počítačem podporované výuky

Výhody počítačem podporované výuky

Kulik a Kulik (1991) uvádějí jako výhodu učení pomocí počítačů například úsporu času a možnost zjednodušení učení. Dalšími výhodami je např. jeho nepřetržitá dostupnost (počítač je vždy na svém místě); motivace žáků grafikou, zvukem; nízké náklady a značné výpočetní možnosti; velká kapacita ukládání na harddisk či jiné výměnné médium (disketa, CD-ROM, DVD, flash disk); vývoj multimédií umožňuje prezentaci učiva pomocí videa, audio nahrávek, počítačové grafiky aj. Výhodou je i možnost připojení více počítačů do sítě ke komunikaci mezi pracovními stanicemi a sběru a vyhodnocování dat např. v centrálním počítači sítě, k rychlému rozšíření kurikula a testových materiálů za nízké náklady.

Nevýhody využívání počítačů ve výuce

Nevýhodou využívání počítačů ve výuce je, že nepřispívá k sociálnímu učení, nedá se zpravidla použít kooperativního učení. Počítač může žákovi poskytnout pouze odstupňovanou pomoc, nemůže ale reagovat na všechny jeho chyby. Navíc ne na všech školách je zatím dostatek počítačů a ne všechny školy jsou připojeny k internetu (v současné době je vybavena a k internetu připojena zhruba polovina škol v ČR, Černochová 2006).

Využívání počítačů ve výuce na českých školách

V ČR již řada škol (od základních po vysoké) přistoupila k implementaci LMS systému. To sice samo o sobě není podmínkou postačující pro efektivní využívání ICT ve výuce, ale z koncepčního hlediska je to krok správným směrem. Nejčastěji užívanými LMS jsou Moodle, eDoceo, WebCT či Microsoft Class Server.

E-learningový systém Moodle na českých VŠ

Moodle je softwarový balík určený pro podporu prezenční i distanční výuky prostřednictvím online kurzů dostupných na WWW. Systém umožňuje či podporuje snadnou publikaci studijních materiálů, zakládání diskusních fór, sběr a hodnocení elektronicky odevzdávaných úkolů, tvorbu online testů a řadu dalších činností sloužících pro podporu výuky. Moodle je software volně šiřitelný na základě GNU licence s otevřeným PHP kódem. Běží na každém operačním systému, který podporuje PHP (Unix, Linux, Windows, Mac OS X, Netware). Všechna data jsou ukládána v jediné databázi. Systém Moodle se úspěšně prosazuje na řadě vysokých školách v ČR i ve světě. Inovace vyvinuté v rámci projektu budou dostupné celé komunitě uživatelů tohoto systému.

Na ČVUT bylo rozhodnuto o implementaci centrálního řídícího systému pro výuku LMS (Learning Management System) resp. CMS (Course Management System). Jeho výběr nebyl jednoduchou záležitostí a v průběhu let škola vystřídala několik systémů. Na některých pracovištích byl používán WebCT, později i levnější ClassServer. Ten však nebyl příliš vhodný pro vysokoškolské prostředí. V současnosti je využíván systém Moodle (stejně jako na UK a dalších VŠ).

Výhled do budoucna

Počítačem podporovaná výuka je již dnes využívána na většině českých vysokých škol ať už pro denní nebo distanční studium. Uznávaným standardem se stává open source LMS Moodle. Situace na středních a základních školách je závislá na materiálních podmínkách dané školy, ale zejména na entuziasmu učitelů, kteří se moderním výukovým technologiím věnují často nad rámec svých pracovních povinností. Všeobecně lze říci, že v ČR byly učiněny významné kroky v zavádění ICT do výuky, ale k tomu, aby se tyto metody staly na našich školách naprosto běžnou součástí výuky vede ještě dlouhá cesta.

4 Tvorba didaktického testu

Pojmu „test" se používá nejen v pedagogice a v psychologii, ale i v jiných od sebe zcela odlišných oborech (např. v lékařství, chemii, matematické statistice). Slovo „test" pochází z latiny (testum = kelímek, v němž alchymisté zkoušeli kovy). Test je definován jako „nástroj pro hodnocení nebo postup, pomocí kterého je získáván vzorek projevů testované osoby v určité oblasti a následně je za použití standardizovaných postupů hodnocen a skórován" (AERA, APA, NCME Standardy 2001). Testy používané ve školství se nazývají didaktické testy[5] (dále jen testy)[6]. Testy využívané při přijímacím řízení na vyšší stupeň školy jsou zvláštním případem testů, v některých státech označované jako testy studijních předpokladů. Jsou to např. americké testy SAT (Scholastic Assessment Test), ACT (ACT assessment) a GRE (Graduate Record Examination), ve Švédsku test SweSAT a u nás test Obecné studijní předpoklady od společnosti Scio, používaný při přijímacím řízení na některé české vysoké školy. Tyto testy patří k testům rozlišujícím nebo-li testům relativního výkonu (norm-referenced tests, NR-testy). Účelem rozlišujících testů[7] je vzájemné porovnání žáků. Úlohy v těchto testech zpravidla nepokrývají důsledně celou vymezenou oblast učiva. Výsledek testování může být uveden ve formě pořadí žáků od nejlepšího k nejhoršímu či např. v percentilech[8]. Druhým hlavním typem didaktických testů jsou testy ověřující nebo-li testy absolutního výkonu (criterion-referenced tests, CR-testy). Ověřující test[9] poměřuje výkon studenta s předem danými kritérii. Výsledky informují o úrovni dosažených znalostí a dovedností testovaných, ale neříkají nic o jejich pořadí. Součástí pravidel hodnocení je tzv. minimální výkon v testu nebo-li hraniční skór (cut-off score)[10]. V běžné školní praxi se obvykle tyto dva druhy testů kombinují.

Celková kvalita testu závisí především na kvalitě úloh, z kterých je sestaven, na jejich počtu a pořadí v testu. Tvorba testových úloh je náročný tvůrčí proces, který vyžaduje teoretické odborné, pedagogické a psychologické znalosti, zkušenosti i intuici. Vytvořit testovou úlohu vyžaduje také určitý talent a znalost jistých zásad, pravidel a doporučení pro psaní testových úloh. V učitelských testech běžně používaných ve výuce i v testech, jejichž výsledky slouží jako informace, na jejichž základě se provádějí důležitá rozhodnutí (přijímací řízení na vyšší stupeň školy), se často setkáváme s úlohami (většinou objektivními), které obsahují řadu nedostatků (např. Byčkovský; Bažantová 2005).

---

[5] Pojem didaktický test (Achievement test), tj. test používaný ve školství, je v literatuře definován různě. Autoři se však shodují v tom, že jde o zkoušku, která objektivně zjišťuje úroveň osvojení učiva u určité skupiny osob. Byckovský (1983) definuje didaktický test jako *„hodnotící nástroj systematického zjišťování (měření) výsledku výuky"*. Jde o testy úspěšnosti.

[6] Klasifikaci daktických testů uvádí např. Byčkovský (1983), Schindler (2006).

[7] NR-test musí být vysoce citlivý, aby dobře rozlišil jednotlivé testované, proto se používají obtížnější úlohy s vysokou citlivostí. Test bývá obtížnější, na jednotlivou úlohu odpoví správně v průměru asi 50 % žáků.

[8] Percentil je v procentech vyjádřený podíl žáků, kteří dosáhli v testu stejného nebo horšího výsledku.

[9] Často se používá ověřujících testů v jednostupňové podobě (např. *mastery tests)*. Úlohy jsou považovány za rovnocenné a hodnotícím kritériem je pouze „množství" – čili dosažení určitého procenta úspěšných odpovědí, aby mohl být daný okruh požadavků uznán za splněný. Optimální CR-test bývá spíše snazší, na jednotlivou úlohu odpoví správně v průměru cca 80 % žáků (Schindler 2006).

[10] Hraniční skór je minimální počet bodů, který student musí získat, aby v testu uspěl.

Chceme-li vytvořit kvalitní vyvážený test, ať s pomocí počítače či bez něj, neměli bychom začít samotným navrhováním testových úloh, protože se nám může stát, že test nebude pokrývat rovnoměrně celé učivo ani zjišťovat požadované úrovně osvojení učiva a bude obsahovat zejména úlohy na zapamatování, které se navrhují poměrně snadno. Při tvorbě testu bychom tedy měli postupovat systematicky, řídit se určitými pravidly. Ovšem ani tato pravidla nezaručují vytvoření dobré úlohy, záleží především na zkušenostech tvůrce úloh. Je zřejmé, že v případě profesionálních testů, které zpravidla vznikají ve specializovaných centrech na tvorbu školských testů bude postup tvorby jiný než u tzv. učitelských testů, které si učitelé vytvářejí sami pro vlastní potřebu a většinou jsou určeny jen na jedno použití (Burjan 2004). Dále se zaměřím na tvorbu profesionálního testu.

Proces tvorby testu zpravidla probíhá ve třech na sebe navazujících etapách:

- **plánování**, kdy navrhujeme projekt testu,
- **konstrukce testu**, tj. tvorba úloh, vytvoření prototypu testu a jeho terénní ověření,
- **analýza testových výsledků**, která se provádí dvakrát - nejdříve při pilotním ověřování, kdy zjišťujeme validitu, charakteristiky jednotlivých úloh a testu jako celku (zejména validitu a reliabilitu testových výsledků), posléze po tzv. ostrém testování, po kterém následuje úprava testu.

4. 1 Vlastnosti kvalitního testu

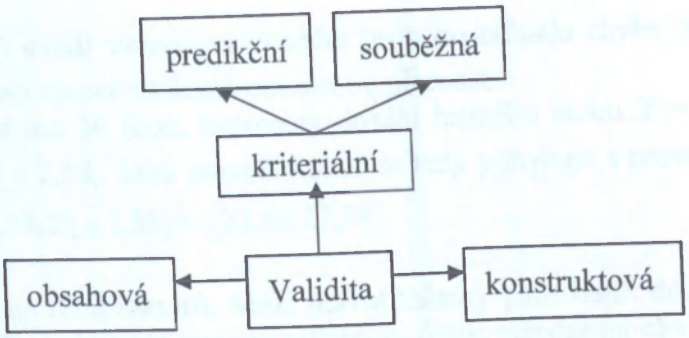
Kvalitní test by měl mít takové vlastnosti, aby jeho výsledky poskytovaly důležité informace k určitým rozhodnutím jako je např. výběr studentů u přijímacího řízení. Jde o tyto vlastnosti:

- validitu (validity)
- reliabilitu (reliability)
- objektivitu (objectivity)
- praktičnost (usability)

Validita

Je test validní? Měří to, co chceme, aby měřil? Do jaké míry skutečně testujeme to, co si myslíme, že testujeme? Všechny tyto otázky znamenají totéž. Validita (adekvátnost, funkčnost či platnost) vyjadřuje míru přijatelnosti závěrů vyvozovaných ze zjištěných testových skórů (dále jen validita). Jde o míru shody (na základě posudků expertů a/nebo na základě empiricky zjištěných údajů) mezi výsledky testu a účelem, pro který byl test vytvořen. Hopkins (1998) definuje validitu měření jako „to, jak dobře měření plní funkci, pro kterou bylo použito. Bez ohledu na ostatní přednosti testu, pokud test není validní, informace, kterou poskytuje, je nepoužitelná...". Validita je nejdůležitější vlastností požadovanou u testů. Na obr. 4-1 je přehled druhů validit: obsahová, kriteriální (predikční a souběžná) a konstruktová (teoretická).

Obsahová validita (content-related validity) vyjadřuje, do jaké míry je test reprezentativním výběrem učiva, jehož osvojování má zjišťovat. Obsahová validita je primárně proces logické analýzy, lze ji zjišťovat pouze expertním posouzením (Hopkins 1998). Zvláštním případem obsahové validity je tzv. **face validita** (face validity), kdy je test posuzován na první pohled, a to buď testovanými či např. učitelem. **Kriteriální validita** (criterion-related validity) měří, jak dobře výsledky testů předpovídají nějaké kritérium (např. studijní úspěšnost, úspěšnost v povolání). Rozlišujeme dva druhy kriteriální validity: predikční (predictive) a souběžnou (concurrent). Příkladem predikční validity je zjišťování vztahů mezi různými prediktory používanými při přijímacím řízení a studijní úspěšností (kritérium), vyjádřenou např. průměrným prospěchem na konci prvního ročníku na vysoké škole. V případě souběžné validity máme kritérium k dispozici již v době testování (např. výsledky z jiného testu). Kriteriální validita se zjišťuje empiricky, např. korelačním koeficientem mezi testovými skóry a kritériem. **Konstruktová** validita (construct validity), nazývaná také jako teoretická validita, vyjadřuje rozsah, ve kterém test odráží určitou abstraktní psychologickou charakteristiku (konstrukt), např. úzkostnost. Konstruktová validita se zkoumá zpravidla expertně.



Obr.4-1 Přehled druhů validit.

Reliabilita

Reliabilita v klasické teorii testu (KTT)

Reliabilita je přesnost a spolehlivost testových skórů. Vyjadřuje, do jaké míry je výsledek testu náhodný, ovlivněn chybu, nevypovídá však nic o správnosti výsledku. Čím je reliabilita vyšší, tím menší vliv má na výsledek testu náhoda. Je nutnou podmínkou pro validitu testu, ne však postačující, naopak validita zaručuje určitý stupeň reliability. Podle klasické teorie testu se naměřený skór X skládá ze dvou nezávislých částí – z pravdivého (jmenovitého) skóru T (true score) a náhodné chyby měření e (measurement error): $X = T + e$.
Pravdivý skór je skór, který žák v průměru získá, když je testován s časovým odstupem znovu za stejných testových podmínek. Ačkoli nikdy neznáme pravé skóry, můžeme zjistit neshody mezi neznámým pravým a známým naměřeným skórem. Rozdílem těchto skórů je *standardní chyba měření* σ_e, což je standardní odchylka chyby měření[11](její střední hodnota je rovna nule, má normální rozdělení).

Reliabilita vyjadřuje, do jaké míry je naměřený testový skór ovlivněn náhodnou chybou, udává se v podobě indexu nebo koeficientu, ale závisí také na skutečné variabilitě (proměnlivosti) měřené vlastnosti. Měřená vlastnost mezi respondenty kolísá, proto se používá rozptyl.

Index reliability testu vyjadřuje korelaci (těsnost vztahu) mezi rozdělením pravdivých a naměřených skórů, značí se ρ_{XT}. Je dán vztahem $\rho_{XT} = \sqrt{1 - \frac{\sigma_e^2}{\sigma_X^2}} = \sqrt{\frac{\sigma_T^2}{\sigma_X^2}}$, kde σ_e^2 je rozptyl odhadu standardní chyby měření, $\sigma_X^2 = \sigma_T^2 + \sigma_e^2$ rozptyl odhadu naměřeného skóru a σ_T^2 rozptyl odhadu pravdivého skóru. Pokud $\rho_{XT} = 1$, rovná se naměřený skór pravdivému skóru. Je-li tedy test vysoce reliabilní ($\rho_{XX} \geq 0{,}9$, což je požadováno u standardizovaných testů), je rozdíl mezi naměřeným a pravdivým skórem relativně malý.

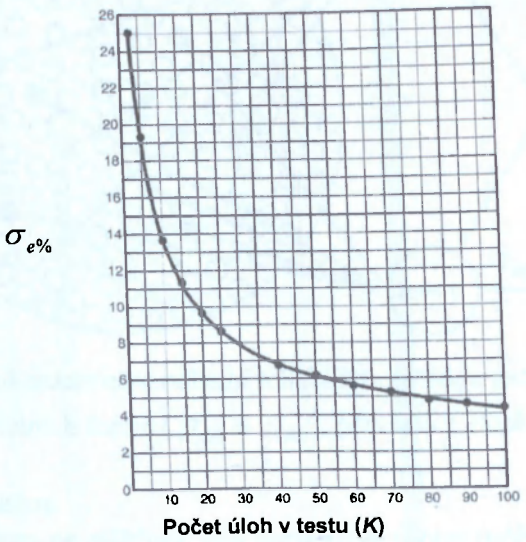
Koeficient reliability vyjadřuje korelaci mezi dvěma soubory naměřených skórů paralelních forem testu, vyjadřuje totéž jako index reliability, ale jiným způsobem, jeho hodnota je větší než hodnota indexu reliability. Značí se ρ_{XX} a je dán vztahem $\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2$. Hodnota koeficientu se teoreticky pohybuje mezi -1 a 1, v praxi mezi 0 a 1.

Standardní chyba měření se odhaduje pomocí vztahu: $\sigma_e = \sigma\sqrt{1 - \rho_{XX}} = \sqrt{1 - \rho_{XT}}$, kde σ je standardní odchylka skórů, ρ_{XX} koeficient reliability a ρ_{XT} index reliability (Hopkins 1998). Lord

[11] Přibližně u 2/3 studentů se liší naměřené skóry od pravého skóru o jednu standardní chybu měření (68% konfidenční interval) nebo méně a jen cca. 5% žáků získá skóry, které se liší o více než dvě standardní chyby (95% konfidenční interval).

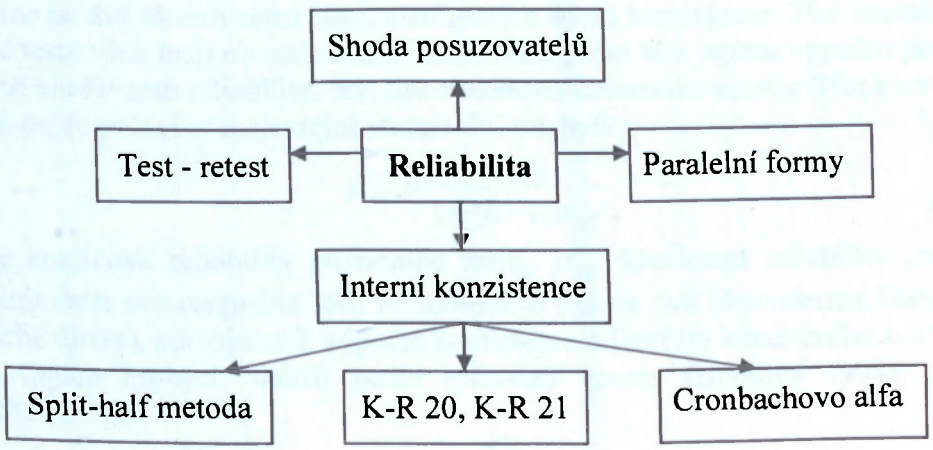
(1959b), Kleinke (1979) uvádí vzorec na výpočet hrubého odhadu chyby měření: $\sigma_e = 0{,}43\sqrt{K}$, kde K je počet úloh. Tento vzorec však má omezenou platnost.
Ilustrativní příklad: Test má 36 úloh, testovaný dosáhl hrubého skóru $X = 25$. Standardní chyba měření je $\sigma_e = 0{,}43\sqrt{36} = 2{,}58$. Jeho pravdivý skór se tedy pohybuje s pravděpodobností přibližně 68% v intervalu $\langle 25-2{,}58;25+2{,}58\rangle = \langle 22{,}42;27{,}58\rangle$.

Reliabilitu testu ovlivňuje řada faktorů. Mezi hlavní faktory patří např. délka testu (viz obr. 4-2), s větším počtem úloh (K, vodorovná osa) se reliabilita (dána standardní chybou měření vyjádřenou v procentech $\sigma_{e\%}$) zvyšuje,), technická kvalita úloh (příliš snadné a příliš obtížné úlohy či méně citlivé reliabilitu snižují), rozptyl úrovní vědomostí testovaných (čím větší rozptyl skutečných úrovní vědomostí testovaných, tím je měření spolehlivější).



Obr. 4-2 Vztah délky testu a reliability testu.

Pro (statistický) odhad reliability testu existuje několik metod (viz obr. 4-3): test – retest, paralelní formy, interní (vnitřní) konzistence a shoda mezi posuzovateli. Metody test-retest a paralelní formy jsou graficky znázorněny na obr.4-4



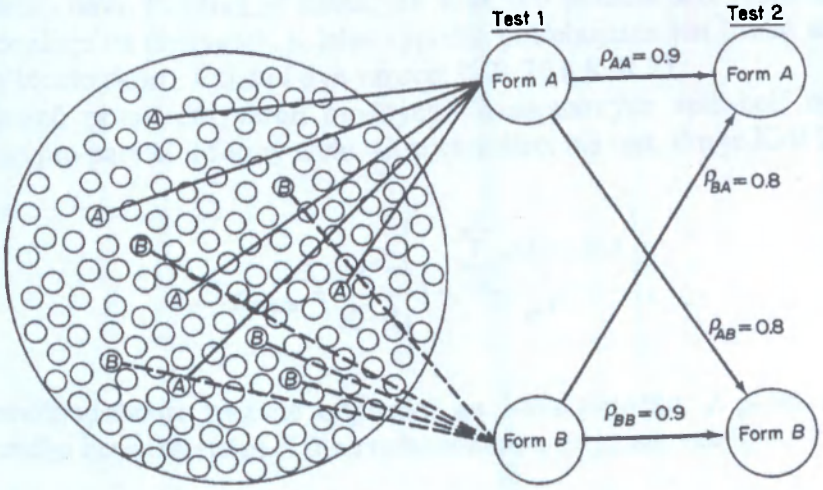
Obr. 4-3 Metody pro odhad reliability.

Test – retest nebo-li stabilita v čase
Tato metoda spočívá v tom, že ten samý test zadáme testovaným po nějaké době znovu. Vypovídá tedy pouze o stabilitě výkonu pro různé administrace jednoho testu. Používá se např. pro měření

síly, rychlosti, hladiny cholesterolu, krevního tlaku apod. Koeficient reliability se počítá pomocí Pearsonova korelačního koeficientu $r_{1,2} = r_{test.retest}$.

Paralelní formy nebo-li ekvivalence

Technika paralelních forem je založená na tom, že jednomu testovanému zadáme dvě varianty testu bez či s časovým odstupem. Reliabilitu lze určit buď Pearsonovým korelačním koeficientem mezi naměřeným skóry dvou paralelních testů nebo pomocí koeficientu reliability. Podmínkou této metody je, aby testy byly zadány stejné skupině testovaných, což je nepraktické. Koeficient reliability paralelních forem bývá nižší než koeficient reliability test – retest.



Obr. 4-4 Grafické znázornění odhadu reliability metodou test-retest (ρ_{AA} a ρ_{BB}) a paralelních forem (ρ_{AB} a ρ_{BA}), převzato z Hopkins 1998.

Interní (vnitřní) konzistence

Metodou interní konzistence se zjišťuje, zda jednotlivé úlohy měří to samé, přičemž každá úloha se chápe jako samostatný testík a test jako soubor paralelních testíků. Test zadáváme pouze jednou. Je vhodná pro situace, kdy lze testové úlohy vybrat jako reprezentativní vzorek z testu. Není tedy vhodná např. pro měření výšky člověka, kdy pro výpočet chyby měření potřebujeme další nezávislé měření.

A. Split-half-metoda

Test rozdělíme na dvě ekvivalentní části, které potom spolu korelujeme. Tím změníme délku testu. Krátké testy však mají obvykle menší reliabilitu, proto se k jejímu výpočtu používá korigovaného koeficientu reliability, tzv. *Spearmanova-Brownova vzorce* (Hopkins 1998). Předpokládá se, že poloviny mají stejné standardní odchylky.

$$\rho_{XX'} = \frac{L\rho_{XX}}{1+(L-1)\rho_{XX}},$$

kde ρ_{XX} je koeficient reliability původního testu, $\rho_{XX'}$ koeficient reliability „nového" testu, který je L-krát delší než originální test. Rozdělíme-li test na dvě ekvivalentní části A a B (např. na sudé a liché úlohy), použijeme k výpočtu koeficientu reliability korelačního koeficientu těchto dvou částí (jejich hrubých skórů) podle *split-half vzorce* (Hopkins 1998), který vychází z předchozího ($L = 2$).

$$\rho_{XX'} = \frac{2\rho_{X_A X_B}}{1+\rho_{X_A X_B}}$$

Jednodušší na výpočet je *Flanaganova metoda* (1937), protože nevyžaduje výpočet korelačních koeficientů a navíc nepředpokládá, že obě poloviny testy (A, B) mají stejné standardní odchylky (Hopkins 1998).

$$\rho_{XX} = 2\left(1 - \frac{\sigma_A^2 + \sigma_B^2}{\sigma^2}\right), \text{ kde}$$

σ_A^2, σ_B^2 jsou rozptyly polovin testu A a B, σ^2 je rozptyl celkových skórů testu. Rozdíly mezi hodnotami koeficientu reliability vypočteného podle Spearman-Brownova vzorce a Flanaganova vzorce jsou nepatrné, Flanaganův koeficient je nepatrně nižší.

B. *Kuder-Richardsonův vzorec*

Dalším způsobem zjišťování reliability je Kuder-Richardsonův vztah (1937), který také měří vnitřní konzistenci úloh. Používá se často, jen však pro binárně skórované úlohy (vynechaná odpověď se považuje za chybnou). K jeho výpočtu potřebujeme jen hrubé skóry jednotlivých úloh a rozptyly těchto skórů. Existují dva vzorce: K-R 20 a K-R 21.

K-R 20 je vlastně průměrem všech možných Flanaganových split-half odhadů reliability, vhodný zejména pro paralelní formy testu. Čím více úloh má test, tím je K-R 20 vyšší.

$$\rho_{KR20} = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} p_i(1-p_i)}{\sigma^2}\right),$$

kde p_i je pravděpodobnost správné odpovědi na i-tou položku, K počet úloh a σ^2 odhad rozptylu celkového hrubého skóru K úloh (s hodnotou n ve jmenovateli).

Ilustrativní příklad: Máme-li test o 3 položkách, pravděpodobnosti správných odpovědí na tyto položky jsou po řadě 0,8; 0,7; 0,5 a standardní odchylka celkového hrubého skóru je $\sigma = 0{,}9$,

potom $\rho_{KR20} = \frac{3}{2}\left(1 - \frac{0{,}8.0{,}2 + 0{,}7.0{,}3 + 0{,}5.0{,}5}{0{,}9^2}\right) = 0{,}35$

K-R 21 je sice méně přesný než *K-R 20* (koeficient reliability vychází o trošku menší), ale snáz se počítá.

$$\rho_{KR21} = \frac{K}{K-1}\left(1 - \frac{\mu(K-\mu)}{K\sigma^2}\right), \text{ kde}$$

μ je průměrný celkový hrubý skór.

C. *Cronbachovo alfa*

Cronbachovo alfa je nejobecnějším (úlohy mohou být různé obtížnosti, nebinární či kombinované) a nejpřesnějším koeficientem reliability, ale jeho výpočet je pracný. Čím je reliabilita vyšší, tím méně se liší *Cronbachovo alfa* od *K-R 21*. Je to dolní odhad skutečné reliability. Je dáno vztahem

$$\rho_\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_i^2}{\sigma^2}\right), \text{ kde}$$

σ_i^2 odhad rozptylu položky i.

Shoda mezi posuzovateli

Shoda mezi posuzovateli se rozlišuje zpravidla pro binární, kardinální, nominální a ordinální data (více Botz; Lienert; Boehnke 2000). U binárních dat se určuje shoda m posuzovatelů posuzujících N objektů binárně (např. ano-ne). Počítá se pomocí Cohenova koeficientu kappa κ (1960). U nominálních dat se určuje shoda m posuzovatelů posuzujících N objektů podle vícestupňového kritéria, přičemž neshody mezi posuzovateli se považují za rovnocenné. Používá se Fleissův vztah

(1971). U ordinálních dat se posuzuje míra shody m posuzovatelů, kteří seřazují N objektů podle určitého kritéria do řady. Míra shody se určuje podle koeficientu W-Kendalla a Babingtona. U kardinálních dat se zjišťuje míra shody m posuzovatelů posuzujících N nezávislých objektů, které zařazují do k kategorií, přičemž jim dávají různé váhy. Oproti nominálním datům se zde zohledňuje míra neshody úsudků posuzovatelů.

Vztah reliability k charakteristikám úloh

Reliabilita je založena na vzájemném vztahu úloh. Pokud úloha vysoce koreluje s ostatními úlohami v testu, tj. má dobrou diskriminační schopnost, je koeficient reliability vysoký. Obtížnost a citlivost úloh jsou tedy svým způsobem ve vztahu s hodnotou koeficientu reliability. Test s vysokou hodnotou alfa by měl mít obtížnost mezi 0,40 a 0,70, citlivost větší než 0,30. Úlohy s dobrou citlivostí mají zpravidla vysoké hodnoty reliability alfa, obtížnost těchto úloh je také dobrá a většina distraktorů funguje správně (Hopkins 1998).

Reliabilita v teorii odpovědi na položku (IRT)

Co se týče přesnosti měření testu, hlavním rozdílem mezi IRT skóry a tradičními testovými skóry[12] je to, že IRT skóry mají rozdílnou přesnost (chybu měření) pro různé úrovně schopnosti (proficiency) testovaných. V KTT je reliabilita dána poměrem rozptylů pravdivého a naměřeného skóru a je závislá na délce testu[13]. Na rozdíl od KTT je přesnost měření v IRT známa pro všechny dílčí skóry. V kontextu IRT jsou pravdivé skóry nepozorovatelné hodnoty θ, které jsou odhadovány specifickou standardní chybou ze vzorků odpovědí na položku. Reliabilita zde nezávisí na délce testu. Reliabilitě testových skórů v KTT odpovídá v IRT množství informace, které test podává svými úlohami. Množství informace o jednotlivých položkách lze matematicky určit tzv. informačními funkcemi (IIF, item information function) zvonovitého tvaru. Informační funkce testu $I(\theta)$ pro danou θ je definována jako součet informačních funkcí $I_i(\theta)$ jednotlivých položek pro tuto θ, protože úlohy jsou na sobě nezávislé (Hambleton; Swaminathan; Rogers 1991):

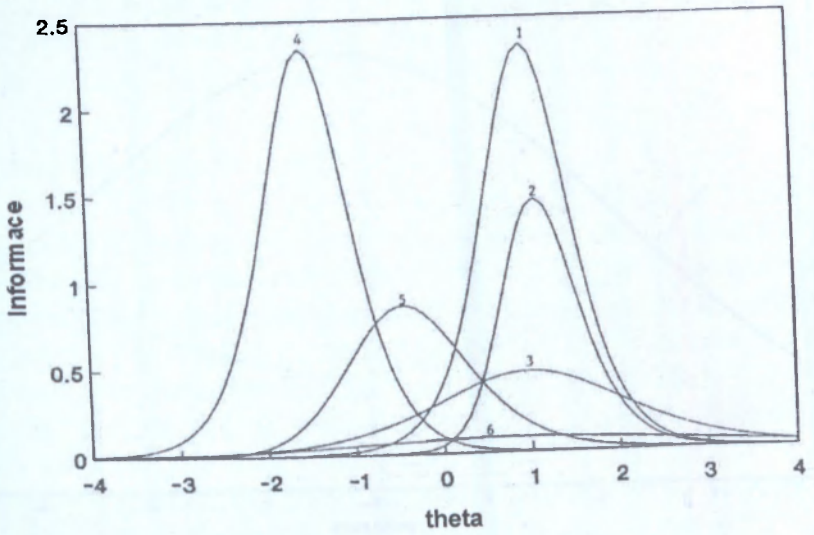
$$I(\theta)=\sum_{i=1}^{n} I_i(\theta), \text{ kde } I_i(\theta)=\frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)}, i=1, 2, .., n,$$

kde $P_i(\theta)$ je charakteristická funkce položky i, $Q_i(\theta)=1-P_i(\theta)$ a $P_i'(\theta)$ je první derivace $P_i(\theta)$. Ze vzorce plyne, že hodnota informační funkce testu závisí na počtu úloh v testu a průměrné hodnotě parametrů citlivosti testových úloh dle zvoleného IRT modelu. Jelikož s počtem položek roste množství informace, měří test jako celek danou schopnost mnohem přesněji než jedna položka. Tvar této funkce závisí na rozložení hodnot parametrů obtížnosti úloh po ose schopnosti a na rozložení a průměrné hodnotě parametrů citlivosti testových úloh.

Úloha měří schopnost s největší přesností, tj. nejlépe rozlišuje mezi testovanými s úrovní schopnosti odpovídající hodnotě parametru b obtížnosti úlohy. To znamená, maximální hodnoty dosahuje informační funkce v blízkosti hodnoty parametru obtížnosti dané položky (viz obr. 4-5). Množství informace poskytované úlohou klesá se vzdalováním úrovně schopnosti θ od obtížnosti úlohy a přibližováním k nule na obou koncích osy schopnosti θ. Je-li množství informace malé (viz obr. 4-5, úloha 6), nedá se schopnost přesně odhadnout a odhady budou široce rozprostřeny kolem skutečné schopnosti. Takové úlohy jsou statisticky téměř nepoužitelné do testu. Je-li množství informace velké, může být hodnota schopnosti θ testovaného odhadována přesně, tzn. všechny odhady budou rozumně blízko ke skutečné hodnotě dané úrovně schopnosti. Parametr citlivosti úlohy značně ovlivňuje maximální množství informace pro odhad schopnosti, která je dána úlohou (viz obr 4-5, informační křivky úloh 1 a 2. Hodnoty $a < 1$ vedou k nízké hodnotě množství informace testu, hodnoty $a > 1{,}7$ vedou k vysoké hodnotě množství informace testu. Protože množství informace testu snižují hodnoty $c > 0$ pro nízké úrovně schopnosti a velké hodnoty c

[12] V KTT je standardní chyba měření konstantní pro všechny dosažené skóry a je specifická pro danou populaci.
[13] Delší test má vyšší reliabilitu než kratší test.

ıbecně pro všechny úrovně schopnosti (u 3-parametrového modelu), je tendence přizpůsobovat data píše 1- či 2-parametrovému modelu (kde $c = 0$, např. Hambleton aj. 1991, Baker 2001).



Obr. 4-5 Ukázka informačních křivek šesti úloh, které se liší množstvím informace, které podávají pro danou úroveň schopnosti θ (upraveno podle Hambleton aj. 1991)

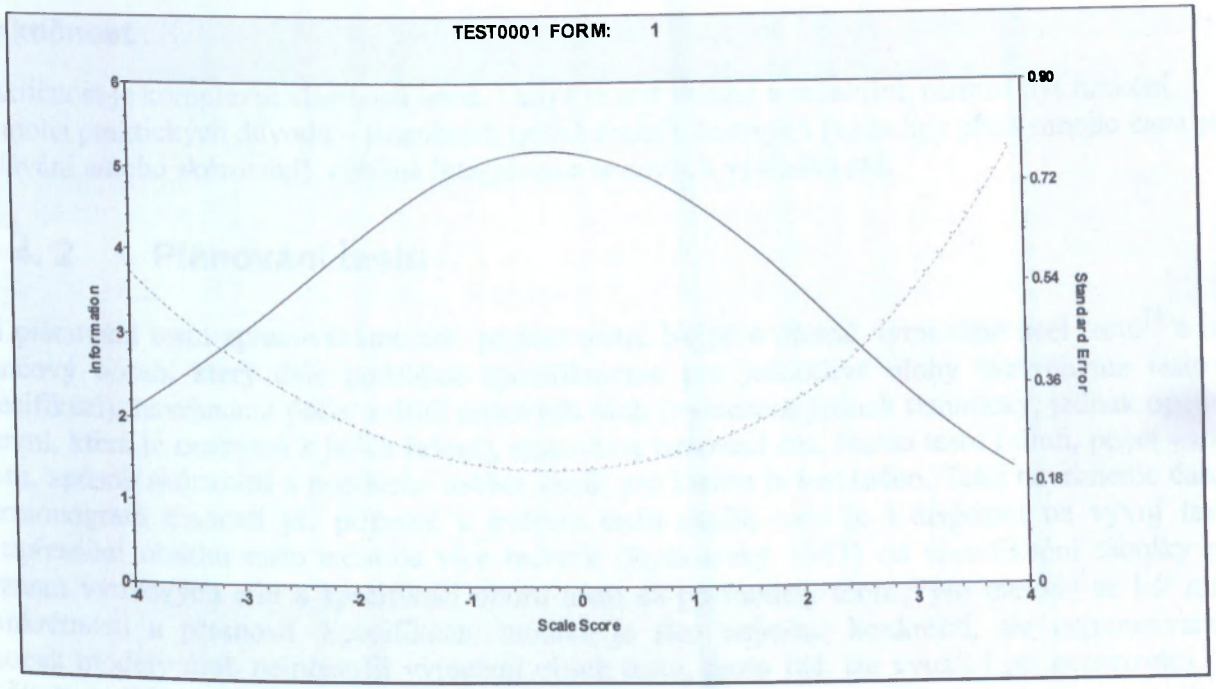
Na obr. 4-5 jsou informační křivky šesti úloh, které se liší množstvím informace, které podávají pro lanou úroveň schopnosti θ testovaných. Úloha 1 podává nejvíce informace pro $\theta = 1$, (maximální ıodnota informační funkce), pro $\theta = -1$ a $\theta = 3$ již žádnou. V úloze 2 se maximum informace oustřeďuje také kolem $\theta = 1$, na obou svých koncích potom podává stejně málo informace. Úloha 3 podává sice nejvíce informace opět pro $\theta = 1$, ale mnohem méně než úlohy 1 a 2. Úloha 4 podává ıejvíce informace pro $\theta = -1{,}5$, atd. Na základě maximálních hodnot informace lze usuzovat, že ılohy 1, 2, 3 jsou těžší než úlohy 4 a 5. Nejméně citlivá je úloha 6, protože podává velmi málo ıformace pro všechny schopnostní úrovně.

Množství informace, které test podává pro hodnotu θ, je v inverzním vztahu k přesnosti (vyjádřena ıtandardní chybou měření), s kterou je schopnost θ odhadována:

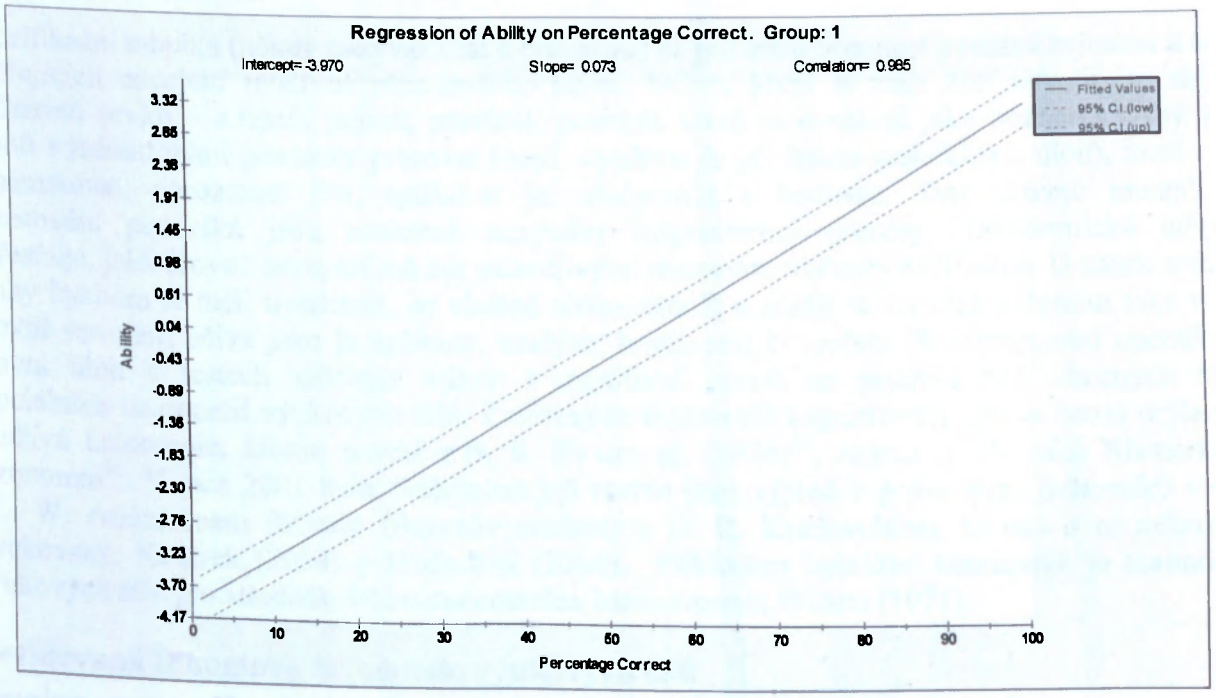
$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}},$$

kde $SE(\hat{\theta})$ je standardní chyba odhadu úrovně θ testovaných testem s informačním přínosem $I(\theta)$ (Urbánek; Šimeček 2001). Ta je koncepčně ekvivalentní k standardní chybě měření v KTT, ıle na rozdíl od chyby měření v KTT umožňuje zobecnění na různé populace. Čím více informace est na dané úrovni schopnosti poskytuje, tím menší je chyba, s níž je úroveň schopnosti ɔdhadována (viz obr.4-6).

Známe-li informaci, kterou test podává na úrovni θ, tj. známe-li chybu měření pro θ můžeme ɔotom sestrojit interval spolehlivosti pro odhad úrovně schopnosti θ daných testovaných (viz obr. 4-7, Emretson; Reise 2000).



Obr.4-6 Ukázka informační funkce testu a standardní chyby měření v testu OSP, varianta A, PedF UK 2006 (výstup z programu BILOG-MG).



Obr. 4-7 Intervaly spolehlivosti pro odhad úrovně schopnosti θ testovaných (svislá osa – ability) v testu OSP, varianta A, PedF UK 2006 (výstup z programu BILOG-MG).

Objektivita

Objektivita je vlastnost testu, která se vztahuje k testovým skórům. Test sestavený výhradně z objektivních úloh (uzavřených či otevřených se stručnou odpovědí) se považuje za objektivní, pokud jeho výsledky nejsou ovlivněny subjektivními názory či postoji hodnotitelů testu. To znamená, pokud je možné vyhodnocovat odpovědi na testové úlohy buď automaticky (pomocí speciálního počítačového software) nebo nezávislými posuzovateli, kteří dojdou ke stejnému výsledku. Problémy s vyhodnocováním jsou u esejů a komplexních úloh, které zcela objektivně hodnotit nelze.

Praktičnost

Praktičnost je komplexní vlastností testu. I když je test validní a reliabilní, nemusí být funkční z mnoha praktických důvodů – finančních (příliš drahý), časových (vyžaduje příliš mnoho času při zadávání a/nebo skórování), obtížná interpretace testových výsledků atd.

4. 2 Plánování testu

Při plánování testu zpracováváme tzv. projekt testu. Nejprve přesně vymezíme účel testu[14] a jeho rámcový obsah, který dále podrobně specifikujeme pro jednotlivé úlohy (navrhneme testovou specifikaci), navrhneme počet a druh testových úloh (vymezené jednak tematicky, jednak operační úrovní, která je nezbytná k jejich řešení), stanovíme testovací čas, formu testu i úloh, počet variant testu, způsob skórování a popíšeme soubor osob, pro kterou je test určen. Také navrhneme časový harmonogram činností při přípravě a ověření testu (kolik času je k dispozici na vývoj testu). K upřesnění obsahu testu existuje více technik (Byčkovský 1983) od specifikační tabulky přes seznam výukových cílů a specifikaci oboru testu až po modely úloh. Tyto metody se liší mírou konkrétnosti a přesnosti. Specifikační tabulka je sice nejméně konkrétní, ale nejpoužívanější, naopak modely úloh nejpřesněji vymezují obsah testu, proto jich lze využít i při generování úloh počítačem. Dále se budeme zabývat dvěma technikami, technikou specifikační tabulky a pozdeji při generování úloh i technikou modelování úloh.

Specifikační tabulka

Specifikační tabulka (někdy nazývána také *blueprint*) se pro svou obecnost používá zejména u testů ověřujících osvojení relativně různorodého učiva. Učivo, které si mají žáci osvojit, se skládá z různých prvků – z faktů, pojmů, pravidel, postupů, které se označují jako poznatky. Aby žáci mohli s jednotlivými poznatky pracovat (např. využívat je při řešení praktických úloh), musí si je zapamatovat, porozumět jim, aplikovat je, analyzovat a hodnotit. Tyto úrovně mentálního zpracování poznatků jsou souhrnně nazývány kognitivními procesy. Taxonomická tabulka upřesňuje, jaká úroveň osvojení má být jednotlivými testovými úlohami ověřována. U každé testové úlohy bychom si měli uvědomit, co vlastně úloha zkouší a snažit se do úloh zahrnout také vyšší úrovně osvojení učiva jako je aplikace, analýza, hodnocení či tvoření. K vymezování operačních úrovní úloh v testech zjišťující výkon v kognitivní oblasti se používá buď obecných nebo speciálních taxonomií výukových cílů. Z obecných taxonomií kognitivních cílů se zatím nejčastěji používá taxonomie, kterou navrhl v B. S. Bloom aj. (1956)[15], známá je ale také Niemerkova taxonomie[16]. V roce 2001 byla zveřejněna její revize jako výsledek práce týmu odborníků v čele s L. W. Andersonem (bývalý Bloomův student) a D. R. Krathwohlem. U nás o ní referovali Byčkovský; Kotásek (2004) a Hudecová (2004). Příkladem speciální taxonomie je taxonomie výukových cílů pro středoškolskou matematiku, kterou navrhl Wilson (1971).

Revidovaná Bloomova taxonomie výukových cílů

Smyslem vymezování cílů je definovat, co učitel chce, aby se žáci naučili. Použití detailních cílů, tj. vysvětlení žákových aktivit může žákům pomoci lépe porozumět významu různých učebních aktivit. Původní Bloomova taxonomie (Bloom et al, 1956) vychází z jednodimenzionální klasifikace kognitivních procesů. Základ tvoří šest hierarchicky uspořádaných kategorií kognitivních procesů (znalosti, pochopení, aplikace, analýza, syntéza, hodnotící posouzení), které se dále dělí na subkategorie. To znamená, že k dosažení vyšší cílové kategorie je nezbytné důkladné

[14] Jaký je hlavní účel testu, k čemu mají sloužit výsledky testování, pro koho jsou výsledky určeny, jde o ověřující či rozlišující test atd.

[15] Bloomova taxonomie se stala jednou z nejpoužívanějších publikací v americké školní praxi a byla postupně přeložena do 22 jazyků.

[16] Polský odborník v oblasti měření a hodnocení výsledků výuky Niemierko vydal v roce 1979 taxonomii výchovně vzdělávacích cílů. U nás o ní informoval např. Byčkovský (1983).

zvládnutí nižší kategorie cílů, tj. příslušné učební látky na nižší úrovni osvojení. Subkategorie jsou charakterizovány obecným popisem požadovaných činností (chování) studenta. První taxonomická kategorie vychází z jisté zobecněné struktury jednotlivých oblastí lidského poznání, zahrnuje konkrétní poznatky, způsoby jejich využívání a abstrakci v příslušné oblasti poznání, přičemž se od studenta požaduje zejména jejich zapamatování, znovuvybavení. Zbylých pět kategorií označených jako intelektové schopnosti a dovednosti se vztahuje k organizaci a reorganizaci poznatků či učební látky, ke způsobům operování s nimi, k jejich aplikaci a syntéze nebo k hodnocení předloženého (popř. zapamatovaného) materiálu či problémů. Každá kategorie a subkategorie je nejprve vymezena verbálně, následně doplněna příklady kognitivních cílů a u subkategorií jsou také uvedeny testové úlohy. Základem revidované Bloomovy taxonomie je narozdíl od původní Bloomovy taxonomie dvoudimenzionální taxonomická tabulka uvedená v tab. 4-1.

Tab. 4-1 **Taxonomická tabulka** (upraveno podle Anderson; Krathwohl 2001)

POZNATKY	KOGNITIVNÍ PROCESY					
	1 Zapamatovat si	2 Porozumět	3 Aplikovat	4 Analyzovat	5 Hodnotit	6 Tvořit
A Poznatky faktické						
B Poznatky konceptuální						
C Poznatky procedurální						
D Poznatky metakognitivní						

Jednu dimenzi tvoří druhy poznatků (faktické, konceptuální, procedurální, metakognitivní), vyjádřené podstatnými jmény, které se má žák naučit. Tato dimenze vznikla z původní Bloomovy první kategorie znalosti. Poznatky jsou osvojovány na různých úrovních kognitivních procesů uspořádaných hierarchicky, jež tvoří druhou dimenzi. Jsou vyjádřeny slovesy (zapamatovat si, porozumět, aplikovat, analyzovat, hodnotit, tvořit). Obě dimenze jsou dále členěny na kategorie a subkategorie. V tab. 4-2 je revidovaná Bloomova taxonomie kognitivních cílů ve srovnání s původní Bloomovou taxonomií.

Tab. 4-2 Revidovaná Bloomova taxonomie kognitivních cílů ve srovnání s původní Bloomovou taxonomií (upraveno podle Byčkovský; Kotásek 2004)

Taxonomie kognitivních cílů			
podle Blooma 1956	podle Andersona a Krathwohla 2001		
	Poznatky	faktické	– terminologie; konkrétní poznatky
		konceptuální	– klasifikace kategorií; zákonitosti a zobecnění; teorie, modely a struktury
		procedurální	– specif. postupy a algoritmy, techniky a metody v příslušném oboru; kritéria, která umožňují vybrat vhodný postup
		metakognitivní	– obecné strategie učení, poznávání a řešení problémů; znalosti kognitivních nároků, které klade řešení různých úloh včetně kontextu a podmínek; sebepoznání
Kognitivní procesy			
Znalosti	Zapamatovat si	- vybavovat si příslušné znalosti z dlouhodobé paměti	
Porozumění	Porozumět	- konstruovat význam sdělení zprostředkovaného ústně, písemně či graficky (interpretování, parafrázování, uvádění příkladu, klasifikování, shrnutí, usuzování, srovnávání, vysvětlování)	
Aplikace	Aplikovat	- používat známé postupy při řešení běžných úloh či v nových situacích	
Analýza	Analyzovat	- rozkládat celek na podstatné části, určovat jejich vzájemné vztahy a jejich vztah ke struktuře celku nebo jeho účelu (rozlišování, strukturování, přisuzování)	
Syntéza	Hodnotit	- vyjadřovat hodnotící stanoviska na základě kritérií a norem (ověřování, testování, monitorování, vyjadřování kritických soudů)	
Hodnocení	Tvořit	- skládat prvky tak, aby vytvářely koherentní nebo funkční celek; reorganizovat prvky do nových struktur a modelů (generování, plánování, vytváření originálních děl)	

Jak uvádějí Byčkovký; Kotásek (2004), umožňuje revidovaná Bloomova taxonomie nejen klasifikaci vzdělávacích cílů, ale i analyzování učebních a vyučovacích aktivit, které byly zvoleny k dosažení cílů, vhodnost hodnotících nástrojů navržených ke zjištění, zda cílů bylo dosaženo a v jaké míře.

Specifikační tabulka je dvourozměrná, její řádky tvoří tematické celky (obsah učební látky), sloupce operační úrovně osvojení a někdy také váhu učiva (počet vyuč. hodin). Do jednotlivých políček tabulky zapisujeme požadovaný počet úloh, jejichž součet musí být roven celkovému počtu úloh v testu (viz např. Byčkovský 1983).
Tabulka vytváří svým složením předpoklady k vytvoření vyváženého testu, zkoušejícího reprezentativní výběr učiva.

Postup při vytváření specifikační tabulky

1) Nejdříve si **téma/ učební látku rozdělíme na dílčí části podle obsahu** (např. podle učebnice), kterým následně přiřadíme určitou váhu, např. podle toho, jaký čas byl dané látce ve výuce věnován nebo podle rozsahu učební látky v učebnici či dle důležitosti učební látky, kterou posuzujeme nejlépe ve spolupráci s dalšími kompetentními osobami.
2) V dalším kroku určíme **počet úloh v testu a počet variant testu**. Aby byl test dostatečně spolehlivý a přesný, tj. měl vysokou míru reliability, měl by obsahovat co největší počet úloh, protože s počtem úloh se jeho reliabilita zvyšuje. Je třeba navrhnout až o 50% víc úloh, než jich

bude v konečné verzi testu. Za spodní hranici se u didaktických testů považuje 10 úloh. Horní hranici určuje především testovací čas, který je zpravidla omezen vyučovací hodinou (45 minut). Jde o hrubý čas. Čistý testovací čas po odečtení doby na zadávání a vybrání testu (zpravidla 5-15 min) je 30-40 minut. Kratší testy většinou trvají 15-25 minut a testy určené k tomu, zda žáci porozuměli výkladu, zpravidla nepřesahují 10 minut. Pro testy výstupní je ale nutné počítat nejméně se 40-60 minutami čistého času. Počet úloh závisí také na druhu testových úloh, z kterých má být test sestaven, na jejich složitosti. Jak uvádí Byčkovský (1983), lze orientačně u SŠ studentů u jednodušších uzavřených úloh a úloh se stručnou odpovědí počítat s časem od 0,5 do 1,5, příp. 2 minut na jednu úlohu. U složitějších úloh jsou časové nároky vyšší (až do 25 minut na úlohu). Při tvorbě úloh musíme také přihlédnout na věk, intelektovou úroveň testovaných a další jejich charakteristiky. Ve stejném čase zřejmě student vysoké školy vyřeší větší počet úloh než student výběrové SŠ, natož pak student běžné SŠ. Ulohy se širokou odpovědí jsou časově náročnější, proto jich lze do testu zařadit mnohem méně než uzavřených úloh či úloh otevřených se stručnou odpovědí. Pokud nelze při testování zamezit opisování či napovídání, je žádoucí vytvořit více variant testu, které však musejí být co do obsahu i co do obtížnosti jednotlivých úloh shodné.

3) K jednotlivým částem učiva stanovíme úroveň osvojení poznatků pomocí taxonomie výukových cílů (např. Anderson; Krathwohl 2001a, tab. 4-2). Určitým vodítkem může být tabulka, kterou navrhl Byčkovský (1983), kde je uvedena vhodnost jednotlivých typů testových úloh k měření různých úrovní osvojení učiva podle Bloomovy taxonomie (tab. 4-3). Kolik úloh v testu má danou úroveň osvojení zkoušet, není stanoveno, je to zcela na úvaze autora testu a povaze učiva, i když vzhledem k rozvíjení myšlení žáků je žádoucí upřednostňovat spíše vyšší úrovně osvojení.

Tab. 4-3 Vhodnost jednotlivých druhů úloh pro měření různých úrovní osvojení učiva (upraveno podle Byčkovský 1983 na základě revidované Bloomovy taxonomie)

	Cílová kategorie (úroveň osvojení) →		zapamatovat	porozumět	aplikovat	analyzovat	hodnotit	tvořit
Vhodnost úlohy ↓	otevřené široké	nestr.	-	-	+	+	++	++
		struk.	-	+	++	++	++	+
	otevřené se struč. odp.	prod.	++	++	++	+	-	-
		dopl.	++	+	+	-	-	-
	uzavřené	dich.	++	++	+	-	-	-
		MC	+	++	++	-	+	-
		přiř.	++	++	+	+	-	-
		uspoř.	+	++	-	-	+	-

++ velmi vhodné + vhodné - málo vhodné nebo nevhodné

Ukázka konkrétní specifikační tabulky (obsahově operační matice) pro test Obecné studijní předpoklady (OSP), který byl použit v roce 2006 při přijímacím řízení na PedF UK, je v tab. 4-5.

Tab. 4-5 Příklad specifikační tabulky pro test Obecné studijní předpoklady, PedF UK, 2006

Tematický celek/ dílčí témata		Počet úloh						
		Kognitivní cíl dle revidované Bloomovy taxonomie (viz tab. 4-2)						
		B1	B2	B3	B4	B5	B6	celkem
Verbální oddíl	slovní spojení				9			9
	porozumění textu		3					3
	antonyma				5			5
Analytický oddíl	logický úsudek						7	7
	analýza textu				6			6
Kvantitativní oddíl	aritmetika -slovní úlohy			6				6
	geometrie/ stereometrie			4				4
	algebra			1				1
	pravděpodobnost			1				1
	grafy			3				3
celkem		0	3	15	20	0	7	45

4. 3 Konstrukce testu

Naplánováním testu má autor ujasněno, jaké učivo na jaké úrovni a kolika testovými úlohami má být zkoušeno. Podle specifikační tabulky jsou v etapě konstrukce navrhovány jednotlivé testové úlohy, které jsou dále posuzovány nejméně dvěma experty (na obsahovou validitu) a podle jejich doporučení upravovány nebo vyřazovány. Posuzovatelé musejí mít nejen znalosti příslušného oboru, ale i dostatek zkušeností s tvorbou didaktických testů. Najít vhodné kompetenty ochotné spolupracovat není snadné. Posuzovatelům předkládá jejich navrhovatel úlohy v konečné grafické podobě a velikosti, v jaké je chce mít v testu, přitom do nich neuvádí správnou odpověď, aby neovlivnil kompetenty. Vymezí vlastnosti, které mají být posuzovány, předá posuzovatelům instrukce pro posuzování a zaznamenávání na speciální formuláře (viz Byckovsky 1983).

Kompetenti posuzují, jak úlohy odrážejí příslušné výukové cíle; jestli opravdu měří jen ty cíle, které autor uvádí. Pro každou úlohu je nakonec vypočítán průměr, střední hodnota a rozpětí hodnocení jednotlivých posuzovatelů. Dále se posuzuje, zda uvedená správná odpověď je opravdu správná, technické kvality úloh, míra důležitosti učiva u každé úlohy, (test by měl obsahovat úlohy přiměřené svému účelu) a obtížnost úloh dle vlastního subjektivního uvážení.

Následuje sestavení prototypu testu spolu s jeho příslušenstvím. V etapě terénního ověření je prototyp testu zadán menšímu vzorku populace, pro níž je určen, a následně jsou výsledky testu (především obtížnost a citlivost jednotlivých úloh; validita a reliabilita testových výsledků) získané z pilotáže statisticky analyzovány, ať již na základě klasické teorie testu (KTT) nebo IRT. Podle těchto výsledků dochází k závěrečné úpravě testu.

Tvorba testových úloh

Navrhování dobrých testových úloh[17] je časově náročná tvůrčí činnost vyžadující vedle odborných znalostí také zkušenosti, dodržování určitých pravidel a doporučení a jistou dávku tvořivosti. Výběr úloh závisí na cíli, který má test plnit, na obsahu učiva, které má být předmětem testování, materiálních a technických podmínkách, ale mnohdy také na tom, jaké druhy úloh má tvůrce testu v oblibě. V některých případech lze skupinu obsahově podobných úloh navrhovat počítačem, jde

[17] Testovou úlohou se rozumí nejen zadání úlohy, ale i předpis pro její vyhodnocování, jehož jasný a podrobný popis je nezbytný zejména u otevřených úloh vyžadujících rozsáhlejší odpověď, které se vyhodnocují vícestupňově (Byčkovsky 1983).

o tzv. generování úloh. K tomu je však nutné použít určitých technik, např. šablon pro tvorbu úloh (Alessi; Trollip, 2001) nebo modelů úloh, které uvádějí např. Hartke (1978) a Haladyna (2004). V zahraničí se tvorbou úloh zabývají i profesionálové z velkých testovacích center (největším na světě je Educational Testing Service v USA), které do vývoje úloh investují nemalé částky. Celý proces tvorby vysoce kvalitních úloh je velmi finančně náročný. Haladyna (2004) odhaduje cenu za vývoj jedné nové úlohy pro vysoce kvalitní testovací programy v USA na 1000 $. Zkušený odborník specializovaný na návrh testových úloh byl schopen denně napsat 5 až 15 úloh dle jejich povahy, což se považovalo za dobrý výkon (Wesman, 1971). Dnes jich pravděpodobně se zvyšujícími se nároky na úlohy (úlohy testující vyšší úrovně osvojení než jen zapamatování) napíše méně. U nás se tvorba úloh považuje zatím spíše za příležitostnou práci než za práci na plný úvazek na profesionální úrovni (Schindler 2006).

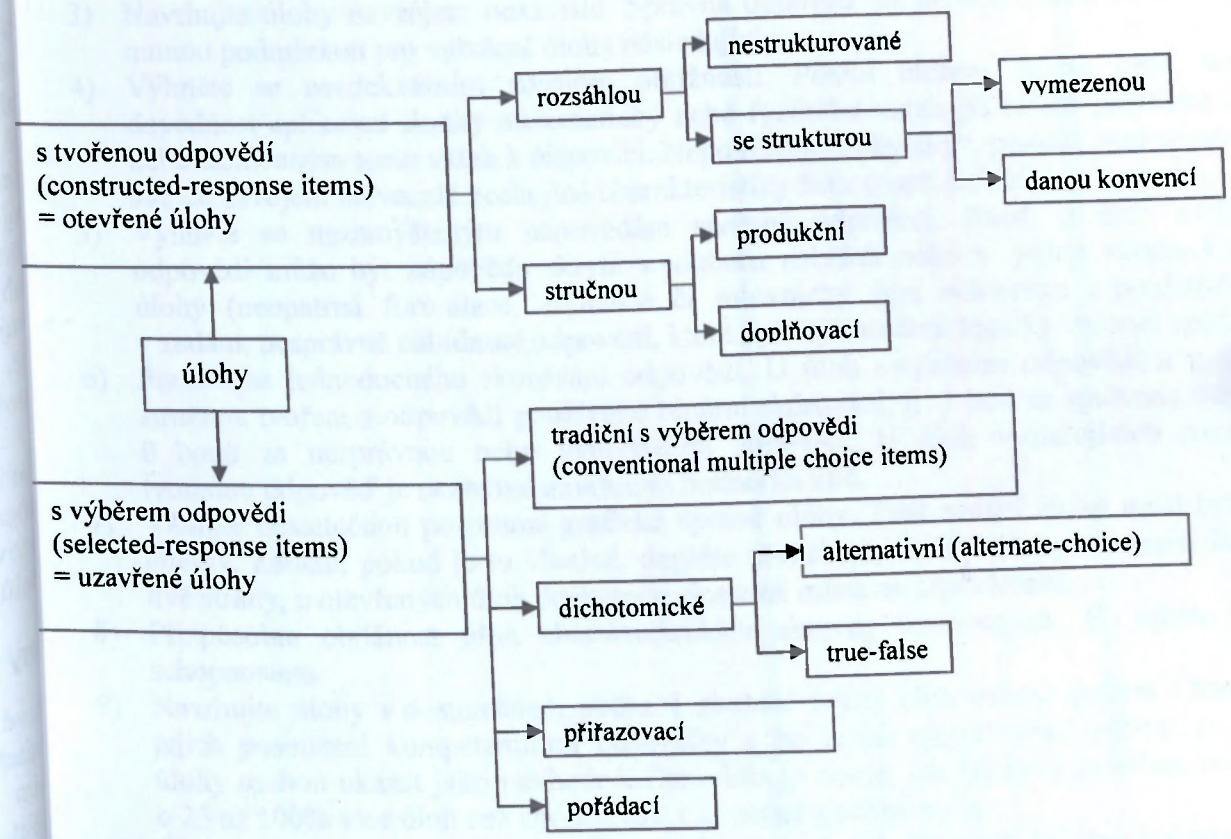
Vzhledem k tomu, že ne všechny původně navržené úlohy se ukáží jako vhodné, je třeba navrhovat mnohem více úloh, než kolik jich má být nakonec do testu zařazeno. Zkušenosti odborníků v testování ukazují, že by měl být počet navrhovaných úloh o 25-100% (v případě použití IRT[18] modelů při tvorbě úloh o 100-200%) vyšší než počet zamýšlených úloh v testu.

Druhy testových úloh (test item types)

Testové úlohy jakožto základní jednotky testu se stejně jako celý test mohou lišit svým obsahem, formátem, způsobem zadávání, způsobem vyhodnocování (skórování) a způsobem zpracování. Jedním z nejzákladnějších rozdílů testových úloh je typ odpovědi. Z tohoto pohledu lze úlohy rozdělit na *úlohy s výběrem odpovědi* (*selected-response items*) a *úlohy s tvořenou odpovědí* (*constructed-response items*). V odborné literatuře se setkáváme s různým dělením úloh. Přehled jednotlivých druhů úloh uvádějí např. Byčkovský (1982), Chráska (1999), Urbina (2004). Úlohy s výběrem odpovědi a úlohy se stručnou tvořenou odpovědí jsou úlohy uzavřené, protože představují omezený počet možností, z kterých testovaný vybírá. Díky jejich objektivnímu vyhodnocování se tyto úlohy označují jako objektivní. Úlohy s tvořenou odpovědí jsou naopak úlohy otevřené, žák v nich odpověď sám vytváří.

Bez ohledu na druh a formu má být součástí každé úlohy zadání, které jasně vymezuje, co se od testovaného požaduje. Součástí úlohy je také způsob jejího hodnocení a seznam povolených pomůcek. Přehled úloh používaných v didaktických testech je na obr. 4-8.

[18] O IRT hovoříme podrobněji dále v kap. 4.



Obr. 4-8 Druhy testových úloh (upraveno podle Byčkovský 1983, Urbina 2004)

Doporučení k tvorbě objektivních úloh uvádějí např. Byčkovský (1983), Haladyna a Roid (1982) či Herman (1988). Pro úlohy, k jejichž řešení je potřeba produktivního myšlení (higher order thinking), uvádí doporučení Haladyna (1997).

Testová úloha se zpravidla skládá z několika částí: z instrukce, výchozího textu, kmenu úlohy, případně nabídek odpovědí (distraktorů a správné odpovědi). Ne všechny úlohy musí obsahovat všechny tyto části, záleží na jejich druhu. **Instrukcí** se rozumí pokyny, co mají testovaní dělat. **Výchozí text** je materiál, který se přímo vztahuje k řešení úlohy, jej jejím východiskem. Většinou jde o text, ale může to být také obrázek, tabulka, graf či schéma. Výchozí text navozuje problém nebo situaci, kterou pak žáci posuzují či interpretují (Schindler a kol. 2006). Za **kmen úlohy** se považuje zadání ve formě otázky, pokynu či neúplného tvrzení. **Nabídky** jsou součástí úloh s výběrem odpovědi. Zahrnují správnou odpověď a distraktory (nesprávné odpovědi).

Obecná doporučení pro návrh všech druhů testových úloh

(upraveno podle Byčkovský 1983)

Při navrhování kvalitních testových úloh bychom měli dodržovat obecná doporučení a zásady platné pro většinu testových druhů úloh.

1) Úlohami testujte především důležité učivo nebo jevy. Testové úlohy by měly odrážet hlavní cíle, kterých má být při výuce dosaženo. Zaměření obsahu testu významně ovlivňuje to, na co žáci při učení kladou důraz. Úlohy vyžadující mechanické zapamatování nevýznamných podrobností, znalosti okrajových fakt a nedůležitých údajů zpravidla vedou žáky k tomu, že se soustředí na pamětní učení. Naopak úlohy na aplikaci přispívají k tomu, že se žáci při svém učení snaží uplatňovat vyšší kognitivní procesy než jen zapamatování.
2) Úlohy formulujte jasně, srozumitelně, stručně, nápaditě, ale úplně. Obdobné formulace unavují a snižují pozornost studenta. Používejte jen běžných cizích slov a jejich počet omezte na minimum, netvořte dlouhá souvětí.

3) Navrhujte úlohy navzájem nezávislé. Správná odpověď na určitou úlohu by neměla být nutnou podmínkou pro vyřešení úlohy následující.
4) Vyhněte se neadekvátním zdrojům obtížnosti. Pokud úlohou chcete např. testovat dovednost aplikovat složitý matematický nebo fyzikální vztah při řesem praktické úlohy, dejte testovaným tento vztah k dispozici. Nepoužívejte „chytáky", protože jimi se nezkouší stupeň osvojení učiva, ale zcela jiné charakteristiky žáka (např. postřeh, vtip apod.).
5) Vyhněte se nezamýšleným nápovědám správné odpovědi. Např. u úloh s výběrem odpovědi může být nápověda skryta v souboru nabídek nebo v jejich vztahu k zadání úlohy (neopatrná formulace, stylizace či mluvnický tvar některého z použitých slov v zadání, nesprávně nabídnuté odpovědi, které bystřejší student logicky vylouči apod.).
6) Používejte jednoduchého skórování odpovědí. U úloh s výběrem odpovědi a u úloh se stručnou tvořenou odpovědí používejte binární skórování, tj. 1 bod za správnou odpověď, 0 bodů za nesprávnou nebo vynechanou odpověď. U úloh vyžadujících rozsáhlejší tvořenou odpověď je nezbytné navrhnout hodnotící klíč.
7) Věnujte dostatečnou pozornost grafické úpravě úlohy. Text zadání úlohy musí být dobře čitelný. Zadání, pokud je to vhodné, doplňte obrázkem. Zadání úlohy se nesmí lámat na dvě strany, u otevřených úloh poskytněte dostatek místa na zápis řesem.
8) Přizpůsobte obtížnost úloh charakteristikám skupiny testovaných, tj. jejich věku a schopnostem.
9) Navrhujte úlohy v dostatečném počtu. I zkušení tvůrci úloh musejí počítat s tím, že po jejich posouzení kompetentními odborníky a po jejich empirickém ověření se některé úlohy mohou ukázat jako nevhodné. Zpravidla je nutné, jak již bylo zmíněno, navrhnout o 25 až 100% více úloh než chceme mít v konečné podobě testu.
10) Úlohy nechte vždy posoudit kompetentními osobami (alespoň jedním kolegou, který vyučuje týž předmět). Je to nejdůležitější doporučení, protože takto lze objevit řadu menších či větších nedostatků v zadání úloh, které unikly jejich tvůrci. Při posuzování úloh používejte speciálních formulářů.

Úlohy s tvořenou odpovědí

Úlohy s tvořenou odpovědí, v literatuře mnohdy označované jako úlohy otevřené (v americké literatuře známé jako *constructed-response items* nebo-li *free-response items*) jsou úlohy vyžadující od studenta tvořenou odpověď. Podle rozsahu požadované odpovědi se otevřené úlohy dále dělí na úlohy s rozsáhlou a úlohy se stručnou odpovědí.

Úlohy s rozsáhlou odpovědí

V otevřených úlohách s rozsáhlou odpovědí se od studenta vyžaduje delší souvislá logicky uspořádaná psaná odpověď nebo řešení, např. pojednání na určité téma, vyřešení zadaného problému, popis postupu, rozsáhlejší výpočet, apod. Někdy je zadáním úlohy přesně vymezena struktura požadované odpovědi. Kmen je záměrně formulován široce, aby umožňoval individuální přístup studenta k odpovědi.

Otevřené úlohy s rozsáhlou odpovědí patří k tradičním úlohám, jsou vhodné zejména při testování komplexních vědomostí či dovedností, jejichž osvojení trvá delší dobu, tedy pro testování vyšších kognitivních úrovní osvojení učiva (syntéza a hodnotící posouzení, také pro analýzu a aplikaci). Podle podrobnosti zadání dále dělíme úlohy s rozsáhlou odpovědí na strukturované a nestrukturované. Hlavní výhodou těchto úloh je jejich snadné navrhování, hlavní nevýhodou jejich velmi obtížné vyhodnocování, které je subjektivně ovlivněno posuzovatelem (navíc může být problémy přečíst rukopis testovaného). Náročné a pracné bývá stanovení jasných kritérií hodnocení a sestavení klíče pro všechna možná řešení. Při skórování úloh s rozsáhlou odpovědí se zpravidla postupuje tak, že za správné a úplné zodpovězení úlohy se přisuzuje určitý počet bodů. Eseje například můžeme hodnotit buď analyticky nebo holisticky (přímo). Při analytickém hodnocení se

posuzují samostatně různé charakteristiky jako je např. obsah, kreativita, styl, syntax, pravopis. Při holistickém posuzování hodnotíme esej jako celek. U obou způsobů je zapotřebí skórovací klíč.

Eseje se dnes již vyhodnocují i automaticky (Automated Essay Scoring) pomocí počítačových programů (viz dále). Příklad šestistupňového klíče pro holistické posuzování eseji z rozřazovacího testu z angličtiny (English Placement Test)[19] používaného na California State University je v tab. 4-6. Eseje hodnotí zpravidla dva nezávislí posuzovatelé, maximální počet bodů je 6. Ve sporné situaci esej posuzuje ještě třetí člověk a jeho skór rozhodne.

Tab. 4-6 Příklad skórovacího klíče pro holistické posuzování eseje

Skór (v bodech)	Hodnocení	Typické charakteristiky posuzovaného eseje
6	**Výborně**	Text eseje jasně a do hloubky vystihuje všechny body zadání, prokazuje jasné a komplexní myšlení. Charakterizuje ji logický, promyšlený a bohatý psaný projev. Hlavní myšlenky jsou soudržné a dobře zdůvodněné. Zřetelně vykazuje obratné používání jazyka (příp. s drobnými chybami).
5	**Velmi dobře**	Text eseje jasně vystihuje zadání, avšak ne všechny body do hloubky. Prokazuje jasné myšlení s určitou mírou složitosti a hloubky myšlení. Charakterizuje ji promyšlený psaný projev. Hlavní myšlenky jsou dobře vyjádřeny. Může obsahovat určité množství chyb ve stylu, vyjadřování a větné skladbě.
4	**Dobře**	Text eseje prokazuje přiměřenou schopnost písemného projevu. Může obsahovat i do očí bijící chyby, které však nemají závažný vliv na sdělovaný význam. Vystihuje sice všechny hlavní myšlenky, ale ne všechny v dostatečné hloubce. Uvedené myšlenky jsou většinou zdůvodněny a ilustrovány. Psaný projev je přiměřeně obratný.
3	**Dostatečně**	Text eseje prokazuje částečnou jasnost myšlení, ale postrádá komplexnost. Psaný projev je částečně promyšlený, ale není úplně přesný. Hlavní myšlenky jsou vyjádřeny jen dílčím způsobem, mnohdy zjednodušeny. I když obsahuje značné množství chyb, částečné ovládnutí jazyka je ještě zřejmé.
2	**Nedostatečně**	Text eseje poukazuje na chybné pochopení tématu, vykazuje určité problémy v jasnosti a v komplexnosti myšlení. Písemný projev je nesouvislý a neucelený. Uvádí banální či zjednodušené obecnosti bez udání důvodů či ilustrací. Vyskytují se časté pravopisné a stylistické chyby, chyby ve větné skladbě.
1	**Zcela nedostatečně**	Text eseje vykazuje zásadní nedostatky ve schopnosti písemného projevu a v porozumění zadání. Zcela postrádá smysluplné, logické uspořádání myšlenek. Nejsou vyjádřeny žádné hlavní myšlenky. Obsahuje závažné chyby (stylistické, ve větné skladbě, pravopisné), které mají zásadní vliv na sdělovaný význam.

Automatické skórování esejů

Automatické vyhodnocování esejů (Automated Essay Scoring, AES) je definováno jako počítačová technologie, která vyhodnocuje a skóruje psanou prózu (např. Shermis & Burstein, 2003). AES systémy se používají převážně pro úsporu času, nákladů, z důvodu reliability a zobecnitelnosti při vyhodnocování psaného textu (Bereiter, 2003; Burstein, 2003; Hamp-Lyons, 2001; Myers, 2003; Page, 2003; Rudner & Liang, 2002). Nejrozšířenějšími komerčními programy na automatické

[19] Eseje se vyhodnocovaly i automaticky pomocí programu *E-rater TM*, rozbor chyb se prováděl pomocí speciálních pomůcek (*Critique Writting Analysis Tools*). Referoval o tom Dr. Paul A. Ramsey z Educational Testing Service v roce 2003 v Intenzivním kurzu Konstrukce a analýza testů pořádaném UK v Praze.

skórování esejů ať pro účely low-stakes (školní testy) či high-stakes testování (důležité testy) jsou: Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), E-rater (více Attali, Burstein 2006) a Criterion, IntelliMetric a MY Access a nekomerční Bayesian Essay Test Scoring System (BETSY). *Project Essay Grader*™ (PEG) byl vyvinut 1966 na žádost College Board. *Intelligent Essay Assessor*™ (IEA) je produktem firmy Pearson Knowledge Analysis Technologies. *E-rater*® (electronic essay rater, používá se již od roku 1999 a bylo provedeno více než 2 miliony vyhodnocení výsledků esejů, Ramsey 2003) a jeho výuková aplikace *Criterion*℠ (systém podporovaný webem) byly vyvinuty společností Educational Testing Service (ETS). Zkrácené odpovědi hodnotí ETS automaticky programem C-Rater, který je vhodný pro všechny vědní obory. Zjišťuje se jím pojmová přesnost. *IntelliMetric*™ je znám jako první nástroj na vyhodnocování esejů založený na umělé inteligenci (Elliott, 2003; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003). Je používán společností College Board. *MY Access!*® je jeho výuková aplikace podporovaná webem. *Bayesian Essay Test Scoring System*™ *(BETSY)* je nekomerčním programem vyvinutým L. M. Rudnerem z USA.

Řada studií dokazuje, že automatické vyhodnocování esejů je stejně přesné jako vyhodnocování lidskými posuzovateli (Attali, 2004; Burstein & Chodorow, 1999; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003; Vantage Learning, 2000a, 2000b, 2001b, 2002, 2003a, 2003b).

Dnes se AES systémy běžně používají jako výukové nástroje (instructional tools) ve školních třídách v USA (např. MY Access!® and CriterionSM) a jako spoluhodnotitelé vedle lidských posuzovatelů při širokoplošném standardizovaném testování (např. ETS používá od roku 1999 e-rater vedle lidských posuzovatelů ke skórování GMAT esejů). AES se stále vyvíjí (více o systémech pro automatické vyhodnocování esejů např. Dikli 2006, Attali; Burstein 2006).

Doporučení pro návrh úloh s rozsáhlou odpovědí
(upraveno podle Hrabal, Lustigová, Valentová 1992)

1) Zařazujte do testu z časových důvodů jen málo širokých úloh.
2) Nepoužívejte otázek z učebnic ani takových, které připouštějí pouhou reprodukci textu učebnice.
3) Specifikujte to, nač se má testovaný soustředit a co považujete za okrajové, pokud toto rozlišení není vlastním cílem úlohy.
4) Poskytněte testovaným dostatek času a sdělte jim předem časový limit.
5) Připravte si ke každé úloze předem detailní vzorovou odpověď s ohodnocením.

Příklad: Matematická úloha vyžadující rozsáhlé řešení
Charakterizujte jednotlivé kuželosečky (uveďte jejich vlastnosti) a načrtněte jejich graf.

Příklad: Esej na téma (z English Placement Test na California State University)
Občas se setkáváme s názorem, že hudba nejen poskytuje lidem zábavu, ale má i vliv na jejich myšlení a chování.
Do jaké míry dokáže hudba podle vašeho názoru posluchače nejen bavit, ale i ovlivňovat? Své názory zdůvodněte nebo doložte příklady ze své četby, svého okolí či vlastní zkušenosti.

Otevřené úlohy se stručnou odpovědí

V úlohách se stručnou odpovědí se od studenta vyžaduje, aby uvedl stručnou odpověď (slovní, numerickou, grafickou aj.) na vyznačené místo. Rozlišujeme úlohy **produkční** (short answer) a **doplňovací** (completion item). Produkční úlohy tvoří kmen ve tvaru úplného tvrzení či otázky, zatímco u úloh doplňovacích je zadání neúplné tvrzení, které má student na vyznačených místech doplnit (fill-in-the-blanks). I tyto úlohy se poměrně snadno navrhují, ale neumožňují na rozdíl od

úloh s výběrem odpovědi uhádnutí správné odpovědi bez příslušných vědomostí. Zpravidla se předpokládá, že vytvořit odpověď je pro studenta náročnější než rozpoznat správnou odpověď z uvedených nabídek. Nevýhodou úloh se stručnou odpovědí je možná víceznačnost pochopení zadání úlohy, časová náročnost a obtížnější skórování. Odpovědi mohou být zřejmě nejen správné a nesprávné, ale i částečně správné. Mnohdy žák odpoví správně, i když jinak, než si představoval tvůrce testu, proto by autor testu měl úlohu formulovat co nejpřesněji, aby se vyhnul nejednoznačnosti zadání úlohy. Odpovědi na tyto úlohy nemůže z výše uvedených důvodů skórovat kdokoli, ale jen osoba s potřebnou kvalifikací. Úlohami se stručnou odpovědí lze testovat většinu znalostí, porozumění i aplikaci. Nehodí se pro analýzu, syntézu a hodnotící posouzení (viz Bloom 1956).

Doporučení pro návrh úloh se stručnou odpovědí (upraveno podle Byčkovský 1983)

1) Úloh používejte jen v případě, že lze odpovědět velmi stručně, nejlépe jedním slovem.
2) Předem uvažte všechny možné správné odpovědi, a je-li jich mnoho, raději úlohu nepoužívejte.
3) Předem připravte klíč pro vyhodnocování odpovědí.
4) Nechte v úlohách dostatek místa pro uvedení odpovědi; úlohu uspořádejte tak, aby byla snadno skórovatelná.
5) Místo na doplnění umisťujte pokud možno na konec věty.
6) Nevyžadujte doslovné znění naučeného textu z učebnice, ale naopak podporujte u studentů vlastní formulace.
7) Dávejte přednost produkčním úlohám před doplňovacími. Chcete-li přece jen použít doplňovací úlohy, dodržujte následující doporučení:
8) Vynechané slovo musí mít podstatnou úlohu ve větě, vynechávejte jen důležité údaje.
9) Z neúplné věty musí být studentovi jasné, co se má doplnit.
10) Pokud má být do úlohy doplněno více údajů, vynechejte pro doplnění zhruba stejné místo.

Příklad doplňovací úlohy:
Housenka vyleze za den 8 *cm*, v noci o 2 *cm* sklouzne. Jestliže byla housenka na jeho vrcholu čtvrtý den, stéblo trávy je _____cm dlouhé.

Příklad produkční úlohy:
Která tři z uvedených písmen sem nepatří? (převzato z Byčkovský, Bazantová 2005)

A	**B**	**X**
C	**D**	**Y**
E	**F**	**Z**

Odpověď: _____

Úlohy s výběrem odpovědi nebo-li úlohy uzavřené

Úlohy s výběrem odpovědi známé také jako objektivní úlohy či úlohy s danými nemennými odpověďmi jsou označovány jako úlohy uzavřené. Představují omezené množství nabídek, z kterých testovaný vybírá. Mezi tyto úlohy (dělení podle např. Urbina 2004) se řadí úlohy **multiple-choice**, **dichotomické** (true-false), **přiřazovací** (matching) a **uspořádací** (ranking).

Multiple-choice úlohy (Conventional Multiple Choice Items) – MC úlohy

Klasické multiple-choice úlohy patří mezi nejpoužívanější druhy úloh v testech. Jedná se o úlohy s nabízenou odpovědí, které se skládají z kmene tvořeného otázkou nebo neúplným tvrzením, příp. krátkým textem, obrázkem, schématem či grafem a několika nabídkami (alternativami) odpovědí. Chybným odpovědím se říká distraktory. Vzhledem k povaze testovaného učiva mají MC úlohy univerzální použití, lze jimi zjišťovat osvojení téměř všech kognitivních úrovní (znalost, porozumění, aplikace, analýza i jednodušší případy hodnotícího posouzení). Skórování úloh je

jednoduché, přesné a rychlé. Pokud se odpovědi zaznamenávají na speciální archy, lze je pomocí počítače s optickým scannerem skórovat automaticky. K výhodám MC úloh patří kromě již zmíněné univerzálnosti použití a snadnému skórování také možnost měnit pomocí distraktorů obtížnost úlohy a menší pravděpodobnost uhádnutí správné odpovědi než u dichotomických úloh.

MC úlohy se v testech vyskytují v několika formách. Základní formou je úloha s jednou správnou odpovědí. Dalšími modifikacemi jsou úlohy s nejlepší (nejpřesnější) odpovědí, úlohy se záporem v kmeni nebo-li úlohy s nesprávnou odpovědí, úlohy s vícenásobnou správnou odpovědí a úlohy situační.

a) Klasické úlohy s výběrem odpovědi (correct-answer type)

Student vybírá jednu správnou odpověď z nabízených 3 až 5 alternativ. Kmen se skládá obvykle z otázky typu: kdo, co, jaký, kam, kde a kdy.

Příklad: Jestliže Honzovi trvá 6 hodin vypracovat seminární práci z biologie (předpokládejme konstantní pracovní tempo), jakou část práce dokončí za 2 hodiny?

(A) $\frac{1}{12}$ (B) $\frac{1}{8}$ (C) $\frac{1}{6}$ (D) $\frac{1}{4}$ (E) $\frac{1}{3}$

Příklad: Housenka vyleze za den 8cm, v noci o 2cm sklouzne. Jak vysoké je stéblo trávy, jestliže housenka byla na jeho vrcholu čtvrtý den?
a) 24 cm b) 26 cm c) 28 cm d) 32 cm

Příklad aplikace v nové situaci: (převzato z testu *SAT*)
Pro všechna přirozená čísla n definujeme **n** následujícím způsobem:
je-li n sudé, pak **n** = n/2,
je-li n liché, pak **n** = 2n.
Je-li y prvočíslo >2, pak **2y** je rovno
(A) y/4 (B) y/2 (C) y (D) 2y (E) 4y

Příklad netypické MC úlohy: (převzato z testu *SAT*)
V měně státu Ug je 15 dopů 1 tif. Je-li 10 dopů 1 dedakop, kolik tifů je 6 dedakopů?
A) 1/15 B) 2/5 C) 3/2 D) 3 E) 4

Distraktory mohou přímo ovlivnit obtížnost úlohy. V následujícím příkladu jsou nabídky od sebe dost vzdálené (heterogenní), a tak je snadné i pro studenta, který toho o americké historii moc neví, zvolit správnou odpověď.

Příklad:
Kdo byl hlavním autorem Prohlášení nezávislosti?
A) Abraham Lincoln
B) Thomas Jefferson
C) Franklin Roosevelt
D) Král Jiří II.
E) Kateřina Veliká

Příklad: těžší varianta
Kdo byl hlavním autorem Prohlášení nezávislosti?
A) George Washington
B) Thomas Jefferson
C) Alexander Hamilton
D) Benjamin Franklin
E) James Madison

V této úloze je odpovědět správně na otázku obtížnější. Distraktory jsou homogenní.
(obě úlohy převzaty z *NBME Constructing Written Test Questions For the Basic and Clinical Sciences*)

b) Úlohy s nejlepší odpovědí (One-Best-Answer Items, best-answer type)

U těchto úloh není jedna nesporně správná odpověď a ostatní chybné, ale jednotlivé nabídky se liší mírou správnosti a úplnosti. Student má označit nejlepší či nejsprávnější odpověď. Otázka je často typu *jak* či *proč*. Tento typ úloh je pro žáky obtížnější než úlohy s jednou správnou odpovědí.

Příklad:
Pacient na pohotovosti si nemůže vzpomenout, jaké prášky bere na srdce. Říká, že mu zní v uších. Jeho tep je 80/min. EKG ukazuje dlouhotrvající PR a QRS intervaly. Které z následujících prášků nejspíše bere?
A) digoxin
B) lidokain
C) phenytoin
D) propranolol
E) chinidin
(převzato z *NBME Constructing Written Test Questions For the Basic and Clinical Sciences*)

c) Úloha se záporem v kmeni

U úloh se záporem v kmeni se od studenta požaduje nesprávná odpověď, která nesplňuje určitý požadavek či postrádá určitou vlastnost. U těchto úloh je nezbytné zdůraznit zápor v kmeni úlohy (např. podtržením, velkými písmeny, tučným písmem apod.), protože by ho žák mohl přehlédnout a následkem toho označit chybnou odpověď, i když patřičné vědomosti má.

Příklad: Poměr dvou kladných celých čísel je 3:2. Které z následujících tvrzení o těchto číslech NEMOHOU být pravdivá?
A) Jejich součtem je liché číslo.
B) Jejich součtem je sudé číslo.
C) Jejich součin je dělitelný 6.
D) Jejich součin je sudé číslo.
E) Jejich součin je liché číslo.
(převzato z testu *SAT*)

d) Úloha s vícenásobnou odpovědí (Complex MC)

V úloze s vícenásobnou odpovědí se od žáka požaduje výběr několika správných odpovědí. Počet možných odpovědí tedy převyšuje počet nabídek. Student musí být předem upozorněn, že existuje více správných odpovědí. Nevýhodou této úlohy je, že jsou možné i částečně správné odpovědi, což může vést k obtížím při skórování. Nedoporučuje se vyhodnocovat úlohy na principu „všechno nebo nic", tj. za úplnou správnou odpověď přidělíme 1 bod, za jinou odpověď 0 bodů, protože se tím snižuje citlivost položky. Vhodnějším způsobem je skórovat úlohu tak, že za každou označenou správnou odpověď a neoznačenou nesprávnou odpověď získá student 1 pomocný bod. Výsledný počet bodů potom vydělíme počtem nabídek v úloze, takže student za jednu úlohu může získat max. 1 bod. Pro zjednodušení skórování lze úlohu snadno převést na úlohu s jednou správnou odpovědí.

Příklad: Kružnice $x^2 + y^2 = 10$ je symetrická podle
A) osy x
B) osy y
C) podle libovolné přímky procházející počátkem
- správná odpověď: A, B i C

Převedení na úlohu s jednou správnou odpovědí.

Příklad: Kružnice $x^2 + y^2 = 10$ je symetrická podle

I. osy x
II. osy y
III. podle libovolné přímky procházející počátkem

A) pouze I
B) pouze II
C) I a II
D) II a III
E) I, II a III
- správná odpověď: E

Příklad: Jestliže $\frac{x}{3} = x^2$, jakou z daných hodnot může mít x?

I. $-\frac{1}{3}$ II. 0 III. $\frac{1}{3}$

(A) jen I
(B) jen II
(C) jen III
(D) jen II a III
(E) I, II a III

(úlohy převzaty z testu *SAT*)

e) Situační úloha

Zvláštní formou úlohy s výběrem odpovědi je úloha označovaná jako situační či interpretační. Jedná se o úlohy s větším počtem nabídek, než je obvyklé, které však nejsou explicitně vyjmenovány ve formě dlouhého seznamu, ale vyplývají z dané situace. Předností těchto úloh je malá pravděpodobnost uhádnutí správné odpovědi bez příslušných vědomostí.

Příklad: Na místo označené hvězdičkou doplňte takovou číslici, aby výsledné šesticiferné číslo bylo dělitelné 11.

76*394

Ačkoli to na první pohled nevypadá, i tato úloha je úlohou s vícenásobným výběrem odpovědi, protože student vybírá jedno z deseti čísel (0-9). Pravděpodobnost uhádnutí správné odpovědi je tedy pouhých 10%.

Testovým úlohám s výběrem odpovědi bývá často vytýkáno, že jsou výrazně snazší než úlohy otevřené produkční. To platí ale pouze pro ty úlohy, které měří znalost konkrétních poznatků. Zkušenosti odborníků totiž ukazují, že dobře navržené MC úlohy testující vyšší úrovně osvojení jako je porozumění a aplikace bývají stejně nebo naopak dokonce obtížnější než úlohy produkční.

Navrhování dobrých testových úloh s výběrem odpovědi je velmi pracné a náročné na čas, vyžaduje praktické zkušenosti a také jistou dávku tvořivosti. Na druhou stranu lze ale jednoduše a rychle zpracovávat a skórovat jejich výsledky. Nejtěžší je návrh atraktivních (přijatelných) distraktorů. Při jejich navrhování bychom měli dodržovat určité zásady a doporučení.

Zásady a doporučení pro tvorbu úloh s výběrem odpovědi
(upraveno podle Haladyna; Downing 2002, Byčkovský 1983)

Obsah úlohy

1. Každá úloha by měla odrážet přesně vymezený obsah, který byl předmětem výuky, a specifickou mentální úroveň osvojení učiva.
2. Testujte jen důležité učivo, vyhněte se triviálnímu či příliš speciálnímu obsahu úloh.
3. Používejte nový, neobvyklý materiál k testování vyšších úrovní osvojení učiva. Obměňte formulace při psaní úlohy, které byly použity ve vyučování nebo se nacházejí v učebnici. Netestujte pouze zapamatování konkrétních poznatků.
4. Nepoužívejte úlohy pro testování příliš rozsáhlých znalostí nebo komplexních řešení.
5. Nevztahujte obsah jedné úlohy na obsah úlohy jiné. Jednotlivé úlohy by měly být navzájem nezávislé.
6. Vyhněte se úloh „chytáků", protože se jimi nezkouší stupeň zvládnutí učiva, ale jiné charakteristiky jako např. postřeh, vtip atd.
7. Používejte jednoduchého srozumitelného slovníku blízkého testovaným.
Vyhněte se nahromadění cizích slov.
8. Nechte úlohy posoudit nezávislými kompetentními osobami s ohledem na jejich správnost a jasnost.

Formát úlohy

Doporučuje se formátovat úlohu raději vertikálně místo horizontálně pro usnadnění čtení, ale výzkumy ukazují, že horizontální uspořádání úlohy (kvůli ušetření místa) nemá vliv na výsledek studenta v testu. Úloha je rozdělena na kmen, pod kterým se nacházejí pod sebou v jednom či dvou sloupcích jednotlivé nabídky seřazené do určitého logického (je-li to možné, např. podle podrobnosti, přesnosti) či numerického sledu (podle velikosti a stejným způsobem, např. intervalem). Jednotlivé nabídky označte písmeny.

Stylová stránka úlohy

1. Upravujte a obměňujte úlohy.
2. Formulujte úlohu jasně, stručně, ale úplně. Dodržujte pravidla českého pravopisu a správnou větnou skladbu, používejte jednoznačných slov. Vyhněte se příliš dlouhé slovní formulaci úlohy. Všechny nabídky by měly být v gramatické a logické shodě s kmenem úlohy.
3. Věnujte dostatečnou pozornost grafické úpravě úlohy.

Psaní kmene úlohy

1. Ujistěte se, že kmen je formulován jasně a stručně. Z kmene musí být jasná podstata úkolu, který má student provést. Čtení příliš dlouhého kmene žáky zatěžuje.
2. Kmen by měl být smysluplný sám o sobě i bez nabídek.
3. Zahrňte základní myšlenku do kmene místo do nabídek odpovědi.
4. Vyhněte se negativní formulaci kmene. Pokud záporu chcete užít, zvýrazněte ho (např. velkými písmeny, tučným písmem či podtržením).
5. Nepoužívejte v kmeni slov nebo údajů, které by mohly sloužit jako nápověda.
6. Z textu kmene úlohy s nejlepší odpovědí musí být studentovi jasné, že se po něm vyžaduje nejpřesnější a ne správná odpověď. Zdůrazněte tuto skutečnost vhodnou formulací, např. „...nejvíce vyhovuje..", „... se dá nejlépe popsat..", apod.

Psaní nabídek úlohy

1. Navrhujte dostatek smysluplných a jednoznačně nesprávných distraktorů (foils, distractors), tj. chybných odpovědí tak, aby byly pro testovaného s nedostatečnými vědomostmi stejně atraktivní jako správná odpověď. Výzkumy v USA ukazují, že stačí 3 distraktory. Nebyla prokázána závislost počtu distraktorů na obtížnosti úlohy, ale úlohy s více atraktivními distraktory byly více citlivé (rozlišující). Nenavrhujte však více jak 5 distraktorů.

2. Ujistěte se, že správná či bezpochyby nejvhodnější odpověď je jen jedna. Úlohy s vícenásobnou správnou odpovědí raději převádějte na úlohy s jednou správnou odpovědí.
3. Vyhněte se slovní asociaci mezi kmenem a správnou odpovědí.
4. Správnou odpověď umisťujte mezi distraktory zcela náhodně.
5. Distraktory se nesmějí obsahem navzájem překrývat nebo jinou formou vyjadřovat totéž, musejí být navzájem nezávislé.
6. Soubor nabízených odpovědí k jedné úloze by měl být homogenní jak svým obsahem, tak i dramatickou strukturou.
7. Zachovejte přibližně stejnou délku formulace nabídek, aby žádná nabídka „nevyčnívala" od ostatních.
8. Nepoužívejte negativní formulace nabídek.
9. Jen opatrně nebo spíš vůbec nepoužívejte nabídek typu: „nic z uvedeného", „vše z uvedeného", „nevím".
10. Vyhněte se ve formulaci úlohy nápovědě. Zdrojem nápovědy mohou být
    - některá slova: vždy, nikdy, absolutně, naprosto, úplně
    - neopatrná stylizace
    - mluvnický tvar některého ze slov použitých v zadání úlohy
    - očividně absurdní, nesprávné nabídnuté odpovědi, které lze logicky vyloučit
    - zřejmá správná odpověď – podstatně delší nebo podrobněji formulovaná
11. Používejte při formulaci distraktorů typické chyby studentů.
12. Humoru používejte velmi opatrně, spíš výjimečně nebo raději vůbec.

Dichotomické úlohy

U dichotomických testových úloh jsou studentovi předkládány dvě alternativy odpovědi, z nichž tu správnou má předem stanoveným způsobem označit (např. podtržením, zakroužkováním apod.). U těchto úloh se někdy rozlišují úlohy alternativní a tzv. true-false úlohy (Urbina 2004). Dichotomické úlohy se snadno navrhují, což může vést k vytvoření triviálních úloh. Naopak jejich nedostatkem je velká pravděpodobnost (50%) uhádnutí správné odpovědi i bez příslušných vědomostí, kterou lze ale zmenšit dostatečným počtem těchto úloh.

a) Alternativní úlohy (Alternate-choice) se skládají z kmene a dvou nabídek. Lord (1977) uvedl logický argument ve prospěch těchto úloh, že většina průměrných a výborných studentů zúží svůj výběr na dvě pravděpodobné možnosti (vyloučí nefunkční distraktory). I další studie tuto variantu úlohy podpořily (Downing, Haladyna, 1992-93). Pokud je test dostatečně dlouhý, má faktor „hádání" malý vliv na celkový výsledek v testu, navíc vytvoření dvou nabídek je snazší než vytvoření tří až pěti, jak je to obvyklé u klasických MC úloh.

Příklad:
Ze 48 sehraných zápasů hokejový tým prohrál dvanáctkrát. Jaký je poměr jeho proher vzhledem k jeho výhrám?
A) 1:4
B) 1:3

b) True-False úlohy (TF: two-choice/ binary choice) se skládají většinou z nějakého tvrzení, z konstatování jistého faktu (píp. otázky), o kterém má student rozhodnout, zda je správné či nikoli. Student vybírá odpověď z nabídek typu: ANO – NE, správně – chybně/ nesprávně, pravda (fakt) – úsudek (mínění).

Příklad 1: (převzato z *Haladyna, Downing, Rodriguez 2002*)
Zaškrtni správnou odpověď.
Hlavním městem Uruguaye je Montevideo. *ANO - NE*

Příklad 2: (převzato z *Hrabal, Lustigová, Valentová 1992*)
Zaškrtni správnou odpověď.
Jistý muž chtěl mít fontánu uprostřed jezera. Vedl tedy trubku z hlubiny u dna, kde je vysoký tlak až nad hladinu, kde je tlak nízký. Bude jeho fontána opravdu fungovat? *ANO - NE*

Modifikací této úlohy je tzv. trs úloh (Multiple True-False Items = MTF items) tvořen zpravidla 3 až 20 nabídkami, kdy je několik true-false úloh obsahově vázáno na společný kmen úlohy. Tento typ úloh je neobvyklý.

Příklad: (upraveno podle *Haladyna, Downing, Rodriguez 2002*)
Jste uznávaný ekologický farmář. Znáte tajemství růstu silných a zdravých rostlin. Které z následujících tvrzení by popisovaly Vaše farmářské metody?

1. Když zasadíte nějaké fazole, ujistíte se, že fazole budou dobře chráněné před sluncem, aby dostaly jen málo světla nebo žádné světlo. *ANO - NE*
2. Když zaséváte semínka, ujistíte se, že budou zalévány a udržovány ve vlhké půdě. *ANO - NE*
3. Zaséváte semínka jen za vhodné teploty. *ANO - NE*
4. Protože víte, jak dochází k opylování, nastříkáte své plodiny insekticidy, abyste zabránili včelám a ostatnímu hmyzu poškozování plodin. *ANO - NE*

Doporučení pro návrh dichotomických úloh
(upraveno podle Byčkovský 1983, Chráska 1999)

1) Tvrzení v úloze formulujte tak, aby bylo jednoznačně správné či nesprávné.
2) Nepoužívejte příliš dlouhých a komplikovaných tvrzení; každé tvrzení by mělo obsahovat pouze jednu hlavní myšlenku.
3) V tvrzeních nepoužívejte dvojího záporu, vyhýbejte se použití negativní otázky.
4) V tvrzeních se pokud možno vyhněte příslovcím typu: často, vždy, nikdy, téměř, zřídka apod., protože většinou usnadňují odpověď.
5) Navrhujte přibližně stejný počet správných a nesprávných tvrzení.
6) Nepoužívejte vět vytržených z učebnice, ani je neobměňujte jejich negací.

Přiřazovací úlohy (Matching items)

Přiřazovací úlohy se skládají ze dvou množin pojmů a z instrukce. Po studentovi se vyžaduje přiřazení 3 – 12 pojmů z první skupiny (tzv. návěstí) k pojmům z druhé skupiny (doplňky) tak, aby správně zachycovalo jejich vzájemný vztah vymezený v instrukci. Soubor doplňků představuje nabídky společné všem úlohám (navěstí). Souborům návěstí a doplňků předchází instrukce, která spolu s návěstím vytváří kmen položky. Soubor doplňků může být početnější než soubor návěstí (tzn. že některé doplňky zůstanou nepřiřazeny) nebo naopak počet doplňků menší než soubor návěstí (tj. jeden doplněk přiřadíme k více jak jednomu návěstí). Dokonalé přiřazení, kdy počet návěstí a doplňků je shodný, se z důvodu ulehčení nedoporučuje.
Přiřazovací úlohy můžeme použít k testování úrovní osvojení učiva jako je zapamatování, porozumění a jednodušší aplikace. Výhodou těchto úloh je snadné skórování, úspora místa oproti tradičním MC úlohám a možnost ověření homogenní učební látky větším počtem jednoduchých

úloh. Jejich předností je také poměrně nízká pravděpodobnost uhádnutí správného řešení ($P = \frac{1}{k^n}$, k je počet prvků návěstí, n je počet doplňků).

Doporučení pro tvorbu přiřazovacích úloh (upraveno podle Byčkovský 1983)

1) Navrhujte homogenní soubory doplňků a návěstí. Porušení homogenity doplňků může někdy vést k usnadnění odpovědi logickým vyloučením nehomogenního doplňku, porušení homogenity návěstí může studenta dezorientovat.
2) Dávejte přednost krátkým doplňkům.
3) Nenavrhujte nadměrné soubory návěstí a doplňků, max. 10 pojmů.
4) Zásadně nerozdělujte úlohu na dvě stránky testového zadání.
5) Návěstí a doplňky uvádějte v navzájem oddělených sloupcích. Návěstí se umisťují zpravidla vlevo a číslují se, zatímco doplňky stojí vpravo a označují se písmeny.

Příklad: (převzato z Byčkovský, Bažantová 2005)
Ke jménům autorů přiřaďte jejich díla.

(1) Karel Čapek	_____	(A) Krysař
(2) Josef Škvorecký	_____	(B) Bylo nás pět
(3) Fráňa Šrámek	_____	(C) Matka
(4) Viktor Dyk	_____	(D) Stříbrný vítr
(5) Karel Poláček	_____	(E) Hubička
		(F) Mirákl
		(G) Zbabělci

Pořádací úlohy (Ordering items)

Pořádací úlohy se skládají z instrukce a souboru pojmů patřící některou svou vlastností do jedné třídy. Student má za úkol uspořádat pojmy podle daného kritéria a daným způsobem (uvedeno v instrukci) do řady, např. podle velikosti, stupně obecnosti, podle významu, chronologicky apod. Uspořádací úlohy jsou zvláštním případem přiřazovacích úloh s dokonalým přiřazením, kde funkci doplňků plní pořadová čísla. Jsou tedy vhodné stejně jako úlohy přiřazovací pro testování osvojení učiva na úrovni zapamatování, pro jednodušší případy porozumění a aplikaci.

Pravděpodobnost uhádnutí správného řešení je poměrně nízká ($P = \frac{1}{k!}$. k je počet prvků).

Příklad: Seřaďte následující reálná čísla podle velikosti od NEJMENŠÍHO (1) po největší (5).

A) 0,60 B) $\frac{10}{15}$ C) - 0,56 D) $-\frac{11}{20}$ E) $\frac{\sqrt{2}}{2}$

1. _____ 2. _____ 3. _____ 4. _____ 5. _____

Tento typ úloh je poměrně pracný a časově náročný na skórování, protože mezi úplně špatnou a zcela správnou odpovědí existuje mnoho částečně správných. Z tohoto důvodu není vhodné skórovat úlohy binárně (za úplně správné pořadí 1 bod, za jiné 0 bodů), pokud obsahují více než 3 až 4 pojmy. Uplatnění binárního skórování totiž značně snižuje citlivost úlohy (tj. rozlišení mezi dobrými, průměrnými a slabými studenty). Při skórování úloh s větším počtem seřazovaných pojmů

doporučuje Byčkovský (1983) použít složitější, ale citlivější způsob skórování, kde se žákovi přidělí skór x podle vzorce $x = \frac{\sum d_{max} - \sum d}{\sum d_{max}}$,

kde d jsou jednotlivé odchylky žáka od správného řešení a d_{max} jsou největší možné odchylky žáka od správného pořadí (vytvoří se obrácené pořadí ke správnému, které se od něj maximálně vzdaluje, a odečtením správného a obráceného pořadí (v absolutní hodnotě) stanovíme největší možné odchylky). Užití vzorce ilustrujeme na příkladě fiktivních odpovědí jednoho žáka.

		správné pořadí	obrácené pořadí	d_{max}	pořadí u žáka	d
C	- 0,56	1	5	4 (=5-1)	2	1 (=2-1)
D	$-\frac{11}{20}$	2	4	2 (=4-2)	1	1 (=2-1)
A	0,60	3	3	0 (=3-3)	5	2 (=5-3)
B	$\frac{10}{15}$	4	2	2 (=4-2)	4	0 (=4-4)
E	$\frac{\sqrt{2}}{2}$	5	1	4 (=5-1)	3	2 (=5-3)
	součet $\sum$			12		6

$x = \frac{12-6}{12} = 0{,}5$. Výkon žáka v dané testové úloze bychom skórovali 0,5 bodu.

Doporučení pro tvorbu pořádacích úloh (upraveno podle Byčkovský 1983)

1) Volte takový soubor pojmů, který jde jednoznačně uspořádat.
2) Vyhněte se zařazování neexistujících pojmů nebo pojmů, které nepatří do jedné třídy.
3) Zařazujte jen pojmy, které lze krátce formulovat.
4) Nezařazujte více jak 6-8 pojmů.
5) Nezapomeňte v instrukci uvést způsob zaznamenávání a kritérium uspořádání pojmů.
6) S ohledem na snazší skórování upravte úlohu tak, aby student uváděl odpověď na předem stanovené místo.

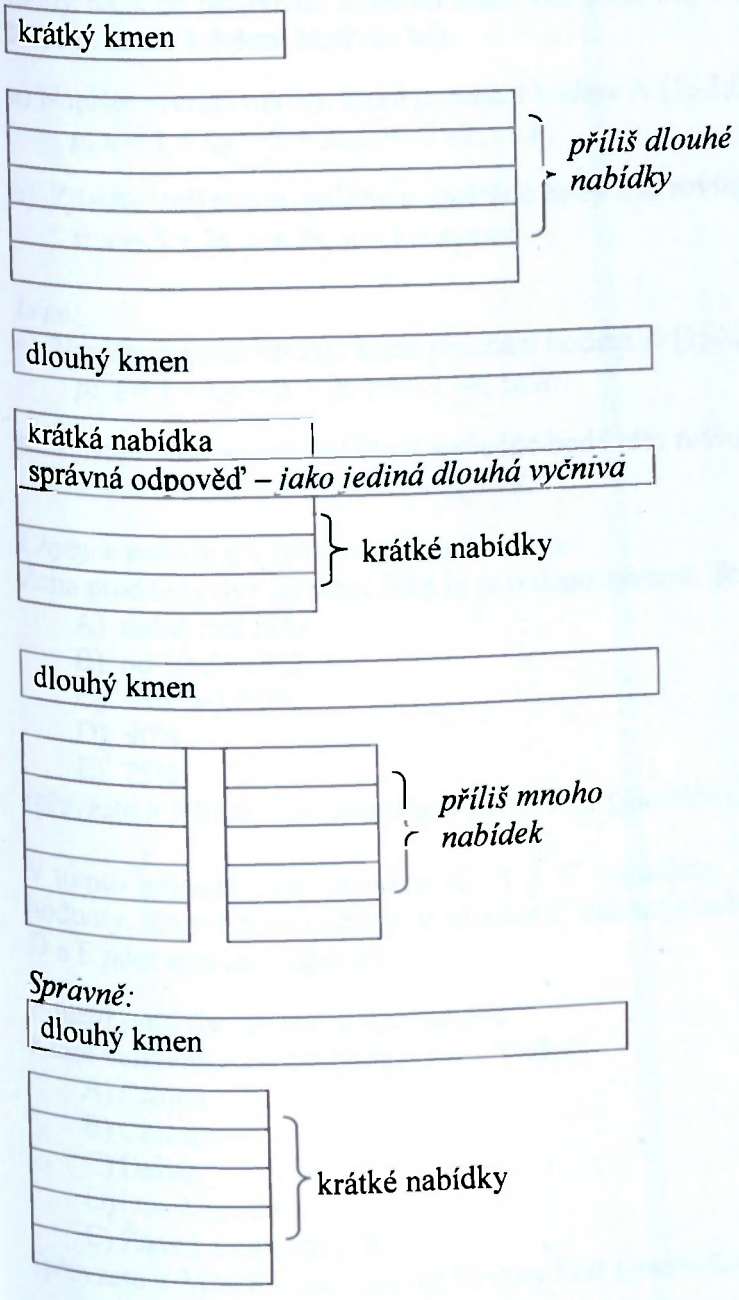
Příklady některých nedostatků u objektivních testových úloh

Ve školní praxi se stále ještě nezřídka setkáváme s nevhodně navrženými úlohami, které mají řadu nedostatků (Byčkovský; Bažantová 2005). Jde především o úlohy objektivní. Důvodem mohou být např. malé praktické zkušenosti či nedostatečná péče věnovaná přípravě úloh, nerespektování doporučení a zásad pro tvorbu úloh. Dále uvádím některé příklady nevhodně navržených úloh.

Příklad nedodržení obsahové homogenity (převzato z Byčkovský; Bažantová 2005)
Dyslexie je specifická
A) porucha rovnováhy
B) porucha čtení
C) porucha matematické schopnosti
D) porucha činnosti štítné žlázy

Lépe:
Dyslexie je specifická vývojová porucha učení, která se projevuje
A) poruchou pravopisu
B) poruchou čtení
C) poruchou psaní
D) motorickou neobratností

Příklad chyb v nabídkách u úloh s výběrem odpovědi (převzato z Byčkovský; Bažantová 2005)



Příklad nápovědy v zadání úlohy - gramatická souvislost s kmenem
(převzato z Byčkovský; Bažantová 2005)

Co je to varmuže?
(A) plněné šišky
(B) ochucené placky
(C) ovocná kaše
(D) syrové bochánky

Příklad nápovědy v zadání úlohy - logická souvislost s kmenem
(převzato z Byčkovský; Bažantová 2005)

Libreto k opeře B. Smetany Hubička napsala
(A) Eliška Krásnohorská
(B) Karolina Světlá
(C) Karel Sabina
(D) Marie Červinková

Příklad úlohy, která se skládá ze dvou navzájem závislých dílčích úloh (správně by měly být dílčí úlohy na sobě nezávislé, a pokud tomu tak není, mělo by být uvedeno správné řešení první části, které je nutné k řešení části druhé).

a) Napište rovnici roviny, která prochází bodem A [1;-3;0] a je kolmá k přímce
p: x = 1 + t,y = 3 + 2t, z = -7 -4t, t∈R.

b) Zjistěte vzájemnou polohu a společné body této roviny a přímky
s: x = 5 + 3s, y = 2s, z = 1 + s, s∈R.

Lépe:

a) Napište rovnici roviny, která prochází bodem A [1;-3;0] a je kolmá k přímce
p: x = 1 + t,y = 3 + 2t, z = -7 -4t, t∈R. [$x + 2y - 4z + 5 = 0$]

b) Zjistěte vzájemnou polohu a společné body této roviny a přímky
s: x = 5 + 3s, y = 2s, z = 1 + s, s∈R.

Chyby v nabídkách vyjádřených numericky.
Žena prodělala dvě infekce. Jaká je pravděpodobnost, že je žena neplodná?

A) méně než 20%
B) od 20 do 30%
C) více než 50%
D) 90%
E) 75%

(převzato z *NBME Constructing Written Test Questions For the Basic and Clinical Sciences*)

V tomto případě jsou nabídky A, B a C vyjádřeny rozmezím, zatímco D a E jsou konkrétní hodnoty. Kromě toho rozmezí v nabídce C zahrnuje nabídky D a E, což téměř automaticky vyřazuje D a E jako správné odpovědi.

Použití nabídky „žádné z nabízených"
Které velkoměsto je nejblíže k New Yorku?

A) Boston
B) Chicago
C) Dallas
D) Los Angeles
E) Žádné z nabízených

(převzato z *NBME Constructing Written Test Questions For the Basic and Clinical Sciences*)

Když student zvolí E, není zřejmé, zda myslí na Philadelphii či Londýn.

Netradiční úlohy

Kromě zmíněných tradičních úloh se objevují v testech i netradiční, tzv. problémové úlohy vyžadující řešení dobře strukturovaných problémů v autentických situacích (např. PISA 2003). Řešením problémových úloh rozumíme schopnost jedince využívat kognitivních procesů k řešení situací, v nichž není okamžitě zřejmý způsob jejich řešení. Rozlišují se problémy dobře strukturované (*well-structured*) a problémy špatně strukturované (*ill-structured*). K dobře strukturovaným patří úlohy, které mají jasný postup, definovatelná omezení a jasný počet řešení, o jejichž správnosti lze rozhodnout. Např. kódovací systém zubů je konečný, dospělý člověk má obvykle 32 zubů. Student buď musí určit kód k názvu zubu nebo naopak. Počet úloh je tedy konečný, obsahuje 64 úloh. Slovo *ill-structured* pochází od Simona (1973) a znamená vícenásobný, složený. Ill-structured problémy zahrnují více jak jednu správnou odpověď nebo jedno lepší řešení než druhé.

V životě jsme konfrontováni i špatně strukturovanými problémy, u kterých není zřejmý nebo jednoznačný postup řešení, mohou mít několik alternativ řešení, o jejichž správnosti nelze rozhodnout. Pokud se problémy týkají situací z reálného života, označují se jako *autentické* (např. Wiggins, 1993).
V následující tabulce jsou uvedeny charakteristiky dobře strukturovaných autentických problémových úloh použitých při mezinárodním výzkumu PISA realizovaném v roce 2003, jejichž řešení vyžaduje znalosti a postupy z několika oblastí.

Tab. 4-7 Charakteristiky řešení tří druhů problémů (upraveno podle Tomášek; Potužníková 2004)

	Rozhodování	**Systémová analýza a projektování**	**Odstraňování chyb**
Cíl	Volba možností za omezujících podmínek	Identifikace vztahů mezi částmi systému a/nebo návrh systému vyjadřující dané vztahy mezi částmi	Rozeznání a oprava chyb v systému či mechanismu
Postupy řešení	Porozumění situaci vymezené několika alternativami a omezeními řešení v zadání úlohy	Porozumění informacím charakterizující daný systém, a požadavkům stanoveným v zadání úlohy	Porozumění hlavním prvkům systému nebo mechanismu a jeho špatnému fungování, porozumění požadavkům úkolu
	Identifikace relevantních podmínek	Identifikace relevantních částí systému	Identifikace příčinných vztahů mezi proměnnými
	Znázornění možných variant	Znázornění vztahů mezi částmi systému	Znázornění fungování systému
	Rozhodování, které řešení je optimální.	Analýza nebo projekt systému, který postihuje vztahy mezi částmi	Nalezení nefunkčního místa v systému a/nebo navržení řešení
	Kontrola a posouzení řešení	Kontrola a posouzení analýzy nebo projektu systému	Kontrola a posouzení řešení
	Sdělení a/nebo zdůvodnění řešení	Sdělení analýzy nebo zdůvodnění navrženého projektu.	Sdělení nebo zdůvodnění zjištěného problému nebo řešení
	Rozhodování	**Systémová analýza a projektování**	**Odstraňování chyb**
Možné zdroje komplexnosti a obtížnosti úlohy	Počet omezujících podmínek	Počet proměnných a povaha vztahů mezi nimi	Počet vzájemně propojených částí systému nebo mechanismu a způsoby jejich vzájemné interakce
	Počet a druh použitých znázornění (verbální, obrazové, číselné)	Počet a druh použitých znázornění (verbální, obrazové, číselné)	Počet a druh použitých znázornění (verbální, obrazové, číselné)

Příklad problémové úlohy

typ problému: rozhodování
forma úlohy: uzavřená s tvořenou odpovědí

DOVOLENÁ (upraveno podle Tomášek; Potužníková 2004)

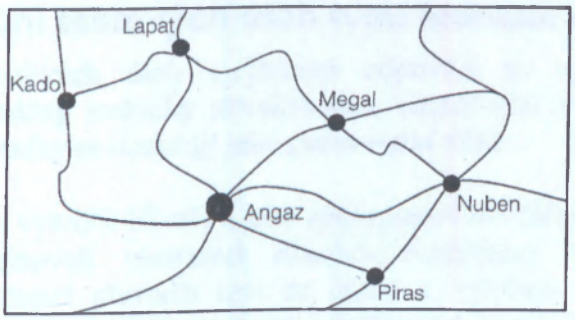
Vaším úkolem je naplánovat nejlepší trasu na dovolenou. Na obrázcích 1 a 2 lze najít jak umístění jednotlivých měst, tak i vzdálenosti mezi nimi.

a) Vypočítej nejkratší vzdálenost po silnici mezi městy Nuben a Kado.
Vzdálenost: ..... kilometrů.

b) Zorka žije v Angazu. Chce navštívit Kado a Lapat. Může ujet **nanejvýš 300 kilometrů denně**, ale svou cestu může přerušovat noclehy v kempech kdekoli mezi městy.
Zorka stráví v každém městě dvě noci, takže bude mít na prohlídku každého města celý jeden den. Sestav pro Zorku plán cesty a do následující tabulky doplň, kde bude Zorka nocovat.

Den	Nocleh
1	Kemp mezi městy Angaz a Kado
2	
3	
4	
5	
6	
7	Angaz

Obrázek 1: Silniční mapa



Obrázek 2: Nejkratší vzdálenost mezi městy v kilometrech.

Kado	550				
Lapat	500	300			
Megal	300	850	550		
Nuben	500		1000	450	
Piras	300	850	800	600	250
	Angaz	**Kado**	**Lapat**	**Megal**	**Nuben**

Hodnocení a poznámky k úloze a)
Úplná odpověď: 1050 km
Nevyhovující odpověď:
Jiný počet kilometrů či odpověď typu Nuben – Angaz – Kado bez udání počtu kilometrů.

Pro správné zodpovězení otázky musí žáci pomocí mapy určit vhodnou trasu a na základě vzdáleností uvedených v tabulce spočítat celkovou vzdálenost. Vyhledání některých vzdáleností vyzaduje čtení tabulky odspodu, na rozdíl od běžného čtení zleva dolů. Při podrobnějším pohledu na práci žáků byly pozorovány různé druhy chyb. Např. odpověď 1100 km naznačuje, že žák počítal vzdálenost po delší trase Nuben – Piras – Angaz – Kado. Jiní žáci zvolili správnou cestu, ale při počtu vzdálenosti udělali početní chybu.

Hodnocení a poznámky k úloze b)

Úplná odpověď:

Den	Nocleh
1	Kemp mezi městy Angaz a Kado
2	Kado
3	Kado
4	Lapat
5	Lapat
6	Kemp mezi městy Lapat a Angaz
7	Angaz

Částečná odpověď:

Jedna chyba. Za chybu považujte, když položka neodpovídá příslušnému dni: např. prohlídka města Lapat třetí den, název města šestý den, žádná položka šestý den.

Nevyhovující odpověď: jiná odpověď

K zadání této otázky je stanoven větší počet omezujících podmínek, které musí být splněny současně. Pro získání částečné dopovědi mohl žák udělat početní chybu, ale musel porozumět zadání otázky a provést základní analytické kroky řešení.

4. 4 Generování testových úloh typu multiple-choice

Navrhování kvalitních testových úloh s výběrem odpovědi je velmi náročné, proto byly v posledních 40 letech hledány techniky zkvalitňující, usnadňující a urychlující tvorbu většího počtu těchto úloh. Tyto techniky se označují jako ***generování úloh***.

Základy ke generování úloh vytvořil Hively a jeho spolupracovníci (Hively; Patterson; Page, 1968), když podrobně charakterizovali vlastnosti různých modifikací úloh v matematice. První algoritmickou teorii umožňující převádět text na úlohy s výběrem odpovědi navrhl Bormuth (1970). Jeho teorie byla pro svoji složitost v praxi málo využívána. O teoriích tvorby úloh psali v 80. letech 20. století Earlier; Roid a Haladyna (1982). Všechny tyto aktivity znamenaly začátek nové éry generování úloh. V současné době je generování úloh považováno za perspektivní vědní obor, který se stále vyvíjí. O generování úloh píší např. Irvine a Kyllonen (2002) či Wainer (2000), podle kterého generování úloh hraje nezbytnou roli především v kontextu počítačového testování, obzvláště počítačového adaptivního testování. Příkladem počítačového programu na automatické generování úloh je *Math Test Creation Assistant* (TCA) vyvinutý odborníky z Educational Testing Service.

Existuje několik užitečných technik generování úloh, které uvádí Haladyna (2004). Jsou to **skelety úloh** (item shells), **modelování úloh** (item modeling), **klíčové charakteristiky/ rysy** (key features), **generování pomocí scénářů** (generic item sets) a **konverze komplexních praktických úloh na soubor úloh s výběrem odpovědi** (CR item format conversion). Těchto pět metod představuje praktické techniky generování úloh, které urychlují jejich vývoj a poskytují systematický základ pro vytváření nových kvalitních MC úloh. Skelety úloh jsou technikou předpisující, která závisí na používání již existujících úloh. Modelování úloh má pevně danou strukturu kmene úlohy a každý model úlohy testuje jen jeden typ obsahu a kognitivní úrovně. Klíčové rysy jsou závislé na komisi expertů a mají systematický přístup. Jsou vhodné ke generování úloh jen ve specifickém kontextu, jakým je např. léčení pacienta či řešení klinických problémů. Sady obecně použitelných úloh se přibližují svým pojetím modelování, ale mají pevně danou strukturu otázky. Jsou vhodné pro všechny typy testování, zejména pro školní testování. Konverze komplexních praktických úloh poskytuje základ převedení komplexních úloh osvojující vyšší kognitivní úrovně na úlohy s výběrem odpovědi, které mají stejný nebo podobný požadavek na kognitivní úroveň osvojení, ale jsou na rozdíl od těchto úloh objektivně skórovatelné.

Techniky generování úloh (podle Haladyna 2004)

Skelety úloh (item shells)

Skelety úloh (item shells) jsou vhodné pro začínající tvůrce úloh s výběrem odpovědi. Jde o kostry úloh obsahující pouze syntaktickou strukturu MC úlohy, která umožňuje vytvořit rychle a snadno velké množství podobných úloh. Tato technika je vhodná pro osvojování faktických, konceptuálních a procedurálních poznatků na úrovních zapamatování, porozumění a aplikace. Nevýhodou je, že sice takto vytvoříme velký počet úloh, ale stejné struktury, přitom v testech se obvykle požaduje různorodost úloh. Jistým řešením je použít různé skelety úloh. Skelety úloh nelze také použít pro všechny vzdělávací obsahy, jsou vhodné především pro medicínu a jazyky. Existují dva způsoby vytváření skeletů.

Skelet úlohy tvoří kmen a několik nabídek, z nichž je zpravidla jen jedna správná. Protože je obtížné navrhnout větší počet smysluplných atraktivních distraktorů, doporučuje se používat dvou až čtyř, také automatické skórování je potom pohodlnější.

Prvním a jednodušším způsobem je přizpůsobit obecný skelet úloh* již úspěšně ověřeným úlohám. K tomu slouží seznam otázek tvořících kmen úloh, jež uvádí Haladyna (2004). Uvádím upravený výběr:

- Jak zní definice …?/ Jak lze nejlépe definovat …?
- Jaký je význam …?
- Co je synonymem k …?
- Co je podobné …?
- Co je charakteristické pro …?
- Co je rozdílné pro…?
- Čím se liší … od … ?
- Jaký je vztah mezi … a …?
- Co by se stalo, kdyby …?
- Co je příčinou/ důvodem …? Co způsobuje...?
- Co je následek/ důsledek …?
- Co je nejdůležitější (nejvýznamnější), popř. nejméně důležité (nejméně významné)?

* *Skelet úlohy:*

Kmen (otázka nebo neúplné tvrzení)

A. správná odpověď
B. atraktivní nesprávná odpověď (distraktor)
C. atraktivní nesprávná odpověď (distraktor)
D. atraktivní nesprávná odpověď (distraktor)

Druhým způsobem je převést velmi úspěšné úlohy na skelety úloh. Je třeba identifikovat typ myšlenkových procesů, vymezit obsah učiva, napsat správnou odpověď a 3-4 smysluplné distraktory. Postupujeme podle následujících kroků:

1. Urči kmen úspěšné úlohy (successfully performing items)
2. Podtrhni klíčová slova nebo fráze představující obsah úlohy.
3. Urči variace pro každé klíčové slovo/ frázi.
4. Vyber jednu variaci pro klíčová slova.
5. Napiš kmen úlohy.
6. Napiš správnou odpověď.
7. Napiš požadovaný počet distraktorů.

Např. Skelety úloh z medicíny (Haladyna 2004)

Porozumění

Jaké jsou hlavní symptomy .....?

☞ Tento skelet pomáhá vytvořit velké množství úloh zabývající se symptomy lidských nemocí.

Předvídání/ predikce

Co je nejběžnější (příčinou nebo symptomem) (problému pacienta)?

☞ Tento obecný skelet umožňuje vytvořit kombinace na základě principu příčina-důsledek. Porozumění pojmů je důležité pro úspěšné řešení těchto úloh.

Aplikace znalostí a dovedností

A) Je diagnostikována pacientova nemoc. Jaké léčení je nejvhodnější, nejefektivnější?

☞ Tento skelet vytváří úlohy na množství lidských nemocí na základě taxonomie nebo typologie nemocí a léčebných alternativ, z nichž jedna je nejlepší.

B) Je poskytnuta informace o problému pacienta. Jak by měl být pacient léčen?

☞ Skelet poskytuje informaci o nemoci či zranění pacienta. Kompletní úloha od testovaného požaduje, aby na základě daných informací určil správnou diagnózu a spravný způsob léčení.

Konkrétní příklad z medicíny:

1. Urči kmen úspěšné úlohy (successfully performing items)

Šestileté dítě bylo přivezeno do nemocnice s pohmožděninami břicha a hrudníku jako následky autonehody. Jaké by mělo vypadat počáteční ošetření?

2. Podtrhni klíčová slova nebo fráze představující obsah úlohy.

Šestileté dítě bylo dovezeno do nemocnice s pohmožděninami břicha a hrudníku jako následek autonehody. Jak by mělo vypadat počáteční ošetření?

3. Urči variace pro každé klíčové slovo/ frázi.

Věk osoby: dítě do 3 let, dítě 3-12, dítě 13-18, mladý dospělí 19 – 31, střední věk 32 –59, stáří přes 60 let

Urazové zranění nebo komplikace: škrábnutí/ řezné rány, pohmožděniny, zlomeniny, vnitřní zranění

Typ nehody: automobilová, doma, na rekreaci, pracovní uraz

4. Vyber věk, úrazové zranění nebo komplikace a typ nehody

Dítě do 3 let odřeniny nehoda na kole

5. Napiš kmen úlohy.

Malé dítě do 3 let je přivezeno do nemocnice s těžkými odřeninami jako následek nehody na kole s matkou. Jak by mělo vypadat počáteční ošetření?

6. Napiš správnou odpověď.

A) Provést vizuální vyšetření.

7. Napiš požadovaný počet distraktorů.

B) Zamezit infekci.
C) Podat lék proti bolesti k uklidnění dítěte.
D) Poslat na laboratorní testy.
E) Vyčistit rány dezinfekčním prostředkem.

Skelet úlohy z chemie pro 8. třídu na téma plyny a jejich charakteristiky (Haladyna 2004):

1. Urči kmen úlohy.

Co je typickou charakteristikou vodíku?

2. Podtrhni klíčová slova nebo fráze představující obsah úlohy.

Co je typickou charakteristikou vodíku?

3. Urči variace pro každé klíčové slovo/ frázi.

Co je typickou charakteristikou (plynu studovaného k tomuto tématu ve vyučování)?

4. Vyber příklad z řady variant.

Kyslík

5. Napiš kmen úlohy.

Co je typickou charakteristikou kyslíku?

6. Napiš správnou odpověď.

B) Je vedlejším prvkem ve vodě.

7. Napiš požadovaný počet distraktorů.

A) Má menší hustotu než vodík.
C) Lze ho po částečně destilovat.
D) Má nižší bod varu než vodík.

Modely úloh (item modeling)

Modelování úloh je technika, při které vytváříme obecný model úlohy, na jehož základě můžeme vytvořit velké množství podobných úloh. Teoretická východiska této metody zveřejnili v roce 1968 Hively, Patterson, Page a Osburn. V praxi bylo modelování úloh použito 1973 v rozsáhlém projektu výuky matematiky a přírodních věd na základních školách v Minessotě týmem v čele s Hivelym. Modelů se využívá při tvoření úloh s výběrem odpovědi a úloh se stručnou tvořenou odpovědí pomocí počítače, kdy je nezbytné generovat větší počet úloh jednoho typu zhruba stejné obtížnosti. Model úlohy nespecifikuje pouze kmen, ale většinou také poskytuje základ pro vytvoření správné odpovědi a distraktorů (v nejpropracovanější formě modelů). Na rozdíl od skeletů úloh je modelování vhodné i pro měření vyšších úrovní kognitivních procesů. Modely lze dobře uplatnit především v oblasti čtení, psaní, matematiky a v oblasti medicíny, tedy v předmětech obsahově měřitelných, kvantitativních. Méně vhodné jsou pro společenské, přírodovědné vědy a porozumění textu. Modelování úloh lze využít jak pro rozsáhlé sumativní testování, tak i pro dílčí formativní testování ve třídě. Modely úloh bývají někdy označovány jako šablony úloh či fazety.

Příklad modelu úlohy s výběrem odpovědi z matematiky z pravděpodobnosti
(upraveno podle Haladyna 2004):

V {krabici} je x červených, y žlutých a z modrých předmětů. Jeden z nich vytáhneme. Jaká je pravděpodobnost, že vybraný předmět bude {červený, žlutý, modrý}?

A. 1/n	[atraktivní distraktor]
B. 1/{x, y nebo z}	[atraktivní distraktor]
C. {x, y nebo z}/{x+y+z}	[správná odpověď]
D. {x, y nebo z}/{součet nevybraných objektů}	[atraktivní distraktor]

Příklad úlohy:
V pytlíku jsou 2 červené, 4 žluté a 6 modrých kuliček. Jednu z nich vytáhneme. Jaká je pravděpodobnost, že vybraná kulička bude žlutá?

A. 1/12
B. 1/4
C. 1/3
D. 2/3

Příklad využití modelování úloh vyžadujících stručnou tvořenou odpověď
(elementární algebra, upraveno podle Haladyna 2004):
Modely úloh pro počítání s absolutní hodnotou:
$|a+b|+|x+y|$ či $|a+b|-|x+y|$

Generační pravidla:
a, b, x, y jsou jednociferná nenulová celá čísla

Příklad úlohy:
Vypočítejte.
$|5-3|-|1-8|=$

Klíčové rysy (Key features)

Myšlenka klíčových rysů se objevila na začátku 80. let 20. století v Kanadě v medicíně (Page a Bordage a kol.) a nebyla dosud v žádném jiném oboru aplikována. Klíčový rys je buď důležitý, rozhodující krok v myšlenkovém procesu při řešení problému (např. léčení pacienta) nebo krok v tomto procesu, ve kterém může nastat chyba, která sníží efektivitu řešení problému (pacientova léčení). Tento krok se nazývá klíčový rys, protože pomáhá rozlišovat mezi uchazeči různých schopností. Klíčové rysy jsou vhodné jak pro MC úlohy, tak i pro úlohy otevřené se stručnou

odpovědí. Používají se pro řešení klinických problémů, kde je nutné zapojit vyšší myšlenkové procesy, nikoli pro měření znalostí a dovedností.
Klíčový rys se skládá zpravidla ze stručného kmene, po němž následuje několik otázek, které má testovaný zodpovědět výběrem z daného dlouhého seznamu možných správných odpovědí. Správných odpovědí je většinou více.

Postup při vývoji key features v medicíně (upraveno podle Haladyna 2004)

1. Jasně definujte specifickou oblast klinických problémů, potíží a diagnóz.
2. Promyslete si detailní plán vyšetření.
3. Předložte klinické situace.
4. Vyberte pro každý problém klíčové rysy (2-5).

Příklad: Problém z medicíny zahrnující 3 navzájem propojené klíčové rysy
Dospělý pacient si stěžuje na bolavou oteklou nohu. Lékař by měl

- zahrnout do své diagnózy trombózu hlavních tepen
- zjistit rizikové faktory pro trombózu hlavních tepen v jeho rodině
- nařídit venogram jako definitivní test pro trombózu hlavních tepen

5. Vyberte konkrétní případ a napište pro něj scénář zahrnující všechny relevantní informace a několik otázek.
6. Vytvořte skórovací klíč.
7. Proveďte pilotní otestování.

Příklad úlohy zahrnující klíčové rysy (Page 1995)
56 letý muž, Pavel, Vás navštíví v ambulantní klinice a stěžuje si na bolest levé nohy, která začala před dvěma dny a od té doby se postupně zhoršuje. Jeho noha je citlivá pod kolenem a oteklá kolem kotníku. Nikdy neměl podobné problémy. Jeho druhá noha je v pořádku.
Otázka 1: Jakou diagnózu zvažujete? Napište až tři možnosti.
Otázka 2: S ohledem na Vaši diagnózu co z jeho rodinné anamnézy byste obzvláště chtěli zjistit? Vyberte ze seznamu až sedm možností.

1. Činnost na počátku symptomů
2. Příjem alkoholu
3. Alergie
4. Angina pectoris
5. Protizánětlivé léčení
6. Kouření cigaret
7. Barva stolice
8. Kašel
9. Bolesti hlavy
10. Zvracení krve (hematemesis)
11. Hormonální léčení
12. Impotence
13. Bolest v lýtkových svalech při chůzi
14. Bolest v dolní části zad
15. Noční pomočování (nokturie)
16. Bušení srdce (palpitace)
17. Pravidelná porucha citlivosti v končetinách
18. Záchvatová noční dušnost
19. Chorobná žíznivost (polydipsie)
20. Dřívější problémy s kolenem
21. Dřívější problémy se zády
22. Dřívější nádorový růst (neoplazie)
23. Dřívější infekce močového traktu
24. Nedávný zubní zákrok
25. Nedávné znehybnění (imobilizace)
26. Nedávné bolesti krku
27. Nedávný operační zákrok
28. Pracovní prostředí v poslední době
29. Zranění na noze
30. Zranění na ruce

Scénář se skládá ze dvou otázek. První se vztahuje na předběžnou diagnózu. Testovaný by měl být schopen předložit bez možnosti výběru 3 smysluplné hypotézy o původu pacientových potíží. Tato část úlohy je tedy otevřená se stručnou tvořenou odpovědí. Druhá část úlohy je ve formátu úlohy s vícenásobným výběrem odpovědi, kdy testovaný má vybrat až 7 alternativ z nabídnutých 30.

Generování pomocí scénářů (Generic item sets, GIS)

Tato technika je vhodná k simulaci komplexního myšlení, tj. pro testování vyšších kognitivních úrovní ve formě objektivně skórovatelných MC úloh (příp. otevřených úloh se stručnou odpovědí). Využívá se konceptu skeletů úloh, ale v mnohem propracovanější formě. Kořeny této techniky spadají do 80. let 20. století (Guttman, Hively). Generování pomocí scénářů je vhodné pro kvantitativní předměty (podobně jako modelování úloh) jako např. pro účetnictví, statistiku, medicínu či farmacii. Pro svou univerzálnost se tato technika, jak uvádí Haladyna (2004), stává stále populárnější. Pomocí těchto scénářů lze generovat velké množství testových úloh k různým účelům – formativnímu, sumativnímu testování či k přípravě na test nebo k domácí přípravě na výuku. Výsledkem je soubor několika spolu souvisejících úloh. Scénář má sice danou strukturu, ale tvůrce úloh má volnost psát zajímavé scénáře a identifikovat v nich faktory, které se dají systematicky měnit. Psaní správné odpovědi je jednoduché, psaní distraktorů zato vyžaduje tvořivost. Klíčovým prvkem GIS je obecný scénář (viněta), jímž je krátký příběh obsahující většinou důležité informace k řešení problému.

Příklad scénáře pro studenty na začátku statistického kurzu – očekávané dovednosti studentů (upraveno podle Haladyna 2004)

Je dána situace, kde je použita korelace mezi dvěma proměnnými. Student má

- určit výzkumnou otázku/ hypotézu
- určit měřitelné náhodné jevy X a Y
- určit statistické proměnné *x* a *y* zastupující X a Y
- zformulovat nulovou hypotézu a alternativní hypotézy
- stanovit testové kritérium a počet stupňů volnosti
- stanovit statistickou hladinu významnosti α
- určit pozorovanou hladinu významnosti, kterou naznačují výsledky, porovnat ji s danou hladinou významnosti a rozhodnout, zda nulovou hypotézu zamítnout či nikoli
- napsat spolu s výsledky závěr s ohledem na nulovou/ alternativní hypotézu
- diskutovat možnost chyb prvního a druhého druhu v tomto problému
- zformulovat závěrečný výrok s ohledem na výzkumnou otázku/ hypotézu

Je vyučováno a testováno 18 běžných statistických testů. S použitím každého testu existují 4 varianty:

a) statistická/ zadaná a pozorovaná hladina významnosti jsou dány
b) statistická/ zadaná hladina významnosti je dána, ale pozorovaná hladina významnosti nikoli
c) není dána statistická hladina významnosti, ale je dána pozorovaná hladina významnosti
d) není dána ani statistická, ani pozorovaná hladina významnosti

Tedy zkoušená oblast zahrnuje 72 možností. Jakmile je vygenerován scénář, lze vytvořit čtyři logické variace k jednotlivým scénářům:

a) *statistická/ zadaná a pozorovaná hladina významnosti jsou dány*
Dva výzkumníci studovali u 42 mužů a žen vztah mezi množstvím spánku každou noc a kaloriemi spálenými na rotopedu. Získali korelaci 0,28. Příslušná pozorovaná hladina významnosti pro dvoustranný test je 0,08. Použili alternativní hypotézu a volili statistickou hladinu významnosti $\alpha = 0,05$.

b) *statistická/ zadaná hladina významnosti je dána, ale pozorovaná hladina významnosti nikoli*
Dva výzkumníci studovali u 1442 mužů a žen vztah mezi množstvím spánku každou noc a kaloriemi spálenými na rotopedu. Získali korelaci 0,11. Příslušná pozorovaná hladina významnosti pro dvoustranný test je 0,08. Použili alternativní hypotézu a volili statistickou hladinu významnosti $\alpha = 0,05$.

c) *není dána statistická hladina významnosti, ale je dána pozorovaná hladina významnosti*
Dva výzkumníci studovali u 12 mužů a žen vztah mezi množstvím spánku každou noc a kaloriemi spálenými na rotopedu. Získali korelaci 0,68. Příslušná pozorovaná hladina významnosti pro dvoustranný test je 0,12. Použili alternativní hypotézu a volili statistickou hladinu významnosti $\alpha = 0,05$.

d) *není dána ani statistická, ani pozorovaná hladina významnosti*
Dva výzkumníci studovali u 42 mužů a žen vztah mezi množstvím spánku každou noc a kaloriemi spálenými na rotopedu. Získali korelaci 0,13. Příslušná pozorovaná hladina významnosti pro dvoustranný test je 0,28. Použili alternativní hypotézu a volili statistickou hladinu významnosti $\alpha = 0,05$.

Úloha z oblasti podnikání založená na obecné tabulce (upraveno podle Haladyna 2004)
Sylvia Vasquez prodává telefony v nákupním středisku u řeky. Pomozte jí dát dohromady, jak se daří jejímu podnikání. Tabulka obsahuje údaje o jejím hypotetickém malém podnikání, je základem pro generování úlohy, přičemž druh zboží, jméno obchodníka, všechna čísla mohou být upravena k vytvoření výhodných, výdělečných a nevýhodných situací.

A	B	C	D	E	F
Typ telefonu	Zakoupeno kusů	Cena za jednotku	Cena prodeje	Počet prodaných kusů	Obdržen obnos
Ekonomický	400	$ 28	$ 36	127	?
Lepší	250	$ 67	$ 84	190	?
Nejlepší	125	$ 125	$ 275	15	?

Vzor kmenů obecné úlohy vycházející z tabulky.

1. Který typ telefonů se prodává nejlépe?
2. Který typ telefonů je {nejvýnosnější, nejméně výnosný} za jednotku?
3. Který typ telefonů je {nejvýnosnější, nejméně výnosný} celkovým prodejem?
4. Jaké je rozpětí zisku za jednotku pro {ekonomický, lepší. nejlepší} typ telefonu?
5. Jaký je hrubý příjem z {ekonomických, lepších, nejlepších} telefonů?
6. Jaký je hrubý příjem ze všech telefonů?
7. S ohledem na současný prodej a rozpětí příjmů jaká objednávka na další měsíc má největší smysl?
8. Jaké je rozpětí příjmů, předpokládáte-li příjmy za poslední měsíc a horní hranici 92%?

Konverze komplexních praktických úloh na soubor multiple-choice úloh

Poslední zmíněnou technikou generování úloh je konverze komplexních úloh na úlohy s výběrem odpovědi. Existuje řada důvodů, proč převádět komplexní úlohy na MC úlohy (upraveno podle Haladyna 2004):

1. Zadávání komplexní úlohy je časově náročnější než zadávání MC úlohy.
2. Komplexní úlohu musí skórovat jeden či dva odborníci podle důležitosti testu. Toto skórování je finančně nákladné. Naopak MC úloha ušetří náklady.
3. Výzkumy naznačují, že MC úlohy mohou často dobře zastoupit komplexní úlohy.

Odborníci oprávněně tvrdí, že komplexní úloha je většinou přesnější v tom, co přesně chceme testovat (více reliabilní), ale zřejmě lze obětovat trochu přesnosti větší efektivitě. S tímto kompromisem můžeme převést velice dobrou komplexní úlohu do MC formátu.

Příklad: (podle Haladyna 2004)
Jana měla dva balíčky stejných chipsů. Dbá o svou postavu, takže se podívala na nutriční hodnoty na jednom z těchto balíčků a zjistila: 1 balíček – 28g, energetická hodnota v balení: 140 kalorií, z toho 70 tuky. Složení: brambory, rostlinný olej, sůl

Celkem tuků	8 g	12%
nasycených	1,5 g	8%
Cholesterol	0 mg	0%
Sodík	160 mg	7%
Celkem sacharidů	16 g	5%
vláknina	1 g	4%
cukry	0 g	
Bílkoviny	2 g	
Doporučená denní dávka (v %)		
Vitamin A 0%		Vitamin C 8%
Vápník 0%		Zelezo 2%

1. Kolik kalorií měla Jana celkem ve dvou balíčkách chipsů?
) 70 B) 140 C) 280
. Její denní dávka sodíku je 2400mg. Měla moc sodíku?
) ano B) ne C) není poskytnuto dostatek informací
3. Její denní dávka gramů tuku je 65. Kolik tuku získala po snězení 2 balíčků chipsů?
) více než denní dávku
B) mnohem méně než její dávku
C) nelze říct z dané informace
4. Kolik procent vitamínu C získala vzhledem k její denní dávce?
A) 0% B) 2% C) 8% D) 16%
5. Co je hlavní složkou v tomto balíčku chipsů?
A) brambory B) rostlinný olej C) sůl

Tento set úloh zkouší studentovu schopnost porozumění čtenému textu či aplikaci znalostí k řešení problému.

4. 5 Administrace testu

Konečná verze testu je připravena k administraci (zadávání). Administrace testu může být buď individuální nebo hromadná. Kromě toho musíme rozlišit administraci testů formou tužka-papír, která umožňuje jak individuální, tak i hromadné testování, a administraci počítačem či on-line, ať již jde o testování neadaptivní či adaptivní, nebo v omezené míře hromadné neadaptivní testování. Adaptivní on-line testování je vždy individuální povahy.

4. 6 Vyhodnocování testu

Testy jsou dle své povahy vyhodnocovány buď ručně nebo automaticky. Pokud testujeme formou tužka-papír na speciálně upravené záznamové archy, je možné pomocí skenerů načíst data týkající se testovaných a jejich odpovědí na testové úlohy a pomocí speciálního software testy vyhodnotit (viz dále). Zatím však ve většině případů počítače nejsou schopny vyhodnocovat odpovědi na otevřené úlohy s rozsáhlejším řešením vícestupňově, ty musí podle předem stanoveného předpisu vyhodnocovat kompetentní posuzovatel či posuzovatelé a výsledek posouzení vložit do paměti počítače.

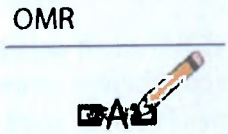
Automatické vyhodnocování

Ke sběru dat z papírových záznamových archů z testování se využívají k optickému snímání různé typy scannerů podle typu záznamových archů a technologií, na kterých jsou založeny. Mezi tyto technologie patří OMR (Optical Mark Recognition), OCR (Optical Character Recognition) a ICR (Intelligent Character Recognition) technologie. Rozlišují se tři základní typy scannerů: OMR scannery[20], OCR/ICR Imaging scannery[21], kombinované scannery[22] pro OMR, OCR i ICR a scannery, které nevyžadují připojení k počítači, tzv. Test Scoring Machines[23]. Dále budou popsány jednotlivé technologie snímání.

[20] Příkladem OMR scannerů a ovladacího software je OpScan, EZData od firmy NCS Pearson či Remark Office 5.0.
[21] Například software Image ScanTools, ScanTools Plus a NCS Accra software za použití speciální nabídky.
[22] Kombinovanými scannery jsou například OpScan *i*NSIGHT, 5000*i* scanner od firmy NCS Pearson, Cognition software od firmy Scantron.
[23] Příkladem je SelfScore Test Scoring Machine od firmy NCS Pearson.

OMR (Optical Mark Recognition)

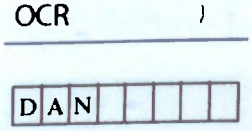
OMR je proces získávání dat (značek a čárových kódů) z určených oblastí na stránce papíru. OMR optický scanner najde pouze značku (většinou se jedná o křížek, nikoli tvar značky) nebo zjistí její nepřítomnost v definované oblasti pomocí změn v odrazivosti povrchu v místě označeném značkovačem (perem, propisovací tužkou či obyčejnou tužkou; pro větší rychlost se doporučují inkoustová pera a propisovací tužky, některá zařízení je dokonce vyžadují). Určitá zařízení používají speciální světlopropustný papír a pracují se změnami v množství světla procházejícího papírem. OMR scannery lze použít pro multiple choice úlohy, příp. dotazníky s Lickertovou škálou. Obvykle bývá vybrané pole (které může mít kulatý, hranatý, oválný či šestiúhelníkový tvar) označeno křížkem a v případě chyby zcela vybarveno a označeno jiné pole. Konkrétní nastavení se však může lišit dle možností použitého systému (scanneru a vyhodnocovacího software, viz obr. 4-9).
Předností OMR, jak uvádí společnost NCS Pearson, je její přesnost (až 99,9%) a rychlost získávání dat (2000–10 000 archů za hodinu dle scanneru). Při přijímacím řízení na PedF UK v roce 2006 byl použit OMR scanner pro snímání dat ze záznamových archů testu OSP (viz obr.9-1).



Obr. 4-9 Ukázka označení správné odpovědi (www.scantron.com)

OCR (Optical Character Recognition)

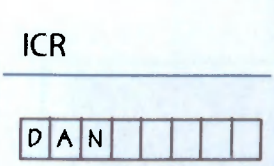
OCR technologie převádí tištěný strojový text do podoby zpracovatelné počítačem (viz obr. 4-10). Zpravidla se text převede optickým scannerem do bitmapového formátu (obrázku) a ten je následně analyzován rozpoznávacím software na základě identifikování šablon (pattern recognition) s využitím prvků umělé inteligence. Dnešní systémy již většinou nejsou omezeny na jeden typ použitého písma, ale dokáží zpracovat i text psaný neznámým fontem. Zpracování tištěného textu psaného latinkou je považováno za vyřešený problém, problémy však mohou nastat při čtení znakového písma obsahující velký počet znaků (např. čínština). OCR je o něco méně přesná než OMR, ale přesnější než ICR. NCS Pearson uvádí při dobře vyladěném systému, správně navržených záznamových arších a při manuální úpravě přesnost až 99,9%. Rychlost získávání dat je jen o trošku nižší než u OMR (1800 až 10 000 archů za hodinu dle typu scanneru). Často se používá v kombinaci s OMR technologií.



Obr. 4-10 Ukázka strojového zápisu písmen (www.scantron.com)

ICR (Intelligent Character Recognition)

ICR je technologií příští generace, která umožňuje rozpoznání ručně psaného hůlkového textu zapsaného po písmenech do jednotlivých políček (viz obr.4-11). Tento text (v bitmapovém formátu) se převádí do strojově čitelného textu. ICR je méně přesná technologie než OMR a vyžaduje manuální úpravu k získání perfektních dat. Nicméně v porovnání s ručním zadáváním dat umožňuje ICR mnohem rychlejší sběr dat. NCS Pearson uvádí při dobře vyladěném systému, správně navržených záznamových arších a čitelném písmu dokonce přesnost až 99,5%.



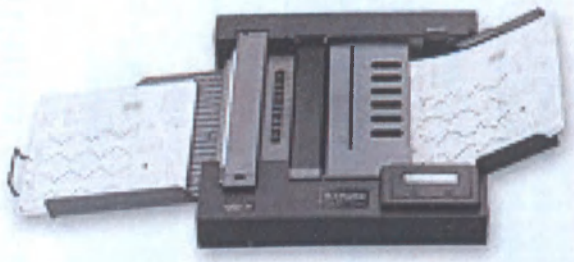
Obr. 4-11 Ukázka zápisu hůlkového písma (www.scantron.com)

Technologií se teprve rozvíjející je ICR, která dokáže rozpoznat již při samotném psaní hůlkové či psací písmo psané speciálním digitálním perem. Přitom využívá informace o individuálních rysech písma (např. rychlosti pohybu, sklonu písma). Je určena pro oblast průmyslu a finančnictví. Přesnost této metody ale zatím není dostačující, jak uvádí Wikipedia.

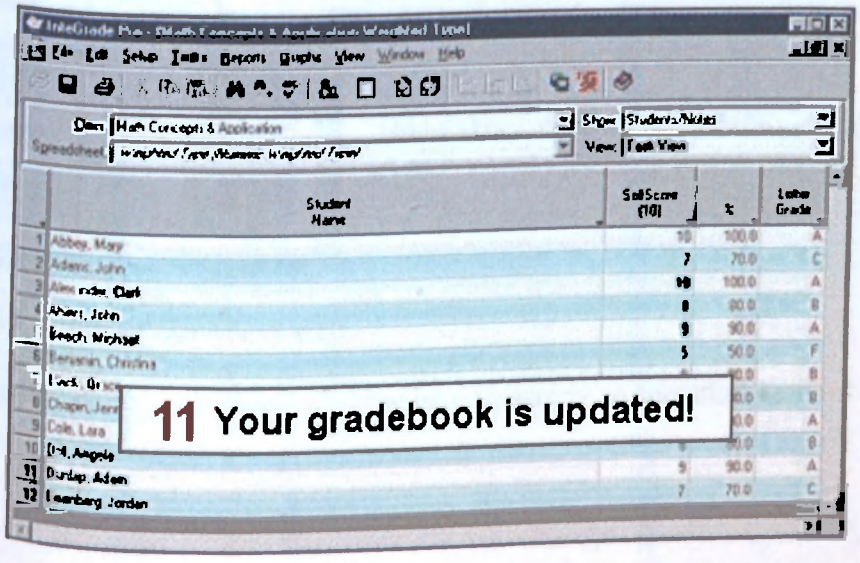
V některých praktických aplikacích je výhodné kombinovat více technologií. OCR, resp. ICR například pro získání základních údajů z hlavičky testu, kterou testovaný vyplní sám ručně, a OMR pro snímání odpovědí na multiple-choice úlohy.

SelfScore Test Scoring Machines

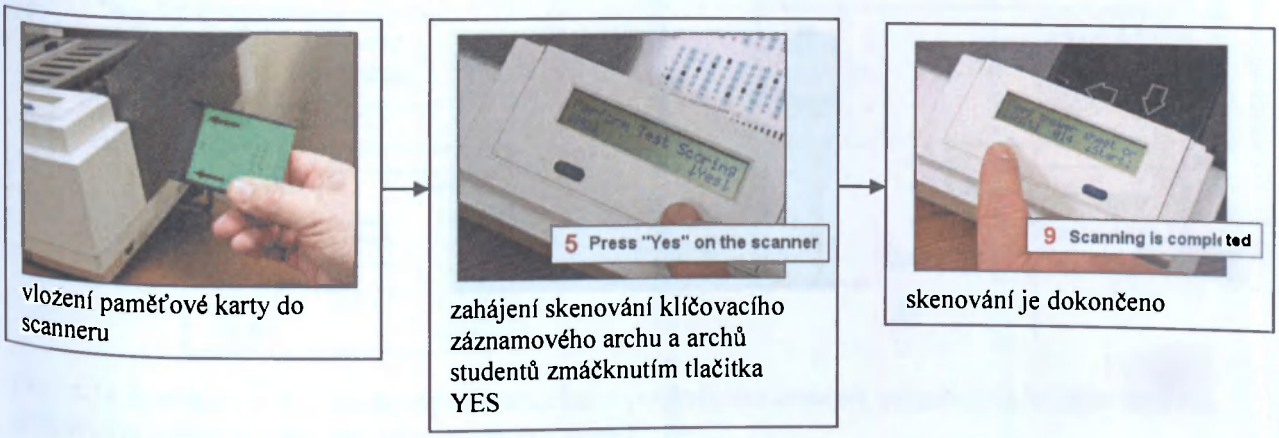
SelfScore je registrovaná známka firmy NCS Pearson pro testovací systém, který se skládá ze scanneru OpScan s integrovanou vyhodnocovací jednotkou (viz obr. 4-12), z InterGrade Pro software a SelfScore záznamových archů. Testy jsou připraveny standardně na počítači a poté jsou potřebné údaje o záznamovém archu (např. označení správné odpovědi, umístění políček pro vyplňování, přečtení čárového kódu atd.) přeneseny přes paměťovou kartu do samostatně pracující skenovací jednotky. Přístroj se ovládá pomocí menu zobrazeného na displeji, kde se také zobrazí, že skenování bylo dokončeno. Výsledné skóry jsou tisknuty již při skenování přímo na určená místa v záznamovém archu a kromě toho jsou ukládány do tzv. gradebook (viz obr. 4-13) a poté přeneseny pomocí paměťové karty zpět do počítače. Do scanneru je vložen i speciální záznamový arch pro výsledky z položkové analýzy (rozložení četností, průměrný skór atd.). Celý proces je znázorněn na obr. 4-14



Obr. 4-12 SelfScore Test Scoring Machine



Obr. 4-13 Ukázka z gradebook (uvedeno je jméno, hrubý skór, relativní skór, známka)



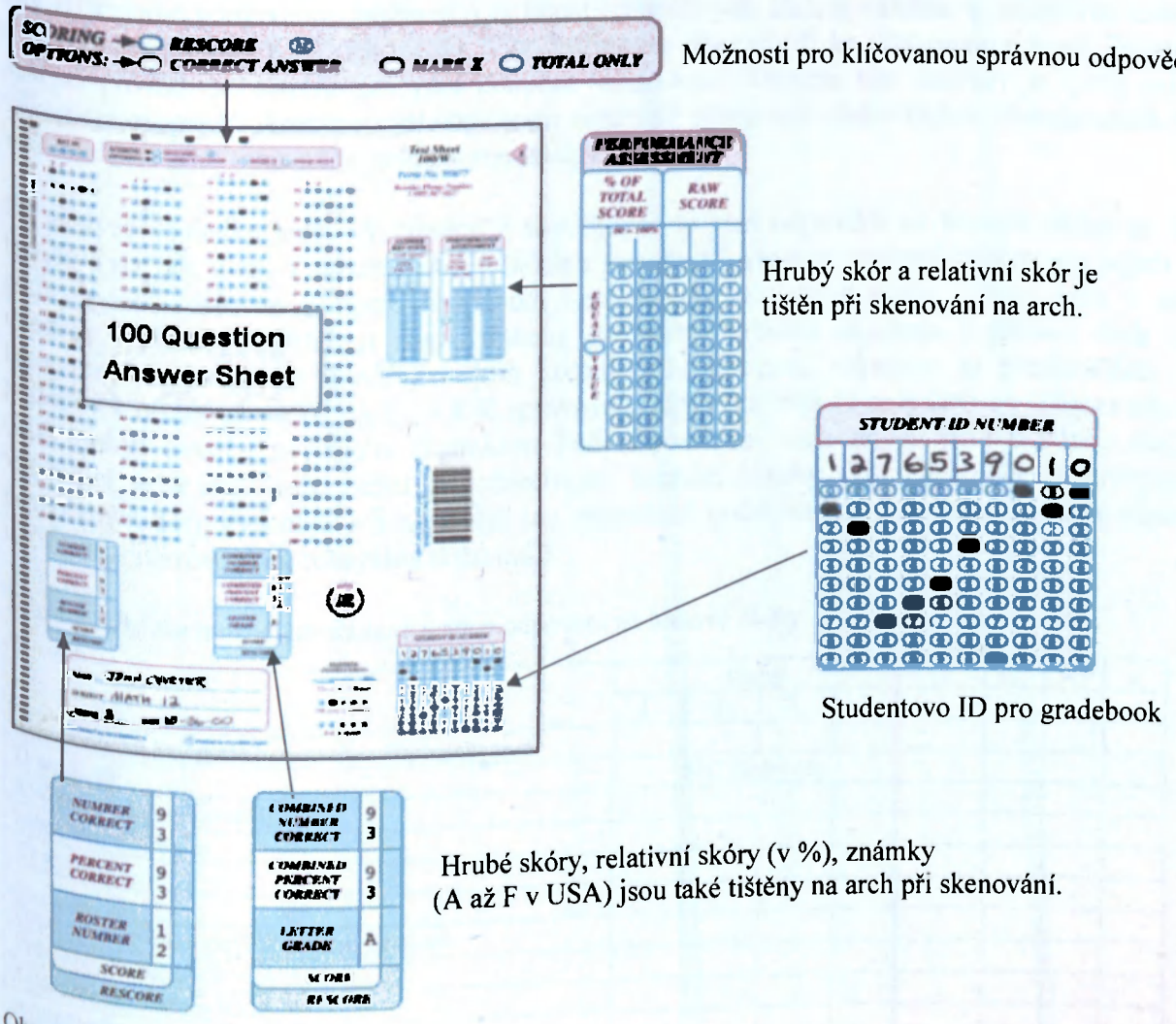
vložení paměťové karty do scanneru

zahájení skenování klíčovacího záznamového archu a archů studentů zmáčknutím tlačitka YES

skenování je dokončeno

Obr. 4-14 Skenovací proces na Selfscore Test Scoring Machines

Záznamový arch je znázorněn na obr. 4-15 Kromě odpovědí testovaných obsahuje také výsledky položkové analýzy pro daného testovaného.



Možnosti pro klíčovanou správnou odpověď

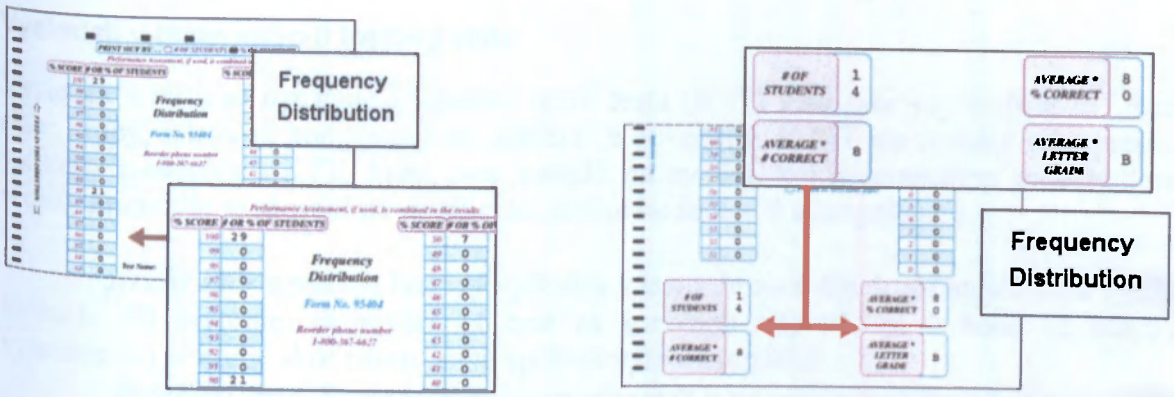
Hrubý skór a relativní skór je tištěn při skenování na arch.

Studentovo ID pro gradebook

Hrubé skóry, relativní skóry (v %), známky (A až F v USA) jsou také tištěny na arch při skenování.

Obr. 4-15 SelfScore záznamový arch.

Souhrnné výsledky položkové analýzy jsou vytištěny na speciální arch (obr. 4-16).



Obr. 4-16 Speciální arch se souhrnnými statistikami (rozložením četností, průměrným hrubým skórem, průměrným relativním skórem, průměrnou známkou).

4. 7 Analýza výsledků testu

Analýzu výsledků testu provádíme zpravidla dvakrát. Poprvé v etapě konstrukce testu po pilotáži, kdy zjišťujeme především obtížnost a citlivost jednotlivých úloh a validitu a reliabilitu testových výsledků (ať pomocí KTT nebo IRT). Tyto informace nám slouží ke zdokonalení testu. Podruhé se analýza provádí po ostrém testování konečné verze testu. Účelem této analýzy je zjistit zejména výkon testovaných. Analýza výsledků testu zahrnuje stanovení statistických charakteristik testu, kontrolu časového omezení a položkovou analýzu.

Při analýze výsledků testu vycházíme z uspořádané matice odpovědí na testové úlohy ve formě tabulky (viz tab. 4-8). V tabulce jsou v řádcích sestupně[24] uvedeni studenti (někdy jen jejich ID[25]) podle jejich skórů, ve sloupcích odpovědi testovaných na testové úlohy. Úlohy jsou v tabulce uvedeny v pořadí, ve kterém jsou zařazeny do testu. Tabulka obsahuje v záhlaví druh úlohy, taxonomickou kategorii vyjadřující druh kognitivního procesu, o kterém se předpokládá, že je potřebný k řešení příslušné úlohy, a klíč správných odpovědí. Někdy se u úloh s výběrem odpovědi (MC úlohy) uvádí i nefunkční distraktor. Pod odpověďmi testovaných jsou uvedeny statistiky vztahující se k testu jako celku či jednotlivým úlohám (hrubý skór (X), počet správných (c), nesprávných (w), vynechaných odpovědí (o), průměrný počet bodů v úloze, směrodatná odchylka, obtížnost, citlivost úloh, reliabilita testu atd.).

Tab. 4-8 Ukázka možné uspořádané matice odpovědí na testové úlohy.

	úlohy				X	c	w	o
	1.	2.	3.	…				
Taxonomická kategorie								
Druh úlohy	MC úlohy (5)							
Klíč správných odpovědí								
nefunkční distraktor								
ID studentů								
…								
Počet správných odpovědí celkem na úlohu								
Průměrný počet bodů v úloze								
s (směrodatná odchylka)								
p (obtížnost úlohy)								
d (citlivost úlohy)								
reliabilita po vynechání úlohy								
změna oproti původní reliabilitě testu								
Označení úlohy: ☺ vhodná ? podezřelá ☹ nevhodná								

[24] Pokud mají studenti stejný skór, je jejich řazení náhodné.
[25] ID je identifikační číslo studenta.

Výsledek v testu nebo-li testový skór

Výsledek v testu se označuje v klasické teorii testu (KTT) jako skór se zkratkou X. Rozlišuje se skór: hrubý, relativní, korigovaný na „hádání" a odvozený. V IRT jde o skóry schopnosti podobné velikostí z-skórům v KTT, které jsou závislé na zvoleném matematickém modelu (proficiency score). Zpravidla se převádí na stupnici se střední hodnotou 0 a rozptylem 1.

Hrubý skór označuje bodový výsledek v testu. Je součtem dosažených bodů v jednotlivých úlohách. Při binárním skórování (1 bod za správnou odpověď a 0 bodů za nesprávnou či vynechanou) je hrubý skór roven počtu správně vyřešených úloh.

Relativní skór X_R se vyjadřuje v procentech jako poměr hrubého (X) a maximálně dosažitelného skóru (X_{max}): $X_R = 100\frac{X}{X_{max}}$.

U testů s úlohami binárně skórovanými se vztah zjednodušuje na $X_R = 100\frac{X}{K}$, kde K je počet úloh v testu.

Korigovaný skór na „hádání" X_{kor} se používá, i když spíše zřídka, u testů sestavených z uloh s výběrem odpovědi: $X_{kor} = R - \frac{W}{a-1}$, kde

R je počet správných odpovědí, W je počet chybných odpovědí, a je počet nabízených odpovědí. Znamená to, že např. v úloze s pěti nabízenými odpověďmi, z nichž pouze jedna je správná, se za chybnou odpověď odečítá z hrubého skóru 0,25 $\left(=\frac{1}{5-1}\right)$ bodu, správná odpověď je skórována 1 bodem a vynechaná 0 body.

Korigovaný skór na „hádání" se používá velmi málo, protože výzkumy dokazují, že testovaný zpravidla nehádá správnou odpověď, ale snaží se vyloučit (pro něj) nesprávné nabídky. Kromě toho příčinou chybné odpovědi nemusí být hádání, ale chyba, které se testovaný dopustil, a to ho znevýhodňuje oproti testovaným, kteří se o odpověď vůbec nepokusili a úlohu vynechali.

Vedle těchto tří skórů existuje ještě celá řada tzv. *odvozených skórů* (derived scores), např. z-skór, percentilový skór a T-skór. Zmíníme pouze dva.

Základním druhem odvozených skórů je ***z-skór***, který vychází z průměrného hrubého skóru a standardní odchylky hrubých skórů:

$$z = \frac{X_i - \overline{X}}{s},$$

kde X_i je průměrný skór skupiny o i testovaných či skór i-tého testovaného, $\overline{X}$ je průměrný hrubý skór a s je standardní odchylka hrubých skórů. Negativní z-skór mají testovaní, jejichž hrubé skóry jsou podprůměrné.

Ilustrativní příklad:
V testu OSP, varianta A (na PedF UK v roce 2006), který obsahoval 45 úloh, byl průměrný hrubý skór v testu $\overline{X} = 24{,}62$, standardní odchylka $s = 6{,}11$. Průměrný hrubý skór těch, kteří úlohu vyřešili správně, byl $X_i = 24{,}89$.

Odpovídající z-skór tedy bude roven $z = \frac{24{,}89 - 24{,}62}{6{,}11} = 0{,}044$. Tedy, výkony testovaných, kteří vyřešili úlohu 1 správně, jsou 1/25 standardní odchylky nad průměrným hrubým skórem.

Percentilový skór vyjadřuje podíl testovaných, kteří dosáhli v testu stejného nebo horšího bodového výsledku než daný testovaný. Např. umístil-li se student na 70. percentilu, znamená to, že 70 % testovaných dosáhlo horšího výsledku a 30 % lepšího výsledku než on. Pokud je znám

i celkový počet testovaných, je možné díky percentilu určit i absolutní pořadí účastníka. Výpočet percentilu je jednoduchý. Testované rozdělíme podle jejich hrubých skórů do určitého počtu intervalů (např. 100), v nichž se nachází přibližně stejný počet testovaných . Pořadí intervalu, v němž se výsledek testovaného nalézá, potom udává jeho percentil. K výpočtu můžeme použít funkci PERCENTRANK v programu Microsoft Excel.

V IRT modelech se umisťují testovaní na stejnou škálu jako úlohy. Jednotky této škály se značně liší od standardních KTT skórů, protože nemusejí být normálního rozložení. Nejčastěji se používají tři typy jednotek: *logit units*, *odds units* a proportion true score.

Výsledek testovaných v testu závisí na mnoha faktorech. Patří k nim například příprava studenta a jeho vlastnosti, úroveň a efektivita výuky, závažnost rozhodnutí vyvozovaných z výsledku testu, charakteristiky testu, objektivita vyhodnocování odpovědí testovaného, vhodné prostředí při testování a kvalita testu samotného.

Stanoveni statistických charakteristik testu

Statistickými charakteristikami testu, které nám umožňují test či jeho varianty zhodnotit (jejich vyrovnanost), jsou především **obtížnost**, **citlivost** a **reliabilita** testu (viz výše v této kapitole)[26]. Je zřejmé, že příliš obtížný či příliš snadný test, který v dané populaci nevyřeší téměř nikdo nebo naopak všichni, nemá téměř žádnou rozlišovací schopnost, a proto ani validitu k jakémukoli účelu. Proto nalézt optimální úroveň obtížnosti testu je také podmínkou validity.

Obtížnost testu v KTT

Za předpokladu, že analyzujeme výsledky testu o k úlohách (i = 1, 2, .., k) zadaného N testovaným, můžeme obtížnost testu p (nebo q)[27] vyjádřit zlomkem

$$p = \frac{\bar{X}}{X_{\max}} \text{ (nebo } q = 1 - \frac{\bar{X}}{X_{\max}}),$$

kde $\bar{X}$ je průměrný skór skupiny testovaných a X_{max} je nejvýše dosažitelný skór v testu. Vynásobíme-li zlomek číslem 100, získáme obtížnost testu vyjádřenou v procentech. p je tedy číslo, které udává průměrný podíl správných odpovědí v souboru testovaných. Čím je obtížnost vyjádřená p (resp. q) větší, tím více (resp. méně) bodů v průměru studenti získali. Pokud je test sestaven pouze z úloh binárně skórovaných (1 za správnou odpověď a 0 za jinou), je $X_{max} = k$.

Uvedenou definici obtížnosti lze použít pro jednotlivé položky i pro celkové hodnocení testu. Obtížnost testu *nemusí* být průměrem obtížností jednotlivých položek. Vztah platí jen pro shodné maximální počty bodů u jednotlivých položek, např. pro binárně skórované úlohy.

Citlivost testu v KTT

Citlivost testu vyjadřuje míru schopnosti testových skórů rozlišovat mezi testovanými s vysokou a s nízkou úrovní měřeného znaku. Míru citlivosti můžeme vyjádřit směrodatnou odchylkou (standard deviation) s či jejím čtvercem nazývaným rozptyl (variance) s^2 nebo variačním koeficientem V souboru naměřených skórů.

Směrodatnou odchylku skórů snadno vypočítáme pomocí vzorce:

$$s = \sqrt{\left(\frac{\sum_{i=1}^{n} X_i^2}{N}\right) - \bar{X}^2},$$

[26] U testů z přijímacího řízení se zkoumá také kriteriální predikční validita testu.
[27] Viz výše v kap. 4.

kde $\overline{X}$ je průměrný skór skupiny testovaných, N počet všech testovaných a X_i je skór testovaného i. Někdy je ve jmenovateli N–1 (standard deviation as a sample).
Čím jsou standardní odchylka či rozptyl menší, tím více se naměřené skóry jednotlivých testovaných hromadí kolem průměrného skóru.

Variační koeficient skórů stanovíme ze vztahu:

$$V = \frac{100s}{\overline{X}}.$$

Variační koeficient v našem případě udává, z kolika procent se podílí směrodatná odchylka souboru naměřených skórů na průměrném skóru.

Míra obtížnosti a citlivosti testu vycházející z klasické teorie testu (KTT) závisí na tom, na jakém souboru testovaných ji měříme. Oproti tomu pomocí teorie odpovědi na položku (IRT) zjišťujeme obtížnost a citlivost položek (pomocí parametrů b a a)[28] a tím i celého testu nezávisle na charakteristikách souboru testovaných, což nám zaručuje větší přesnost. Stanovení odhadů parametrů obtížnosti a citlivosti v IRT vyžaduje použití specializovaného software (např. ConQuest, BILOG-MG).

Kontrola časového omezení testu

K vypracování testu mají mít studenti přiměřený čas, tj. převážná většina studentů má mít dostatek času na vypracování testu, avšak zbytečně dlouhý čas umožňuje opisování.

Bez ohledu na to, že zkušený konstruktér testu se snaží odhadnout přiměřenou dobu potřebnou k vypracování testu většinou studentů, často se může stát, že studenti nemají na zpracování testu dostatek času. Proto se provádí kontrola časového omezení. Ta je založena na předpokladu, že pokud student na konci testu vynechává odpovědi k úlohám, je to proto, že tyto úlohy z časových důvodů nestihl řešit. Tyto úlohy se označují jako nedosažené.

Uplatňuje se hrubé pravidlo, že test se považuje za časově přiměřený, pokud byl dokončen, bez ohledu na správnost odpovědí, alespoň 80 % testovaných (Swineford 1974).

Vyrovnanost variant testu

Vyrovnanost několika variant testu je podrobněji popsána a ukázána na příkladech testu OSP na PedF UK v Praze z roku 2006 ve výzkumné části.

Vyrovnanost variant testu pomocí KTT

Ukazatele obtížnosti a citlivosti testu umožňují zhruba posoudit statistickou vyrovnanost několika testových variant. Jednotlivé varianty testu by měly být vyvážené nejen svým obsahem, počtem a druhy úloh, ale také s ohledem na statistické charakteristiky: obtížnost a citlivost. Varianty testu se považují za vyrovnané (paralelní)[29], pokud obsahují úlohy podobné obtížnosti (p) a citlivosti (r). Rozložení skórů u statisticky vyrovnaných variant testu by mělo být téměř shodné (křivky rozložení skórů by se měly překrývat).

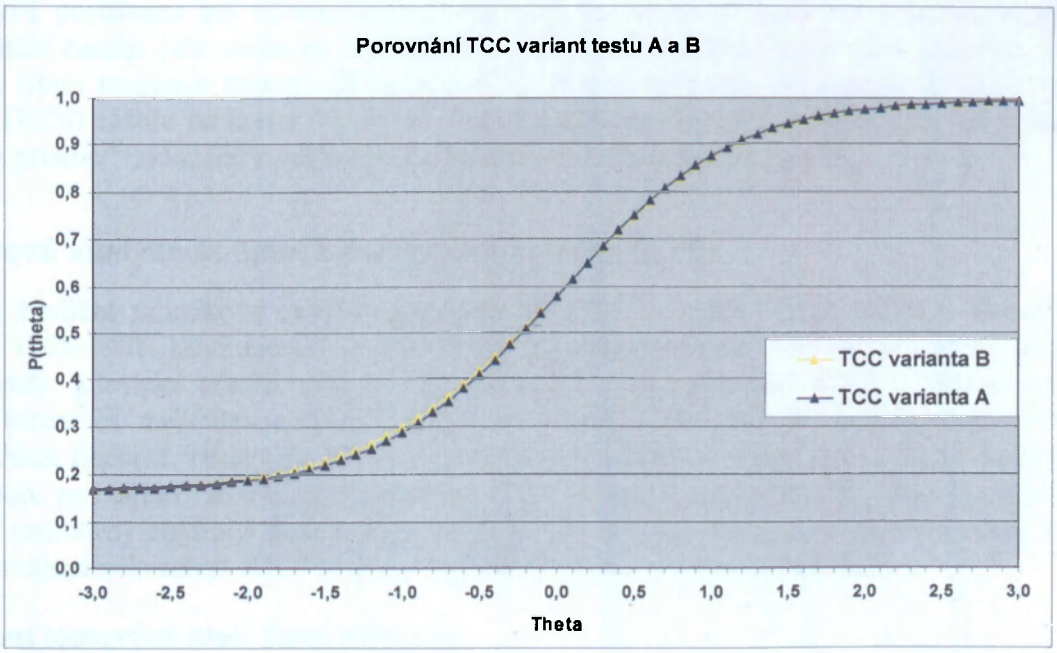
Vyrovnanost variant testu pomocí IRT

K posouzení vyrovnanosti několika variant testu se při použití teorie odpovědi na položku (IRT) využívá charakteristických funkcí/ křivek[30] (test characteristic curve, TCC) a informačních funkcí[31] (test information function, TIF) variant testu. Varianty testu se považují za vyrovnané (paralelní), pokud jejich charakteristické (viz obr. 4-17) a informační funkce (viz obr. 4-18) jsou téměř

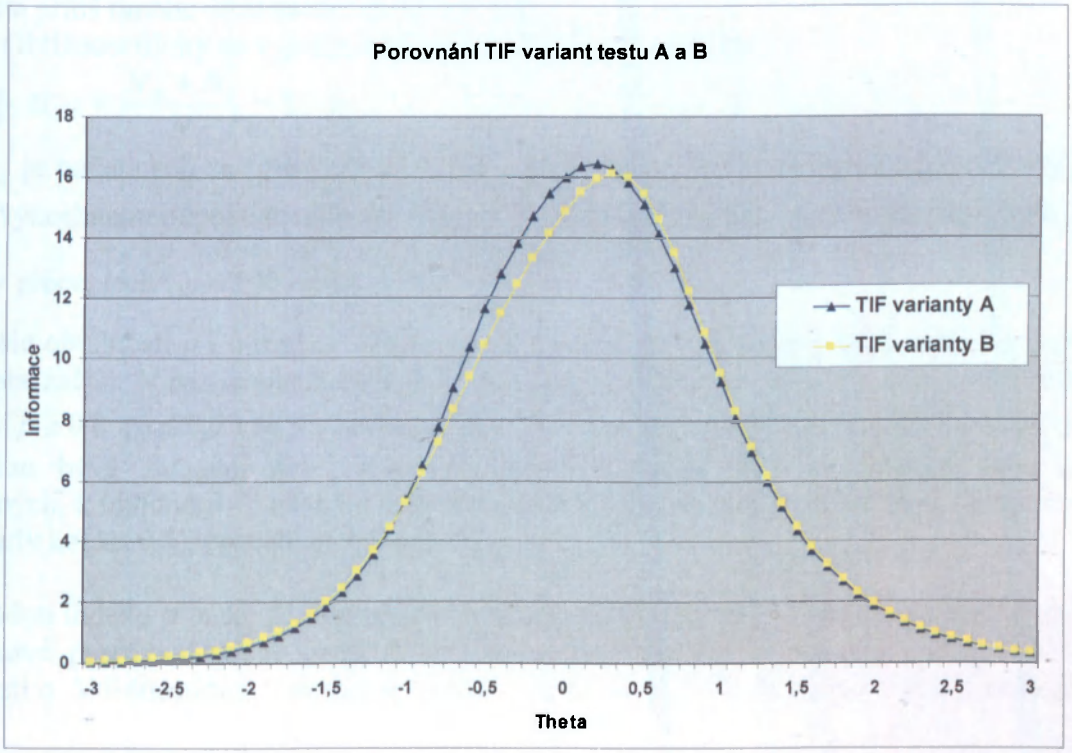
[28] Viz dále u položkové analýzy.
[29] Za předpokladu, že skóry testových variant mají přibližně symetrické normální rozložení.
[30] Bude podrobně vysvětleno dále.
[31] Viz výše v kap. 4 u reliability testu.

identické (překrývají se). Porovnání informačních funkcí pro danou úroveň θ lze provádět také pomocí výpočtu relativní efektivity (relative efficiency) jednoho testu ve srovnání s druhým jako odhad na úrovni θ: $RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)}$,

kde $RE(\theta)$ udává relativní efektivitu a $I_A(\theta)$ a $I_B(\theta)$ jsou informační funkce testů A a B. Jestliže je např. $I_A(0) = 16$ a $I_B(0) = 14{,}4$ pro $\theta = 0$, potom $RE(\theta) = 1{,}11$. To znamená, že na úrovni schopnosti $\theta = 0$ test A funguje, jako by byl o 11% delší než test B. Test B by bylo tedy potřeba prodloužit, příp. test A zkrátit o 11% úloh k získání stejné přesnosti měření jako testem A v $\theta = 0$ (Hambleton aj. 1991).



Obr. 4-17 Charakteristické křivky dvou vyrovnaných variant testu A a B (světlejší křivka)



Obr. 4-18 Informačních funkce dvou téměř vyrovnaných variant testu A a B (světlejší křivka)

4. 8 Statistická analýza úloh (položková analýza)

I když se některé věci v testování za posledních 50 let velmi změnily, např. výpočetní možnosti (dnes již prakticky neexistují žádná omezení), účel položkové analýzy zůstal stejný. Stále se položkovou analýzou míní zjišťování statistických charakteristik úloh (četnost vynechaných odpovědí, obtížnost, citlivost), četnost a druhy chyb, které přispívají k odhalení úloh nevyhovujícím účelu testu a úloh s technickými nedostatky. Kromě toho ukazatele obtížnosti a citlivosti úloh pomáhají při navrhování vyrovnaných testových variant. U úloh s výběrem odpovědi se zjišťuje ještě funkčnost jednotlivých distraktorů.

V dnešní době se uplatňuje vedle tradiční položkové analýzy (již se běžně provádí na počítači), která bývá používána při vývoji testů založených na klasické teorii testů (classical test theory, KTT), stále častěji (ale zatím ne u nás) také položková analýza vyplývající z teorie odpovědi na položku (item response theory, IRT). Největší světová testovací organizace Educational Testing Service (ETS) začala na konci 20. století uplatňovat k analýze svých testů ještě jinou metodu, tzv. grafický přístup[32] založený na odhadu souboru křivek odpovědi na položky.

Položková analýza pomocí klasické teorie testu (KTT)

Základy tradiční položkové analýzy založené na KTT položili Alfred Binet a Theodore Simon v letech 1905-1908, když hledali testové úlohy vhodné obtížnosti pro různá věková období dětí a adolescentů. Klasická teorie testů (KTT) vychází z teorie pravých skórů[33] (true score) a chyby měření (error of measurement), tj. z teorie reliability testových skórů (přesnost měření). KTT předpokládá lineární vztah ($X = T + E$) mezi pozorovatelným skóre testu (X) a latentními skóry, které spolu navzájem nekorelují, tj. pravým (T) a chybovým skórem (E),testu (Komenda, 2003), přičemž průměrný chybový skór v populaci testovaných je roven nule a chybové skóry paralelních testů navzájem nekorelují. KTT využívá ukazatelů obtížnosti a citlivosti úloh.

Obtížnost testových úloh (item difficulty)

Požadovaná obtížnost úloh se odvozuje od účelu testu. Většinou je žádoucí, aby rozlišující test (např. test studijních předpokladů) obsahoval úlohy různé obtížnosti, nejvíce středně těžkých úloh a ne příliš mnoho úloh snadných či těžkých.

Obtížnost úlohy se vyjadřuje (Chráska 1999) buď podílem

$$p = \frac{N_S}{N} \text{ či } q = \frac{N_{ch} + N_0}{N} = 1 - p,$$

kde N_S je počet osob se správnou odpovědí, N_{ch} je počet osob s chybnou odpovědí a N_0 je počet osob s vynechanou odpovědí. $N = N_S + N_{ch} + N_0$ je celkový počet osob ve skupině. Někdy se p a q udává v procentech ($p = 100\frac{N_S}{N}$, $q = 100\frac{N_{ch} + N_0}{N}$ = 100-p).

Ukazatelé obtížnosti p i q mohou nabývat hodnot od 0 do 1. Čím je p nižší nebo-li q vyšší, tím je úloha obtížnější. V rozlišujících testech se považují úlohy s $p \leq 0{,}2$ ($q \geq 0{,}8$) za příliš obtížné, úlohy s $p \geq 0{,}8$ ($q \leq 0{,}2$) za příliš snadné a do testů se bez zvláštních důvodů do testů nezařazují. Výjimkou bývá zařazení dvou nebo tří velmi snadných úloh na začátku testu k motivaci testovaných, k uklidnění. V běžných případech se doporučuje zařazovat do testu úlohy s p mezi 0,3 a 0,7 (tedy kolem 0,5), protože bývají citlivější.

Nevýhodou indexu p je to, že čím je jeho hodnota vyšší, tím snazší úlohu označuje a naopak. Pro výzkumové nesrovnalosti se proto někdy při popisu obtížnosti testových úloh uvádí hodnota obtížnosti q. V Educational Testing Service (ETS) koncem 60. let minulého století z tohoto důvodu

[32] Grafický přístup jakožto zatím ojedinělý vypouštíme.
[33] Pravý skór získáme rozdílem skóru testu a chybového skóru.

začali používat k rozlišení obtížnosti úloh ukazatel delta Δ (Bažantová; Byčkovský 2006), který je založen na převodu klasického indexu obtížnosti na standardní z-skóry podle vztahu $\Delta = 13 + 4z$ (pro $p = 0{,}34$ je $z = -1$, tedy $\Delta = 9$; pro $p = 0{,}84$ je $z = 1$, tedy $\Delta = 17$; pro $p = 0{,}50$ je $z = 0$, tedy $\Delta = 13$).

Údaj o obtížnosti položky přináší mnoho informací nejen o testovaném souboru, ale i o kvalitě testu (o uspořádání úloh v testu). Zvlášť u časově omezených testů se úlohy řadí podle obtížnosti od nejlehčích po nejtěžší. Položky, které vyřešila správně většina studentů, či naopak úlohy, které nevyřešil nikdo, jsou z hlediska citlivosti testu nevýznamné a zpravidla se z testu vyřazují.

Obtížnost uzavřených úloh

U úloh uzavřených, kde jsou odpovědi nabízeny, může student i bez potřebných znalostí dojít ke správné odpovědi náhodně, hádáním, přičemž vysoký počet nabídek snižuje možnost dospět hádáním ke správnému řešení úlohy. Koeficient obtížnosti (snadnosti) je tím zkreslen, a proto Guilford (1936) navrhl korigovaný ukazatel obtížnosti. Za předpokladu, že distraktory jsou stejně atraktivní, se hodnota skutečné obtížnosti p_{corr} zahrnující korekci na hádání určí ze vztahu: $p_{corr} = p - \frac{n_X}{N(a-1)}, \left(\frac{1}{a-1} \le p \le 1\right)$, kde a je počet nabídek, n_X je počet chybných odpovědí, $p = \frac{N_s}{N}$. Tento korigovaný ukazatel obtížnosti (přísnější než p) se ale používá velmi omezeně, protože zkušenosti ukazují, že studenti hádají méně, než se předpokládá (Lord 1952). Účelné je ho použít při srovnávání obtížnosti úloh s různým počtem nabídnutých odpovědí.

Citlivost (diskriminace) testových úloh (item discrimination, item fairness)

Má-li test měřit, jak si jednotliví studenti osvojili určité učivo nebo jaké mají předpoklady ke studiu na vyšším vzdělávacím stupni, měly by výsledky v testu dostatečně rozlišovat mezi lepšími a horšími studenty. K tomu by měla přispívat každá testová úloha tím, že ji bude správně řešit více „lepších" než „horších" studentů. Jak tento požadavek úloha plní, lze vyjádřit různými ukazateli citlivosti úloh. Někdy bývá citlivost úloh označována jako diskriminační (rozlišovací) schopnost úlohy. Nízkou citlivost mají úlohy, které správně vyřeší stejný počet „lepších" i „horších" žáků. Naopak úlohu s vysokou citlivostí řeší úspěšně „lepší" žáci, zatímco „horší" žáci nikoli.

Většinou se citlivost vyjadřuje rozdílem ve výkonu (počtu správných odpovědí) mezi kontrastními skupinami, tj. skupinami „nejlepších" a „nejhorších" testovaných. Testované seřadíme sestupně podle jejich výkonu (hrubého skóru) v testu a rozdělíme je na 2-4 části, z nichž vezmeme skupinu studentů s nejnižším počtem dosažených bodů a skupina s nejvyšším počtem dosažených bodů. Je-li počet testovaných větší než 60–80, považuje se za optimální vzít za kontrastní skupiny 27 % nejlepších a 27 % nejhorších studentů.

Ukazatelé (koeficienty) citlivosti

Nejjednodušším ukazatelem citlivosti, který se používá zejména v učitelských testech, je **index ULI** (upper-lower index), označovaný též ***d*** či ***D***:

$$d = \frac{N_U - N_L}{0{,}5N},$$

kde N_U je počet správných odpovědí u lepší poloviny studentů, N_L počet správných odpovědí u horší poloviny studentů, N je celkový počet studentů

$$\text{nebo } d = p_U - p_L,$$

kde p_U je obtížnost úlohy pro skupinu „lepších" studentů, p_L obtížnost úlohy pro skupinu „horších" studentů. Uvedený vztah platí pro rozdělení testovaných podle dosaženého počtu bodů na poloviny. Index citlivosti d může nabývat hodnot od -1 do 1. Rozdělíme-li testované na více skupin, např. 5, změní se v prvním uvedeném vzorci pro výpočet d ve jmenovateli číslo 0,5 na 0,2.

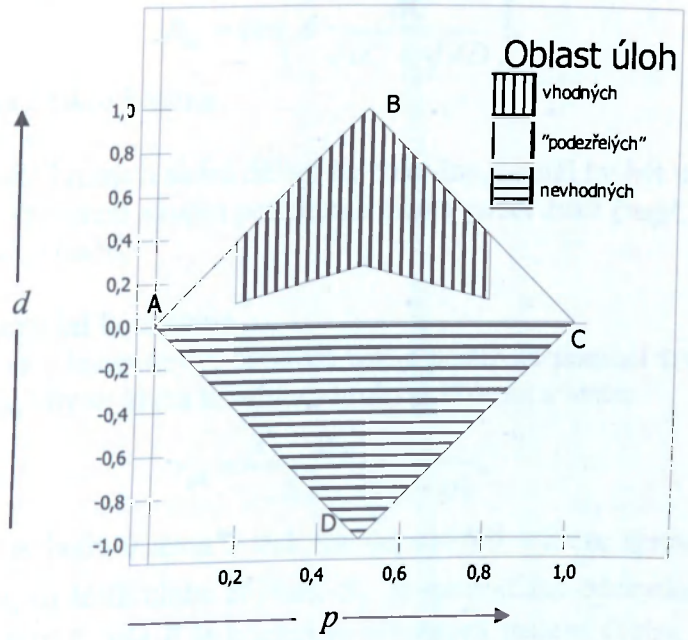
Neřeší-li úlohu nikdo z testovaných nebo ji naopak vyřeší všichni testovaní, je $d = 0$. Teoreticky, pokud úlohu vyřešili všichni horší studenti a žádný z lepších studentů, je $d = -1$. Je žádoucí, aby každá z úloh měla ***d*** s co nejvyšší kladnou hodnotou. Úlohy s hodnotami, kde d je záporné či nulové, se považují za nevhodné.
Doporučené hodnoty ukazatele citlivosti d jsou shrnuty v tab.4-9 (Byčkovský 1983).

Tab. 4-9 Doporučené hodnoty ukazatele citlivosti d

d	hodnocení úlohy
$\geq 0{,}40$	velmi dobrá úloha
0,30 – 0,39	dobrá úloha
0,20 – 0,29	docela dobrá úloha
0,10 – 0,19	méně významná úloha, obvykle k vylepšení
$\leq 0{,}10$	špatná úloha, k vylepšení či vyřazení z testu

Vztah mezi ukazatelem obtížnosti *p* a citlivosti *d*
Hodnoty indexu citlivosti d jsou vázány na hodnoty indexu obtížnosti p (viz obr. 4-19).
Maximální hodnoty $d = 1$ lze dosáhnout pouze při $p = 0{,}5$ a úlohu vyřeší správně všichni „lepší" a žádný z „horších" studentů (v opačném případě je $d = -1$).

U úloh rozlišujících se vyžaduje, aby úlohy s obtížností p mezi 0,3 a 0,7 měly $d \geq 0{,}25$ a úlohy s hodnotou obtížnosti p mezi 0,2 a 0,3 či 0,7 a 0,8 aspoň $d \geq 0{,}15$ (viz svisle vyšrafovaná oblast). U testů ověřovacích stačí, aby bylo d kladné. Úlohy s obtížností $p \in (0{,}2;0{,}8)$ a citlivostí $d \leq 0$ se automaticky považují za nevhodné (viz vodorovně vyšrafovaná oblast).



Obr. 4-19 Obor hodnot citlivosti úloh d v závislosti na obtížnosti p (Byčkovský 1983)

Ačkoli se hodnota d ve srovnání s jinými koeficienty citlivosti obtížněji interpretuje, její výhodou je vedle snadného výpočtu její nezávislost na dodržení specifických podmínek (které v praxi nebývají vždy splněny) nutných pro použití jiných koeficientů. U velkých souborů se k vyjadřování citlivosti používá korelačních koeficientů: bodově-biseriálního r_{pb} a biseriálního r_b (pokud jedna z korelovaných veličin je dichotomická), kterými se vyjadřuje korelace mezi řešením úlohy a výsledkem v testu, tetrachorického koeficientu a koeficientu φ (když obě korelované veličiny jsou dichotomické). I když použití těchto koeficientů vede ke stejným výsledkům (např. Englehart 1965, Oosterhof 1976, Beuchart; Mendoza 1979), dává se přednost r_{pb} a r_b s tím, že se u těchto korelačních koeficientů používá i opravy na autokorelaci (např. Henryssen 1971).

Čtyřpolní a tetrachorický koeficient

Výpočet tetrachorického koeficientu citlivosti je sice pracnější než výpočet koeficientu ULI, ale je většinou spolehlivější. Pro každou úlohu je třeba sestavit tzv. čtyřpolní tabulku (2x2, tab. 4-10), kde jsou uvedeny počty žáků z „horší" (H) a „lepší" (L) skupiny, kteří úlohu zodpověděli správně, a počty těch, kteří odpověděli chybně, příp. neodpověděli vůbec.

Tab. 4-10 Čtyřpolní tabulka

	správně	chybně	součet
lepší žáci	A	B	A + B
horší žáci	C	D	C + D
součet	A + C	B + D	A + B + C + D

Z uvedené tabulky vyplývá, že A je počet lepších žáků se správnou odpovědí v úloze, B s chybnou či neuvedenou odpovědí. C je počet horších žáků, kteří v úloze odpověděli správně, D chybně nebo neodpověděli. Na základě této tabulky (hodnot A, B, C a D) lze vypočítat dva různé koeficienty, koeficient čtyřpolní a koeficient tetrachorický.

Čtyřpolní koeficient (Chráska 1999) se vypočítá vztahem:

$$R_{4f} = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Tetrachorický koeficient (Chráska 1999) citlivosti (přesněji řečeno kosinus odhadu tetrachorického koeficientu) se vypočítá podle následujícího vzorce:

$$R_{tet} = \cos\left(\pi \frac{BC}{\sqrt{BC} + \sqrt{AD}}\right),$$

kde *cos* je goniometrická funkce kosinus.

Pokud jsou testovaní podle hrubých skórů děleni na poloviny, neměl by být u vyhovujících úloh R_{tet} nižší než 0,15. Pokud k vytvoření skupin použijeme menší počet žáků (např. 27%), potom musíme vypočítané R_{tet} posuzovat přísněji.

Bodově biseriální a biseriální koeficient

Často se u velkých souborů testovaných citlivost úlohy posuzuje pomocí tzv. bodově biseriálního koeficientu korelace[34] r_{pb}, kdy se úloha koreluje s hrubým skórem v testu:

$$r_{pb} = \frac{\overline{X_S} - \overline{X_N}}{S_X} \sqrt{\frac{p}{(1-p)}},$$

kde $\overline{X_S}$ je průměrný počet bodů v testu u těch, co odpověděli v úloze správně, $\overline{X_N}$ je průměrný počet bodů v testu u těch, co řešili úlohu chybně. S_X je směrodatná odchylka vypočítaná ze všech testových výsledků, $p = 0{,}01P$, kde P je koeficient obtížnosti testové úlohy. Vyhovující úloha by měla mít bodově biseriální koeficient citlivosti minimálně 0,20. Koeficient r_{pb} je modifikací Pearsonova korelačního koeficientu[35] pro případ, kdy jedna z korelovaných veličin je dichotomická (nabývá pouze hodnot 0 a 1).

[34] Hopkins (1998) ho označuje jako standardní index (standard index). Hambleton; Swaminathan; Rogers (1991) ho označují jako klasický diskriminační index úlohy (classical item discrimination). Popham (1978) ho považuje za pravděpodobně nejběžnější index diskriminace úlohy.

[35] Pearsonův korelační koeficient je dán vzorcem: $r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sqrt{(y_i - \bar{y})^2}}}$.

Biseriální koeficient lze získat z bodově biseriálního jeho vynásobením číslem závislém pouze na obtížnosti úlohy (např. Lord; Novick 1968, Millman; Greene 1988). V literatuře se uvádí, že při aplikaci na *stejný soubor dat je biseriální koeficient o 25% vyšší než bodově biseriální koeficient* (např. Magnusson 1967, Millman; Greene 1988). Jinými slovy, oba koeficienty spolu vysoce korelují, s tím, že r_b má vždy větší hodnotu než r_{pb}. Akceptovaná minimální hodnota r_{pb} je většinou +0,30, r_b minimálně +0,40 (v učitelských testech +0,15, resp. +0,20, Hills 1976). Zatímco r_{pb} může dosáhnout max. hodnoty 1, hodnota koeficientu r_b může být vyšší než 1.

Analýza nenormovaných odpovědí

Vedle zjišťování obtížnosti a citlivosti testových úloh se dále provádí, jak uvádí Byčkovský (1983), analýza tzv. nenormovaných odpovědí, tj. odpovědi chybných a vynechaných. Při této analýze se zkoumají nesprávné odpovědi (a jejich četnost) a četnost vynechaných odpovědí.

Četnost vynechaných odpovědí

Zvláštní pozornost se při položkové analýze věnuje četnosti vynechaných odpovědí. To, že testovaný úlohu vynechá, nemusí vždy znamenat, že nemá vědomosti potřebné k jejímu řešení, ale že mu na řešení úlohy již nezbyl čas. Vynechané úlohy se mnohdy nedají přesně specifikovat, protože pokud se nejedná o adaptivní testování (kde je posloupnost úloh dána), může student řešit úlohy v libovolném pořadí. Může se tedy stát, že úlohy na konci testu nemusel student řešit jako poslední. Protože v některých případech může testovaný vynechat úlohu proto, že neporozuměl dobře zadání úlohy, doporučuje se zkontrolovat každou uzavřenou úlohu, kterou neřešilo více než 20 % testovaných a každou otevřenou úlohu, kterou vynechalo 30-40% testovaných. Mimo to by položková analýza neměla být prováděna u úloh, které nedosáhlo více než 20 % testovaných (Byčkovský 1983).

Četnost a druhy chyb

U uzavřených úloh, kde nesprávné odpovědi (tzv. distraktory) nabízíme, je jejich analýza poměrně jednoduchá. Zjišťujeme, zda mezi distraktory není jeden nebo několik, které

- volí jen velmi málo nebo nikdo z testovaných,
- volí více lepších než horších žáků,
- korelují kladně s testovými výsledky,
- jsou voleny výrazně více než správné odpovědi.

V posledních třech případech je vhodné zjistit i to, zda klíčovaná odpověď je skutečně správnou odpovědí či zda neexistuje více než jedna správná odpověď.

Analýza nesprávných odpovědí u otevřených úloh (s tvořenou dopovědí) je poněkud složitější, protože nejprve musíme typické chyby klasifikovat, což vyžaduje také zkušenosti a intuici, pokud student neuvede postup řešení. Doporučuje se chyby rozdělit do dvou kategorií, na chyby hlavní (základní) a chyby vedlejší. Základními chybami jsou takové, které jsou způsobené neznalostí učiva. Za vedlejší chyby se naopak považují méně významné chyby jako jsou drobné početní chyby či chyby vzniklé např. přehlédnutím, nepřesností, špatnou čitelností textu apod. Jako hrubé kritérium při posuzování přiměřenosti úloh lze použít požadavek, aby četnost vedlejších chyb v odpovědích na úlohu nebyla vyšší než četnost chyb hlavních. Častých chyb v otevřených úlohách zjištěných analýzou můžeme využít v uzavřené úloze jako distraktorů.

Učelem statistické analýzy testových úloh je tedy rozdělit úlohy na vhodné, nevhodné a podezřelé (které mohou být zatíženy technickými nedostatky). Podezřelé úlohy se dále zkoumají a případně upravují či v krajním případě vyřazují. Shrnutí charakteristik podezřelých úloh je v tab. 4-11 (upraveno podle Byčkovský 1983).

Tab. 4-11 Charakteristiky podezřelých úloh

obtížnost		$p < 0{,}2$ nebo $p > 0{,}8$ nebo zjištěná hodnota p se výrazně liší od hodnoty odhadnuté kompetenty
citlivost		$d < 0$ pro všechny hodnoty p $d < 0{,}15$ pro $0{,}2 < p < 0{,}3$ nebo $0{,}7 < p < 0{,}8$ $d < 0{,}25$ pro $0{,}3 \leq p \leq 0{,}7$
neuvedené odpovědi	uzavřené úlohy	$N_0 > 0{,}2\ N$[36]
	otevřené úlohy	$N_0 > (0{,}3\text{-}0{,}4)\ N$[37]
nesprávné odpovědi	uzavřené úlohy	málo atraktivní distraktory (téměř nikdo je nevolí) nebo distraktory, které volí více „lepších" než „horších" studentů
	otevřené úlohy	počet vedlejších chyb je větší než počet chyb hlavních

Položková analýza pomocí teorie odpovědi na položku (IRT)

Položková analýza založená na KTT má svá omezení. Nejdůležitějšími z nich je závislost charakteristik položek na souboru testovaných, kterým byly položky zadány, a to, že KTT nahlíží na položky výhradně v kontextu konkrétního testu, tj. položky nejsou od celku testu oddělitelné (položky jsou korelovány s celkovým skórem). Nelze předpokládat, jak testovaný v úloze odpoví. Výše uvedené nedostatky překonávají testy vyvinuté na základě teorie odpovědi na položku (item response theory, IRT), která byla na prakticky aplikovatelné úrovni zpracována v posledních 20 letech a ve světě se používá při vývoji nástrojů pro širokoplošné testování.

IRT uvažuje o položkách a jejich vlastnostech samostatně, nezávisle na souboru testovaných. IRT modely zahrnují vztah mezi charakteristikami položek a úrovní měřeného latentního rysu/ schopnosti (která charakterizuje testované a nezávisí na charakteristikách položek). Tento vztah lze matematicky popsat tzv. charakteristickou křivkou nebo-li funkcí položky (viz dále).

Stanovení položkových parametrů (tj. kalibraci) a (citlivosti), b (obtížnosti), c („hádání") předchází zpravidla klasická položková analýza, pomocí které vyřadíme úlohy s velmi malou citlivostí (hodnota ukazatele citlivosti blízko nuly nebo záporná), abychom zaručili konvergenci při kalibraci parametrů. K výběru dobrých položek se využívá jejich informačních funkcí (informace úloh se liší pro různé úrovně schopnosti θ testovaných), pomocí kterých můžeme vybrat s velkou přesností úlohy odpovídající námi zvolené úrovni schopnosti θ. Úlohy s vyššími hodnotami parametru a poskytují více informace o skórování testovaných a tím větší přesnost. Chceme-li například vytvořit test s takovým hraničním skórem, který vytřídí 50% testovaných, vybereme úlohy s vysokými hodnotami a a s hodnotami b blízko nuly (průměrná schopnost θ). Podle informačních funkcí položek vybereme tedy takové, které podávají maximální informaci pro hodnoty $\theta = 0$. Nakonec zkontrolujeme náš výběr pomocí informační funkce testu, která je výsledným součtem informačních funkcí jednotlivých položek, a křivky standardní chyby. Standardní chyba by měla být pro oblast okolo $\theta = 0$ co nejmenší.

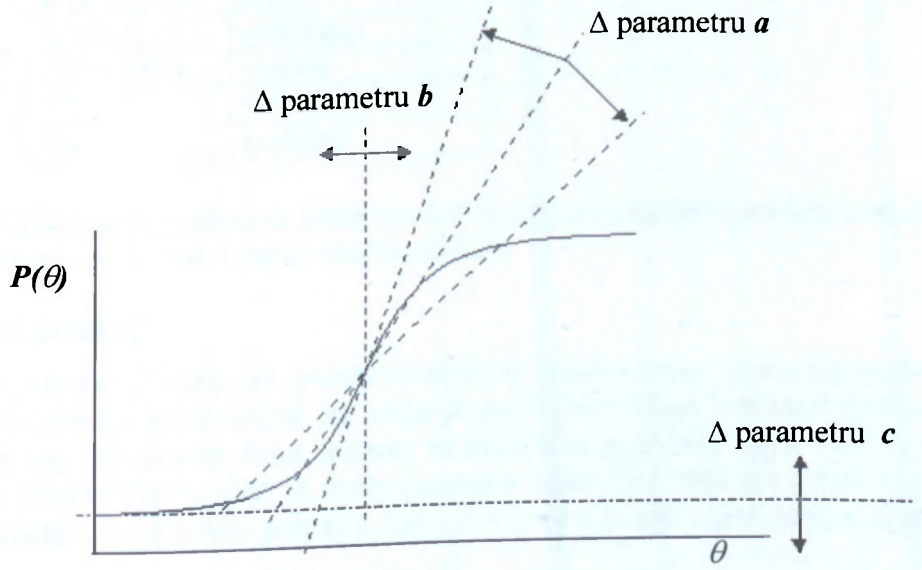
Problematika teorie odpovědi na položku (IRT) je velmi rozsáhlá a přesahuje rámec této disertační práce, proto dále uvádíme pouze přehled jejích základních myšlenek. Velmi povrchně je nastíněna otázka odhadu položkových parametrů a odhadu úrovně schopnosti testovaných, protože vyžaduje poměrně složitý matematický aparát, který je nutné řešit za pomocí software (viz např. Embretson; Reise 2000, Hambleton; Swaminathan; Rogers 1991).

[36] N_0 je četnost vynechaných odpovědí, N celkový počet odpovědí.
[37] N_0 je četnost vynechaných odpovědí, N celkový počet odpovědí.

IRT a její předpoklady

Teorie odpovědi na položku (IRT) popisuje vztah charakteristik úloh (položkové parametry) a charakteristik testovaných (latentní rysy/ schopnosti) pomocí pravděpodobnosti správné odpovědi.Bylo vyvinuto množství různých IRT modelů pro dichotomická a polytomická data. Teorie odpovědi na položku (IRT) pro binárně (dichotomicky) skórované úlohy je založena na dvou základních předpokladech (Hambleton; Swaminathan; Rogers 1991):

1) odpověď testovaného na příslušnou testovou položku lze předpovědět či vysvětlit souborem latentních rysů nebo-li schopností testovaného (označených řeckým písmenem theta θ). Latentní rysy nejsou přímo měřitelné, ale předpokládá se, že se projevují v chování testovaného a ovlivňují jeho odpovědi. Jsou na testu nezávislé.
2) vztah mezi odpovědí testovaného v testové položce a jeho schopnostmi může být matematicky popsán funkcí pravděpodobnosti správné odpovědi na položku $P(\theta)$, tzv. **charakteristickou křivkou/ funkcí položky** (item charakteristic curve/ function, ICC). Tato křivka esovitého tvaru zachycuje, jak při rostoucí úrovni schopnosti θ roste pravděpodobnost správné odpovědi. Tvar a polohu křivky (viz obr. 4-20) určují podle zvoleného modelu jeden až tři parametry – obtížnost (b), diskriminační schopnost (citlivost) úlohy (a) a pseudonáhodný parametr hádání (c).



Obr. 4-20 Ukázka charakteristické křivky položky (upraveno podle Chong 2006)

Parametry položek

Obtížnost položky

Obtížnost položky je vyjádřena parametrem b (difficulty parametr nebo threshold), který teoreticky může nabývat hodnot od - ∞ do + ∞, ale v praxi se jeho hodnota pohybuje mezi -3 a 3 (Baker 2001). Čím větší je b. tím obtížnější je úloha. Graficky je obtížnost úlohy dána polohou charakteristické křivky (ICC) vzhledem k ose schopností θ. Jde o bod na ose schopnosti, pro který je pravděpodobnost správné odpovědi rovna 0,5, resp. (1+c)/2 u 3-parametrového modelu. Čím je ICC položena více doprava vzhledem k vodorovné ose θ , tím těžší je úloha. Sečteme-li charakteristické funkce položek v celém testu, můžeme výslednou charakteristickou funkci testu použít k předpovídání skórů testovaných s danou úrovní schopnosti θ. Je-li test složen z relativně obtížných úloh, je charakteristická funkce testu posunuta doprava a testovaní mají tendenci k nižším očekávaným skórům než je tomu u relativně snadných položek. Příklad bude uveden dále u 1-parametrového modelu.

Citlivost (diskriminační schopnost) položky

Diskriminační schopnost položky je dána parametrem *a* (discrimination parameter), který teoreticky může nabývat hodnot od - ∞ do + ∞ (Baker 2001), ale v praxi se jeho hodnota pohybuje obvykle mezi 0 a 2,8 (Baker 2001). Čím větší je hodnota *a*, tím lépe úloha rozlišuje mezi testovanými nalevo a napravo od své polohy. Graficky se citlivost úlohy projevuje strmostí ICC v jejím prostředním úseku. Čím větší sklon má křivka (čím je strmější), tím má úloha lepší rozlišovací schopnost, tím je citlivější. Strmost křivky, a tím také parametr *a* dosahuje své maximální hodnoty v bodě, ve kterém se úroveň schopnosti θ rovná obtížnosti položky. To znamená, *b* označuje bod na ose schopnosti θ, v kterém úloha nejlépe rozlišuje mezi testovanými. Negativní parametr *a* citlivosti značí něco chybného v úloze. Buď se jedná o úlohu s technickými nedostatky nebo jde o dezinformaci zpravidla mezi studenty s vysokou úrovní schopnosti. Baker (2001) uvádí doporučené hodnoty parametru *a* (viz tab. 4-12).

Tab. 4-12 Doporučené hodnoty parametru citlivosti *a* pro logistické modely položek.

a (logistický)	*a* (normální)	citlivost úlohy
0	0	žádná
0,01-0,34	0,006-0,2	velmi nízká
0,35-0,64	0,206-0,376	nízká
0,65-1,34	0,382-0,788	přiměřená
1,35-1,69	0,794-0,994	vysoká
≥ 1,7	> 1	velmi vysoká
+ ∞	+ ∞	perfektní

Pro převod z logistického modelu na model podobný normální ogivě je potřeba hodnoty vydělit číslem 1,7 (normal ogive model value, scaling factor).

Uhádnutelnost položky

Uhádnutelnost položky je dána tzv. pseudonáhodným parametrem *c* (guessing parameter, pseudo-chance-level parameter), který udává, jak velká je pravděpodobnost uhádnutí správné odpovědi na všech úrovních osy schopnosti. Jeho hodnota se teoreticky pohybuje mezi 0 a 1, v praxi většinou mezi 0 a 0,35 (Baker 2001). Čím je tento parametr vyšší, tím výše na svislé ose $P(\theta)$ je dolní asymptota charakteristické křivky položky. Když $b < 0$ a $a < 1$, pak *c* není zřejmé (Baker 2001).

IRT modely

Vztah mezi úrovní latentního rysu/ schopnosti θ a pravděpodobností správné odpovědi $P(\theta)$ na dichotomicky skórovanou položku lze popsat více či méně přesně třemi různými unidimenzionálními modely[38] nebo-li logistickými funkcemi (Hambleton; Swaminathan; Rogers 1991). Modely se snaží specifikovat očekávaný vztah mezi pozorovatelnými odpověďmi testovaných na úlohy a nepozorovatelnými latentními rysy (schopnostmi), které řídí jejich odpovědi.

1-parametrový logistický model nebo-li Raschův model

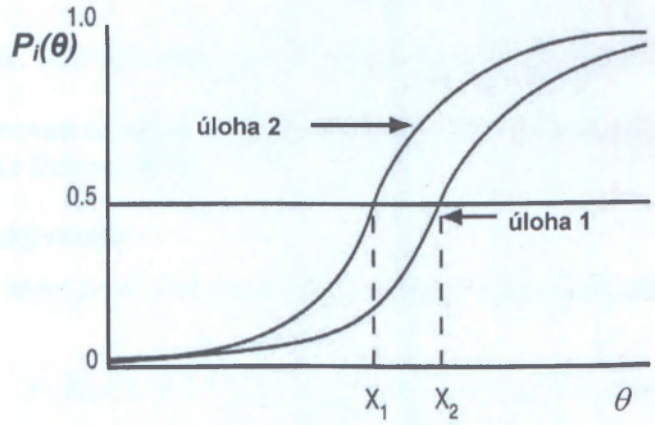
Nejjednodušším a současně nejrozšířenějším IRT modelem, nazývaným podle dánského matematika Raschův model, je **1-parametrový model**, který obsahuje pouze parametr obtížnosti. Tento model je tedy vhodný pro testy složené z přibližně stejně citlivých úloh.

Má tvar $$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}, \quad i = 1, 2, .., n, \text{ kde}$$

[38] Vedle unidimenzionálních modelů (jsou nejjednodušší, pracují pouze s jedním latentním rysem) existují pro binární data také multidimenzionální modely, v kterých dvě nebo více úrovní latentního rysu ovlivňují výkon testovaného (více viz např. Embretson; Reise 2000). Řada modelů vznikla také pro polytomické formáty odpovědí na položky či pro škály.

$P_i(\theta)$ je pravděpodobnost, že náhodně vybraný testovaný se schopností θ vyřeší úlohu i správně; nabývá hodnot od 0 do 1
D je konstanta rovna 1,7, pomocí které se distribuční funkce normálního rozdělení (ogiva) převádí na výhodnější logistickou funkci (protože obě funkce mají velmi podobný průběh)
b je parametr obtížnosti úlohy (viz výše)
θ je úroveň schopnosti (latentního rysu) testovaného

Obr. 4-21 zobrazuje charakteristické křivky dvou úloh, které se liší jen s ohledem na obtížnost. Úloha 2 je snazší než úloha 1, protože bod X_1 leží vzhledem k ose θ blíže k nule než X_2. ICC úlohy 1 leží více vpravo.



Obr. 4-21 ICC úloh 1 a 2 s rozdílnou obtížností

Vodorovnou osu tvoří úroveň schopnosti testovaného θ, svislou pravděpodobnost správné odpovědi na úlohu (upraveno podle Urbina 2004).

2-parametrový logistický model nebo-li Lordův model

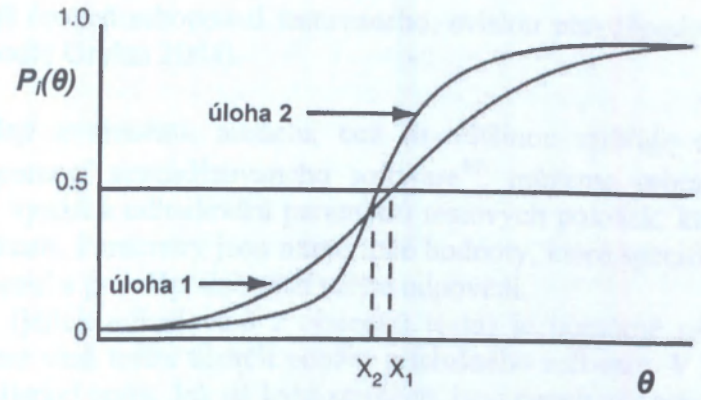
2-parametrový model poprvé zavedl v 50. letech 20. stol. Lord, v 60. letech 20. století se jím zabýval Birnbaum. Tento model, zřejmě zobecněním 1-parametrového modelu, uvažuje vedle obtížnosti položky (b) také její citlivost (a). Používá se u otevřených úloh.

Má tvar $$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \; i = 1, 2, .., n,$$ kde

a je parametr vystihující diskriminační schopnost položky[39].

Na obr. 4-22 jsou charakteristické křivky dvou úloh, které se liší v obtížnosti a citlivosti. Úroveň schopnosti asociovaná s 50% pravděpodobností správné odpovědi je trochu vyšší u úlohy 1 (x_1) než u úlohy 2 (x_2). Kromě toho stoupání (strmost) těchto dvou křivek, které ukazuje poměr změny ve schopnostech a změny v pravděpodobnosti správné odpovědi, se liší. Úloha 2 jakožto strmější v prostředním úseku je citlivější než úloha 1. Křivky, které se protínají jako v tomto případě, jsou nežádoucí.

[39] Ostatní proměnné a konstanta D jsou shodné s 1-parametrovým modelem.



Obr. 4-22 ICC úloh 1 a 2 s rozdílnou obtížností a citlivostí.

Vodorovnou osu tvoří úroveň schopnosti testovaného θ, svislou pravděpodobnost správné odpovědi na úlohu (upraveno podle Urbina 2004).

3-parametrový logistický model

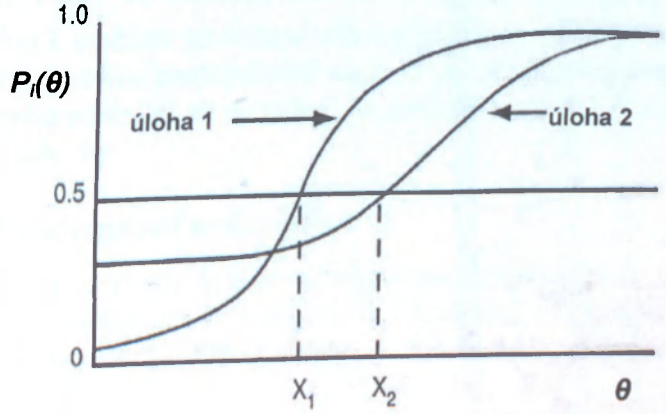
3-parametrový model, který je vhodný pro úlohy s výběrem odpovědi, navrhl Birnbaum. Tento model je dán tvarem

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \; i = 1, 2, .., n, \text{ kde}$$

c je parametr hádání a nabývá hodnot od 0 do 1, ale ze své povahy by neměl být vyšší než $1/k$, kde k je počet nabídek pro položku

Kromě parametrů a a b zahrnuje 3-parametrový model také parametr hádání c, jehož hodnota vyjadřuje pravděpodobnost dosažení správné odpovědi při tzv. slepém hádání nezávisle na schopnosti θ (u úlohy se 4 nabízenými odpověďmi je pravděpodobnost uhádnutí 0,25).
U tohoto modelu již díky parametru c není dolní asymptotou charakteristické křivky úlohy hodnota 0 jako u 1- a 2-parametrického, ale c. Tím je parametr obtížnosti položky určen bodem na škále schopnosti, v němž $P(\theta) = (1 + c)/2$. Parametr a je stále úměrný strmosti ICC v bodě $b = 0$, přičemž zde je tato strmost rovna $a(1 - c)/4$.

Obr. 4-23 zobrazuje ICC křivky dvou úloh, které se liší třemi parametry: a, b a c. ICC úlohy 1 je strmější než ICC úlohy 2, tj. úloha 2 zřejmě nerozlišuje mezi jedinci různých úrovní θ tak dobře jako úloha 1. Dle ICC úlohy 2 lze usoudit, že i testovaní nízkých úrovní θ jsou schopni správně uhádnout odpověď na úlohu 2, parametr c je u této úlohy vyšší než u úlohy 1 (blíže k 0 na svislé ose). Navíc 50% pravděpodobnost úspěchu je asociovaná s vyšší úrovní schopnosti (X_2) u úlohy 2. Úloha 2 je tedy obtížnější než úloha 1. Úloha 2 je proto zřejmě z hlediska měřitelných charakteristik méně vhodná než úloha 1.



Obr. 4-23 ICC úloh 1 a 2 s rozdílnou obtížností, citlivostí a různým parametrem hádání.

Vodorovnou osu tvoří úroveň schopnosti testovaného, svislou pravděpodobnost správné odpovědi na úlohu (upraveno podle Urbina 2004).

Pokud data odpovídají zvolenému modelu, což se většinou zjišťuje statisticky χ^2 testem či grafickou metodou pomocí specializovaného software[40], můžeme sebraná data o testových a položkových skórech využít k odhadování parametrů testových položek, které rozmístí testované a úlohy na škále schopnosti. Parametry jsou numerické hodnoty, které specifikují formu vztahu mezi naměřenými schopnostmi a pravděpodobností určité odpovědi.
Stanovení parametrů (jejich odhadování z výsledků testu) je poměrně náročné, neboť vyžaduje iterativní postupy, které však může ulehčit použití příslušného software. V IRT se toto odhadování parametrů nazývá **kalibrací testu**. Jak již bylo zmíněno, jsou parametry schopnosti (charakterizující testované) nezávislé na testových položkách, s jejichž pomocí jsou kalibrovány, a položkové parametry nezávislé na pravděpodobnostním rozdělení schopnosti ve skupině testovaných. Techniku kalibrace testu navrhl jako první koncem 60. let 20. století A. Birnbaum, jak uvádí Komenda (2003), a od 70. a 80. letech byl potom navrhován výpočetní software. Dnes se k odhadování IRT parametrů používají programy jako je např. TESTFACT, BILOG-MG, XCALIBRE, MULTILOG, PARSCALE či RUMM.

Odhadování parametrů testové položky a schopnosti testovaného

V IRT modelech závisí pravděpodobnost správné odpovědi na schopnostech testovaného a na parametrech, které úlohu charakterizují. Při použití těchto IRT modelů lze oproti KTT z odpovědí na úlohy získat odhady parametrů testových položek, a to nezávisle na znalostech testovaných. Tzn. odhady parametrů testové položky nezávisejí na úrovni schopnosti studentů, kteří tuto položku řeší. Číselné hodnoty parametrů jsou tedy vlastností položky a ne skupiny testovaných, která na položku testu odpovídá. Specifikují formu vztahu mezi měřenými schopnostmi a pravděpodobností určité odpovědi na položku. Nejpoužívanějšími technikami odhadování parametrů testové položky i schopnosti jsou *metoda maximální věrohodnosti* (maximum likelihood procedure) a *Bayesova metoda odhadování parametrů*. Dále nastíníme pouze metodu maximální věrohodnosti.

Odhadování parametrů testové položky nezávisle na znalostech testovaných

K určení parametrů testové položky použijeme např. *metodu maximální věrohodnosti* (Joint maximum likelihood procedure), která je aplikovatelná na všechny tři parametrické modely. Předpokládáme, že parametry schopnosti jsou známy. Jak uvádí Komenda (2003), je základním krokem najít charakteristickou funkci položky, která by nejpřesněji vystihla empirické četnosti správných odpovědí. Za matematický model vyrovnávací křivky je třeba si zvolit jeden ze tří IRT modelů, např. dvouparametrový. Nejprve se zvolí počáteční číselné hodnoty parametrů b a a a vypočítají se pro ně hodnoty pravděpodobnosti správné odpovědi $P(\theta_i)$ pro každou úroveň schopnosti. Potom se vyhodnotí shoda mezi empirickými hodnotami $p(\theta_i)$ a právě vypočítanými modelovými hodnotami $P(\theta_i)$ ve všech hodnotách θ. Najdou se korekce stávajících hodnot obou parametrů, které povedou k lepšímu přiblížení teoretické křivky. Tento iterativní postup se opakuje tak dlouho, dokud nejsou korekce zanedbatelně malé. Tím odhadování končí, číselné hodnoty obou parametrů vypočítané jako poslední se považují za konečné odhady b a a určující tvar teoretické charakteristické křivky položky.

Odhadování parametrů schopnosti testovaného

V IRT je účelem aplikace testu zjistit, kde se zkoušený umisťuje na ose schopnosti θ. Test k měření neznámé latentní vlastnosti (schopnosti) je tvořen n položkami, z nichž každá měří určitou stránku této vlastnosti. Při odhadování neznámé hodnoty parametru schopnosti testovaného budeme

[40] Četné statistické metody k ověření vhodnosti modelu uvádějí např. Hambleton; Swaminathan; Rogers 1991, Orlando; Thissen 2000 aj.s

předpokládat, že číselné hodnoty parametrů testových položek jsou známy. Z toho plyne, že hodnoty známých položkových parametrů a parametrů schopnosti mají společnou metriku. K odhadu parametru schopnosti se využívá vedle známých hodnot položkových parametrů také vektor odpovědí testovaných (složený z 0 a 1 kvůli dichotomického skórování).

Opět použijeme iterativní metodu maximální věrohodnosti. Nejprve zvolíme počáteční hodnotu schopnosti zkoušeného a vypočítáme pravděpodobnost jeho správné odpovědi (pomocí známých položkových parametrů) na každou položku testu. Vypočítané pravděpodobnosti se mají co nejvíce shodovat s vektorem odpovědí testovaného. Potom provedeme korekci odhadu schopnosti. Tím získáme přesnější hodnotu parametru schopnosti. Postup opakujeme tak dlouho, dokud korekce není zanedbatelně malá. Výsledkem je číselný odhad parametru schopnosti studenta. Tento proces se provádí odděleně pro každého testovaného. Postupné přiblížení k hledané hodnotě parametru schopnosti jedince se vypočítává podle vzorce:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_{i=1}^{n} a_i \left[u_i - P_i(\hat{\theta}_s)\right]}{\sum_{i=1}^{n} a^2{}_i P_i(\hat{\theta}_s) Q_i(\hat{\theta}_s)}$$, kde $\hat{\theta}_s$ je odhadnutá schopnost testovaného v kroku (iteraci) s, a_i je diskriminační parametr položky i, u_i je odpověď testovaného na položku i ($u_i = 1$ pro správnou odpověď, $u_i = 0$ pro chybnou odpověď), $P_i(\hat{\theta}_s)$ je pravděpodobnost správné odpovědi, $Q_i(\hat{\theta}_s) = 1 - P_i(\hat{\theta}_s)$ je pravděpodobnost chybné odpovědi na položku i pro zvolený IRT model ICC a na úrovni schopnosti θ v iteraci s.

Odhad schopnosti studenta nezávisí na tom, jaké položky se pro tento odhad použily, pokud všechny položky měří tu samou latentní vlastnost a hodnoty všech položkových parametrů mají společnou metriku. Necháme-li tedy testovaného řešit dva paralelní testy po 15 položkách s různou průměrnou obtížností a jeho odpovědi využijeme k odhadu jeho schopnosti, měli bychom dostat v obou případech stejnou hodnotu odhadu schopnosti.

Pohled do historie IRT

I když u nás ještě teorie odpovědi na položku (item response theory, IRT) není příliš známa[41], její pojetí a metodologie byly vyvinuty před více než tři čtvrte stoletím. Uplné počátky IRT spadají na počátek 20. století, kdy vznikly práce Bineta a Simona (1911). Ti používali k popisu funkčního vztahu mezi poměrem správné odpovědi na položku a chronologickým věkem testovaných (po roce od 5 do 15 let) tabulku pro umístění položek v inteligenčním testu (Binet-Simonova škála). Později Terman (1916) přepracoval Binet-Simonovu škálu, graficky znázornil poměr správné odpovědi jako funkci věku a datům přizpůsobil pravidelnou křivku známou dnes jako charakteristická křivka položky (ICC). Po mnoho let byl tento přístup pomocí charakteristických křivek (ICC) považován pouze za alternativu položkové analýzy.

Pojmový základ IRT položil 1925 Thurstone ve svém díle „*A Metod of Scaling Psychological and Educational Tests*". V něm popisuje techniku pro umístění úloh z testu mentálního vývoje dětí od Bineta a Simona (1905) na věkové škále. Thurstonovi kolegové a studenti pozdeji vypilovali teoretické základy IRT. 1936 odvodil Richardson vztah mezi IRT parametry a klasickými charakteristikami. Slo o prvni způsob získávání odhadů IRT parametrů. Nové postupy výpočtu maximálně věrohodných odhadu IRT parametrů popsal později Lawley (1943-44) z University of Edinburgh. Ten v roce 1943 ve své práci ukázal, že mnoho z pojetí klasické teorie testu lze vyjádřit ve formě parametrů charakteristické křivky položky. Lawley definoval pravý skór z hlediska položek v testu a ukázal, že klasický koeficient reliability může být také vyjádřen jako funkce těchto položkových parametrů (dnes známa jako informační funkce).

[41] O IRT informovali zatím jen Komenda (2003), Denglerová (2003, 2005), Urbánek; Šimeček (2001), Jelínek; Květoň; Denglerová (2006).

Na Lawleyho navázal v 50. letech Lord (1952), jehož práce se stala v posledních 50 letech 20. století hybnou silou jak pro vývoj teorie založené na položkách, tak pro její aplikace. Lord systematicky definoval, rozvíjel a zkoumal tuto teorii a zároveň vyvíjel počítačové programy, které by umožnily její uvedení do praxe. Jeho snaha vyvrcholila zveřejněním knih (s M. Novickem, 1968; 1980) zabývajícími se praktickými aplikacemi IRT. Práce Lawleyho a Lorda položily základní koncepty teorie založené na položkách, dnes známé jako teorie odpovědi na položku (Baker 2001). K položení základů IRT přispěl 1960 významně i G. Rasch. Teze jeho publikace *Probabilistic models for some intelligence and attainment tests* (např. Urbánek; Šimeček 2001) měnily dosavadní náhled na povahu měření u stále většího počtu odborníků z oblasti testů. Rozšíření IRT však přesto zůstávalo až do konce 80. let 20. století v pozadí, zejména kvůli komplexním základním statistickým výpočtům, a v oblasti hodnocení ve vzdělávání dominovala klasická teorie testu založená na testových skórech. V roce 1985 vydal F. Baker knihu s názvem *The Basics of Item Response Theory*. K rozvoji IRT přispěl tím, že spojil teoretické poznatky o IRT se softwarem pro osobní počítače APPLE II a IBM. Tento software umožňoval čtenářům experimentovat s teoretickým pojetím teorie.

II VÝZKUMNÁ ČÁST

5 Výzkumné problémy

Počítače se již desítky let využívají na celém světě k různým činnostem spojeným s testováním: k analýze testových výsledků a jejich interpretaci, k optickému snímání záznamů testovaných, k administraci testů i vytváření testových variant. Tyto činnosti vyžadují nejen mnoho času, ale některé z nich, např. adaptivní testování a zjišťování IRT parametrů se bez počítačů neobejdou. Významné je zejména využití počítačů k položkové analýze pomocí IRT. Jak jsem již dříve uvedla, touto problematikou se naši odborníci zabývají pouze ojediněle (např. Komenda a Mazuchová, Denglerová, Jelínek, Květoň, Šimeček, Urbánek), a to zatím jen teoreticky. V souvislosti s touto skutečností a s vývojem uplatnění počítačů při testování uvedenými v teoretické části formulujeme následující výzkumné problémy, které dále specifikujeme výzkumnými otázkami.

A **Jaký je současný stav využívání počítačů při testování?**

B **Čím se liší využití metod KTT a IRT při tvorbě testu a analýze jeho výsledků (teoreticky a při konkrétní aplikaci)?**

Problém upřesňujeme následujícími výzkumnými otázkami.

Ba Jaké jsou rozdíly zmíněných metod při analýze výsledků testovaných
Bb Jaké jsou rozdíly zmíněných metod při posuzování vyrovnanosti variant
Bc Jaké jsou rozdíly zmíněných metod při položkové analýze

6 Využívání počítačů k testování

Testování je jednou z důležitých metod hodnocení výsledků učení. Využíva se v průběhu výuky (k formativnímu hodnocení) a na jejím konci (k sumativnímu hodnocení) nebo při přijímacím řízení na jiný druh nebo vyšší stupeň školy. Jelikož výsledky testů pro sumativní hodnocení a testů používaných při přijímacím řízení mohou mít velký vliv na budoucnost studenta, je důležité navrhovat, administrovat a vyhodnocovat testy velice pečlivě. Konvenční tvorba a administrace, vyhodnocování a interpretace výsledků testů vyžadují mnoho práce, času a důsledné dodržování určitých pravidel. Náročnost těchto činností může být výrazně snížena použitím počítačů. To však klade určité nároky na hardware a software.

Hardware pro samostatnou pracovní stanici by měl obsahovat počítač, zpravidla propojený s jinými počítači do lokální sítě či sítě pro dálkový přenos dat. Tato pracovní stanice může být také terminálem pro počítač pro větší počet uživatelů. Dále dostatečně velkou paměť pro softwarové aplikace a dostatečně velkou kapacitu ukládání - na pevné disky, diskety, USB disk, CD-ROM; příslušenství pro komunikaci s počítačem, příslušenství umožňující zobrazení textu, grafiky a zvuku

(např. grafickou kartu, video a zvukovou kartu), tiskárnu a příslušenství pro komunikaci dat mezi stanicí a řídícím počítačem. Hardware musí být doplněn dvěma základními složkami **software**: vhodným operačním systémem a aplikačním software (testware) pro tvorbu (generování), administraci, vyhodnocování a případně i interpretaci výsledků testů.

Využívání počítačů k tvorbě, skórování, vyhodnocování a analýze testu není ve světě ani u nás žádnou novinkou. Zpočátku byly počítače využívány jako zaznamenávací zařízení, monitorovaly studentův pokrok, třídily a skladovaly testová data. Používaly se také při širokoplošném testování k analýze výsledků testů. Později se pozornost obrátila k tvorbě testu. Počítačem podporované testování je však stále ve svém vývoji. Zcela novou etapu výkonových testů (v porovnání s testy zadávanými formou tužka-papír) představují v současné době ve světě testy zadávané počítačem, z nichž některé existují pouze v elektronické podobě, jiné jsou jen alternativou k „papírovým" testům o stejné délce i obtížnosti. Obzvláště efektivní se zdají být v posledních letech počítačové adaptivní testy (computer adaptive tests), kdy počítač vybírá úlohy pro testovaného z relativně velké banky úloh podle jeho odpovědi na úlohy předešlé (viz kap. 7). Pokud testovaný odpoví správně, dostane úlohu obtížnější, pokud chybně, je mu zadána úloha snadnější.

6. 1 Tvorba počítačového testu

V posledních 30 letech jsou počítače využívány ke konstrukci testu, a to dvojím způsobem.
První a častější metodou je zaznamenání testových úloh vztahujících se k nějakému učivu do počítače a jejich skladování podle obsahu a obtížnosti v počítači v přesně stejné podobě, jak se objeví v testu. Základem tvorby nebo-li generování testu pomocí počítače je banka úloh (pool, item bank), ze které podle našeho požadavku počítač vytvoří test nebo několik jeho variant. Bankou úloh se rozumí velký soubor úloh tříděných podle příslušnosti k určitému tematickému celku, operační úrovni (viz kapitola 4), určité obtížnosti a citlivosti.
Počítač vybírá, seskupuje a předává tyto úlohy testovaným buď předem stanoveným nebo interaktivním způsobem, kdy je výběr úlohy závislý na výkonu zkoušeného v předchozí úloze. V druhém případě jde o tzv. adaptivní testování.
Druhou metodou je ukládání do počítače pouze pravidel pro generování testových úloh místo hotových úloh. Tento způsob je vhodný pro ověřovací testy.

U tradičně zadávaných testů forma tužka-papír učitel rozhodne, jaké úlohy a v jakém pořadí v testu použije. Testy potom vytiskne na počítači, rozmnoží a rozdá studentům, kteří je vyplňují. V počítačovém světě jsou testy buď baleny do elektronických souborů nebo generovány, když je potřeba. Jednotlivé úlohy mohou být na počítači sestavovány do testové formy mnoha způsoby. Např. v jednom testu se mohou úlohy objevit v pořadí, ve kterém byly uloženy v počítači. V pozdějším testu mohou být vybírány náhodně. Jiný způsob je uspořádat úlohy náhodně pro každého testovaného tak, že každý odpovídá na stejné otázky, ale v různém pořadí. Nebo je možné vybírat úlohy z velkého souboru, takže každý testovaný obdrží v testu různý set úloh. Počítač umožňuje také zařazení pestřejších, různorodějších úloh.

Generování úloh počítačem

Při generování úloh, jak již bylo zmíněno, počítač ukládá místo úloh v přesně takové podobě, v jaké se objeví v testu, pouze jejich obecný formát či šablonu. Tedy např. místo vytvoření 10 rozdílných úloh na výpočet obsahu trojúhelníka uloží tvůrce úloh do počítače obecný tvar úlohy (obr. 6-1) a nechá počítač doplnit různá čísla pro každou novou úlohu (Aiessi, Trollip 2001).

Vypočítej obsah trojúhelníka, jehož strany jsou dlouhé ____ cm, ____ cm, ____ cm.

Odpověď: _____ cm^2.

(2 body)

Obr. 6-1 Obecný formát úlohy na výpočet obsahu trojúhelníka

Ačkoliv tento způsob vytváření testových úloh značně zamezuje opisování, může současně snižovat reliabilitu testu, což ale na rozdíl od rozlišujících testů není u ověřovacích testů, které učitelé vytvářejí více, tak významné.

Myšlenku velkých bank úloh můžeme vidět i v širším kontextu. Přestože si jednotliví učitelé či školy mohou vytvářet vlastní banky úloh, je prospěšné sdílet banku úloh s jinými učiteli či jinou školou pomocí síťově zapojených počítačů k centrálnímu serveru. Učitel má tak možnost sestavit si snadno více paralelních testů. Poté, co jsou úlohy vybrány, mohou být na jednotlivých školách vytištěny pro namnožení nebo být studentům zadávány on-line. Výhodou sdílené banky úloh, do které úlohy vložilo několik učitelů, je tedy rychlý přístup k většímu množství úloh, než u individuální banky úloh. Učitelé mohou kdykoli vytvářet různé testy, což minimalizuje možnost vyzrazení úloh. Pokud by takto velké banky úloh byly pečlivě spravovány, mohla by být shromážděna vhodná data k vyhodnocení každé úlohy. Špatné úlohy by mohly být z banky na základě položkové analýzy eliminovány a dobré ponechány, a tím by se zlepšila kvalita testů všech učitelů. Pomocí laserové tiskárny můžeme do testů určených k vytištění integrovat i složitější grafiku a vytvářet testy ve strojově čitelné formě, které potom vyplněné může dále zpracovávat optický scanner (viz kap. 4). Optický scanner je dnes již schopen zaznamenat i poznámky psané rukou, i když zatím s velkou chybovostí. Pokud jsou testy vyhodnocovány automaticky (strojově), je snadné nechat počítač provést statistickou analýzu, a tím aktualizovat banku úloh vyřazením úloh nevhodných. V současné době se používají různé typy scannerů spolu se software využívajícím rozdílné technologie (OMR, ICR, OCR). Nebezpečím velkých bank, kdy úlohy vkládá větší počet lidí, je však důvěryhodnost vložených úloh, zda skutečně testují dané učební cíle, zda jsou kvalitní a dostatečně vhodné. Pro neformální testování například není rozsáhlá banka úloh zapotřebí.

6. 2 Administrace počítačového testu

Administrace testu na počítači probíhá tak, že testovaný sedí před počítačem a odpovídá na otázky, které se objevují na obrazovce. Počítač přebírá mnoho z role učitele a zadavatele testu. Samozřejmě ale zůstává lidský faktor stále nezbytný k určení obsahu testu a řízení testování. Výhodou zadávání testu na počítači oproti běžnější formě testů tužka-papír je *individualizace* dovolující testovaným začít řešit test, když jsou připraveni a ne ve stanovenou dobu. Testovaný se může také okamžitě po dopsání testu dovědět svůj výsledek. Obsah testu tak může být přizpůsoben individuálním potřebám testovaných. Počítačové testy mají výhody i pro učitele. Všechny odpovědi jsou v počítači ukládány a mohou sloužit k zlepšení banky úloh. Počítač dovoluje seskupovat data o jedincích či skupinách, ukazuje testovanému čas, který mu zbývá do konce testu, ukládá šablony odpovědí či zaznamenává, kolikrát testovaný použil „pomoc". Speciální způsob administrace vyžaduje adaptivní testování (viz kap. 7).

Důležité faktory související s tvorbou a administrací testu

V tab. 6-1 uvádíme faktory charakterizující testy při tvorbě a zadávání testu.

Faktory související s tvorbou testu	Faktory související s administrací testu
- účel testu - význam testu - cíle - délka testu - typy a zápis úloh - hraniční skór - časový limit - způsob sdělení výsledku z testu - prezentovaný výsledek	- obsluha testovacího programu - kontrola testové situace uživatelem - zabezpečení programu - tři fáze zadávání testu: Před testem Během testu Po testu

Tab. 6-1 Důležité faktory související s tvorbou a administrací testu

Faktory související s tvorbou testu

Prvním krokem při tvorbě jakéhokoli testu na počítači je stejně jako u testu v papírové podobě potřeba jasně stanovit jeho účel a vymezit obsah, jaký má test pokrývat. Protože ten samý test může být použit k různým účelům, měl by být interpretován vzhledem ke svému účelu.

Uvědomování si míry významu testu může v testovaném vyvolávat obavu z testu. U testování formou papír-tužka je zkoušenému dovoleno mazat odpověď a měnit ji. Mnoho počítačových testů to ale neumožňuje. Pokud test má důležité důsledky pro testované jako např. získání licence, přijetí na vyšší stupeň školy, zvyšuje se u testovaných úzkostnost, pokud nemohou své odpovědi měnit. Ke zmírnění obav z testování by měl proto počítačový program, který test zadává, vyhovět potřebám a očekáváním zkoušených, tj. být dostatečně flexibilní.

Nejdůležitějším rysem testu je specifikace jeho cílů, ať jde o test zadávaný formou tužka-papír či počítačový test. Tvůrce testu musí mít ujasněno, co má test testovat. Ke stanovení cílů se obvykle používá specifikační tabulka (viz kap. 4).

Dále musíme stanovit délku testu. Ta je dána počtem úloh, které test obsahuje. Běžné testy ve třídě jsou většinou omezeny vyučovací hodinou či charakteristikami testovaných. Čím mladší testovaní, tím kratší test. Čím více úloh, tím více času je potřeba na jejich řešení, ale s počtem úloh roste i reliabilita testu. Motivace, úroveň schopnosti čtení a dokonce i fyzické prostředí ovlivňují délku testu.

Jiným důležitým aspektem je způsob zapsání úloh do počítače. Úlohy musejí být do počítače zapsány tak, aby testovaný musel odpovídat pouze na úlohy a nezabývat se ještě přemýšlením o tom, jak má odpověď uvést. Chybný způsob zapsání správné odpovědi může vést ke ztrátě bodů pro testovaného, což není správné. Např. má-li student napsat jméno současného prezidenta ČR a napíše Václav Klaus místo Klaus (odpověď, kterou tvůrce testu a tedy i počítač očekává), pravděpodobně bude jeho odpověď skórována automaticky jako nesprávná.

Počítačový program může v sobě zahrnovat okamžitou zpětnou vazbu pro testovaného ve formě sdělení jeho výsledku (prospěl, neprospěl či jeho skór), ale nemusí, i když většinou jej testy poskytují. Počítač sbírá data, zaznamenává hrubý skór (bodový výsledek v testu), jednotlivé odpovědi na úlohy, čas, změny provedené v odpovědích a většinou hned po dopsání počítačového testu se testovaný z monitoru dovídá svůj výsledek. Forma sdělení výsledku z testu je různá.

Kromě informace pro testované poskytuje počítač i informace pro učitele (výsledky všech testovaných a souhrnné statistiky testu).

Neméně důležitými aspekty při tvorbě testu jsou také stanovení časového limitu (time limit) a hraničního skóru (cut-off-score, passing score) v testu. Hraniční skór odděluje od sebe testované, kteří v testu uspěli od těch, kteří v něm neuspěli. Může být i vícestupňový, záleží na účelu testu.

Počítač je schopen vedle tradičních testů realizovat i nové formy testů, které by bez použití počítače nebyly možné. Nejběžnějšími typy úloh v multimediálním prostředí jsou úlohy s výběrem odpovědi či se stručnou tvořenou odpovědí, protože je jednoduché na takové úlohy odpovídat (stačí jedno kliknutí myši, zmáčknutí tlačítka na klávesnici či napsání pár slov) a je snadné je také automaticky skórovat. Např. lze jednoduše pomocí myši uspořádat pojmy do určitého sledu (uspořádací úlohy) či přiřadit pojmy k jiným (přiřazovací úlohy). Skórování širokých otevřených úloh (extended items) je mnohem obtížnější. I když jsou většinou tyto úlohy stále skórovány manuálně, je dnes díky vývoji počítačové technologie již omezeně možné jejich automatické skórování (automated essay scoring, viz kap. 4).

Faktory související s administrací testu na počítači

Poté, co jsme navrhli obsah testu, stanovili jeho účel, navrhli jednotlivé úlohy i s jejich pořadím v testu, určili, jaké výsledky a v jaké formě budou testovaným či učiteli k dispozici, je třeba se zamyslet nad tím, jak test zadávat. Jak test bude vypadat na obrazovce, jak se bude obsluhovat, jaké možnosti budou přístupné pro testované, jaké pro učitele a jaká bezpečnostní opatření budou učiněna proti neočekávaným událostem. Anderson a Trollip (1981) uvádějí tři důležité principy: snadnou obsluhu programu, maximální kontrolu uživatele a ochranné zabezpečení programu.

Obsluha testovacího programu
Uživatel programu byl měl mít snadný přístup k potřebným informacím. Testovaný musí vědět, jak program ovládat, jakým způsobem zapisovat odpovědi atd. Program by měl být snadno ovladatelný, aby se testovaný mohl soustředit jen na obsah odpovědí.

Kontrola testové situace uživatelem (testovaným nebo učitelem)
Testování vyvolává zpravidla u testovaného úzkost, kterou je možné snížit, může-li testovaný testovou situaci kontrolovat. Má-li např. možnost zvolit si pořadí, v jakém bude jednotlivé úlohy řešit, nebo své odpovědi změnit. Další pomocí je, pokud má testovaný kdykoli možnost si přečíst instrukce k testu.

Zabezpečení programu
Program musí obsahovat ochranné sítě a mechanismy, které ztěžují náhodné činnosti testovaného, které by mohly způsobit vymazání dat. Protože důsledky vymazání dat mohou být značné, je vhodné nainstalovat do programu navíc ještě ochranné sítě, které fungují, i když byly ochranné mechanismy překonány.

Tři fáze administrace testu na počítači

Testování na počítači je zpravidla třífázový proces, který se skládá z fáze před tím, než zkoušený vyplňuje test, z fáze během testování a fáze po testování. Jak učitel, tak testovaný mají ve všech třech fázích odlišné role a vyžadují rozdílné informace. Pro testovaného je nejdůležitější první a druhá fáze.

První fáze: Před testem

Role učitele

Učitel musí udělat rozhodnutí o tom, kdo bude mít přístup k testu, kolik úloh bude test obsahovat, stanovit hraniční skór, časový limit a pořadí, v kterém se úlohy budou objevovat na obrazovce počítače. Ve většině testových situací je důležité se ujistit, že test řeší správní lidé. To lze udělat tak, že testovaný napíše své jméno či identifikační číslo, které program zkontroluje podle předem nadefinovaného seznamu a podle něj mu zadá správný test. Kontrola proti opisování je složitější. Pokud všichni testovaní sedí v řadě u monitorů, nastává stejná situace jako u tradičně zadávaných testů typu tužka-papír. Některé testovací programy dovolují učiteli zvolit délku testu či pořadí úloh, zda se úlohy budou na obrazovce objevovat náhodně či podle výběru testovaného či v předem stanoveném pořadí. Je vhodné, aby si test zkusil vyplnit nejdříve sám učitel, než ho zadá testovaným. Všechny odpovědi ale musí následně vymazat, pokud se ukládají, aby tím nedošlo k znehodnocení souhrnných statistik testu.

Role testovaného

Testovaný by měl být dopředu seznámen s ovládáním programu, mít možnost si vyzkoušet testovací prostředí (práce s klávesnicí, myší) včetně spouštění audio a video souborů, pokud je program nespouští automaticky ještě před začátkem testu. Na začátku testu potřebuje testovaný získat tři typy informací. Za prvé jasné pokyny, jak používat počítač a testovací program. Za druhé časový limit, kdy se začne čas odpočítávat a jaké pomůcky jsou povoleny. A za třetí, pokud to situace vyžaduje, detaily o zkoušce, jaké učivo bude zkoušeno, počet otázek či zda je stanoven hraniční skór.

Druhá fáze: Během testu

Role učitele

Během testování se toho po učiteli moc nevyžaduje, zejména pokud nejsou počítače propojeny do sítě. Učitel se může jen připravit na to, kdyby byl problém během zadávání testu. Pokud učitel přistihne žáka při opisování, může mu test zablokovat, aniž by se jeho dosavadní odpovědi ztratily, protože se během vyplňování testu ukládají. Druhou situací může být náhodné přerušení testu pro selhání počítače nebo při výpadku proudu. V této situaci potřebuje učitel testovanému znovu otevřít přístup k testu. Je žádoucí, aby program obsahoval ochranné mechanismy, které by i v tomto případě zamezily ztrátě odpovědí, protože pro testovaného je samotná situace přerušení dost stresová, natož aby musel všechny otázky zodpovídat znovu. Jsou-li počítače zapojeny do sítě, je užitečné poskytnout učiteli na jeho obrazovce údaje o testovaných – jejich jména, variantu testu, která jim byla zadána, počet již zodpovězených otázek, zbylý čas atd. Pokud učitel potřebuje s testovaným komunikovat (vyskytne-li se např. v nějaké otázce chyba), tak by to měl provádět tichou formou, např. e-mailem nebo chatem.

Role testovaného

Testovanému musí být dány k dispozici následující informace: text každé úlohy; úlohy vynechané či označené pro přezkoušení; čas, který zbývá do konce testu (pokud je stanoven). Tyto údaje by měly být nezávislé na programu, sám testovaný by si měl moci tyto informace kdykoli vyvolat z menu v programu.

V některých aplikacích je vhodné vyžadovat po testovaném, aby na otázku odpověděl ihned po objevení se na monitoru, zatímco u jiných aplikací je lepší dát mu možnost volby odpovědět či neodpovědět. Poskytnutí testovanému možnost neodpovědět na otázku hned, když se objeví na obrazovce, je to samé jako dovolit mu prohlédnout si všechny otázky, než na ně začne odpovídat. rady tradičních testů je listování povoleno. Časově příznivější je pro testovaného, když má možnost označit si otázky, ke kterým se chce vrátit a nemusí listovat celým testem, než na tyto úlohy narazí.

Změnit odpověď může žák okamžitě po dopsání odpovědi, když je ještě na obrazovce, či později, až všechny ostatní zodpoví (druhá nabídka není běžná u většiny testů). V adaptivních testech se po testovaném však vyžaduje okamžitá odpověď.

Testovaným by měla být poskytnuta okamžitá zpětná vazba, zvláště pokud jejich odpověď obsahuje chybu formátu. Např. pokud nabídky odpovědí u multiple-choice úlohy jsou označeny písmeny A až D a testovaný napíše jako odpověď číslo, okamžitě se mu objeví informace, že musí zadat písmeno, ne číslici (viz obr. 6-2).

Housenka vyleze za den 8cm, v noci o 2cm sklouzne. Jak vysoké je stéblo trávy, jestliže housenka byla na jeho vrcholu čtvrtý den?

Napiš písmeno, pak stiskni **RETURN**.

A. 24 *cm*
B. 26 *cm*
C. 28 *cm*
D. 32 *cm*

► 2

Musíte napsat písmeno, ne číslo.
Stiskněte **RETURN**, abyste mohli znovu odpovědět.

Obr. 6-2 Příklad chyby formátu

Užitečným prvkem je poskytnutí testovanému prostor na poznámky buď on-line či na papíře. U počítačového testu se zpravidla upřednostňuje on-line verze poznámek. Poznámky v počítači plní dvojí roli. Zaprvé umožňují shromažďovat informace o obsahu zkoušky či fungování testovacího software, zadruhé poskytují testovanému prostor vyjádřit své rozčilení či frustraci z testu. Tyto poznámky jsou poslány učiteli či jiné osobě odpovědné za opravování testu. Někdy poznámky testovaného mohou přispět k zamyšlení se nad úlohou, zda nemá více možných řešení.

Pokud je test časově omezen, musí být po uplynutí časového limitu přerušen, i kdyby testovaný ještě nebyl hotov. Program má dávat testovanému opakovaná varování, že se blíží konec časového limitu. Po ukončení testu je čas vynulován pro dalšího testovaného.

Třetí fáze: PO TESTU

Role učitele

Pro učitele a někdy ani pro testované nekončí testovací proces dopsáním testů. Učitel má k dispozici výsledky testovaných, údaje o testu a poznámky zkoušených k testu. Tyto informace mu jsou zprostředkovány pomocí menu či seznamu nabídek v učitelském programu. Přístup k těmto údajům by měl být spravován zákonem, chráněn heslem, bezpečně uzamčen či zakódován. Testovaný smí vidět pouze své výsledky testu, popř. souhrnnou informaci o výkonu celé skupiny. Zpravidla nejsou tyto údaje skladovány navždy, ale jen pár let s ohledem na typ informací. Vždy je vhodné, aby originální data zůstala nedotčena a manipulovalo se pouze s kopiemi. Většinou jsou výsledky také vytištěny, protože školní systém nebo testovací centrum vyžaduje trvalý tištěný záznam o všech testech.

Role testovaného

Nejdůležitější informací pro testovaného je po vyplnění testu jeho výsledek a vhodná zpětná vazba. Zpravidla je mu obojí poskytnuto, buď v elektronické podobě na monitoru či je mu dokonce

umožněno si svůj výsledek v testu vytisknout. Testovaný má samozřejmě přístup pouze ke svým výsledkům, nikoli k výsledkům ostatních.

Simulace

Zvláštním typem testu, který vyžaduje použití počítače, je **simulační test**. Simulace mohou být použity i v hodnotící fázi vyučování. Simulační testování má velký význam. Člověk není testován jen ze znalostí, ale také z praktických dovedností. Obtížnost při používání simulací k testování zpravidla tkví v automatizaci hodnotícího procesu. Relativně jednoduché je napsat simulaci, která dovoluje testovanému létat s letadlem za použití vhodného přístroje, ale mnohem složitější je naprogramovat simulaci tak, aby hodnotila výkon testovaného automaticky (viz např. Clauser: Margolis, aj. 1997a, 1997b, Trollip 1979). Automatizované hodnocení výkonu v simulacích vyžaduje dva rozdílně důležité kroky. V prvním kroku musíme stanovit hranice akceptovatelného výkonu, jak má správně podaný výkon vypadat. V druhém kroku musejí být v počítači naprogramovány hodnotící postupy. To může být také složité, obzvlášť jestli simulace má poskytnout feedback, proč něco bylo chybně uděláno. Oproti multiple-choice testům, které je snadné vyhodnocovat, protože oblast očekávaných odpovědí je malá, je ve většině simulací rozsah uživatelských akcí velmi velký a často plynulý. Např. hodnocení, jak lékař ošetřuje pacienta na pohotovosti, zabere delší čas a zahrnuje množství různých ač akceptovatelných řešení, např. je přípustné zkontrolovat pacientovi puls před měřením krevního tlaku a naopak. Další obtíží používání simulačních testů je rozhodnutí, jaký stupeň přesnosti či správnosti má být u validního testování požadován. Validita simulace je primárně funkce schopnosti predikovat výkon v reálných situacích, takže vyžaduje vysoký stupeň podobnosti s reálnou situací.

Příklad použití simulace při testování (Alessi, Trollip 2001)

Představme si simulační test, který určuje přístup do workshopu zabývajícím se spravedlivými pracovními záležitostmi, kterým čelí manažeři jako je diskriminace, sexuální obtěžování, msty atd. Program vystavuje uživatele několika situacím, s kterými se jako manažeři setkávají na pracovišti a vyžaduje po nich vhodné reakce, které by ve své situaci podnikli. Uživatel v roli manažera může vést virtuální rozhovory se svými zaměstnanci a nechat si od nich poradit. V každém případě simulační úloha obsahuje vhodné i méně vhodné lidi, vhodné i méně vhodné otázky, které jim má testovaný položit, a vhodné a nevhodné jednání, které má podniknout. Simulační test je úspěšně dokončen, pokud uživatel podnikl vhodná jednání, aniž by promluvil s nevhodnou osobou nebo položil nevhodné otázky. Je vyžadován perfektní výkon, protože společnost může být předmětem právní žaloby, jestliže manažér nejedná správně v tomto typu situací. Před předáním případu jsou představeni zaměstnanci. Testovaný může získat více informací o všech, klikne-li myší na jejich fotky. Tyto informace zahrnují záznam o práci, produktivitu, postupy v kariéře atd. Po představení všech lidí je prezentováno zadání úlohy. Typickým scénářem úlohy je např.:

> *Stephen zavolal a řekl, že chce s Vámi hovořit, protože se cítí být diskriminován. Každý rok byli dva lidé, kteří byli nejúspěšnější v prodeji, posláni na prestižní konferenci. Stephen říká, že vždycky chtěl jet na tuto konferenci a očekával, že letos konečně pojede, protože byl v prodeji druhý nejlepší. Nicméně Jack, jeho supervizor, se rozhodl poslat Jeannine a Billa. Jeannine měla nejlepší výsledky, ale Bill byl až třetí nejlepší. Stephen naznačil, že si myslí, že nebyl poslán kvůli tomu, žeje nejstarší a půjde za několik let do důchodu.*

Testovaný může nyní dělat rozhovory s některými nebo se všemi zaměstnanci oddělení. Obr. 6-3 znázorňuje seznam otázek, které může svému supervizorovi Jackovi položit. Tento seznam se mění v závislosti na osobě, s kterou je rozhovor veden. Některé otázky jsou vhodné, jiné nikoli, záleží jen na zkoušeném, které zvolí.

Jack

Klikni na otázku, kterou chceš Jackovi položit:

„Jaký důvod jste udal Stephenovi, že jste ho nevybral pro tu konferenci?"
„Řekl Vám Stephen, že se chystá jít za pár let do důchodu?"
„Je Stephen tak starý, aby šel do důchodu?"
„Můžete změnit obraty prodeje tak, aby podporovaly Vaše rozhodnutí?"
„Jaké je srovnání Stephenova prodejního obratu s Jeanniným a Billovým?"
„Proč jste se rozhodl neposlat Stephena na konferenci?"
„Jak jste rozhodoval o tom, kdo pojede na konferenci?"
„Existují jiné záležitosti, které ovlivnily Vaše rozhodnutí neposlat Stephena?"

Zavři

Obr. 6-3 Seznam otázek pro supervizora

Další obrázek (obr. 6-4) ukazuje Jackovu odpověď na první otázku.

„Jaký důvod jste dal Stephenovi, že jste ho nevybral pro tu konferenci?"

Řekl jsem mu, že vyslat Billa by bylo pro oddělení lepší, protože ve firmě zůstane pracovat déle než Stephen.

Zavři

Obr. 6-4 Jackova odpověď

Poté, co zkoušený dokončí dotazování zaměstnanců, může si zvolit, jaké kroky dál podnikne. Seznam kroků pro tento případ je na obr. 6-5. I zde neobsahuje seznam jen vhodné akce.

Jaké kroky z následujících nyní podniknete?

A. Podpořit Jackovo rozhodnutí a poslat Jeannine a Billa.
B. Požádat Stephena, aby odešel do důchodu dříve.
C. Disciplinárně potrestat Jacka.
D. Poslat Stephena na konferenci s Jeannine a Billem.
E. Upozornit Stephena, že není týmový pracovník.

Posuď odpověď/di.

Obr. 6-5 Kroky, které testovaný má možnost zvolit.

Zkoušený již může obdržet feedback k vybraným činům. Obr. 6-6 až 6-8 demonstrují různé stupně poskytované zpětné vazby. U zaměstnanců, které nebyli dotázáni, není poskytnuta žádná zpětná vazba. Tato informace záměrně chybí, aby nebylo zřejmé, jak případ příště dokončit. V našem příkladu udělal zkoušený nějaké nevhodné volby a neukončil test uspokojivě (viz obr. 6-8).

Zpětná vazba k vybraným krokům:

A. *Požádat Stephena, aby odešel do důchodu dříve.*
To je nevhodný krok, protože jestli změníte délku, podmínky Stephenova zaměstnání, protože on vznesl stížnost týkající se diskriminace, mohl by vznést stížnost ohledně nezákonného protiopatření.

B. *Disciplinárně potrestat Jacka.*
To je vhodný krok, protože Jack věděl, že obrat z prodeje je jediným kritériem pro výběr k účasti na konferenci. Nicméně když nevybral Stephena, odchýlil se od tohoto postupu. A navíc naznačil, že Stephenův plán odejít do důchodu byl motivačním faktorem poslat Billa. Vzhledem k tomu, že Jack svými činy nedodržel zásady firmy proti věkové diskriminaci a vystavil ji tak riziku, učinit disciplinární kroky proti němu je vhodné.

Zavři

Obr. 6-6 Zpětná vazba případu

Hovořil jste s následujícími osobami:

Jack	Bylo vhodné mluvit s Jackem.
Jeannine	Nebylo to vhodné mluvit s Jeannine.
Bill	Nebylo to vhodné mluvit s Billem.

Zavři

Obr. 6-7 Zpětná vazba k rozhovorům se zaměstnanci

Vámi položené otázky nebyly všechny vhodné. Níže jsou uvedeny ty, které jste položil spolu se zpětnou vazbou k jejich přiměřenosti.

Otázky pro Jacka:

„Jaký důvod jste dal Stephenovi, že jste ho nevybral pro tu konferenci?"
To je vhodná otázka. Jackova odpověď Vám pomůže potvrdit informaci, kterou Vám dal Stephen.

„Je Stephen tak starý, aby šel do důchodu?"
To je nepřiměřená otázka. Otázka je pro Vás ohledně toho, zda Jack udelal vhodné rozhodnutí, nepodstatná. Otázky tohoto typu jsou v rozporu se zásadami firmy o posuzování rozhodnutí zaměstnanců výhradně podle legitimních kriterii.

Více

Obr. 6-8 Zpětná vazba k položeným otázkám

Účelem tohoto simulačního počítačového testu je zaručit, že manažeři budou dobře připraveni na spravedlivé jednání se zaměstnanci, dříve než budou přijati a zúčastní se workshopu pro manažery.

Zabezpečení testu

Při použití počítačového testu je nezbytné zajistit jeho bezpečnost. Dále jsou popsány dva aspekty bezpečnosti, jednak testu samotného, jednak přenosu testů přes web, intranet či sítě LAN (lokální síť) a WAN (síť pro dálkový přenos dat).

Bezpečnost testu

Čím je test důležitější, tím lepší musí být jeho zabezpečení. Jedná se především o minimalizování možnosti úniku obsahu testu na veřejnost. Je třeba zabránit testovaným, aby si cokoli během testu zaznamenávali. Potom mohou případně zveřejnit pouze ty informace o testu, které si zapamatovali. Dále je zapotřebí připravit tak rozsáhlou banku úloh, aby procento shodných otázek ve dvou různých testech bylo zanedbatelné, což u počítačového testování je obvyklé. Také je vhodné ve vztahu k bezpečnosti nepoužívat stejné testové otázky více než jednou. Další bezpečnostní opatření se týkají fyzické bezpečnosti testu v tištěné podobě. Testy musejí být uloženy natolik bezpečně, aby nedošlo k úniku informací mimo pověřené osoby. Pokud jsou testy uloženy v elektronické podobě, je vhodné je zabezpečit heslem a vhodným šifrováním a pořídit jejich zálohu. Je třeba věnovat zvýšenou pozornost útoku hackerů, tedy interně i externě zabezpečit počítačovou síť. V některých případech mohou být v testovacích místnostech umístěny kamery a citlivé mikrofony, které snižují možnost podvádění při samostatném testu. Je třeba mít na zřeteli, že u elektronických testů mohou být škody způsobené prolomením ochrany testovacích programů či dat mnohem vážnější než při zcizení jedné kopie testu v tištěné podobě.

Zabezpečení sítě, v které probíhá testování

V rámci veřejné internetové sítě není snadné zajistit bezpečnou komunikaci. Ani nákladná bezpečnostní opatření zahrnující firewalls (kombinace hardware a software vytvořených k zabránění neautorizovaného přístupu) nezaručují stoprocentní ochranu. Doporučuje se tedy zpřístupnit testy on-line pouze pro nezbytnou dobu a mimo ni je uchovávat off-line. Off-line se rozumí buď na počítači nepřipojeném k internetu nebo na záznamovém médiu, které lze z počítače vyjmout (CD-ROM, USB disk, diskety, externí harddisk aj.). I v případě, že je databáze testů dobře zabezpečena, je třeba řešit identifikaci testovaných, zejména pokud se k testu mohou testovaní připojit z domova. Bohužel nejspolehlivější metody jako např. analýza hlasu jsou také nejnákladnější. Nejbezpečnější metoda ať už pro testování na webu či formou tužka-papír je přítomnost proktora pro každého testovaného, který dohlíží na průběh testování. Je třeba zvážit výhody zlepšení bezpečnosti oproti nákladům. Slabinou použití webu pro testování je to, že je prakticky nemožné zabránit někomu, aby si cokoli z monitoru vytiskl. Je možné vybírat otázky z velké banky úloh, měnit pořadí správné odpovědi u MC úloh či generovat početní úlohy dynamickým způsobem. Přesto však testování přes internet nelze u důležitých testů doporučit.

Zabezpečení ochrany osobních údajů

Nejdůležitější, co je třeba zabezpečit, je soukromí (privacy) a bezúhonnost (integrity) testových dat. Soukromí dat znamená, že výsledky testů se spolu s osobními údaji o testovaném nedostanou do nepovolaných rukou. Bezúhonností dat se rozumí, že výsledky v testu jedné osoby nebudou zaměněny s jinými či nedojde k jejich ztrátě. Je tedy zapotřebí přiřadit každému testu i každé osobě identifikační kód a provádět bezpečnostní zálohování veškerých důležitých dat.

Výhody počítačového testování

Běžné počítačové testování umožňuje *větší standardizaci*. Časování zobrazování úloh na počítači umožňuje precizní kontrolu nad tím, co testovaný vidí a slyší. Administrace testu na počítači umožňuje novými způsoby, které nejsou u manuálně administrovaných testů možné, standardizovat podmínky, pokyny a postupy administrace.

Počítačové testování také poskytuje *větší bezpečnost při testování*. Neexistují žádné papírové kopie nebo klíče správných odpovědí, které by někdo mohl odcizit či okopírovat nebo jinak zneužít. Testy administrované na počítači mohou zahrnovat několik úrovní bezpečnostních opatření (heslo, kódování aj.), aby bylo cizím osobám zabráněno v přístupu k testovým materiálům, bankám úloh či klíčům správných odpovědí. Pořadí testových úloh může být také náhodně změněno, je-li to požadováno, aby student neměl potřebu sledovat monitor jiného testovaného.

Podoba testu v tištěné podobě má své silné stránky, ale i určitá omezení. Může snadno zobrazovat text a obrázky, s většími náklady i fotografie. V rámci tištěných testů není však možné provádět časování (timing), změny pořadí zobrazení úloh v testu, animaci či pohyb. Tištěnou stránku nahrazuje v počítačovém formátu zobrazení na monitoru. Nabízí *větší variabilitu zobrazení* testu. Počítač oproti tomu umožňuje použití *nových typů úloh* založených např. na animacích, komplexní grafice, zvuku.

Počítačové verze běžných standardizovaných testů nabízejí významnou *úsporu času při administraci i vyplňování testu.* Vyplnění správných odpovědí na úlohy testu v počítači vyžaduje méně času ve srovnání se skenovatelnými papírovými záznamovými archy. Počítače zjednodušují respondentům zadávání opovědí na některé typy úloh jako je např. přiřazování slov textu či označení obrázku (myší). V budoucnu bude jednou zřejmě možné vyplňovat test i ústně a komunikovat s počítačem pouze pomocí mikrofonu.

Počítačově zadávané testy umožňují nejen rychlejší zpracování, ale také eliminují některé tradiční typy chyb jako např. špatné přiřazení odpovědi k otázce. Zobrazení pouze jediné otázky v daný moment namísto celého testu, jak je obvyklé u testů papír-tužka, umožňuje studentovi soustředit se pouze na jeden problém. Pomáhá tedy získat přesnější výsledek u studentů, kteří mají potíže s koncentrací nebo čtením. Chyby spojené s vyhodnocováním testů jsou díky počítači téměř odstraněny. Jde zejména o chybné identifikace správných odpovědí ať už selháním lidského faktoru nebo chybou vzniklou při skenování záznamových archů. Pokud by přesto došlo k chybě, např. chybnému zadání klíče správných odpovědí do počítače, je velmi snadné tuto chybu dodatečně odstranit a přepočítat výsledky v testu. Na druhé straně je množství vizuální informace dostupné žákovi v jednom okamžiku na monitoru obvykle limitováno zobrazovacími možnostmi monitoru. To by mohl být problém při práci s delšími pasážemi textu, které není možné zobrazit na jednu obrazovku, a tím jsou testovány do jisté míry i paměťové schopnosti studenta, pokud text neumožňuje scrolování. Měření času potřebného na jednotlivé úlohy nabízí dodatečné informace, které je možné analyzovat, např. jak rychle žák čte, jak rychle analyzuje komplexní úlohy či grafické a zvukové informace.

Použití počítačů eliminovalo chyby také při skórování testových výsledků, umožnilo snadné získání dílčích skórů v testu, které mají také svou vypovídající hodnotu. Čas potřebný pro zpracování dat a přípravu výsledků klesl z dnů až týdnů na řádově minuty. To výrazně zvýšilo praktický přínos testu pro hodnocení procesu výuky. Elektronická verze výsledků zjednodušuje jejich přenos a archivaci. Výsledky v testu lze tak s téměř nulovou chybovostí a nízkými náklady přenést do centrálního počítače (úložiště), kde jsou k dispozici pro položkovou analýzu a jiné analýzy či čistě archivní účely.

Počítačové testování také umožňuje snadné vyhodnocování rozdílů mezi skupinami studentů v závislosti na typu školy, pohlaví, sociálnímu zázemí atd.

Nevýhody počítačového testování

Nevýhodou počítačového testu je to, že úlohy musejí být typu multiple-choice, přiřazovací či uzavřené se stručnou tvořenou odpovědí, aby mohly být počítačem skorovány.

Asi největší nevýhodou jsou logistické problémy – nedostatek počítačů a velké finanční náklady. Počítačové testy jsou nákladnější jak pro tvůrce a zadavatele testu, tak pro testované (zpravidla o polovinu) než jejich papírová verze.

Některé studie také ukazují, že čtení z monitoru počítače trvá déle než z tištěného materiálu a kromě toho je pro studenty obtížnější najít v textu na obrazovce chyby (např. Bugbee; Bernt 1990).

Současný stav využívání počítačů při testování

Na rozdíl od dnešní situace ve vyspělých zemích světa, kde se různé typy testů včetně těch počítačových používají velmi často a kde existuje bohatá nabídka testů jak od státních, tak soukromých institucí, v České a Slovenské republice není testování, natož počítačové na profesionální úrovni zcela běžné, i když jisté kroky již v tomto směru byly učiněny (vznik instituce *CERMAT* a soukromé firmy *Scio* a EXAM zabývající se testováním). Chybí zde vysokoškolské instituce, které by se specializovaly na pedagogické měření a instituce, které by ve velkém tvořily profesionální školské testy podle potřeb škol.

Také je nedostatek odborných knih a časopisů v češtině a slovenštině zabývajících se teorií a praxí pedagogického měření. Většina dostupné literatury týkající se vývoje v testování včetně současných trendů je v angličtině. Odborníků, kteří ovládají teorii testu (alespoň klasickou, když už ne moderní IRT), není u nás stále mnoho. Navíc chybí peníze, za které by příslušné instituce mohly vyvíjet testy na profesionální úrovni a za které by školy mohly od nich testy kupovat. Oproti tomu v USA se tvorbou testových úloh zabývají velká testovací centra a soukromé firmy, které do vývoje úloh investují nemalé částky. Haladyna (2004) uvádí, že vývoj jedné kvalitní úlohy stojí cca. 1000 dolarů.

Největšími světovými organizacemi, zabývajícími výzkumem, ale i tvorbou a distribucí širokého spektra testů (často i počítačových) pro potřeby školské praxe, jsou Educational Testing Service (ETS) v USA, NFER ve Velké Británii, CITO v Holandsku a ACER v Austrálii. Přehled nejvýznamnějších organizací a firem zabývajících se testováním (zčásti i počítačovým) uvádíme v příloze 1. Počítačové testy stále ještě nenabízí mnoho společností, protože jejich vývoj, především potom údržba mnohonásobně rozsáhlejší banky úloh než v případě papírové verze testu je velmi finančně náročná. I pro testované je počítačový test dražší než papírový. ETS, například, představila v roce 1992 počítačovou verzi testu Graduate Record Exam (GRE) a začal pracovat na počítačové verzi testu SAT, která, pokud je nám známo, nebyla dosud uvedena do praxe. V současné době je za tvorbu a administraci testu SAT zodpovědná jak ETS, tak jiná významná organizace College Board. V lednu 2001 uskutečnila College Board pokus a zadala na 20 vybraných středních školách v USA počítačovou verzi testu SAT. Cílem bylo zjistit, zda by bylo možné administrovat test počítačově v budoucnu pro všechny studenty, kteří by o něj projevili zájem (test byl ovladatelný pouze myší). Počítačové testy jsou dnes součástí ETS nové generace učitelských testů (Praxis Series, National Council Licensure Examination pro dětské sestry) dostupné pouze v počítačové verzi. Přehled nejvýznamnějších dnes dostupných počítačových testů, z nichž některé ještě existují paralelně i v papírové verzi, je v příloze 2.

Oproti situaci u nás, zaujímá v zahraničí teorie pedagogického měření významné místo mezi moderními pedagogicko-psychologickými disciplínami. Každá větší fakulta humanitního zaměření má svůj *Department of Educational Evaluation and Measurement* (Burjan 1999). Vychází velké množství odborných knih a časopisů (např. *Journal of Educational Measurement* (*JEM*)[42], *Practical Assessment, Research and Evaluation*[43]; *Journal of Technology, Learning, and Assessment*[44] či *Journal of Educational Computing Research*[45]), které se věnují problematice testů a každý rok se koná řada odborných konferencí a seminářů věnovaných testováním (pořádaných např. organizacemi AERA, NCME či AEA Europe). Českých a slovenských organizací zabývajících se na profesionální úrovni teorií a praxí pedagogického měření a školskými testy existuje jen velmi

[42] *JEM* vychází čtvrtletně a věnuje se problematice testování, hodnocení a měření výsledků vzdělávání. Je dostupný na: <http://eris.knue.ac.kr/e-jem/jem.htm>

[43] Jde o americký prakticky orientovaný elektronicky časopis věnovaný problematice testování, hodnocení, měření výsledků vzdělávání. Je dostupný na: < www.ericae.net/pare>

[44] Další americký elektronický časopis dostupný na: <www.jtla.org>.

[45] *Journal of Educational Computing Research* je časopis věnovaný testování pomocí počítače (návrh a vývoj hardware a software pro použití ve vzdělávání), počítačem podporovanou výukou a výzkumem v této oblasti. Je dostupný na: <http://baywood.com/journals/>

málo. Na Slovensku je od roku 1994 činná soukromá firma *EXAM*, v ČR působí od roku 1996 soukromá společnost *Scio*, s.r.o., od roku 2002 soukromá firma *Centrum moderního vzdělávání* (*Centre for Modern Education*)[46] a od roku 1999 státní instituce *CERMAT*[47].

Za zmínku stojí ještě dvě americké firmy, které se zabývají mino jiné vývojem software pro tvorbu, administraci, skórování a analýzu počítačových testů, a některé internetové stránky, které se věnují problematice testování pomocí počítače. V prvním případě jde o velkou firmu *Assessment Systems Corporation* (viz příloha 1) specializující se mimo jiné na vývoj software pro tvorbu testů, bank úloh, elektronické testování, adaptivní testování a software pro analýzu výsledků testů a firmu *Question Mark*, výrobce software pro počítačové testování žáků (viz www.questionmark.com). Rozsáhlý zdroj literatury o problematice testování najdeme například na internetových stránkách americké společnosti ERIC (Clearinghouse on Assessment and Testing, viz http://ericae.net/).

Příklady současného využívání počítačů při testování v ČR

Možností využívání počítačů k testování je u nás především testování pomocí CD-ROM či webového rozhraní (internetu). Uvádíme některé zajímavé příklady.

Testování počítačové gramotnosti pomocí hry CASTLE QUEST[48]

Castle Quest je počítačová hra určená k testování počítačové gramotnosti. Toto testování slouží k ověření přínosu projektu *Internetové kluby ČH@VE* pro děti, které navštěvují kluby ČH@VE. Vstupní testování bylo zahájeno v listopadu 2006, srovnávací test proběhne na konci projektu. Testování dosud nebylo dokončeno a vyhodnoceno.

Projekt vede společnost Erudis, o.p.s., financován je z Evropského sociálního fondu, státního rozpočtu ČR a rozpočtu hl. m. Prahy. Projekt podporuje rozvoj informační a počítačové gramotnosti u dětí 2. stupně ZŠ, které často mají ztížený přístup k moderním informačním technologiím, zejména počítači a internetu. Projekt dále podporuje rozvoj moderních forem vzdělávání s využitím ICT(e-learning).

Pro účely testování bylo vytvořeno celkem šest testů ve třech úrovních obtížnosti pro běžné nebo speciální základní školy. Každý test má 30 otázek. Před začátkem testování jsou děti rozděleny do skupin podle jejich aktuální úrovně počítačové gramotnosti (nejnižší, střední, nejvyšší). Lektor poté každému žáku dané skupiny přidělí unikátní kód ze seznamu kódů odpovídající úrovně obtížnosti. Kód tedy slouží nejen pro vstup do hry, ale také k rozpoznání typu a úrovně testu, kterým žák prošel.

Vývoj testovací aplikace (počítačové hry) bude dál pokračovat. Je plánováno vytvoření webového rozhraní, pomocí kterého si každý učitel bude moci vytvářet vlastní testy, a to pro jakýkoliv předmět. Vylepšovány budou také grafické prvky hry, aby byla pro děti zajímavější a dobrodružnější.

Jiným příkladem využívání počítačů k testování jsou LMS systémy (Learning Management Systems) na elektronickou podporu výuky. K nejčastěji užívaným LMS systémům u nás patří Moodle, eDoceo, WebCT či Microsoft Class Server.

e-learning – systém Moodle na českých VŠ

Moodle[49] je softwarový balík určený pro podporu prezenční i distanční výuky prostřednictvím online kurzů dostupných na WWW. Systém umožňuje či podporuje snadnou publikaci studijních materiálů, zakládání diskusních fór, sběr a hodnocení elektronicky odevzdávaných úkolů, tvorbu

[46] Tato firma vytvořila např. portál *Škola za školou*.
[47] Dřívější *Centrum pro reformu maturitní zkoušky*.
[48] Více na http://www.internetoveklubv.cz/.
[49] Více na webových stránkách http://moodle.cz/ nebo http://moodle.org.

online testů a řadu dalších činností sloužících pro podporu výuky. Moodle je software volně šiřitelný na základě GNU licence s otevřeným PHP kódem. Běží na každém operačním systému, který podporuje PHP (Unix, Linux, Windows, Mac OS X, Netware). Všechna data jsou ukládána v jediné databázi.

Systém Moodle se úspěšně prosazuje na řadě vysokých školách (např. na UK v Praze, ČVUT) v ČR i ve světě. Britská Open University se například rozhodla vybudovat rozsáhlý systém kurzů s využitím systému Moodle. Inovace vyvinuté v rámci projektu budou dostupné celé komunitě uživatelů tohoto systému. V ČR se o lokalizaci české verze Moodle stará David Mudrák z Katedry informačních technologií a technické výchovy Pedagogické fakulty Univerzity Karlovy v Praze.

Na Karlově univerzitě se systém Moodle systematicky využívá již několik let a v současné době obsluhuje několik desítek výukových kurzů. Nelze říci, že by se zde *e*-learningové kurzy staly masovou záležitostí, nicméně již nyní existují takřka v rámci každé fakulty a nejde zdaleka jen o kurzy úzce spjaté s ICT. Namátkou vybíráme ze současné nabídky kurzů:

- Pedagogické aspekty *e*-learningu
- Lineární algebra a geometrie
- Základy molekulárně-genetické diagnostiky
- Kapitoly z řecké rétoriky a poetiky

I studentům a zaměstnancům ČVUT jsou *e*-learningové kurzy a studijní materiály zpřístupněné prostřednictvím Moodle, některé z nich jsou však dostupné i pro anonymní uživatele (např. studijní materiály pro kurzy matematiky)[50]. Nástroje kurzu poskytují rozsáhlé možnosti sledování a zaznamenávání činnosti uživatelů a obsah kurzů je možné snadno přenášet na jiné servery se systémem Moodle.

UP Olomouc se vydala cestou implementace vlastního LMS systému[51] „na míru" potřebám distanční vzdělávání na UP. Systém UNIFOR je přístupný prostřednictvím webového rozhraní, je spravován firmou www.net-university.cz. V současnosti je tento systém využíván k distančnímu studiu v univerzitním prostředí, pro další vzdělávání již aprobovaných učitelů, ale i pro studium pracovníků mateřských škol.

LMS systémy se využívají i na českých středních školách. Například Střední odborná škola a Střední odborné učiliště strojírenské a elektrotechnické v Brně [52]využívají počítačový systém EDUBASE pro podporu výuky a testování. Primárním cílem bylo zvýšit úroveň a kvalitu vazby „výuka –ověřování znalostí" spojením výukových materiálů s elektronickým testováním znalostí žáků v jediném systému. Škola hodlá využít zvýšit přirozenou soutěživost žáků v úrovni znalostí ve všech vyučovaných předmětech a očekává zefektivnění výuky pomocí dobře organizovaných materiálů pro podporu výuky v elektronické formě.

Na počítačem administrované testování přechází v souladu s rozhodnutím IAA (Institute of Internal Auditors) Český institut interních auditorů[53]. Od února 2008 začne pořádat zkoušky *Certifikovaný interní Auditor* (CIA ®) a další specializované zkoušky prostřednictvím počítačových testů. Zkoušky bude celosvětově provádět společnost Pearson VUE.

Využívání počítačů k analýze výsledků testů v ČR a SR

U nás v současné době využívá počítač k analýze výsledků testů pouze soukromá firma *Scio* a státní instituce *CERMAT*. Na Slovensku se jedná o soukromou firmu *EXAM*. Kromě toho nabízí *Scio* i on-line testování. Následuje jejich stručný popis.

---

[50] Více na http://ocw.cvut.cz/moodle/course/view.php?id=27.
[51] Více na http://elearning.upol.cz, http://elearning.upol.cz/unifor.html.
[52] Více na http://www.sos-soubrno.cz/download/edubase.pdf.
[53] Více na http://www.ciia.cz/.

Centrum pro zjišťování výsledků výuky (CERMAT)

CERMAT založily v roce 1999 tři resortní ústavy: Ústav pro informace ve vzdělávání (ÚIV), Výzkumný ústav pedagogický (VÚP) a Národní ústav odborného vzdělávání (NÚOV) za účelem přípravy společného základu maturitní zkoušky po obsahové a organizační stránce (s pověřením MŠMT). CERMAT se zabývá výhradně tvorbou testů zadávaných formou tužka a papír. Odpovědi zapisují studenti do záznamových archů. Ty se dále předávají externí formě na optické snímání dat pomocí scannerů. Analýza výsledků testů se provádí počítačově s pomocí programu RESTAN založeném na klasické teorii testu.

Scio

Scio je soukromá organizace (www.scio.cz s.r.o.), která se transformovala od školního roku 2004/05 ze Scio, o.p.s. (od roku 2001), původní Scio nadace (od roku 1996). Hlavním motivem vzniku bylo zefektivnění a zkvalitnění školního i individuálního vzdělávání založené na využití moderních technologií, zejména internetu. Pomáhá školám všech typů s přijímacími zkouškami a s využitím moderních technologií (zejména internetu) ve výuce. Zabývá se tvorbou i vyhodnocováním testů. Analýzu výsledků testů provádí nejprve při pilotáži (pretesting) a na jejím základě sestavují konečné verze testu. Další analýzy (jak celkové statistiky, tak položkovou analýzu) dělají po každém testování a jejich výsledky archivují společně s testy. K analýze používají vlastní programy, které vyvinuli v programu R a v Accessu po vzoru amerického programu Testan. Aktuálně také testují programy pro analýzu založenou na IRT – MULTILOG a PARSCALE a uvažují o jejich pořízení. Testy zatím s výjimkou on-line testování zadávají ve formě tužka-papír. Pro snímání dat ze záznamových archů používají optický scanner a OMR software TestChecker a TC Verify. Firma nabízí také *on-line testy*, avšak v podobě klasických tištěných testů, které jsou v elektronické verzi na webu a po vyplnění odpovědí se vyhodnocuje skór a percentil podle percentilových tabulek. Takto si mohou své znalosti otestovat žáci 5. a 9. tříd ZŠ v předmětech český jazyk a matematika a v testu Obecné studijní předpoklady, studenti SŠ potom v angličtině, němčině, biologii, českém jazyce a literatuře, dějepisu, fyzice, matematice, chemii, základech společenských věd a v testu Obecné studijní předpoklady. Adaptivní testování zatím nenabízejí.

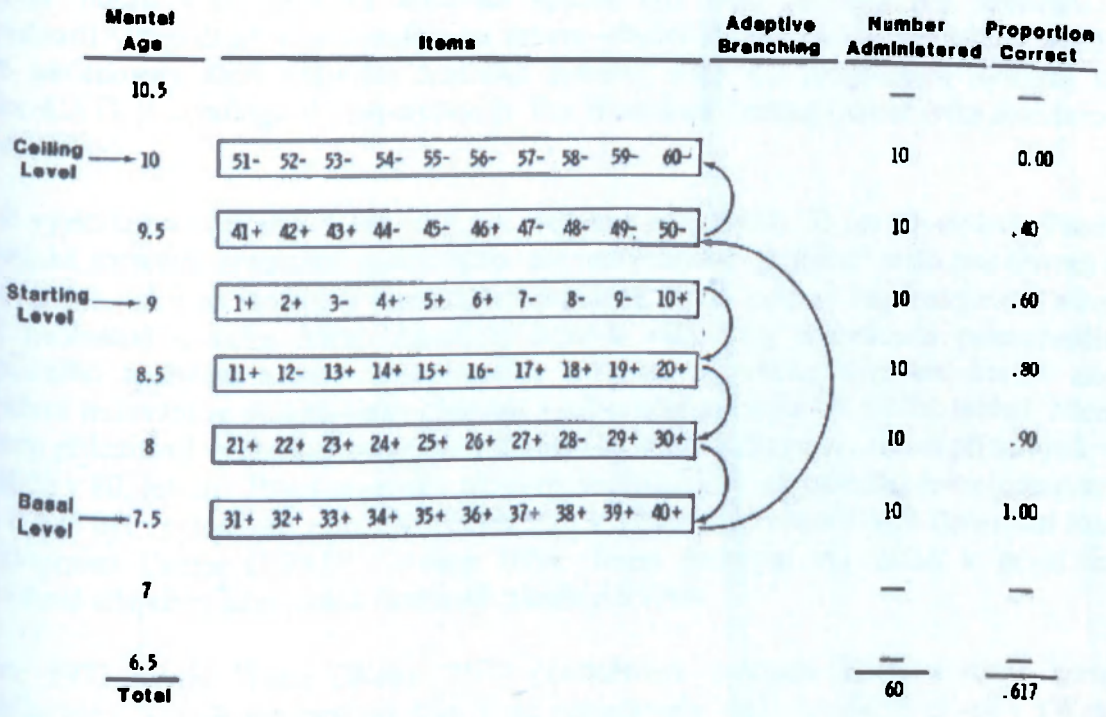
EXAM

Soukromá firma EXAM působí jako jediná svého druhu na Slovensku od roku 1994. Systematicky se na profesionální úrovni věnuje problematice školských testů. Testy se zadávají formou tužka-papír. Testovaní zapisují odpovědi na testové úlohy do záznamových archů, které jsou následně snímány jedním z průmyslových scannerů FUJITSU (např. 3093 GX a FI-4340C), které firma vlastní. Tyto skenery vytvářejí obrazové soubory ve formátu tiff. Ty jsou dále zpracovávány pomocí speciálních OMR programů (Form Reader od ruské firmy ABBYY a Tests Checker od české firmy Lightcomp). Kromě toho používá EXAM ruční čtečky čárových kódů. Výstupem OMR software jsou datové soubory, které jsou následně zpracovávány v různých programech na analýzu výsledků testů (*Reagan*, který si naprogramovali sami, a americký komerční software Testan). K analýze výsledků testů používají výhradně klasickou teorii testu (KTT). Testování na počítači zatím EXAM neposkytuje, ale do budoucna se počítá s on-line testováním.

7 Adaptivní testování

Adaptivní testování je testovací metodologie, která zrovna tak jako například simulace, vyžaduje počítač pro své zadávání, i když adaptivní přístupy, dvou- (nejjednodušší a nejstarší) a víceúrovňové (fixní větvené modely - pyramidové, skokové a stratifikované) existovaly již před nástupem počítačů (Jelínek, Květoň, Denglerová, 2006; Weiss 1973). Adaptivní testování se objevilo již na počátku 20. stol.

První adaptivní IQ test (tzv. Binet IQ test, později zvaný jako Stanford-Binet IQ test) vytvořil Alfred Binet (Binet & Simon, 1905). Test se používá v moderní verzi dodnes. Binetův test se skládal ze setu testových úloh seřazených podle chronologického věku. Administrace tohoto testu byla zcela adaptivní. Binet zařadil úlohy pro určitou věkovou mentální úroveň, pokud cca. 50% dětí daného věku odpověděly úlohu správně. V původní verzi zahrnoval test 9 věkových úrovní (od 3 do 11 let). Tyto úlohy tvořily Binetovu banku úloh pro adaptivní test. Úlohy byly zadávány po deseti individuálně školeným psychologem, který okamžitě odpovědi vyhodnocoval, a podle nich zadával testovanému úlohy vyšší (když odpověděl většinu z nich správně) či nižší věkové úrovně (když většinu chybně). Testování bylo ukončeno, pokud byly u testovaného identifikovány jak základní („basal"), tak stropní (horní, „ceiling") věková úroveň. Stropní úroveň definoval Binet jako věkovou úroveň, na které testovaný nezodpoví ani jednu úlohu správně; základní úroveň naopak jako tu, na které odpoví všechny úlohy správně. Konečný skór testovaného v Binetově testu je založen na podskupině úloh, které zodpověděl správně. Zadávání Binetova testu je ilustrováno na obr. 7-1.



Obr. 7-1 Schéma zadání Binetova testu (Weiss 1973).

Test začal zadáním úloh pro startovací věkovou úroveň 9 (starting level). Testovaný odpověděl úlohy 1, 2, 4, 5, 6 a 10 správně (označeno +) a úlohy 3, 7, 8 a 9 chybně (-). Z 10 zadaných úloh jich odpověděl 6, tedy 60% správně. Protože všechny úlohy nebyly shodně zodpovězeny buď chybně či správně, nedala se tato věková úroveň považovat ani za základní, ani za stropní, a test pokračoval. V tomto okamžiku si mohl testovaný vybrat, zda chce zadat úlohy z vyšší či nižší mentální úrovně. V našem příkladu si testovaný vybral nižší úroveň (zde: úroveň 8,5) pro určení jeho základní

úrovně. Na této úrovni vyřešil testovaný 80% (8 z 10) úloh správně. Dále si vybral věkovou úroveň 8, na které zodpověděl 90% otázek správně. A posléze úroveň 7,5, na které vyřešil všechny úlohy správně a která se tedy stala jeho základní úrovní. Test pokračoval hledáním stropní úrovně. Nejbližší vyšší úroveň jeho startovní byla úroveň 9,5. Testovaný odpověděl z 10 úloh 4 správně (tedy 40%), takže mu byly zadány ještě úlohy z úrovně 10. Tím, že testovaný odpověděl všechny úlohy chybně, byla tato úroveň stanovena jako jeho stropní.

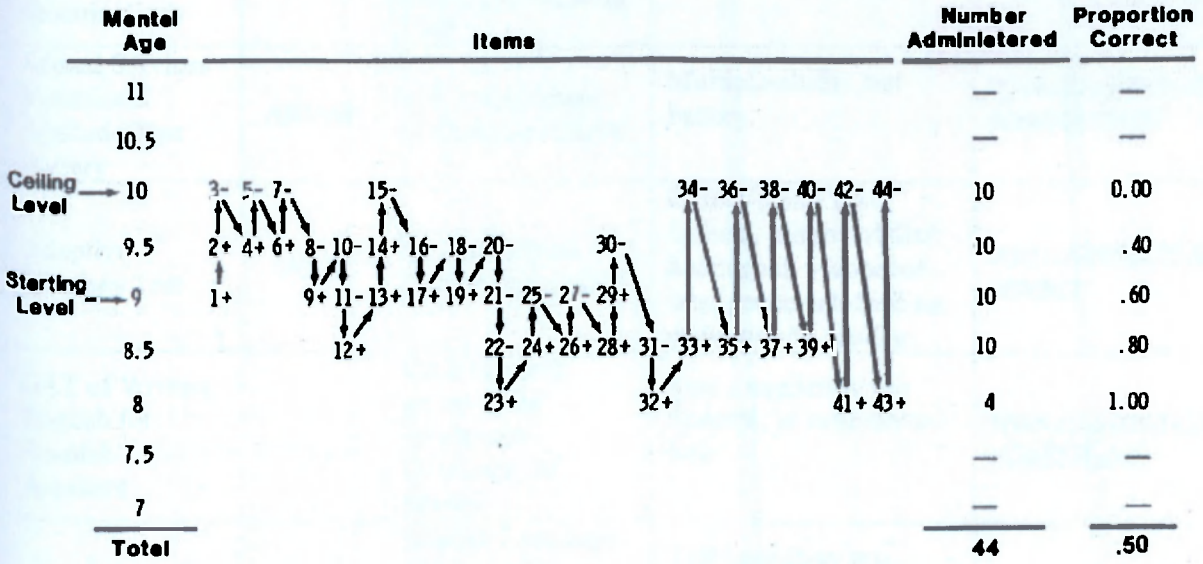
1917 byl vytvořen pro americkou armádu test *Army Alpha* (revize původního testu Examination a, který obsahoval 10 subtestů) o 8 subtestech. Tento test měl již známky podobnosti ke kognitivním částem moderního testu *Armed Services Vocational Aptitude Battery* (ASVAB) používaného v současnosti americkou armádou. Army Alpha a Army Beta (určen pro negramotné a ne anglicky mluvící brance) byly společně prvním širokoplošným testováním IQ (2 milióny mužů byly jím otestovány; Wainer 2000). Testovací vojenské programy se staly rozsáhlejšími za 2. světové války. V roce 1939 následovalo přepracování Army Alpha, tzv. Army General Classification Test (AGCT), test, který obsahoval 4 části. Posledním předchůdcem nynějšího testu ASVAB byl Armed Forces Qualification Test (AFQT).

Paralelně s vojenským testováním probíhal od počátku 20. století vývoj přijímacích testů na americké univerzity, kdy byla založena organizace *College Board*. Úspěch vojenského testovacího programu ovlivnil College Board a ta začala s vývojem testu studijních předpokladů Scholastic Aptitude Test (SAT). 1926 byl SAT, který se skládal z 9 (1928 z 8 a 1929 ze 7) subtestů, poprvé zadán. V roce 1934 se profesor Benjamin Wood z Kolumbijské univerzity spojil s inženýry z IBM, aby společně vytvořili mechanický přístroj na skórování testů. Vynález prvního takového přístroje se připisuje středoškolskému učiteli B. Johnsonovi. Organizace *Educational Testing Service* a *College Board* vyvinuly CAT testovací systém pro IBM počítače pro testování základních dovedností v angličtině a matematice na úrovni střední školy (Abernathy, 1986; Ward aj., 1986). Poté následovaly další adaptivní testovací systémy např. od Assessment Systems Corporation (MicroCAT), Psychological Corporation či The Waterford Testing Center (více Bunderson, Inouye, Olsen, 1989).

Větší výzkum provedl až F. Lord ve 2. pol. 60. let a na počátku 70. let 20. století. Pracoval jak na teoretické struktuře hromadně zadávaného, ale individuálně „ušitého" testu pro úroveň schopností testovaného, tak i na mnohých praktických detailech. První pokusy implementovat adaptivní testy byly neobratné a/ nebo drahé. Americká armáda však brzy rozpoznala potencionální výhody adaptivního testování a podpořila finančně rozsáhlý teoretický výzkum. Studie zaměřené na adaptivní testování se systematicky objevují v odborném tisku od 70. let 20. století. Nicméně první reálnou příležitostí vyzkoušet adaptivní testování byla až dostupnost cenově příznivých výkonných počítačů v 80. letech. První vojenský prototyp počítačového adaptivního testu (computer adaptive test, CAT) byl vyvinut pro Apple III počítače ve výzkumném centru Naval Personnel Research and Development Center (NPRDC) v roce 1984. Tento prototyp byl určen k první širokoplošné počítačové adaptivní administraci subtestů z testu ASVAB.

V roce 1973 navrhl Weiss (Weiss 1973) počítačovou variantu Binetova testu, kterou nazval stratifikovaný nebo-li stradaptivní test. Poté následovaly další stradaptivní testy (Weiss, 1979). Weissův test používal stejnou strukturu banky úloh jako Binetův test. úlohy byly uspořádány po deseti do věkových mentálních úrovní dle obtížnosti (tzv. strata = vrstva). Podobně jako v Binetově testu používá stratifikovaný test proměnlivou startovací úroveň, a tím dovoluje začít na jakékoli úrovni obtížnosti přiměřené každému testovanému. Stratifikovaný test se od Binetova liší v tom, že je zadána vždy pouze jedna úloha a skórována. Další úloha je zadána na základě předchozí odpovědi. Pokud je odpověď správná, je testovanému zadána úloha z následující těžší vrstvy. Když testovaný odpoví chybně, bude mu zadána úloha z nejbližší nižší vrstvy. Tento proces pokračuje tak dlouho, dokud není splněno kritérium k ukončení testu. Test je ukončen, když jsou všechny úlohy nebo pět po sobě následujících úloh v určité vrstvě zodpovězeny chybně. Obr. 7-2 zobrazuje příklad

záznamu odpovědí v stradaptivním testu. V tomto testu byla zvolena za startovací úroveň mentální úroveň 9. První úloha (1) byla zadána a zodpovězena správně (+), a tak další úloha byla zadána z úrovně 9,5. Po správné odpovědi (2+) byla potom zadána úloha z úrovně 10. Protože tato úloha byla zodpovězena chybně (3-), byla následně zadána opět úloha z nejbližší nižší úrovně 9,5, která byla vyřešena správně (4+). Proces pokračoval tímto způsobem až do zadání 31. úlohy. Úloha 30 byla zodpovězena chybně, ale protože z úrovně 9 bylo zadáno již deset úloh, musela být 31. úloha zadána z úrovně nižší, tedy 8, 5. Protože z úrovně 10 byly všechny úlohy zodpovězeny chybně (jako poslední úloha 44), byla mentální úroveň 10 identifikována jako stropní úroveň. Sloupec s hodnotami poměrů správných odpovědí (proportion correct) poukazuje na typické výsledky stratifikovaného testu. Jak očekáváno, tyto poměry se zvyšují (od 0 do 1) se snižující se obtížností úlohy (mentální úrovně). Celkový poměr správných odpovědí je na optimální úrovni roven 0,5.



Obr. 7-2 Příklad záznamu odpovědí v stratifikovaném testu (Weiss 1973)

Adaptivní testování vzniklo pro účely výkonových testů a je v nich také v současnosti nejvíce rozvíjeno, zejména v dichotomně skórovaných testech, i když adaptivní testy se objevují i v psychologii v oblasti diagnostiky (testy osobnosti, Jelínek; Květoň; Denglerová 2006). Adaptivním testováním se zejména v USA zabývá mnoho expertů, např.: Drasgow a Olson-Buchanan (1999), Sands, Waters a McBride (1997), Wainer (2000), Weiss (1983) a mnoho dalších a stále se v této oblasti vedou výzkumy. U nás se adaptivnímu testování věnuje jen několik odborníků (např. Denglerová, Jelínek, Květoň).

V posledních letech se staly počítačové adaptivní testy široce používané, řada testovacích programů v USA, ale i v Evropě je zařadila do své nabídky. Jde např. o počítačovou adaptivní verzi testu *Graduate Management Admission Test* (GMAT), testu studijních předpokladů pro uchazeče o doktorské studium *Graduate Record Examination* (GRE)[54] či o test pro získání licence zdravotní sestry *National Council Licensure Examinations* (NCLEX)[55] používané v USA. Počítačové adaptivní verze některých dalších amerických testů jsou zatím ve výzkumných fázích, např. testy studijních předpokladů *ACT* (American College Testing Program), *SAT* (Scholastic Assessment Test) či test pro udělení lékařské licence *USMLE*^{TM} (Medical Licencing Examination). Výčet některých aktuálně dostupných počítačových adaptivních testů je uveden v tab. 7-1.

[54] V roce 1993 poprvé zveřejnila největší světová testovací organizace ETS počítačovou adaptivní verzi testu GRE. Použití testu GRE v papírové verzi ETS pozvolna redukuje.

[55] Nursing Boards zcela přešla již 1994 od papírové verze testu NCLEX k počítačovému adaptivnímu testu.

Tab. 7-1 Přehled některých počítačových adaptivních testů

název testu	zkratka	kdo ho vytvořil	popis testu	internetový odkaz
Graduate Management Admission Test	GMAT	ETS (USA)	Test pro potřeby Graduate Management Admission Council.	www.review.cz www.mba.com/Takethe GMAT
Graduate Record Examination	GRE	ETS (USA)	Test studijních předpokladů používaný v USA při přijímacím řízení na postgraduální studium.	www.ets.org/portal/site/ets
National Council Licensure Examinations	NCLEX	NCSBN (National Council of State Boards of Nursing; USA)	Test pro udělení licence pro zdravotní sestry.	www.ncsbn.org/nclex
Armed Services Vocational Aptitude Test Battery	ASVAB	U.S. Department of Defence (USA)	Multiple-ability test battery.	www.usmilitary.com/ placementtests
Adaptive Matrices Test	AMT	Dr. Schuhfried GmbH (Rakousko)	Částí Vienna Test System, mimoverbální hodnocení všeobecné inteligence založené na deduktivním úsudku.	www.schuhfried.at/eng/ wts/amt
CAT of Written English for Spanish Speakers		CAT research group at the Autonoma University of Madrid	Test z angličtiny pro Španěle, je zadáván on-line.	www.iic.uam.es/pdfs/ eCatPDF.pdf
Computerized Adaptive Test of English	CATE	English Language and Learning Support of the Information and Learning Resource Services at Middlesex University (Velká Británie)	Test angličtiny pro uchazeče o studium, kteří nemají angličtinu jako rodný jazyk.	www.ilrs.mdx.ac.uk/lang

Počítačové adaptivní testování

Tradiční počítačový (či papírový) test fixní délky předkládá všem testovaným bez rozdílu v jejich výkonnosti stejný počet otázek. Výsledný skór tak závisí obvykle na počtu správně zodpovězených otázek. Čím větší znalosti testovaný má, tím více otázek by měl zodpovědět správně. Pro některé testované jsou některé otázky v testu příliš snadné, jiné příliš obtížné. Správné odpovědi na snadné úlohy, resp. chybné na obtížné úlohy ale nepodávají mnoho informace o jeho znalostech a dovednostech, protože většina testovaných vyřeší tyto úlohy správně, resp. chybně. tak proč tyto úlohy zadávat? Počítačový adaptivní test (CAT) umožňuje testovaným řešit jen úlohy odpovídající jejich schopnostem.

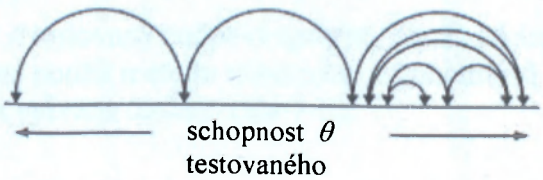
Adaptivní počítačový test (computer adaptive test, CAT) je test, při kterém počítač vybírá úlohy pro testovaného z relativně velké banky úloh podle jeho odpovědi na úlohu předešlou. Pokud testovaný odpoví správně. dostane úlohu obtížnější, pokud chybně, je mu zadána úloha snadnější. Adaptivní testování vyžaduje aparát. který by dovedl smysluplným způsobem popsat úlohy a rozdíly mezi nimi, určit efektivní pravidla pro aktuální výběr úloh k zadání a dospět k výslednému skóru, aniž by byl závislý na konkrétním souboru zadaných úloh (Wainer; Mislevy 2000). Nejvhodnějším matematickým aparátem se ukazuje být teorie odpovědi na položku (IRT, viz kapitola 4), a proto je

na ní také založena většina současných adaptivních testů (např. Goldstein, Wood, 1989; Lord, 1980; Van der Linden, Hambleton, 1997; Wainer, 2000; Embretson, Reise 2000; Baker, Kim 2004). Na tvorbu CAT byly vyvinuty speciální software, např. nejnovější je program FastTEST Professional Testing System Version 2.0 (Fast TEST Pro) z roku 2006 od americké Assessment Systems Corporation (www.assess.com).

Předpokladem počítačového adaptivního testování je dostatečně velká banka úloh, která obsahuje min. 100 různorodých úloh různých obtížností, vytvořená pro danou úroveň schopnosti θ (theta) a danou tématickou oblast. Jednotlivé úlohy banky je zapotřebí kalibrovat, tj. odhadnout pro každou úlohu jednotlivé parametry (obtížnosti, citlivosti) v závislosti na používaném IRT modelu (více viz kapitola 4). Toto odhadování musí probíhat na dostatečně velkém souboru osob, i když charakteristiky úloh nejsou na tomto souboru závislé a měření schopnosti testovaného lze interpretovat i mimo populaci, pro kterou byl test standardizován (Hambleton 1991). Uvažování v rámci IRT je většinou unidimenzionální, proto bývá potřeba při budování banky úloh řešit problém multidimenzionality, např. vyvážením obsahu (content balancing, více např. Kingsbury; Zara 1991, Leung; Chang; Hau 2003) či rozdělením obsahu podle témat do jednotlivých subtestů (multiple scales, více např. Gialluca; Weiss 1979).

Při zadávání adaptivního testu vybírá počítač na základě předem zjištěných parametrů (kalibrací úloh v bance) takové úlohy, které o daném testovaném s určitou odhadovanou úrovní schopnosti θ podávají maximální množství informace. Nejcitlivější úloha rozlišuje mezi jedinci, u kterých se úroveň θ vyskytuje v blízkosti hodnoty obtížnosti dané úlohy.

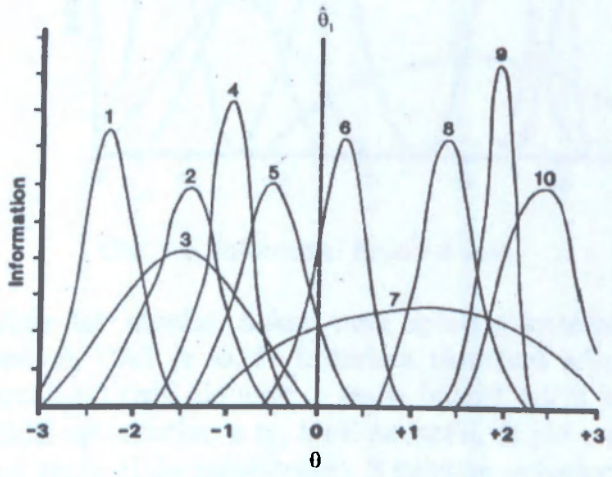
Počítačový adaptivní test založený na IRT pracuje následovně. Cílem testu je zjistit, co testovaný ví o daném tématu. Jinými slovy chceme co možná nejpřesněji odhadnout úroveň jeho schopnosti θ. Nejprve počítač vytvoří počáteční odhad schopnosti θ testovaného, který buď může být pro všechny testované shodný (průměr schopností předešlých testovaných) nebo může být stanoven pro každého testovaného zvlášť na základě nějaké dostupné informace o něm (např. výkon v předchozích testech, známka, informace od učitele). Odpověď testovaného je poté okamžitě skórována a podle množství informace, kterou úloha podává na aktuální úrovni jeho schopnosti θ počítač vybírá (s určitou tolerancí) z banky úloh úlohu s maximálním množstvím informace.[56] Ta je obvykle vybírána podle tzv. pravidla kroku (*step-rule*). Odpoví-li testovaný na první úlohu správně, je původní odhad jeho schopnosti θ zvýšen o určité číslo (často o 0,5 či 1), když chybně, je odhad snížen o stejné číslo. Tento postup se opakuje do té doby, dokud testovaný nezíská vzorek odpovědí (response pattern) skládající se minimálně z jedné chybné a jedné správné odpovědi. Poté se pro výpočet nového odhadu θ, který je založen na všech předchozích odpovědích, použije metoda maximální věrohodnosti[57] (maximum likelihood estimation). Po zadání a skórování každé další úlohy je odhad θ testovaného opět upraven a na jeho základě vybrána další ještě nezadaná úloha, která poskytuje největší informaci. Proces počítačového adaptivního testování (zpravidla konvergentního) znázorňuje obr. 7-3.



Obr. 7-3 Proces výběru úloh (upraveno podle Alessi; Trollip 2001)

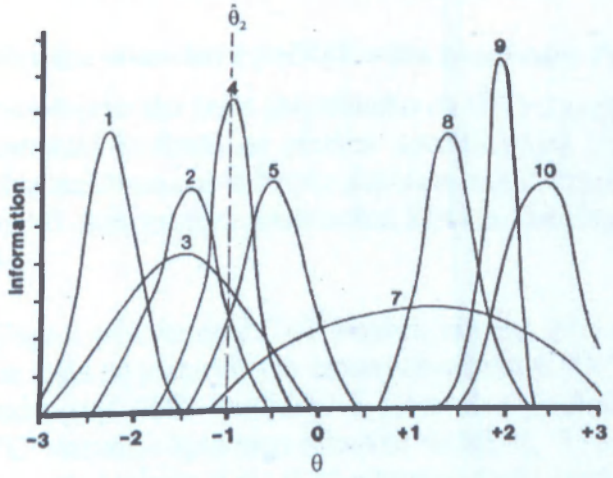
[56] Množství informace se stanovuje pomocí informační funkce z IRT (více v kapitole 4).
[57] Metodou maximální věrohodnosti se odhaduje schopnost jedince jako maximální hodnota určité pravděpodobnostní funkce (Hambleton 1991). Jiná běžně požívaná metoda pro odhad θ testovaného je Bayesova metoda odhadu.

Obrázky 1-3 objasňují výběr úloh podle „maximální informace“ v CAT. Na obr. 7-4 vidíme kromě informačních křivek 10 úloh počáteční odhad schopnosti $\theta = 0$ pro hypotetického testovaného (viz svislá čára). Vodorovná osa je osa schopnosti θ testovaného, svislá osa určuje množství informace. Hodnoty informace jsou vypočteny pro všechny úlohy na této úrovni θ. Z obr. 7-4 je zřejmé, že úloha 6 podává největší množství informace ze všech 10 úloh pro úroveň schopnosti $\theta = 0$ (viz svislá čára). Proto je tato úloha počítačem vybrána, zadána testovanému a poté okamžitě skórována.



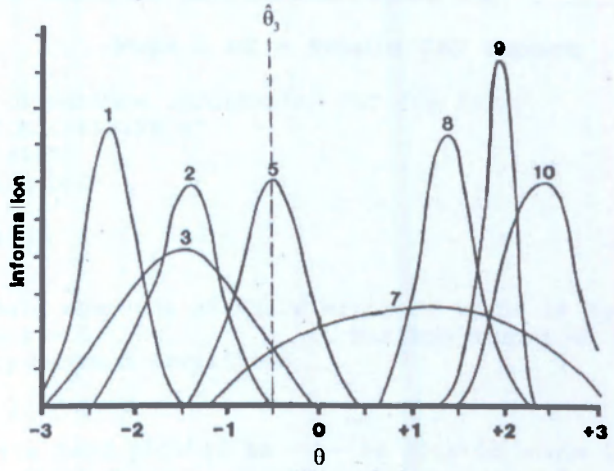
Obr. 7-4 Informační funkce 10 úloh

Na základě tohoto skóru (zde: chybně) je určen nový odhad $\theta = -1$ (zde: použit krok o velikosti 1). Podle množství informace je dále vybrána úloha 4 (obr. 7-5), protože poskytuje pro $\theta = -1$ nejvíce informace, a skórována.



Obr. 7-5 Informační funkce 9 úloh

Za předpokladu, že úlohu 4 testovaný zodpoví správně, čímž získáme vzorek jedné chybné a jedné správné dopovědi, můžeme použít metodu maximální věrohodnosti k dalšímu odhadu θ. Výsledkem je $\theta = -0,5$. Dále tedy byla vybrána úloha 5 (obr.7-6).



Obr. 7-6 Informační funkce 8 úloh

Tento proces pokračuje tak dlouho, dokud není splněno kritérium pro ukončení testu. Jednou z důležitých charakteristik CAT je to, že kritérium ukončení adaptivního testování se může lišit podle různých cílů testování (zda jde např. o testy, jejichž cílem je roztřídit testované na ty, kteří uspěli v testu, tj. zvládli určité učivo a ty, kteří neuspěli, či jde o testy pro výběr testovaných pro přijetí na vyšší stupeň školy či do zaměstnání). S každým odhadem θ je spojena standardní chyba odhadu (standard error of estimate, SEM), protože pokaždé, kdy počítačový program počítá odhad schopnosti θ, je málo pravděpodobné, aby byl tento odhad naprosto přesný. Avšak je možné udat interval, v kterém se odhad bude pohybovat. Tento interval se zmenšuje, je-li zadáno více úloh, což je zřejmé, protože odhad se zpřesní, když se nashromáždí více informací. Program ukončí zadávání úloh v okamžiku, když chyba odhadu je přijatelně malá, tedy když je jisté, že daný odhad je dostatečně blízko k reálné úrovni schopnosti testovaného.

Ukázka výstupu adaptivního testování z počítačového programu FastTEST Pro

Výstup je z počítačového adaptivního testu (založeného na IRT) z programu FastTEST Pro (Weiss 2006) vytvořeného k testování studentů na předem specifikované hladině přesnosti - minimální standardní chyba měření je stanovena na 0,20 pro ukončení testu. Celkem bylo zadáno 40 úloh. Test byl ukončen až po zadání 40 úloh, protože pozorovaná SEM se stabilizovala na hodnotě 0,22 počínaje úlohou 38.

První stránka výstupu (Page 1 of a Sample CAT Report, viz obr. 7-7) ukazuje graf vývoje výkonu testovaného v adaptivním testu po jednotlivých zadaných otázkách. "X" znamená počáteční odhad θ a přerušované čáry představují 95% konfidenční interval (vypočítán jako plus či mínus dvě standardní chyby k θ). "C" označuje správnou odpověď (correct), "I" chybnou odpověď (incorrect). Za počáteční odhad byla vzata hodnota θ = - 0,24 a testová úloha poskytující maximální informaci pro tuto hodnotu θ byla zadána (item 1) a zodpovězena správně. Za následující úlohu byla vybrána ta (item 2), která podávala maximální informaci pro θ = 4 (vysoká hodnota θ byla stanovena k získání co nejrychleji vzorku smíšených odpovědí). Ta byla také vyřešena správně, takže byla zadána další stejně obtížná úloha, která však již byla zodpovězena chybně (item 3). Tím jsme získali vzorek smíšených odpovědí (správně/nesprávně) a od tohoto okamžiku se odhad mohl začít počítat metodou maximální věrohodnosti (θ = 2,52). Úloha poskytující pro θ = 2,52 nejvíce informace (item 4) byla dále zadána a skórována jako správná. Výsledný odhad θ byl tedy zvýšen na θ = 2,77 se standardní chybou 0,68. Z tabulky můžeme vypozorovat, že po správné odpovědi se hodnota odhadu θ zvýší, chybná odpověď naopak znamená snížení hodnoty θ. Rozdíly mezi odhady θ po správných odpovědích se snižují s větším počtem zadaných úloh. Standardní chyba se zpravidla snižuje, další úlohy zlepšují odhad θ. Konečný odhad θ byl θ = 1,55 se standardní chybou SE = 0,22.

Page 1 of a Sample CAT Report

Item-By-Item Report of Maximum Information CAT for file:
C:\RAWFILES\CATSESSION1.PREVIEW_42
Workspace : jbm_122005
Session name : catsession1
Session ID : 7
Module name : cattest1
Module ID : 6

This test will terminate when the standard error of theta is equal to or less than 0.200
Minimum number of items = 5 Maximum number of items = 4
Theta was estimated by maximum likelihood.

Examinee Name : John Q. Public

The standard error band plotted as ---- is plus or minus 2.00 standard errors.
X = Initial theta value C = Correct answer I = Incorrect answer

Item	Theta	SE	-3.....-2.....-1.....0.....+1.....+2.....+3
0	-0.24*	1.00*	----------X----------
1	4.00'	1.00*	. ---------->
2	4.00'	1.00*	. ---------->
3	2.52	0.84	. ----------I-----
4	2.77	0.68	. ----------C---
5	2.38	0.61	. ----------I------
6	2.09	0.61	. ----------I----------
7	1.49	0.89	----------I----------
8	0.36	1.00	----------I----------
9	0.88	0.63	----------C----------
10	1.13	0.56	----------C----------
11	1.34	0.49	. ----------C----------
12	1.44	0.46	. ----------C----------
13	1.55	0.43	. ----------C----------
14	1.67	0.41	. --------C---------
15	1.54	0.38	. --------I---------
16	1.60	0.36	. --------C--------
17	1.70	0.35	. --------C--------
18	1.76	0.34	. --------C--------
19	1.65	0.32	. ------I------
20	1.52	0.31	. -------I------
21	1.40	0.30	. -------I------
22	1.27	0.30	. ------I------
23	1.30	0.28	. ------C-----
24	1.32	0.28	. ------C-----
25	1.36	0.27	. -----C------
26	1.40	0.27	. -----C------
27	1.31	0.26	. ------I-----
28	1.34	0.25	. -----C-----
29	1.37	0.25	. -----C-----
30	1.40	0.24	. ----C-----
31	1.43	0.24	. -----C-----
32	1.46	0.24	. -----C-----
33	1.50	0.24	. ----C-----
34	1.53	0.23	. -----C----
35	1.55	0.23	. -----C-----
36	1.59	0.23	. ----C-----
37	1.62	0.23	. -----C-----
38	1.58	0.22	. ----I-----
39	1.53	0.22	. -----I----
40	1.55	0.22	. ----C--

*Arbitrarily assigned value (libovolně přidělená hodnota)

The final theta estimate based on 40 items was 1.55 with a standard error of 0.22, resulting in a 2.00 standard error band of 1.12 to 1.99
This test was terminated when the maximum number of items was reached.

Obr. 7-7 Část první stránky výstupu z programu Fast TEST Pro

Druhá stránka výstupu (Page 2 of a Sample CAT report, obr. 7-8) obsahuje odpovědi na multiple choice úlohy s 5 nabídkami (answer, od 1 do 5) testovaného, správnou odpověď (corrrect answer), označení (correct?), zda je odpověď správně (Y) či chybně (N), odhad θ (Theta) a její 95% konfidenční interval (SE Band), kompletní identifikátor (Item Identifier) pro každou zadanou úlohu. Konečný odhad θ (Theta) založený na 40 úlohách byl 1,55 se standardní chybou (SE) 0,22 vyplývající z 95% konfidenčního intervalu (SE Band) od 1,12 (-2,0 SE) do 1,99 (+ 2,0 SE). Test byl ukončen po zadání maximálního počtu úloh.

Page 2 of a Sample CAT Report

Item	Answer	Correct Answer	Correct?	Theta	SE	SE Band -2.0 SE	SE Band +2.0 SE	Item Identifier
0				-0.24*	1.00*	-2.24*	1.76*	
1	1	1	Y	4.00*	1.00*	2.00*	6.00*	
ENGLK001|INSTRUCTIONS|OPENING								
2	1	1	Y	4.00*	1.00*	2.00*	6.00*	
ENGLK001|INSTRUCTIONS|ENGLISH1								
3	4	3	N	2.52	0.84	0.84	4.20	
ENGLK001|INSTRUCTIONS|ENGLISH2								
4	1	1	Y	2.77	0.68	1.41	4.12	
ENGLK001|INSTRUCTIONS|ENGLISH3								
5	1	4	N	2.38	0.61	1.16	3.59	ENGLK001|ITEM13
6	1	2	N	2.09	0.61	0.86	3.32	ENGLK001|ITEM184
7	2	4	N	1.49	0.89	-0.29	3.28	ENGLK001|ITEM152
8	4	3	N	0.36	1.00	-1.64	2.36	ENGLK001|ITEM182
9	4	4	Y	0.88	0.63	-0.38	2.13	ENGLK001|ITEM90
10	3	3	Y	1.13	0.56	0.01	2.25	ENGLK001|ITEM287
11	5	5	Y	1.34	0.49	0.35	2.33	ENGLK001|ITEM94
12	1	1	Y	1.44	0.46	0.53	2.36	ENGLK001|ITEM183
13	5	5	Y	1.55	0.43	0.69	2.41	ENGLK001|ITEM51
14	2	2	Y	1.67	0.41	0.85	2.50	ENGLK001|ITEM60
15	3	2	N	1.54	0.38	0.79	2.30	ENGLK001|ITEM98
16	3	3	Y	1.60	0.36	0.88	2.33	ENGLK001|ITEM261
17	2	2	Y	1.70	0.35	1.00	2.40	ENGLK001|ITEM179
18	2	2	Y	1.76	0.34	1.09	2.44	ENGLK001|ITEM86
19	2	3	N	1.65	0.32	1.00	2.29	ENGLK001|ITEM260
20	5	2	N	1.52	0.31	0.90	2.14	ENGLK001|ITEM65
21	1	5	N	1.40	0.30	0.80	2.01	ENGLK001|ITEM69
22	3	1	N	1.27	0.30	0.68	1.86	ENGLK001|ITEM77
23	3	3	Y	1.30	0.28	0.73	1.87	ENGLK001|ITEM97
24	4	4	Y	1.32	0.28	0.77	1.88	ENGLK001|ITEM194
25	2	2	Y	1.36	0.27	0.82	1.90	ENGLK001|ITEM175
26	4	4	Y	1.40	0.27	0.87	1.93	ENGLK001|ITEM278
27	2	5	N	1.31	0.26	0.80	1.83	ENGLK001|ITEM74
28	3	3	Y	1.34	0.25	0.83	1.85	ENGLK001|ITEM66
29	1	1	Y	1.37	0.25	0.87	1.87	ENGLK001|ITEM258
30	2	2	Y	1.40	0.24	0.91	1.89	ENGLK001|ITEM73
31	1	1	Y	1.43	0.24	0.95	1.91	ENGLK001|ITEM277
32	3	3	Y	1.46	0.24	0.98	1.93	ENGLK001|ITEM144
33	2	2	Y	1.50	0.24	1.02	1.97	ENGLK001|ITEM88
34	4	4	Y	1.53	0.23	1.06	1.99	ENGLK001|ITEM151
35	4	4	Y	1.55	0.23	1.09	2.02	ENGLK001|ITEM79
36	1	1	Y	1.59	0.23	1.13	2.05	ENGLK001|ITEM99
37	1	1	Y	1.62	0.23	1.17	2.08	ENGLK001|ITEM76
38	4	5	N	1.58	0.22	1.13	2.02	ENGLK001|ITEM270
39	5	4	N	1.53	0.22	1.10	1.97	ENGLK001|ITEM269
40	1	1	Y	1.55	0.22	1.12	1.99	ENGLK001|ITEM80

Obr. 7-8 Část druhé stránky výstupu z programu Fast TEST Pro

Adaptivní testování má dvě hlavní výhody. Za prvé odhad schopnosti je obvykle přesnější než u tradičních testovacích postupů a za druhé průměrný počet úloh, které musíme zadat, abychom dosáhli odhadu reálné schopnosti testovaného při zachování či dokonce zlepšení přesnosti měření, je výrazně menší (někdy až čtyřikrát, Wainer 2001). Výzkumy ukazují, že počítačové adaptivní testy bývají průměrně až o 50% kratší než jejich klasické verze ve formě tužka-papír (Embretson, Reise 2000). Když jsou žáci velmi dobří, víte již před zadáním testu, že odpoví všechny lehké

otázky správně. Podobně slabí žáci vyřeší všechny těžké úlohy chybně či je nebudou řešit vůbec. To řeší adaptivní test, který zadává pouze ty úlohy nutné k vytvoření odhadu schopnosti testovaného. Velmi dobrému žákovi počítač zadá velmi málo snadných úloh, pokud vůbec nějaké, a testovaný s malou znalostí tématu nikdy neuvidí těžké úlohy. Adaptivní testování je tedy efektivní a kromě toho zůstávají studenti v průběhu celého testování motivováni (dostávají jen úlohy odpovídající jejich úrovni schopnosti). Navíc banka úloh, z které jsou úlohy zadávány, je natolik rozsáhlá, že pouze malá část je nakonec zadána, a tím je zajištěna možnost opakované administrace bez nežádoucího efektu učení se úloh nazpaměť (úlohy se nestanou obecně známými).

Adaptivní testování má samozřejmě i některá omezení. Nejvýznačnějším z nich je to, že adaptivní test založený na IRT může měřit jen jeden typ schopnosti (předpoklad IRT). Tj. není vhodné použít jediný adaptivní test k měření znalostí jak z historie, tak ze zeměpisu či z algebry a současně z geometrie. Chceme-li testovat více jak jednu oblast znalostí (knowledge domain), dva či více adaptivních testů mohou být sloučeny do jediného. Testovací program musí potom ukládat cestu ke všem datům pro každý set úloh zvlášť. Druhým omezením je to, že získat spojitou a přesnou charakteristickou křivku úlohy vyžaduje, aby každá úloha byla zadána několiksetkrát testovaným s velkým rozsahem schopnosti, než může být spolehlivě použita v adaptivním testování. Otázkou také je schopnost adaptivních testů poskytnout zpětnou vazbu. Většina adaptivních testů ukončuje testování ze své podstaty rychle, což znamená, že není zadáno testovanému mnoho úloh. To může vést k příliš včasnému ukončení testování, aniž by byl otestován celý specifický obsah.

8 Závěry k výzkumnému problému A

Výzkumný problém A jsme formulovali:

A Jaký je současný stav využívání počítačů při testování?

Závěry k problému A uvádíme na základě zjištění v kapitolách 4, 6 a 7.

Testování je jednou z důležitých metod hodnocení výsledků výuky. Jelikož výsledky testů používaných při přijímacím řízení (a nejen jich) mohou mít velký vliv na budoucnost testovaného, je důležité navrhovat, administrovat, vyhodnocovat testy a interpretovat jejich výsledky velice pečlivě. Tyto náročné činnosti si můžeme výrazně usnadnit použitím počítačů.

Využívání počítačů k výše jmenovaným činnostem s výjimkou administrace není ve světě a jistým způsobem ani u nás žádnou novinkou. Počítačové testování je však stále ve svém vývoji. Zcela novou etapu výkonových testů, v porovnání s testy zadávanými formou tužka-papír, představují v současné době ve světě testy zadávané počítačem, z nichž některé existují pouze v elektronické podobě, jiné jsou jen alternativou k testům tužka-papír o stejné délce i obtížnosti. Obzvláště efektivní se zdají být v posledních letech počítačové adaptivní testy (computer adaptive tests), kdy počítač vybírá úlohy pro testovaného z relativně velké banky úloh podle jeho odpovědi na úlohy předešlé.

Adaptivní testování má dvě zásadní výhody. Za prvé odhad schopnosti testovaného je obvykle přesnější než u neadaptivního testování a za druhé průměrný počet zadaných úloh, k dosažení odhadu schopnosti testovaného při zachování či dokonce zlepšení přesnosti měření, je výrazně menší. Výzkumy ukazují, že počítačové adaptivní testy bývají průměrně až o 50 procent kratší než jejich „papírové" verze (Embretson, Reise 2000). Adaptivní testování je tedy efektivní a kromě toho zůstávají studenti v průběhu celého testování motivováni, protože dostávají jen úlohy odpovídající jejich úrovni schopnosti (tj. ne pro ně příliš snadné či obtížné). Navíc banka úloh, z které jsou úlohy zadávány, je natolik rozsáhlá, že pouze malá část je nakonec zadána, a tím je zajištěna možnost opakované administrace bez nežádoucího efektu učení se úloh nazpaměť. Samozřejmě má

i adaptivní testování některá omezení. Nejvýznačnějším z nich je to, že adaptivní test založený na IRT může měřit jen jeden typ schopnosti (předpokladem IRT je unidimenzionalita). Z toho plyne, že nemůžeme použít jediný adaptivní test k měření znalostí jak z historie, tak ze zeměpisu či z algebry a současně z geometrie. Druhým značným omezením je to, že získat spojitou a přesnou charakteristickou křivku úlohy vyžaduje, aby každá úloha byla zadána několiksetkrát testovaným s velkým rozsahem schopnosti, než může být spolehlivě použita v adaptivním testování. To je také jeden z důvodů, proč počítačové adaptivní testování není ve světě příliš rozšířeno.

Jiným typem testu, vedle testu adaptivního, který vyžaduje použití počítače, je test simulační. Simulační testování má velký význam, protože člověk není testován jen ze znalostí, ale také z praktických dovedností.

Testy zadávané počítačem se stávají velmi populární a široce akceptované, i když je velmi finančně náročné, proto počítačové testy stále nenabízí mnoho organizací zabývající se problematikou testování. Možnost distribuce zadání a výsledků pomocí lokální sítě, případně internetu nabízí velký potenciál při snižování administrativních nákladů spojených s testováním. Využití internetu však skýtá velká rizika týkající se bezpečnosti testování, a proto se internet k důležitému testování nedoporučuje.

Při tvorbě testovacího programu je důležité nahlížet na testování jako na třífázový proces: na fázi před testem, během testování a na fázi po vyplnění testu. Učitel i testovaní mají v každé fázi odlišné role a požadují různé informace. Testovací program musí mít uživatelsky příjemné ovládání. Pro testované jsou nejdůležitější první dvě fáze. Před testem potřebuje testovaný dostat co nejvíce informací o testovacím programu, a to způsobem, který neposílí ještě více jeho úzkost z testování. Během testu by mělo být testovaným umožněno soustředit se pouze na řešení úloh a nikoli také na postupy obsluhy testovacího programu. Při tvorbě testových úloh, které jsou součástí počítačového testu, musíme dodržovat stejná doporučení kladená na kvalitní úlohy jako v případě tradičních „papírových" testů.

Počítačové testování má oproti testování formou tužka-papír řadu výhod. Běžné počítačové testování umožňuje větší standardizaci. Časování zobrazování úloh na počítači umožňuje precizní kontrolu nad tím, co testovaný vidí a slyší. Administrace testu na počítači umožňuje standardizovat podmínky, pokyny a postupy administrace.

Počítač umožňuje použití nových typů úloh založených např. na animacích, komplexní grafice, zvuku.

Počítačové verze běžných standardizovaných testů nabízejí významnou úsporu času při administraci i vyplňování testu. Vyplnění správných odpovědí na úlohy testu v počítači vyžaduje méně času a redukuje chyby z nepozornosti oproti jejich vyplnění na záznamovém archu. Umožňují také rychlejší zpracování testových výsledků a eliminují některé tradiční typy chyb jako např. špatné přiřazení odpovědi k otázce.

Použití počítačů eliminovalo chyby také při skórování testových výsledků a umožnilo snadné získávání dílčích skórů v testu, které mají také svou vypovídající hodnotu. Čas potřebný pro zpracování dat a přípravu výsledků klesl z dnů až týdnů na řádově minuty. To výrazně zvýšilo praktický přínos testu pro hodnocení procesu výuky. Elektronická verze výsledků zjednodušuje jejich přenos a archivaci.

Zobrazení pouze jediné otázky v daný moment namísto celého testu, jak je obvyklé u testů „papírových", umožňuje studentovi soustředit se pouze na jeden problém, což oceni zejména ti studenti, kteří mají problémy s koncentrací nebo se čtením. Na druhé straně je množství vizuální informace dostupné žákovi v jednom okamžiku na monitoru obvykle limitováno zobrazovacími

možnostmi monitoru. To může být problém při práci s delšími pasážemi textu, které není možné zobrazit na jednu obrazovku. Počítačové testy umožňují tvůrci testu vložit do testu experimentální úlohy určené k pilotáži, které nejsou zahrnuty do skórů testovaných. Počítačové testování také poskytuje větší bezpečnost při testování (znesnadňuje např. opisování, protože testovaní obdrží zcela odlišné varianty testu složené ale ze stejně obtížných úloh).

Nevýhodou skórování testu na počítači je to, že testové úlohy musejí být uzavřené, otevřené se stručnou odpovědí či formy esej. Zatím totiž ve většině případů počítače nejsou schopny vyhodnocovat odpovědi na otevřené úlohy s rozsáhlejším řešením vícestupňově, ty musí podle předem stanoveného předpisu vyhodnocovat kompetentní posuzovatel či posuzovatelé a výsledek posouzení vložit do paměti počítače.

Pokud jde o dostupné počítačové testy ve světě, stále ještě je nenabízí mnoho společností, protože jejich vývoj, především potom údržba mnohonásobně rozsáhlejší banky úloh, která je zapotřebí oproti papírovým verzím testu, je velmi finančně náročná. I pro testované je počítačový test dražší než „papírový". Takové testy nabízí vedle největší světové organizace zabývající se mimo jiné tvorbou testů, Educational Testing Service (test GRE, GMAT, PRAXIS I, TOEFL) také další americké instituce, například College Board (test CLEP), která v roce 2001 zadala na 20 vybraných středních školách v USA počítačovou verzi SAT, nejznámějšího testu studijních předpokladů v USA. Bohužel se nám nepodařilo zjistit, proč počítačové zadávání testu SAT nepokračovala i v následujících letech.

Ve světě, především však v USA existují firmy, které se zabývají vývojem software pro tvorbu testů, bank úloh, elektronické testování, adaptivní testování a software pro analýzu výsledků testů (např. *Assessment Systems Corporation* či *Question Mark*). Kromě jiného téměř veškerá dostupná literatura o využívání počítačů při testování je v angličtině, českých odborných článků je velmi málo.

Pokud jde o možnosti využívání počítačů k testování u nás, jde především o testování pomocí CD-ROM či webového rozhraní (internetu). Zajímavým příkladem je počítačová hra Castle Quest, určená k testování počítačové gramotnosti dětí, či LMS systémy pro elektronickou podporu výuky. K nejčastěji užívaným LMS systémům u nás patří Moodle, eDoceo, WebCT či Microsoft Class Server.

9 Charakteristika aplikace KTT a IRT u testu OSP použitého při přijímacím řízení

Při přijímacím řízení na Pedagogickou fakultu v Praze byl mimo jiné v roce 2006 použit test Obecné studijní předpoklady (OSP) navržený firmou Scio. Test Obecné studijní předpoklady psalo v červnu 2006 v řádném termínu celkem 1279 uchazečů o dvouoborové studium všeobecně vzdělávacích předmětů ve 45 různých kombinacích a 417 uchazečů o jednooborové studium speciální pedagogiky (330) a učitelství pro mateřské školy (87) na Pedagogické fakultě Univerzity Karlovy v Praze. Variantu A psalo 451 uchazečů, variantu B 439, variantu C 467 a variantu D 339 uchazečů.

Charakteristika testu Obecné studijní předpoklady

Test Obecné studijní předpoklady (OSP) je test o pěti variantách (A-E)[58] od společnosti Scio, který byl v červnu 2006 použit při přijímacím řízení na Pedagogické fakultě UK v Praze. Ukázka testu OSP, varianty A je v příloze 3. Test se svými úlohami i způsobem skórování velmi podobá americkému testu studijních předpokladů SAT. Test OSP obsahuje celkem 45 úloh

[58] Variantu E psalo v náhradním termínu jen 73 uchazečů, proto nebyla do analýz zahrnuta.

rozdělených do tří oddílů: verbálního o 17 úlohách, analytického o 13 úlohách a kvantitativního o 15 úlohách. Všechny úlohy jsou svým druhem multiple choice úlohy s pěti nabídkami a právě jednou správnou či nejlepší odpovědí. Nebyly povoleny žádné pomůcky, ale testovaní měli možnost si dělat poznámky na volné archy, které byly součástí testu. Verbální oddíl zahrnuje úlohy na porozumění textu, na vztah mezi slovy, na slova opačného významu. Analytický oddíl vyžaduje od studentů logické uvažování a analýzu textu, kvantitativní oddíl obsahuje matematické úlohy (viz tab. 9-1). Z tabulky je patrné, že úlohy byly zaměřeny většinou na vyšší kognitivní cíle (dle revidované Bloomovy taxonomie, viz kap. 4).

Tab. 9-1 Specifikační tabulka pro test OSP, PedF UK v Praze, červen 2006

Tematický celek/ dílčí témata		Počet úloh						
		Kognitivní cíl dle revidované Bloomovy taxonomie						
		B1	B2	B3	B4	B5	B6	celkem
Verbální oddíl	slovní spojení				9			9
	porozumění textu		3					3
	antonyma				5			5
Analytický oddíl	logický úsudek						7	7
	analýza textu				6			6
Kvantitativní oddíl	aritmetika -slovní úlohy			6				6
	geometrie/ stereometrie			4				4
	algebra			1				1
	pravděpodobnost			1				1
	grafy			3				3
Celkem úloh		0	3	15	20	0	7	45

Na test byl vymezen čas 45 minut, přičemž testovaným bylo dovoleno řešit úlohy v libovolném pořadí. Co se týče skórování, za každou správně vyřešenou úlohu byl započítán 1 bod, za každou nesprávně vyřešenou úlohu se čtvrtina bodu odečítala (studenti však byli instruováni tak, že jim bude část bodu odečtena, ne jak velká, což není správné), vynechaná odpověď se skórovala 0 body. Bodovým výsledkem v testu byl tedy hrubý skór korigovaný na „hádání". Kromě toho byl vypočítán pro každého testovaného také percentilový skór. Jelikož programy ITEMAN a BILOG-MG, které jsme použili k řešení našeho výzkumného problému, vyžadují dichotomické skórování (1 bod za správnou, 0 bodů za chybnou či vynechanou odpověď), byl použit i nekorigovaný hrubý skór k porovnání analýz výsledků testu založených na klasické teorii testu (KTT) a teorii odpovědi na položku (IRT).

Sběr dat

Odpovědi studentů v testu OSP včetně jejich identifikačních kódů byly ze záznamových archů snímány optickým scannerem KODAK i60 (viz obr. 9-1) a převáděny pomocí programu ABBY Form Reader 7.0 (s roční licencí) od ukrajinské firmy ABBY software do elektronické podoby. V letošním akademickém roce se bude ke sběru dat kvůli nevyhovujícím licenčním podmínkám ukrajinské firmy používat Remark Office OMR Software Version 6.0 od americké firmy Principia Products division of Gravic, Inc. s neomezenou licencí.

Obr. 9-1 Optický scanner

Metody zpracování dat

K analýze výsledků výše zmíněného testu Obecné studijní předpoklady jsme použili software ITEMAN a LERTAP 5, jejichž teoretickým základem je klasická analýza testu, a BILOG-MG vycházející z teorie odpovědi na položku. Těsnost vztahu mezi korigovaným a nekorigovaným skórem či mezi ukazateli obtížnosti a citlivosti v KTT a IRT jsme zkoumali korelační analýzou (Pearsonovým korelačním koeficientem) v programu MS Windows EXCEL. Většina použitých tabulek a obrázků vznikla úpravou či shrnutím výstupů z těchto programů pro snazší porovnání.

10 Analýza celkových statistických charakteristik variant testu OSP

V dnešní době se analýza výsledků testu provádí jak podle tradiční analýzy (nejjednodušší a zatím nejrozšířenější), která se používá při vývoji testů založených na klasické teorii testu (KTT), tak ve světě (zatím ne v ČR a SR) stále častěji také podle teorie odpovědi na položku (IRT) a jejích aplikací (Hambleton aj. 1991). V obou případech se dnes používá specializovaný software(viz tab. 10-1), aplikace IRT bez použití počítače není myslitelná. Největší světová testovací organizace Educational Testing Service začala na konci 20. století interně uplatňovat k položkové analýze svých testů ještě jinou metodu, tzv. grafický přístup založený na odhadu souboru křivek odpovědi na položky, který nevyžaduje žádné omezující předpoklady jako IRT. Tento přístup jakožto dosud ojedinělý ze svých úvah prozatím vypouštíme.

Tab. 10-1 Porovnání samostatných programů na analýzu výsledků testu

Program	Platforma	Max. počet úloh	Max. počet testovaných
Classical Item Analysis (Klasická analýza testu)			
Integrity	Webové rozhraní	500	liší se
ITEMAN	Windows	750	neomezený
Lertap 5	Windows/Macintosh	255	65 535
Scrutiny!	Windows	1 000	neomezený
TestFACT	Windows	1 000	neomezený
Rasch Analysis (1-parameter IRT) (Raschova analýza, 1-parametrový IRT model)			
Quest	DOS/Macintosh	400	10 000
RASCAL	Windows	750	neomezený
RSP	DOS	96	neomezený
RUMMFOLD*ss*	Windows	100	5 000
WINSTEPS*	Windows	30 000	10 miliónů
BIGSTEPS*	DOS	3000	20 000
2- and 3-Parameter IRT Analysis (2 a 3-parametrová IRT analýza)			
BILOG-MG	Windows	1,000 až neomezený	neomezený
MSP	Windows	100	32 000
MULTILOG	Windows	neomezený	neomezený
PARELLA	DOS	60	300
PARSCALE	Windows	neomezený	neomezený
XCALIBRE	Windows	750	neomezený

Porovnání provádíme jak na teoretické úrovni, tak při konkrétní aplikaci na souboru dat získaných variantami testu OSP. Sjednocujícími hledisky při porovnání obou metod jsou testové varianty

s úlohami binárně skórovanými (1 za správnou, 0 za chybnou a vynechanou odpověď) a použití speciálního software (ITEMAN, LERTAP 5, BILOG-MG) k provedení analýz testu.

Jedním z nejdůležitějších rozdílů mezi testováním založeném na KTT oproti IRT je efektivita administrace (časová úspora), nahlížení na test a zakládání bank úloh (vytváření kalibrovaných bank úloh, z kterých lze vybrat skupinu úloh pro každého testovaného individuálně). Zatímco ukazatelé funkce testu v KTT a zvláště skórování testu vycházejí z toho, že všem testovaným je zadána celá banka úloh (tedy stejný soubor úloh), umožňují IRT metody, aby testovaným byly zadány naprosto odlišné soubory úloh nebo rozdílný počet úloh, které však měří stejný latentní rys (schopnost). IRT totiž uvažuje o položkách a jejich vlastnostech samostatně z hlediska latentního rysu (schopnosti), který mají položky měřit, a to nezávisle na testu a na souboru testovaných. Test považuje za soubor samostatných položek (viz např. Hambleton; Swaminathan; Rogers 1991, Baker 2001). V IRT je měřený rys odhadován jako úroveň schopnosti (theta), je součástí IRT modelu popisujícího odpovídání testovaných na položky testu. Odhady schopnosti jsou nezávislé na použitých úlohách v testu. To umožňuje, aby jedinci se stejným skórem dosáhli různých úrovní schopnosti.

Oproti tomu KTT nahlíží na položky a jejich vlastnosti v kontextu konkrétního testu, je orientovaná na test, položky nejsou od celku testu oddělitelné (položky jsou korelovány s celkovým skórem) a jsou závislé na souboru testovaných, kterým byly zadány. Odhad měřeného rysu (pravý skór, který je z praktických důvodů často transformován na z-skór, Baker 2001) testovaného vyplývá přímo ze skóru celého testu a nedovoluje žádné úvahy o odpovědích testovaných na položku. Nelze tedy předpokládat jako v případě IRT, jak testovaný v úloze odpoví.

Obtížnost a citlivost testu

Míra obtížnosti a citlivosti testu vycházející z klasické teorie testu (KTT) tedy závisí na tom, na jakém souboru testovaných ji měříme. Oproti tomu pomocí teorie odpovědi na položku (IRT) zjišťujeme obtížnost a citlivost položek (pomocí parametrů b a a, graficky podle charakteristické funkce testu), a tím i celého testu (sečteme-li charakteristické funkce úloh) nezávisle na charakteristikách souboru testovaných, což nám zaručuje větší přesnost a umožňuje vytvářet počítačové adaptivní testy, jejichž délka není konstantní a závisí na zvolené přesnosti měření. Charakteristickou funkci testu můžeme použít k předpovědění skórů testovaných s danou úrovní schopností θ. Je-li test složen z relativně obtížných úloh, je charakteristická funkce testu posunuta doprava a testovaní mají tendenci k nižším očekávaným skórům než je tomu u relativně snadných položek.

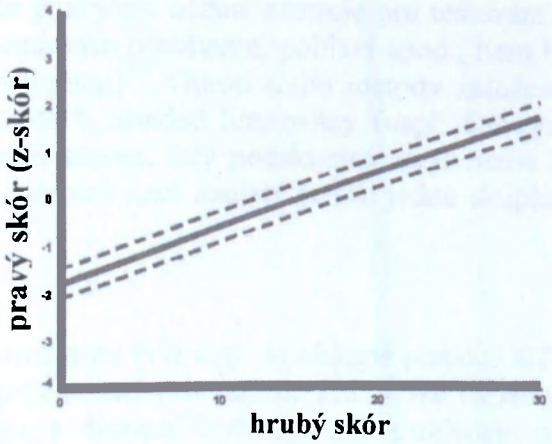
Stanovení statistických charakteristik v KTT lze provádět i ručně (velmi pracné) bez pomoci počítačového software, i když jeho použití je výhodou, zatímco stanovení odhadů parametrů obtížnosti a citlivosti v IRT použití specializovaného software vyžaduje.
KTT předpokládá lineární vztah mezi pravděpodobností správné odpovědi na úlohu a mírou rysu, který má úloha měřit. Míra tohoto vztahu se vyjadřuje korelací (v případě dichotomických úloh bodově biseriální korelací). Jde tedy o lineárně regresní model vztahu, který však není příliš realistický, zejména pro testované s velmi nízkou či velmi vysokou úrovní měřeného rysu (je jen hrubou aproximací). IRT naopak předpokládá nelineární vztah mezi pravděpodobností správné odpovědi a mírou latentního rysu, který je aproximován logistickou křivkou (tzv. charakteristickou křivkou). Jde tedy o nelineární regresní model vztahu (např. Hulin; Drasgow; Parsons 1983).

Reliabilita a přesnost měření

Pojetí reliability v KTT je zhruba analogické k pojetí informace v IRT ve smyslu, že vyšší hodnoty udávají lepší přesnost měření (nebo-li menší chybu měření). V KTT je koeficient reliability vysoký pro úlohy s dobrou citlivostí a obtížností (Hopkins 1998). V IRT jsou vyšší hodnoty informace

asociovány s vyššími hodnotami parametrů citlivosti (a) a malými hodnotami parametru hádání (c). Obdobně jako je standardní chyba odhadu určitého skóru θ v IRT v inverzním vztahu k informaci, je standardní chyba měření v KTT v inverzním vztahu s reliabilitou testu (viz kap.4). Standardní chyba v IRT je koncepčně ekvivalentní k standardní chybě měření v KTT, ale na rozdíl od ní umožňuje zobecnění na různé populace. Čím více informace test na dané úrovni schopnosti poskytuje, tím menší je chyba, s níž je úroveň schopnosti odhadována.

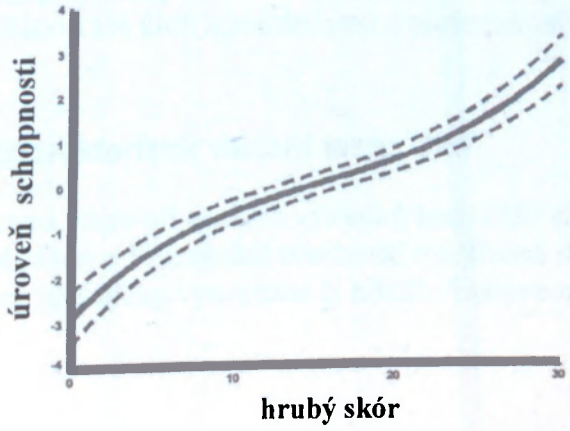
Zásadním rozdílem mezi IRT a KTT je však to, že KTT předpokládá stejnou standardní chybu pro všechny skóry v testu a reliabilitu testových výsledků. Chyba je závislá na souboru testovaných, ale ne na úrovni měřeného rysu (vyplývá z její definice). Konfidenční interval je stanoven pro všechny testované společně a to za předpokladu, že chyba měření má normální rozložení a stejné pro všechny skóry (viz obr. 10-1).



Obr. 10-1 Regrese hrubého skóru (upraveno podle Embretson; Reise 2000)

Z obr. 10-1 je patrné, že odhadovaný pravý skór je odvozen jako lineární transformace z hrubého skóru (lineární regrese). Interval spolehlivosti je také znázorněn pro všechny skóry přímkou (protože stejný interval je aplikován na všechny skóry).

Zatímco v KTT se používá jediné číslo (např. koeficient reliability alfa či standardní chyba měření založená na této reliabilitě) k stanovení přesnosti měření testu, je v IRT požadována plynulá funkce, tzv. informační funkce (test information function, TIF; viz kap. 4) k vyjádření srovnatelných dat s tím, že se přesnost mění pro různé skóry (viz obr. 10-2). V IRT je tedy standardní chyba proměnlivá pro všechny skóry a není závislá na souboru testovaných. Umožňuje zobecnění na různé populace. Interval spolehlivosti pro odhad úrovně schopnosti testovaného je stanoven na základě známé standardní chyby.



Obr. 10-2 Regrese hrubého skóru (upraveno podle Embretson; Reise 2000)

Z obr. 10-2 lze usoudit, že vztah mezi skórem vázaným na úroveň schopnosti a hrubým skórem není lineární a konfidenční interval se stále více rozšiřuje pro extrémní skóry. Standardní chyba je nejmenší, jsou-li úlohy optimálně vhodné pro danou úroveň schopnosti a parametry citlivosti a jsou vysoké.

Dalším rozdílem mezi reliabilitou v KTT oproti IRT je to, že v KTT závisí reliabilita na délce testu. Zatímco v KTT mají delší testy větší reliabilitu, mohou mít v IRT kratší testy vyšší reliabilitu než delší (v případě adaptivních testů), což vyplývá z nezávislosti vlastností položek na složení celého testu. V KTT je vztah reliability a délky testu vyjádřen Spearman-Brownovým vzorcem (viz kap.4). Jestliže například test s reliabilitou 0,86 je zkrácen na dvě třetiny své délky, zmenší se jeho reliabilita na 0,80.

Další předností IRT je to, že poskytuje účinné nástroje pro testování specifických chyb v úlohách (jejich zaujatost např. vůči etnickým menšinám, pohlaví apod.; item bias) pomocí tzv. DIF funkce úlohy (differential item functioning)[59]. Oproti tomu metody založené na KTT jsou v hodnocení specifických typů chyb v úlohách zásadně limitovány (např. Drasgow 1987). V podstatě takové metody nemohou rozlišit mezi situací, kdy podskupiny mají různé průměry a test je zaujatý, a situací, kdy se průměry liší, ale test není zaujatý (takže jedna skupina má skutečně vyšší průměr v testu).

Vyrovnanost variant testu

Ukazatele obtížnosti a citlivosti testu (viz kap. 4) získané pomocí KTT umožňují zhruba posoudit statistickou vyrovnanost několika testových variant. Jednotlivé varianty testu by měly být vyvážené nejen svým obsahem, počtem a druhem úloh, ale také s ohledem na statistické charakteristiky: obtížnost a citlivost. Varianty testu se považují za vyrovnané (paralelní)[60], pokud obsahují úlohy podobné obtížnosti (p) a citlivosti (r). Rozložení skórů u statisticky vyrovnaných variant testu by mělo být téměř shodné, tj. Gaussovy křivky rozložení skórů by se měly překrývat.

K posouzení vyrovnanosti několika variant testu pomocí IRT se využívá charakteristických funkcí/křivek[61] (test characteristic curve, TCC) a informačních funkcí variant testu. Varianty testu se považují za vyrovnané (paralelní), pokud jejich charakteristické a informační funkce jsou téměř identické (se překrývají). Porovnání informačních funkcí pro danou úroveň θ lze provádět také pomocí výpočtu relativní efektivity (relative efficiency) jednoho testu ve srovnání s druhým jako odhad na úrovni θ: $RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)}$,

kde $RE(\theta)$ udává relativní efektivitu a $I_A(\theta)$ a $I_B(\theta)$ jsou informační funkce testů A a B. Jestliže je např. $I_A(0) = 16$ a $I_B(0) = 14{,}4$ pro $\theta = 0$, potom $RE(\theta) = 1{,}11$. To znamená, že na úrovni schopnosti $\theta = 0$ test A funguje, jako by byl o 11% delší než test B. Test B by bylo tedy potřeba prodloužit, příp. test A zkrátit o 11% úloh k získání stejné přesnosti měření jako testem A v $\theta = 0$ (Hambleton aj. 1991).

Analýza statistických charakteristik variant testu OSP

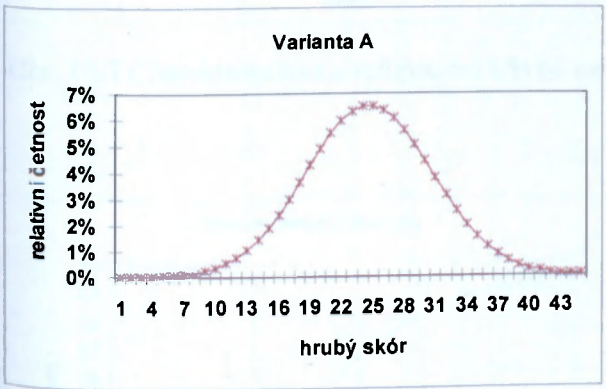
Statistickými charakteristikami, které při analýze výsledků testu OSP sledujeme, jsou především obtížnost, citlivost a reliabilita testu. Na základě obtížnosti a citlivosti dále posuzujeme, zda jsou jednotlivé čtyři varianty testu statisticky vyrovnané či nikoli. Analýza zahrnuje také kontrolu časového omezení.

[59] Jak identifikovat zaujatost úloh uvádějí např. Hambleton aj. (1991), Baker (2001), Embretson; Reise (2000) atd.
[60] Za předpokladu, že skóry testových variant mají přibližně symetrické normální rozložení.
[61] Bude podrobně vysvětleno dále.

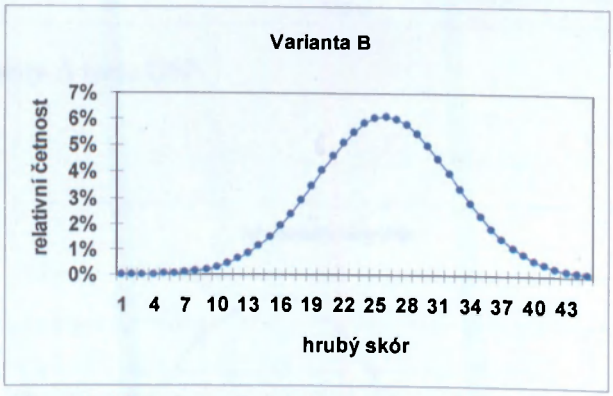
Obtížnosti variant A, B, C, D testu OSP v KTT a IRT

Obtížnost testových variant se v KTT, jak již bylo zmíněno v kap. 4, vyjadřuje průměrným podílem správných odpovědí v souboru testovaných, v IRT parametrem obtížnosti p závislém na zvoleném modelu, jehož hodnoty se v praxi pohybují od -3 do +3 (Baker 2001).

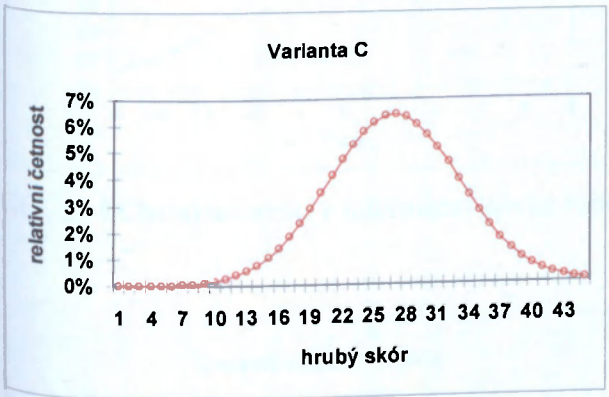
Relativní průměrné skóry variant A, B, C, D jsou po řadě 0,55; 0,58; 0,60 a 0,57 (viz p v tab. 10-2). Testové varianty jsou tedy průměrně obtížné (hodnoty leží mezi 0,5 a 0,6). Je to zřejmé i z grafů Gaussových křivek rozložení skórů všech čtyř variant testu OSP založených na relativních četnostech (viz obr. 10-3 až 10-6), maximální hodnoty křivek se nacházejí přibližně uprostřed škály hrubého skóru.



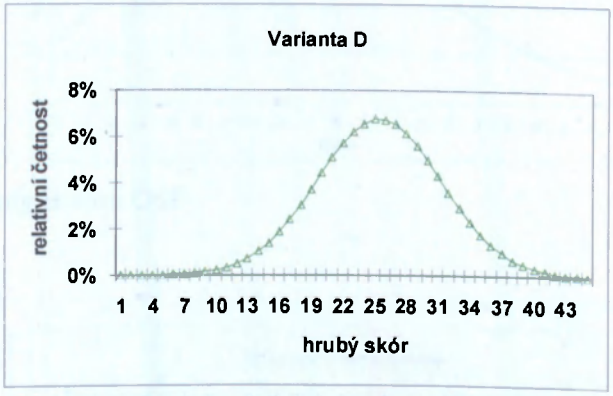
Obr. 10-3 Rozložení skórů (varianta A)



Obr. 10-4 Rozložení skórů (varianta B)

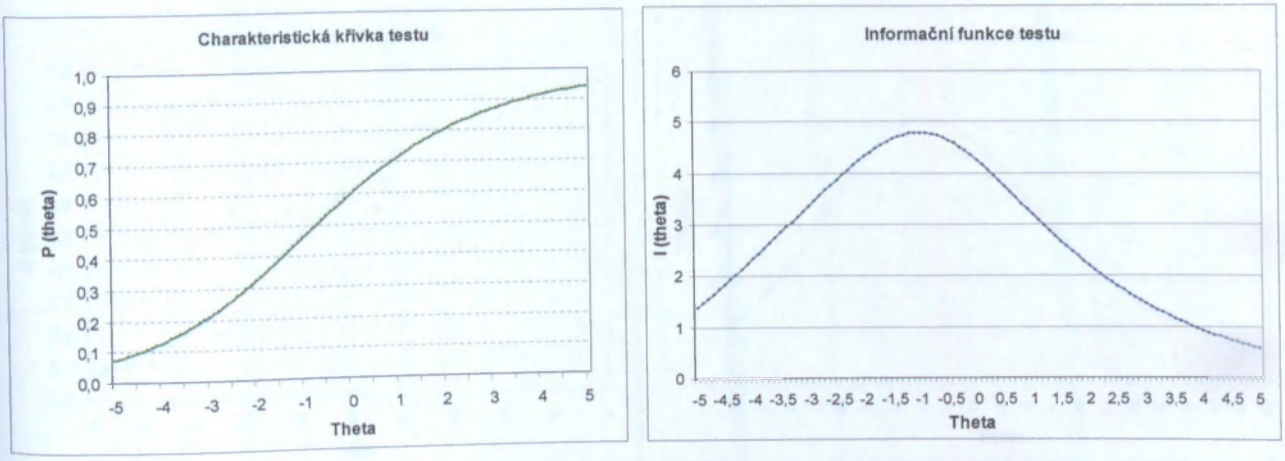


Obr. 10-5 Rozložení skórů (varianta C)

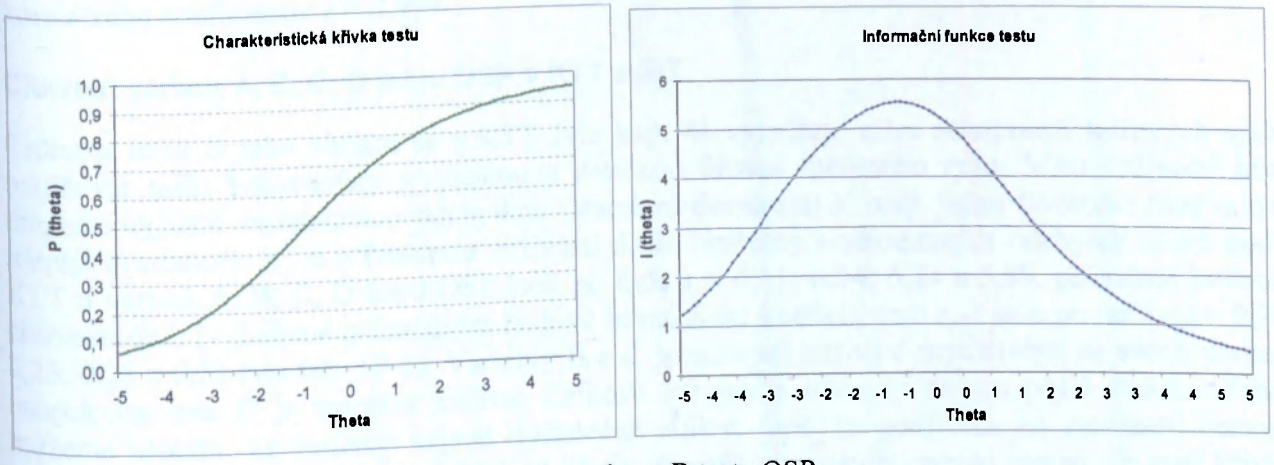


Obr. 10-6 Rozložení skórů (varianta D)

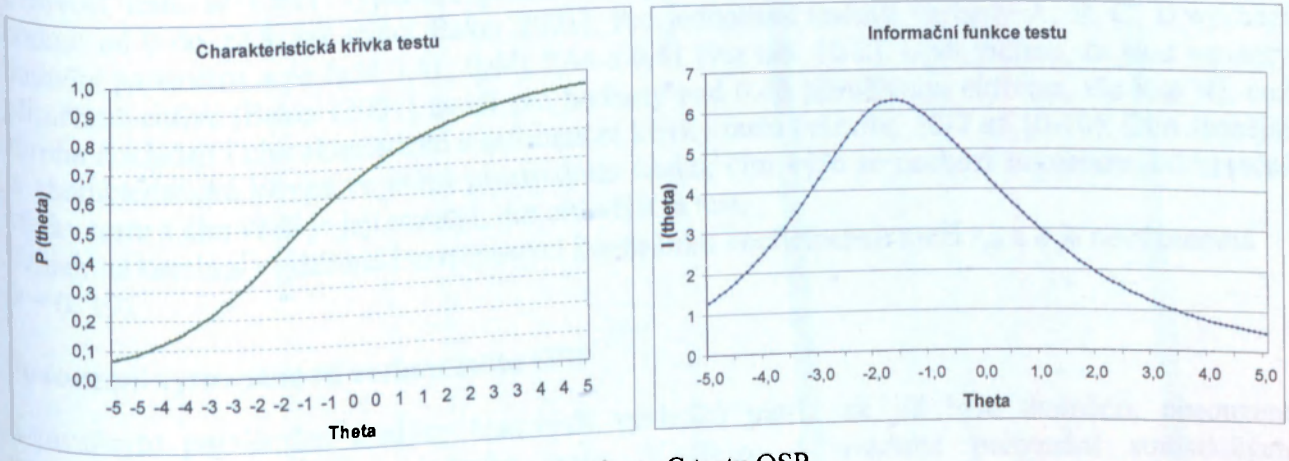
Podíváme-li se na obtížnost variant přesně, zjistíme, že varianta C byla pro testované nejsnadnější a varianta A nejtěžší. 50ti procentní pravděpodobnosti správné odpovědi v KTT (p = 0,5) odpovídá v IRT hodnota parametru b = 0. Malá pravděpodobnost správné odpovědi (těžký test) odpovídá velké kladné hodnotě IRT parametru obtížnosti, velká pravděpodobnost správné odpovědi (snadný test) velké záporné hodnotě IRT parametru obtížnosti. Podle IRT jsou hodnoty parametru obtížnosti po řadě rovny hodnotám b = -0,51; -0,67; -0,76 a -0,61 (viz tab. 10-2). Dojdeme tedy ke stejným výsledkům jako podle KTT, nejsnadnější je opět varianta C, potom B, D a za nejtěžší je považována varianta A. Graficky tuto skutečnost znázorňují charakteristické a informační funkce jednotlivých variant testu (viz obr. 10-7 až 10-10). Bod odpovídající 50% pravděpodobnosti správné odpovědi leží u všech variant (viz charakteristické křivky) přibližně kolem nuly (mezi 0 a 1), s čímž koresponduje poloha maxima informačních funkcí variant zhruba uprostřed škály theta (v kterém test také nejlépe rozlišuje). Z toho plyne, že varianty testu jsou průměrně obtížné. Čím více je informační křivka testu posunuta doprava do oblasti vyšších úrovní schopnosti theta θ, tím těžší je test. Je tedy zřejmé, že varianta C je nejsnadnější a varianta A je nejtěžší.



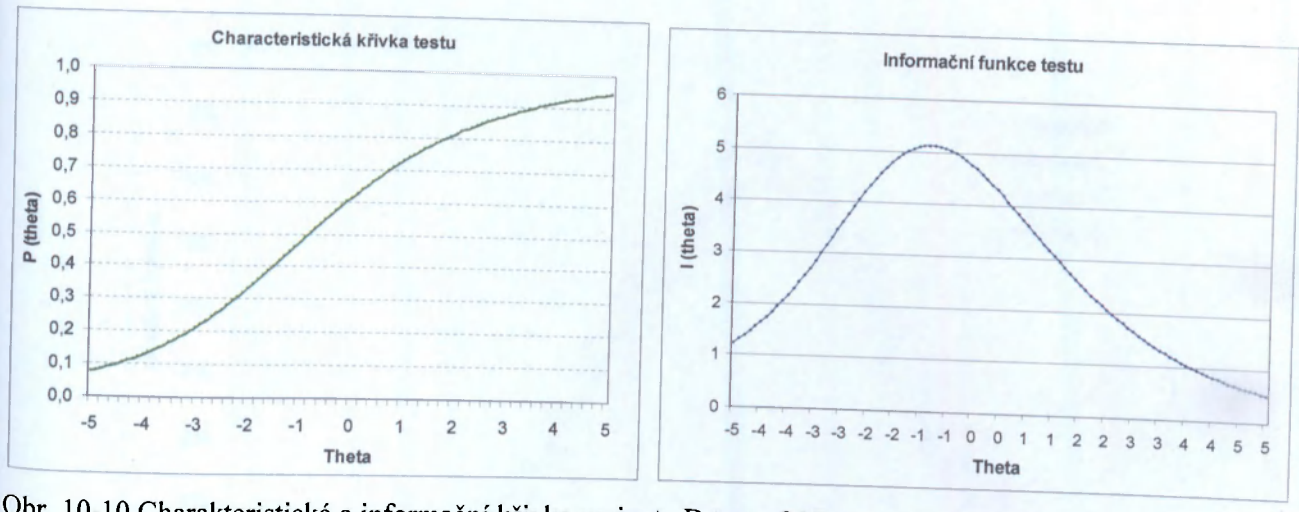
Obr. 10-7 Charakteristická a informační křivka varianty A testu OSP



Obr. 10-8 Charakteristická a informační křivka varianty B testu OSP



Obr. 10-9 Charakteristická a informační křivka varianty C testu OSP



Obr. 10-10 Charakteristická a informační křivka varianty D testu OSP

Vztah mezi obtížností testu v KTT a IRT je těsný, což ukazuje vysoká hodnota Pearsonova korelačního koeficientu: $r = 0{,}999$.

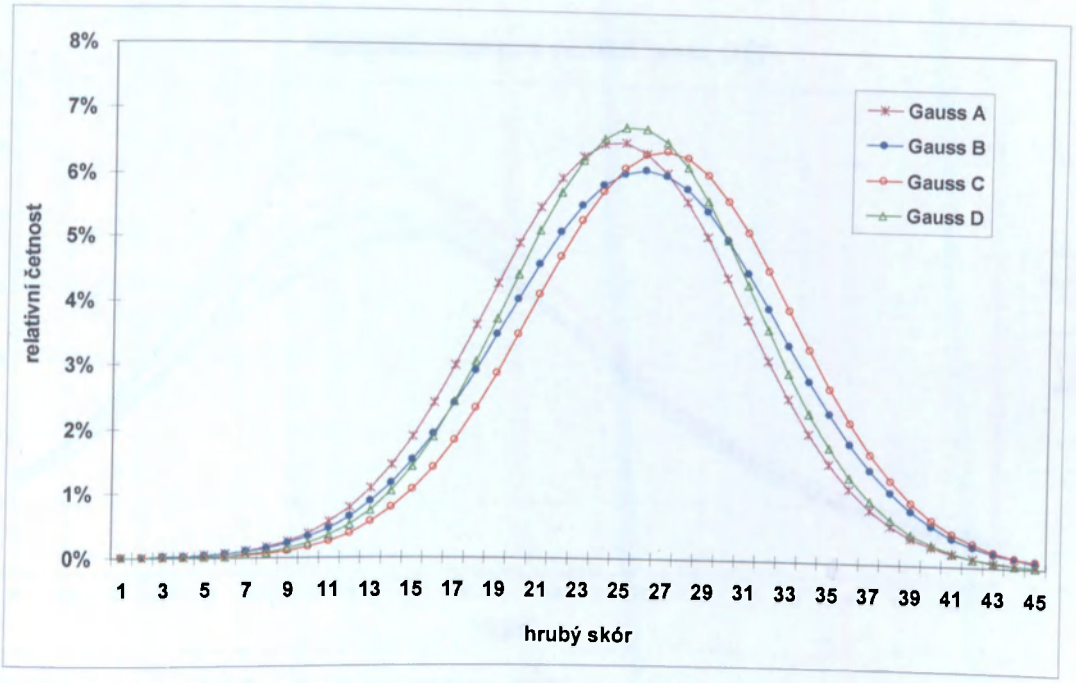
Citlivost variant A, B, C, D testu OSP v KTT a IRT

Citlivost testu či jeho variant se v KTT (viz kap. 4) vyjadřuje míru schopnosti testových skórů rozlišovat mezi testovanými s vysokou a s nízkou úrovní měřeného rysu. Míru citlivosti testu můžeme vyjádřit směrodatnou odchylkou (standard deviation) s, resp. jejím čtvercem nazývaným rozptyl (variance) s^2 a průměrnou citlivostí úloh. Hodnoty směrodatných odchylek skórů podle KTT u variant A, B, C, D testu OSP jsou po řadě $s = 6{,}11$; 6,54; 6,24 a 5,89, průměrné hodnoty citlivosti úloh (vyjádřené průměrným bodově biseriálním koeficientem r_{pb}) jsou po řadě $r_{pb} = 0{,}24$; 0,26; 0,26 a 0,23 (viz tab. 10-2). Varianty B a C jsou stejně citlivé a nejcitlivější ze všech variant, naopak varianta D je nejméně citlivá. Celkově ale jejich hodnoty nejsou příliš velké, z čehož můžeme usoudit, že varianty nejsou dostatečně citlivé. Opět se podíváme na rozložení četností skórů (Gaussovy křivky, viz obr. 10-3 až 10-6). Z grafů je na první pohled patrné, že mají křivky přibližně stejný rozptyl i tvar.

Citlivost testu se v IRT vyjadřuje průměrným parametrem a citlivosti úloh (nabývá zpravidla hodnot od 0 do +2,8, jak uvádí Baker 2001). Pro jednotlivé testové varianty A, B, C, D vychází hodnoty parametru a po řadě 0,41; 0,44; 0,44 a 0,41 (viz tab. 10-2). Opět vidíme, že jsou varianty přiměřeně citlivé (Baker (2001) uvádí pro hodnoty nad 0,40 přiměřenou citlivost, viz kap. 4), což zhruba dokládají i charakteristické a informační křivky testu (viz obr. 10-7 až 10-10). Čím strmější je charakteristická křivka ve svém prostředním úseku, čím výše se nachází maximum informační křivky testu a čím větší je její rozptyl, tím citlivější je test.
Vzájemná korelace vyjádřená Pearsonovým korelačním koeficientem mezi r_{pb} a a je nevýznamná ($r = 0, 19$).

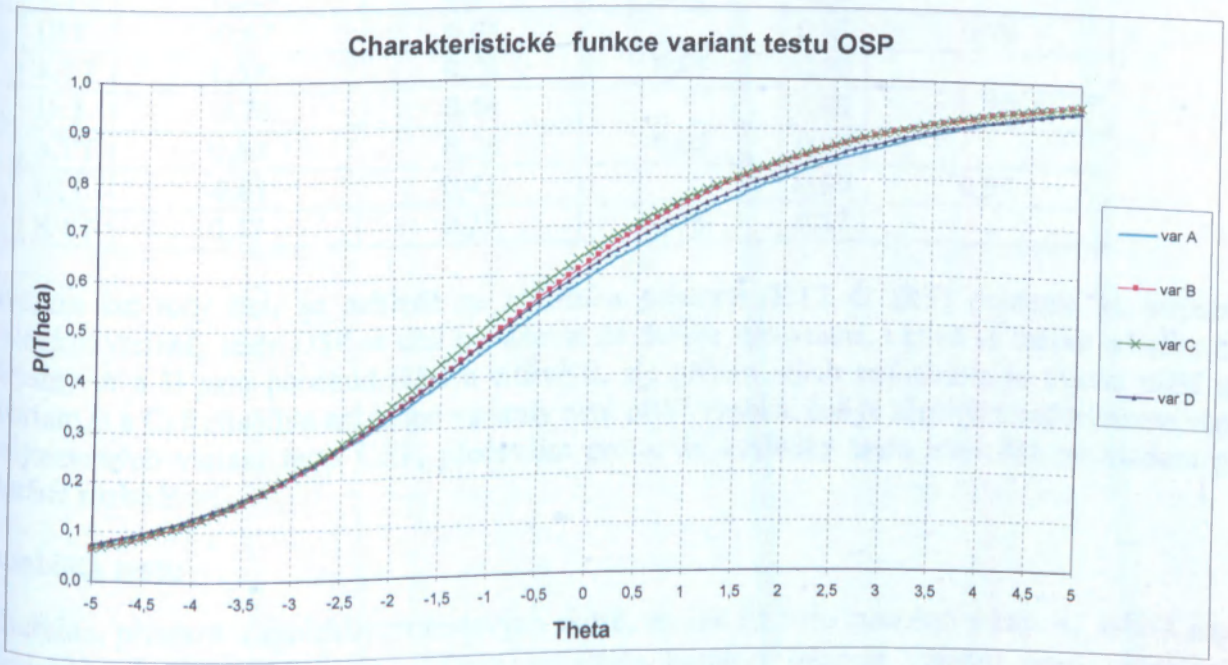
Posouzení vyrovnanosti variant testu OSP

K závažným problémům analýzy testových výsledků patří, jak již bylo zmíněno, posouzení vyrovnanosti všech variant jednoho testu. V příloze 4 uvádíme porovnání statistických charakteristik podle KTT variant A až D testu OSP získaných z programu LERTAP 5 a ITEMAN. Při srovnání jednotlivých statistických charakteristik si můžeme všimnout, že většina z nich je vyrovnaná. Tuto skutečnost graficky znázorňují i Gaussovy křivky rozložení skórů dle relativních četností jednotlivých variant testu (viz obr. 10-11), které se sice zcela nepřekrývají (ideální případ), ale dají se považovat za podobné.

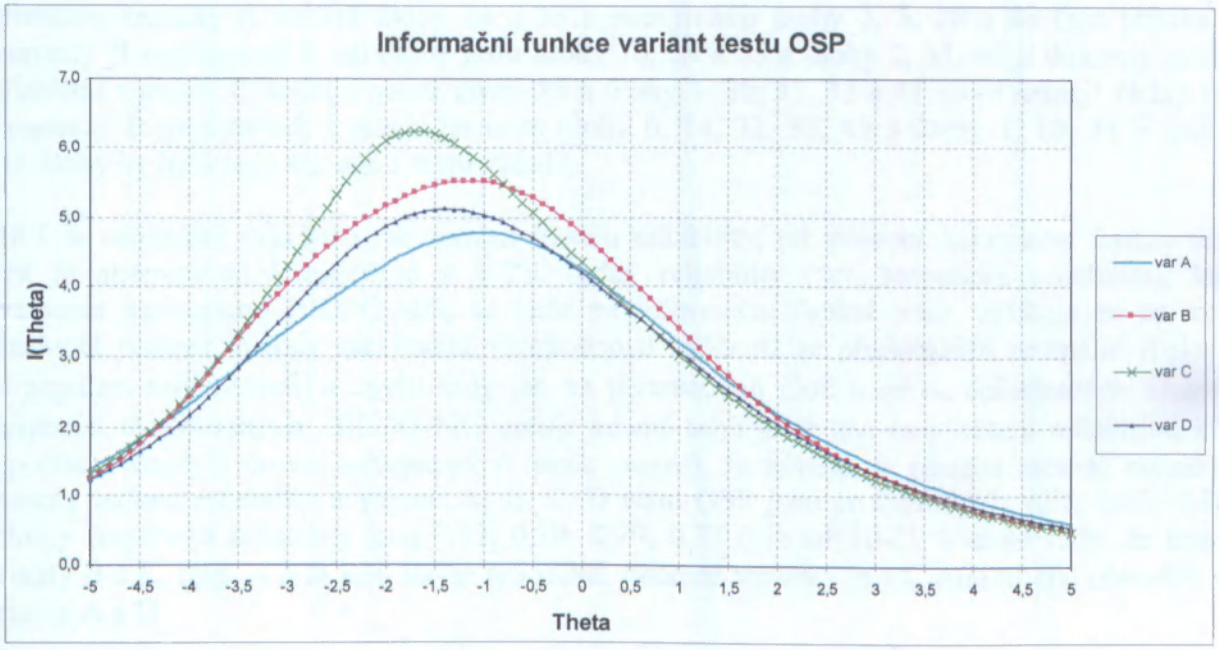


Obr. 10-11 Rozložení skórů variant testu OSP založené na relativních četnostech v KTT

I z analýzy pomocí IRT (jak z vypočtených parametrů citlivosti, tak z charakteristických a informačních křivek) plyne, že jsou jednotlivé varianty téměř paralelní. Nejvíce se však od ostatních vychyluje varianta C (viz obr. 10-12 a 10-13). Křivky se téměř překrývají.



Obr. 10-12 Charakteristické křivky všech variant testu OSP



Obr. 10-13 Informační křivky všech variant testu OSP

Tab. 10-2 Obtížnost, reliabilita a průměrná citlivost úloh variant A, B, C, D testu OSP

		obtížnost (p^{62}/ b)	citlivost (r_{pb}/a)	reliabilita (α)	ρ_{XT}	empir. reliabilita
A	IRT	-0,51	0,41		0,80	0,77
	KTT	0,55	0,24	0,78	0,88	
B	IRT	-0,67	0,44		0,82	0,79
	KTT	0,58	0,26	0,81	0,90	
C	IRT	-0,76	0,44		0,82	0,79
	KTT	0,60	0,26	0,80	0,89	
D	IRT	-0,61	0,41		0,80	0,77
	KTT	0,57	0,23	0,76	0,87	

Závěrem lze tedy říci, že nehledě na použitém přístupu (KTT či IRT) dojdeme ke stejnému výsledku: varianty testu OSP se dají považovat za zhruba vyrovnané, i když se trošku od sebe liší. Varianty A a D jsou poněkud těžší a citlivější, ale přitom jejich realiabilita je trochu nižší než u variant B a C. Reliabilita ani jedné varianty není příliš vysoká, což je zřejmým nedostatkem všech analyzovaných variant testu OSP, především proto, že výsledky testu mají být podkladem pro závažné rozhodnutí.

Reliabilita testu

Reliabilita, přesnost a spolehlivost testových skórů, se, jak již bylo zmíněno v kap. 4, udává např. v podobě indexu reliability ρ_{XT}, který vyjadřuje korelaci (těsnost vztahu) mezi rozdělením pravdivých a naměřených skórů, nebo častěji pomocí nejpřesnějšího koeficientu reliability Cronbachovo alfa. Vysoce reliabilní test by měl mít $\rho_{XT} \geq 0{,}95$ nebo $\alpha \geq 0{,}9$.

U variant A, B, C, D testu OSP je reliabilita vyjádřená v KTT pomocí koeficientu Cronbachovo alfa po řadě 0,78; 0,81; 0,80; 0,76 (viz tab. 10-2) a hodnoty indexu reliability ρ_{XT} jsou po řadě 0,88; 0,90; 0,89; 0,87. Je také patrné, že se od sebe příliš neliší. Hodnoty nejsou sice tak vysoké, ale na druhou stranu nejsou zase tak nízké, takže se dají považovat za celkem uspokojivé. Co se týče velikosti směrodatných chyb měření, jsou po řadě 2,86; 2,87; 2,81 a 2,87. Nejvyšší reliabilitu (ať již vyjádřenou pomocí indexu reliability či koeficientu alfa) má varianta B, nejnižší varianta D.

[62] Jde o relativní průměrný skór.

Reliabilitu varianty A snižují úlohy 24 a 33 a neovlivňují úlohy 3, 5, 26 a 38 (viz příloha 5), u varianty B nepřispívají k reliabilitě testu úlohy 10, 29 a 35 a úlohy 2, 33, 40 ji dokonce snižují. Reliabilitu varianty C snižuje pouze úloha 35 a úlohy 1, 30, 31, 33 a 41 na ní nemají žádný vliv. U varianty D nepřispívají k reliabilitě testu úlohy 6, 14, 32, 35, 43 a úlohy 1, 19, 34 ji snižují. Tyto úlohy by bylo tedy vhodné z testu vyřadit.

V IRT se reliabilita vyjadřuje jak pomocí indexu reliability, tak pomocí informační funkce testu, která je alternativou k reliabilitě v KTT. Index reliability (tzv. teoretická reliabilita), který dostaneme z programu BILOG-MG, se váže na informační funkci testu, aplikuje se na skóry odhadnuté pomocí metody maximální věrohodnosti (přičemž se předpokládá normální rozložení θ v populaci testovaných) a závisí tedy jen na parametrech úloh a ne na odhadnutých úrovních schopností θ testovaných. BILOG-MG určuje kromě toho ještě tzv. empirickou reliabilitu, která se počítá z rozptylů úrovní schopností θ (scale scores). Je závislá na použité metodě odhadu θ. Hodnoty indexu reliability u variant A, B, C, D testu OSP jsou po řadě 0,80; 0,82; 0,82; 0,80 a hodnoty empirické reliability jsou 0,77; 0,79; 0,79; 0,77 (viz tab.10-2). Vidíme tedy, že testové varianty B a C, resp. A a D jsou stejně reliabilní, přičemž varianty B a C jsou trochu přesnější než varianty A a D.

K určení přesnosti testu, a to pro všechny úrovně θ, můžeme také použít informační funkce testu, která je definována nezávisle na skupině testovaných, ale je značně ovlivněna parametry úloh. Reprezentuje chybu měření pro všechny úrovně schopnosti θ . Test měří schopnost s největší přesností, tj. nejlépe rozlišuje mezi testovanými s úrovní schopnosti odpovídající hodnotě parametru b obtížnosti testu. Z informačních funkcí variant testu OSP (viz obr. 10-13) je tedy zřejmé, že varianta A měří s největší přesností (poskytuje nejvíce informace) kolem úrovně schopnosti $\theta = -1$ (přibližně v rozmezí θ od -2,5 a 0,5), varianta B kolem $\theta = -1,13$ (přibližně v rozmezí θ od -3 a 0,5), varianta C $\theta = -1,63$ (přibližně v rozmezí θ od -3 a 0) a varianta D kolem $\theta = -1,38$ (přibližně v rozmezí θ od -2,5 a 0). Jelikož test nebyl obsahově homogenní (skládal se tří naprosto odlišných částí, které neměřily jeden rys), nebyl splněn jeden z předpokladů IRT (unidimenzionalita), a proto je reliabilita jednotlivých variant snížena.

Vztah mezi indexem reliability v KTT a IRT je těsný, což ukazuje hodnota korelačního koeficientu $r = 0,94$.

Kontrola časového omezení

Ani jedna z variant testu OSP nebyla časově přiměřená, protože ji dokončilo méně než 80% testovaných (viz příloha 6), což se považuje za hrubé pravidlo (viz kap. 4). Předpokládáme, že studenti řešili úlohy v pořadí, v jakém jsou uvedeny v testu, a tudíž poslední úlohy, pokud je vynechali, můžeme považovat za nedosažené. U varianty A posledních 11 úloh (dvě třetiny úloh kvantitativního oddílu) nezodpovědělo od 31% (úloha Q 35) do 67% (úloha Q42) testovaných, u varianty B se jedná dokonce o 12 posledních úloh v testu, četnost vynechání úlohy se pohybuje od 21% (úloha Q 34) po 75% (úloha Q 45). Varianty C a D na tom nejsou o nic lépe. U varianty C lze považovat za nedosažené také posledních 12 úloh s výjimkou úloh 34 a 36, kdy procento testovaných, kteří úlohu vynechali, nepřesahuje 20%. Četnost vynechání ostatních zmíněných úloh je od 34% (úloha Q 42) do 65% (úloha Q 40). U varianty D se jedná o celý kvantitativní oddíl (tedy úlohy 31 až 45) s výjimkou úlohy 33. Úlohy nezodpovědělo od 20% (úloha 37) do 63% (úloha 45) testovaných. Samozřejmě se nedá vyloučit možnost, že se testovaní pokoušeli úlohy řešit, ale byly pro ně příliš obtížné (viz příloha 6), a proto je nakonec vynechaly, protože nechtěli riskovat ztrátu bodu.

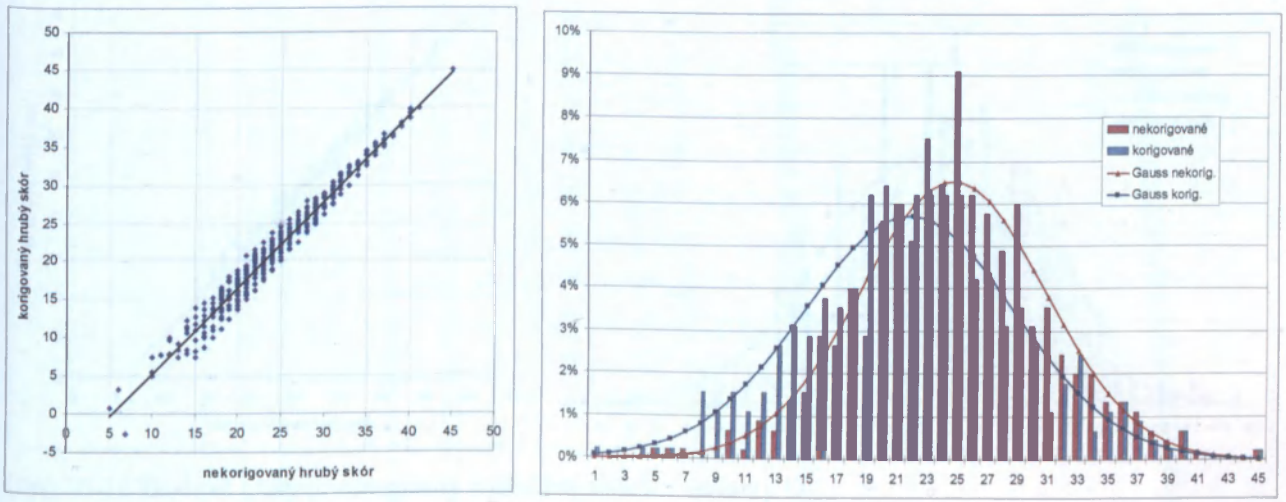
Vztah mezi korigovaným a nekorigovaným hrubým skórem dle KTT v testu OSP

Na souboru dat získaných variantami testu OSP jsme chtěli také zjistit, jaký je vztah mezi nekorigovaným hrubým skórem a skórem korigovaným na hádání. Oba skóry jsme získali pomocí programu LERTAP 5. Pro studium síly vztahu mezi korigovaným a nekorigovaným skórem měřenými současně na všech testovaných jsme jednak použili korelační analýzu, jednak jsme zjišťovali, jak se mění rozložení skórů (maximální, minimální, střední hodnota, rozptyl směrodatná odchylka).

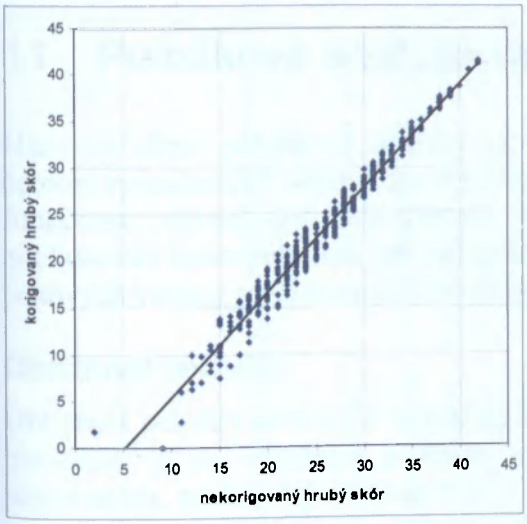
Z rozptylů bodů v bodových grafech a z vysokých hodnot korelačních koeficientů u jednotlivých variant testu OSP (viz obr. 10-13 až 10-16) lze usoudit, že korigovaný skór na „hádání" a nekorigovaný skór spolu navzájem vysoce korelují, tj. že existuje těsný vztah mezi korigovaným a nekorigovaným skórem. U varianty A je korelační koeficient mezi korigovaným a nekorigovaným skórem $r = 0, 982$, u variant B a C je $r = 0, 985$ a u varianty D je $r = 0, 980$. Podívejme se nyní na rozložení obou typů skórů u jednotlivých variant (viz obr. 10-13 až 10-16). Rozložení relativních četností korigovaných skórů (Gaussovy křivky) je ve všech variantách oproti nekorigovaným skórům rovnoměrnější (skóry mají větší rozptyl a směrodatné odchylky, viz tab.10-3). Citlivost testu při použití korekce hrubých skórů je tedy trochu vyšší (s = 6,98), tj. test lépe rozlišuje mezi testovanými s nízkými a vysokými skóry. Průměrný, minimální a maximální skóry jsou ale nižší s výjimkou maxima u varianty A.

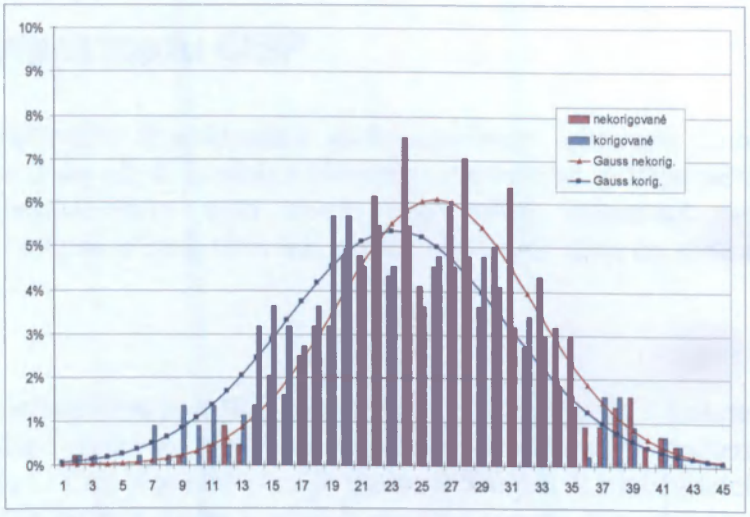
Tab. 10-3 Statistické charakteristiky testu OSP, variant A - D

	A		B		C		D	
	nekorig	korig	nekorig	korig	nekorig	korig	nekorig	korig
n	451	451	439	439	467	467	339	339
min	6,00	-2,50	2,00	0,00	6,00	-2,50	8,00	2,50
medián	25,00	21,50	26,00	23,25	27,00	24,50	25,00	22,50
průměr	24,62	21,74	25,92	23,26	26,86	24,33	25,42	22,63
max	45,00	45,00	42,00	41,50	44,00	43,75	44,00	43,75
s.odchylka	6,05	6,98	6,54	7,43	6,24	7,21	5,89	6,82
rozptyl	37,32	48,79	42,74	55,14	39,00	52,05	34,74	46,49
šikmost	0,12	0,03	-0,04	-0,08	-0,24	-0,41	0,10	0,04
špičatost	0,24	0,31	-0,15	-0,15	0,42	0,71	0,05	0,19

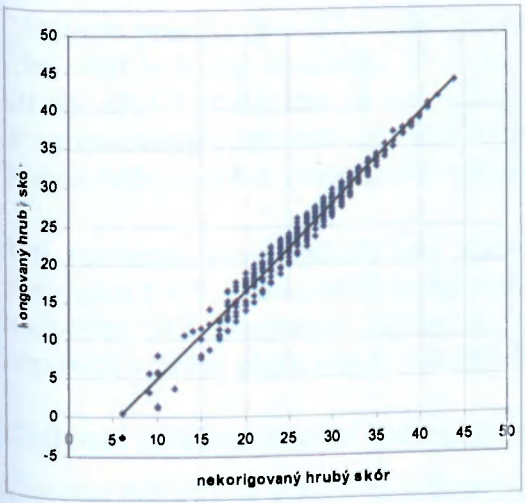


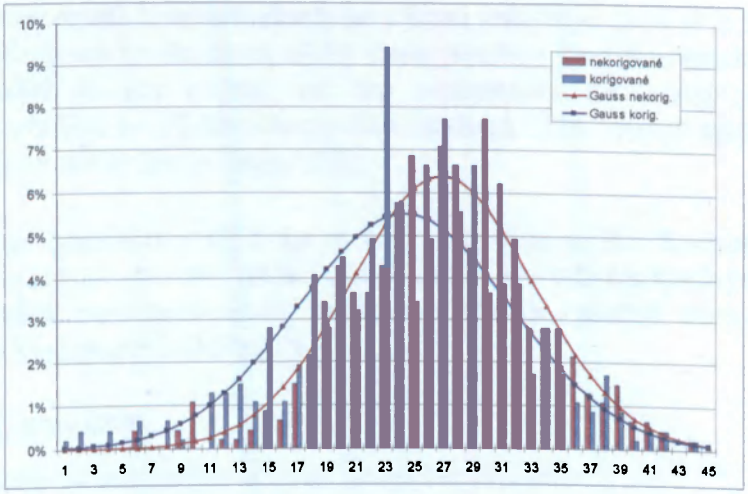
Obr. 10-13 Bodové grafy a histogramy rozložení skórů u varianty A



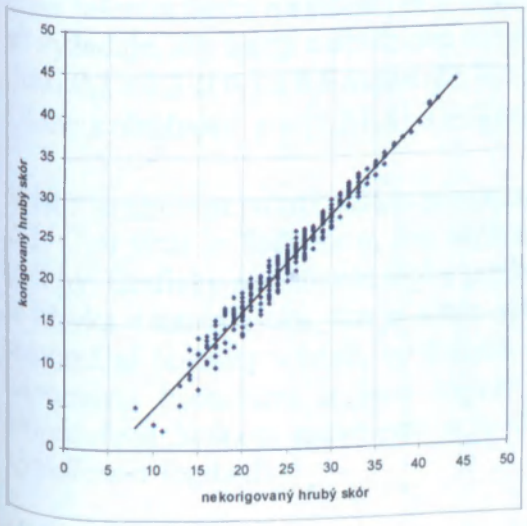


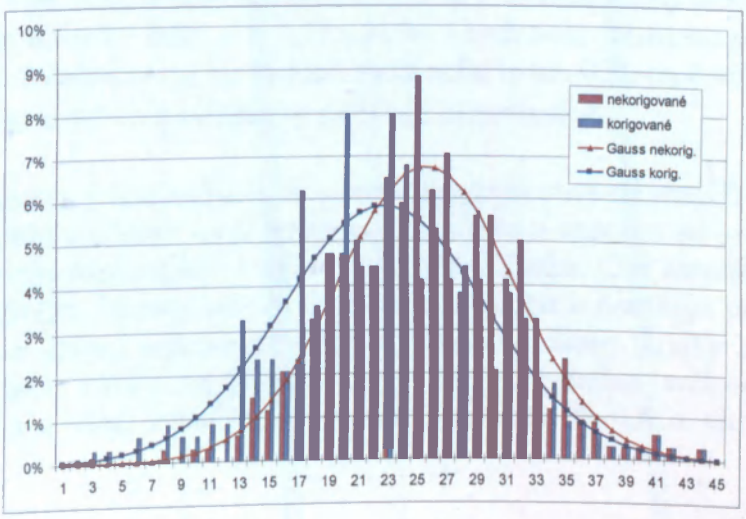
Obr. 10-14 Bodové grafy a histogramy rozložení skórů u varianty B





Obr. 10-15 Bodové grafy a histogramy rozložení skórů u varianty C





Obr. 10-16 Bodové grafy a histogramy rozložení skórů u varianty D

11 Položková analýza variant testu OSP

Hlavním cílem položkové analýzy je zjišťování charakteristik úloh: obtížnosti, citlivosti úloh, četnost vynechaných odpovědí a četnost a druhy chyb. U úloh s výběrem odpovědi se zjišťuje ještě funkčnost jednotlivých distraktorů. Charakteristiky úloh slouží jako zdroj informací pro zdokonalení testových úloh, při zakládání a úpravě bank úloh a k výběru vhodných úloh do testů a testových variant požadovaných vlastností.

Obtížnost položky

Obtížnost položky se v KTT vyjadřuje koeficientem p, v IRT parametrem b (viz kap. 4). Ukazatel obtížnosti p se vyjadřuje podílem počtu osob se správnou odpovědí a celkovým počtem testovaných, může nabývat hodnot od 0 do 1. Čím je p nižší, tím je úloha obtížnější. V rozlišujících testech se považují úlohy s $p \leq 0{,}2$ za příliš obtížné, úlohy s $p \geq 0{,}8$ za příliš snadné. Za optimálně obtížné se považují úlohy s p mezi 0,3 a 0,7 (tedy kolem 0,5), protože bývají citlivější.

Obtížnost položky je v IRT vyjádřena parametrem b, jeho hodnota se v praxi pohybuje mezi -3 a 3. Čím větší je b, tím obtížnější je úloha. Graficky je obtížnost úlohy dána polohou charakteristické křivky (ICC) vzhledem k ose schopností θ. Jde o bod na ose schopnosti, pro který je pravděpodobnost správné odpovědi rovna 0,5 (u 1- a 2-parametrového modelu). Čím více je tento bod umístěn doprava vzhledem k vodorovné ose θ, tím je úloha těžší.

50ti procentní pravděpodobnosti správné odpovědi v KTT ($p = 0{,}5$) odpovídá v IRT hodnota parametru $b = 0$. Malá pravděpodobnost správné odpovědi (těžké úlohy) odpovídá velkým kladným hodnotám IRT parametrů obtížnosti, velká pravděpodobnost správné odpovědi (snadné úlohy) odpovídá velkým záporným hodnotám IRT parametrů obtížnosti.

Citlivost (diskriminační schopnost) položky

Citlivost položky se v KTT vyjadřuje například ukazatelem d nebo bodově-biseriálním koeficientem r_{pb} (příp. biseriálním r_b). Ukazatel d vyjadřuje, zda úlohu vyřešilo správně více lepších než horších testovaných (vzhledem k jejich hrubému skóru). Koeficientem r_{pb} se vyjadřuje korelace mezi řešením úlohy a výsledkem v testu. Oba mohou nabývat hodnot od 0 do 1. U úloh rozlišujících se vyžaduje, aby úlohy s obtížností p mezi 0,3 a 0,7 měly $d > 0{,}25$ a úlohy s hodnotou obtížnosti p mezi 0,2 a 0,3 či 0,7 a 0,8 aspoň $d \geq 0{,}15$. Koeficient r_{pb} by měl být vyšší nebo roven 0,30 (r_b 0,40). Úlohy s obtížností $p \in (0{,}2;0{,}8)$ a citlivostí $d \leq 0$ se automaticky považují za nevhodné.

V IRT je citlivost položky dána parametrem a, jehož hodnota se v praxi pohybuje obvykle mezi 0 a 2,8. Čím větší je hodnota a, tím lépe úloha rozlišuje mezi testovanými nalevo a napravo od své polohy. Graficky se citlivost úlohy projevuje strmostí ICC v jejím prostředním úseku. Čím strmější je křivka v tomto úseku, tím je úloha citlivější. Strmost křivky, a tím také parametr a dosahuje své maximální hodnoty v bodě, ve kterém se úroveň schopnosti θ rovná obtížnosti úlohy. Úlohy se zápornými hodnotami a jsou stejně jako v KTT nežádoucí, značí něco chybného v úloze. Doporučená hodnota parametru a je 0,4 a více, vyjádřeno v normální metrice, či 0,6 a více, vyjádřeno v logicech.

Uhádnutelnost položky

Uhádnutelnost položky je v IRT dána tzv. pseudonáhodným parametrem c, který udává, že testovaný bez jakékoli znalosti odpovědi vyřeší úlohu s výběrem odpovědi správně pouhým uhádnutím. Čím je tento parametr vyšší, tím větší je pravděpodobnost správné odpovědi pouhým hádáním. V KTT něco podobného vyjadřuje skór korigovaný na hádání.

V tab. 11-1 uvádíme shrnutí charakteristik úloh v KTT ve srovnání s parametry úloh v IRT.

Tab. 11-1 Porovnání charakteristik úloh v KTT a IRT

	obtížnost	citlivost	uhádnutelnost
KTT[63]	$p \in \langle 0;1 \rangle$ či q[64] $\in \langle 0;1 \rangle$ - poměr (či procento) testovaných se správnou odpovědí	obvykle r_{pb} / r_b [65] či $d \in \langle 0;1 \rangle$ - korelace úlohy s celkovým skórem[66]	- vyjadřuje se pomocí skóru korigovaného na „hádání" u testů složených z úloh s výběrem odpovědi: $X_{kor} = R - \frac{W}{a-1}$ [67]
	- velké hodnoty p, resp. malé hodnoty q značí **snadné** úlohy - za úlohy optimální obtížnosti se běžně považují úlohy s $0,3 \leq p \leq 0,7$, resp. $0,3 \leq q \leq 0,7$	- velké hodnoty r či d značí úlohy s lepší rozlišovací schopností mezi testovanými - požaduje se $d \geq 0,30$ u dobrých úloh - $r_{pb} \geq 0,30$ či $r_b \geq 0,40$	- používá se velmi málo, protože výzkumy dokazují, že testovaný zpravidla nehádá správnou odpověď, ale snaží se vyloučit (pro něj) nesprávné nabídky - kromě toho příčinou chybné odpovědi nemusí být hádání, ale chyba, které se testovaný dopustil, a to ho znevýhodňuje oproti testovaným, kteří se o odpověď vůbec nepokusili a úlohu vynechali
IRT[68]	$b \in \langle -3;3 \rangle$ (typické hodnoty v praxi, teoreticky však může nabývat hodnot od - ∞ do + ∞)	$a \in \langle 0;3 \rangle$ (obvyklé rozpětí v praxi, teoretický obor hodnot parametru a je od - ∞ do + ∞)	$c \in \langle 0;0,35 \rangle$ (obvyklé rozpětí v praxi, teoreticky se jeho hodnota pohybuje mezi 0 a 1)
	- velké hodnoty b značí **obtížné** úlohy - hodnoty parametrů obtížnosti b se rovnají hodnotám schopnosti θ, jsou tedy měřeny na stejné škále	- čím větší hodnotu má a, tím lépe úloha rozlišuje mezi testovanými - u 2- a 3-parametrového modelu se vyžaduje hodnota parametru $a \geq 1,7$	- c udává, jaká je pravděpodobnost uhádnutí správné odpovědi na všech úrovních osy schopnosti θ
	- b je bod na ose schopnosti θ, pro který pravděpodobnost správné odpovědi je rovna 0,5, u 3-parametrového modelu $(1+c)/2$	- úloha nejlépe rozlišuje mezi testovanými v bodě na ose θ odpovídajícímu parametru obtížnosti b - parametr a je proporcionální ke stoupání ICC[69] v $\theta = b$.	- když $b < 0$ a $a < 1$, pak c není zřejmé (pro široký interval θ by se dolní část grafu charakteristické křivky přibližovala hodnotě c)

Četnost vynechaných odpovědí

Zvláštní pozornost se při položkové analýze věnuje četnosti vynechaných odpovědí. To, že testovaný úlohu vynechá, nemusí vždy znamenat, že nemá vědomosti potřebné k jejímu řešení, ale že mu na řešení úlohy již nezbyl čas. Vynechané úlohy se mnohdy nedají přesně specifikovat, protože pokud se nejedná o adaptivní testování (kde je posloupnost úloh dána), může student řešit úlohy v libovolném pořadí. Může se tedy stát, že úlohy na konci testu nemusel student řešit jako poslední. Protože v některých případech může testovaný vynechat úlohu proto, že neporozuměl dobře zadání úlohy, doporučuje se zkontrolovat každou uzavřenou úlohu, kterou neřešilo více než 20 % testovaných. Položková analýza se provádí pouze u úloh, které řešilo méně než 80% testovaných.

[63] Hodnoty ukazatelů obtížnosti a citlivosti jsou převzaty z Byčkovský (1983).
[64] $q = 1 - p$
[65] Zatímco r_{pb} může dosáhnout max. hodnoty 1, hodnota koeficientu r_b může být vyšší než 1.
[66] Případně korigovaná korelace úlohy s celkovým skórem, kdy do celkového skóru není daná úloha zahrnuta.
[67] kde R je počet správných odpovědí, W je počet chybných odpovědí, a je počet nabízených odpovědí.
[68] Hodnoty pro parametry a, b, c jsou převzaty z Baker (2001).
[69] ICC je charakteristická křivka položky (item caracteristic curve), viz teoretická část, kap. 4

Funkčnost distraktorů v KTT

V poslední době se klade větší důraz i na analýzu toho, jak u úloh s výběrem odpovědi fungují jednotlivé distraktory. Zjišťuje se, zda mezi distraktory není jeden nebo několik, které

- volí jen velmi málo nebo nikdo z testovaných
- volí více lepších než horších žáků
- korelují kladně s testovými výsledky
- jsou voleny výrazně více než správné odpovědi

V posledních třech případech je vhodné zjistit i to, zda klíčovaná odpověď je skutečně správnou odpovědí či zda neexistuje více než jedna správná odpověď.

Kvalitu všech nabízených odpovědí, tj. správné odpovědi a distraktorů můžeme hodnotit také pomocí grafické metody. Graf znázorňuje relativní četnost volby jednotlivých nabídek ve vztahu k hrubému skóru testovaných. I ideálním případě jsou u nejslabších testovaných tyto četnosti podobné. S rostoucím hrubým skórem pak četnost správné odpovědi roste, kdežto četnost distraktorů klesají.

Charakteristiky úloh u variant testu OSP

Podívejme se nyní na obtížnost, citlivost úloh a na funkčnost distraktorů u jednotlivých variant testu Obecné studijní předpoklady. Ve variantách testu OSP se vyskytovaly jak úlohy s málo či vůbec nevolenými distraktory či distraktory, které volilo více lepších než horších testovaných (korelují kladně s jejich hrubým skórem), tak úlohy, v kterých byl distraktor volen více než správná odpověď. Přehled některých takových úloh uvádíme v tab. 11-2. U těchto úloh nás zajímalo znění zadání úloh a nabízené odpovědi. Položkovou analýzu jsme neprováděli u úloh zařazených ke konci testu, které řešilo méně než 75 procent testovaných (viz výše). Ve variantách A a B se jedná celkem o posledních 11 úloh, ve variantě C o posledních 11 úloh s výjimkou úlohy 36 a variantě D o 12 úloh s výjimkou úloh 37 a 42.

Funkčnost distraktorů

Tab. 11-2 Přehled úloh s nefunkčními distraktory

	Úlohy s distraktorem, který		
varianta	nebyl volen nikým nebo jen málo testovanými	který volilo více lepších než horších testovaných	který byl volen častěji než správná odpověď
A	1, 2, 3, 13, 32, 34	26, 32, 33	24
B	1, 2, 3, 6, 7, 13, 15, 32	7	29, 33
C	1, 2, 7, 12, 28, 30, 32	31, 36	
D	1, 2, 5, 9, 13, 15, 21, 33	19, 24	

Ve všech variantách byly úlohy zařazené na začátku testu velmi snadné (hodnota *p* se pohybuje mezi 0,76 a 0,98 s výjimkou úlohy 3 ve variantě D, viz příloha 6). Tyto úlohy byly nejspíš zařazeny na začátek testu k motivaci. Z textu úloh je zřejmé, že ty nejméně volené nebo vůbec nevolené distraktory může testovaný snadno vyloučit. I u ostatních úloh tohoto typu šlo zjevně o nevhodné distraktory. Úlohy, v kterých jeden či více distraktorů volilo více lepších než horších testovaných, tak jako úlohy, kdy testovaní volili častěji distraktor než správnou odpověď, nám připadají bez technických nedostatků. Domníváme se, že to vypovídá pouze o nedostatečných znalostech testovaných a že distraktory jsou naprosto funkční. Dále uvádíme ukázky částí výstupů z programu ITEMAN včetně textů vybraných podezřelých úloh s nefunkčním distraktorem a grafů četností jednotlivých nabídek z programu LERTAP 5.

Varianta A

Úloha 1

Vyberte slovo nebo dvojici slov, která se **nejlépe** hodí do příslušné věty jako celku.

Povrch Evropy je výškově i tvarově _____ , neboť vznikl_____ horotvornými pochody v různých geologických dobách.

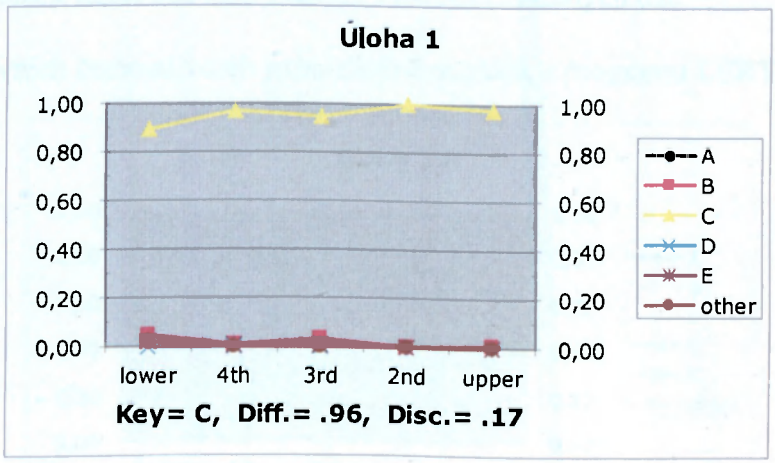
(A) různý – stejnými
(B) podobný – dávnými
(C) rozmanitý – různými
(D) zanedbaný – zbytečnými
(E) zvláštní – rychlými

Výstup z programu ITEMAN (OSP, varianta A, úloha č. 1)

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	0-1	.96	.06	.16	A	.01	.03	.01	-.19	
					B	.02	.03	.01	-.16	
					C	.96	.92	.99	.16	*
					D	.00	.00	.00		
					E	.01	.01	.00	-.15	
					Other	.00	.00	.00	-.12	

Ve výstupu k úloze 1 (viz výše) najdeme kromě hodnoty jejího indexu obtížnosti (p = 0,96), indexu diskriminace (d = 0,06), korigovaného bodově biseriálního koeficientu (r_{pb} = 0,16), také správnou odpověď C (*) a málo volené distraktory D (0,00) a E (0,01). Z textu úlohy i z grafu (viz níže) je zřejmé, že distraktory D a E může testovaný snadno vyloučit jako zcela nesmyslné.

Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Úloha 33

Vaším úkolem je porovnat dvě hodnoty uvedené v rámečku

V loteriích *Superšance* i *Ultrašance* vyhrává pouze ten, kdo uhodne všechna tažená čísla.
V loterii *Superšance* se losuje 6 čísel ze 40, v loterii *Ultrašance* se losuje 37 čísel ze 40.

pravděpodobnost výhry v *Superšanci*	pravděpodobnost výhry v *Ultrašanci*

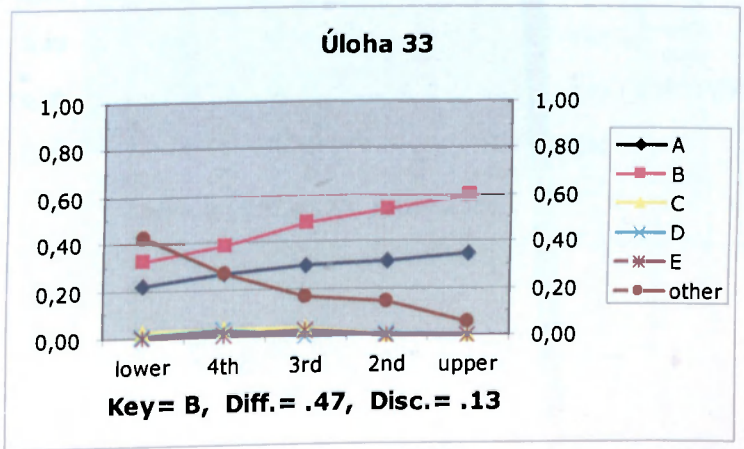
(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

Výstup z programu ITEMAN (OSP, varianta A, úloha č. 33)

	Item Statistics			Alternative Statistics				
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
33	2-3	.47	.37	.17	A	.29	.27	.31	-.09	
					B	.47	.23	.60	.17	*
					C	.02	.03	.01	-.15	
					D	.01	.01	.01	-.07	
					E	**.01**	**.01**	**.01**	**-.02**	
					Other	.21	.00	.00	-.53	

Správnou odpověď B zvolila téměř polovina testovaných, a to častěji ti z lepší skupiny (0,60). Za zmínku stojí distraktor E, jehož použití v tomto typu úloh je poněkud diskutabilní. Logicky uvažující testovaný by měl variantu E vyloučit, neboť možnosti A až D pokrývají všechny situace a právě jedna z nich musí u tohoto typu úloh vždy nastat. Nicméně 1% testovaných tuto variantu zvolilo a navíc použití distraktoru E u této a podobných úloh umožnilo stejný počet variant odpovědí u všech otázek testu, což zjednodušilo i analýzu získaných dat.

Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Varianta B

Úloha 6

Z těchto pěti možností vyberte dvojici, v níž se vztah mezi členy **nejvíce** blíží vztahu v zadané dvojici.

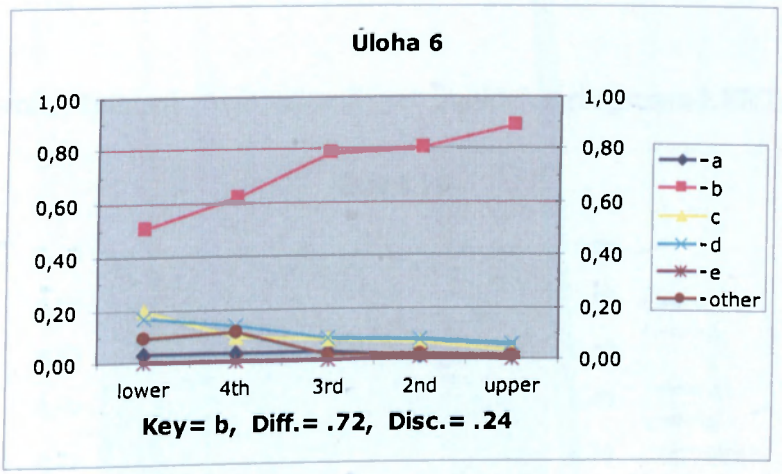
ALFA : OMEGA
(A) prvotní : pokročilý
(B) první : poslední
(C) beta : delta
(D) základní : podstatný
(E) sklep : okap

Výstup z programu ITEMAN (OSP, varianta B, úloha č. 6)

	Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
6	0-6	.72	.44	.24	A	.03	.05	.00	-.16	
					B	.72	.49	.93	.24	*
					C	.09	.18	.03	-.34	
					D	.11	.17	.03	-.25	
					E	**.00**	**.00**	**.00**	**.00**	
					Other	.05	.00	.00	-.27	

Distraktor E nezvolil nikdo z testovaných, neboť ho lze snadno vyloučit jako zcela nesmyslný podobně jako v úloze 1 u varianty A.

Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Úloha 29

Ke každé otázce vyberte tu **nejlepší** z nabízených odpovědí. Pouze jedna odpověď je správná.

Naši předkové vyzkoušeli snad všechny možné kombinace co do počtu partnerů a partnerek. Od polygamie, tedy mnohoženství přes bigamii, monogamii až k polyandrii neboli mnohomužství. Ve zvířecím světě převládá jasně polygamie. Pouze dvě procenta savců žijí v párech. Výjimku představují ptáci, kteří tvoří věrné páry v devadesáti procentech.
Které z následujících tvrzení vyplývá z uvedeného textu?

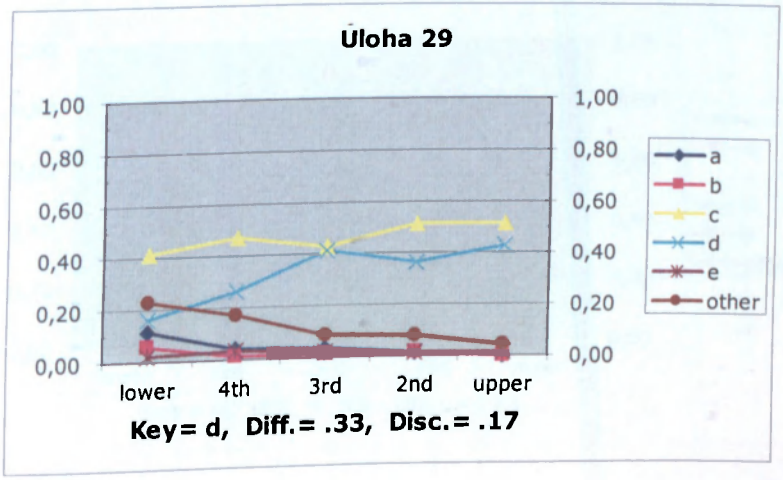
(A) Devadesát procent ptáků žije v polygamii.
(B) Lidé jsou monogamní tvorové.
(C) Devadesát osm procent savců je polygamních.
(D) U většiny zvířat převládá mnohoženství.
(E) Žádní savci nežijí v bigamii.

Výstup z programu ITEMAN (OSP, varianta B, úloha č. 29)

	Item Statistics			Alternative Statistics				
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
29	1-12	.33	.31	.17	A	.04	.08	.02	-.22	
					B	.02	.04	.00	-.18	
					C	**.47**	**.46**	**.46**	**-.17**	
					D	.33	.20	.50	.17	*
					E	.02	.02	.01	-.10	
					Other	.12	.00	.00	-.38	

Téměř polovina testovaných zvolila nesprávnou odpověď C (0,47) což je více než správných odpovědí D (0,33). Nicméně otázku správně odpověděli častěji testovaní z lepší skupiny. Distraktor C splnil svou úlohu velmi dobře. Ti, kdo jej volili, pravděpodobně nesprávně negovali výrok „Pouze dvě procenta savců žijí v párech", jehož negace zní „98% savců žije v polygamii nebo polyandrii".

Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Varianta C

Úloha 2

Vyberte slovo nebo dvojici slov, která se **nejlépe** hodí do příslušné věty jako celku.

Vystupování proti _____ komunistů přineslo téměř všem odpůrcům nejen nepřátelství čelných představitelů režimu, ale i _____ jejich rodin.

(A) hlasům – blahobyt
(B) praktikám – perzekvování
(C) názorům – nesváry
(D) životu – životům
(E) procesům – oblíbenců

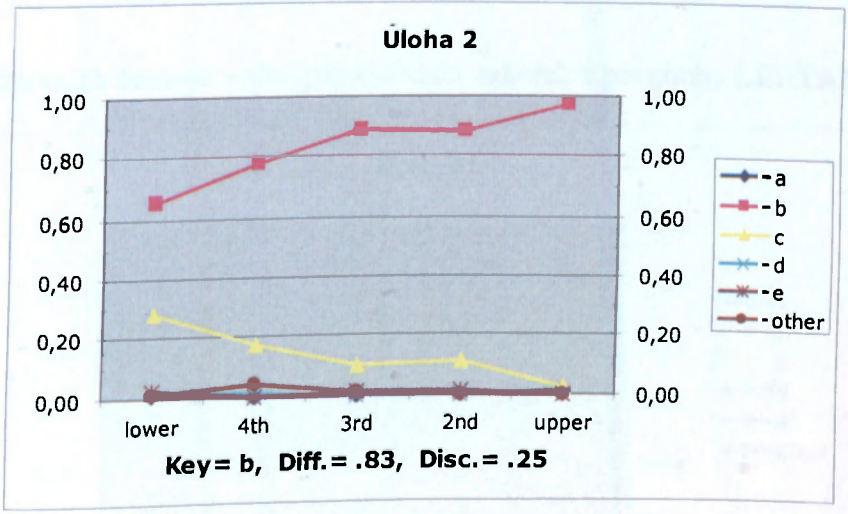
Výstup z programu ITEMAN (OSP, varianta C, úloha č. 2)

Seq. No.	Scale -Item	Item Statistics: Prop. Correct	Disc. Index	Point Biser.
2	0-2	.83	.31	.28

Alt.	Alternative Statistics: Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
A	**.00**	**.01**	**.01**	**-.13**	
B	.83	.63	.95	.28	*
C	.13	.29	.05	-.45	
D	**.01**	**.02**	**.00**	**-.13**	
E	**.01**	**.02**	**.00**	**-.14**	
Other	.01	.00	.00	-.14	

Distraktory A, D a E zvolilo méně než 1% z testovaných, neboť je lze snadno vyloučit jako nesmyslné. Zajímavé je, že celých 13 % (0,13) zvolilo odpověď C (prakticky nikdo z lepší skupiny 5%), která se též nejeví jako příliš smysluplná. Protože v opačném případě by tato otázka ztratila v rámci testu svůj smysl úplně.

Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Úloha 31

Vaším úkolem je porovnat dvě hodnoty uvedené v rámečku.

Je dána kružnice k se středem S a poloměrem r.
Dále jsou dány body A a B, které leží na kružnici k
a délka úsečky AB je rovna $3/2\ r$

úhel ASB	90°

(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

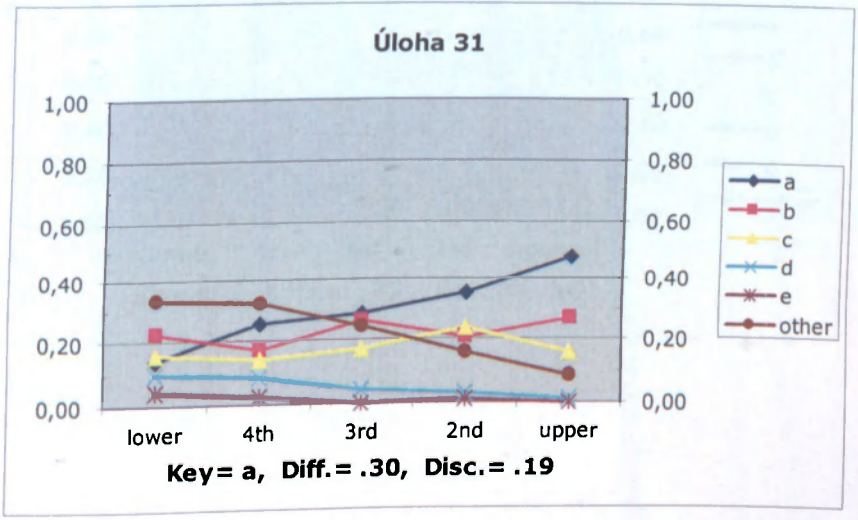
Výstup z programu ITEMAN (OSP, varianta C, úloha č. 31)

Seq. No.	Scale -Item	Item Statistics: Prop. Correct	Disc. Index	Point Biser.
31	2-1	.30	.34	.21

Alternative Statistics: Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
A	.30	.15	.49	.21	*
B	**.23**	**.21**	**.21**	**-.14**	
C	**.17**	**.15**	**.18**	**-.13**	
D	.05	.08	.03	-.16	
E	.01	.04	.01	-.12	
Other	.23	.00	.00	-.43	

Správnou odpověď A zvolilo 30% testovaných, a to častěji ti z lepší skupiny, což je v pořádku. Distraktory B a D splnily svoji funkci, D volilo jen 5% testovaných a o nefunkčnosti distraktoru E platí totéž, co pro úlohu 33 varianty A. Distraktor C volilo sice více testovaných z lepší skupiny (0,18) než z horší (0,15), přesto ale na základě textu úlohy můžeme říct, že šlo o dobrý distraktor. Za zmínku stojí 23% testovaných (other), kteří na tuto otázku raději neodpověděli vůbec. Je to obdobné množství jako u úlohy 33 varianty A (a dalších úloh tohoto typu). Z toho lze usuzovat, že zhruba čtvrtina testovaných považovala tento typ úloh za obtížný nebo si řešením nebyla jista.

Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Varianta D

Úloha 13

K danému slovu vyberte to, které se nejvíce blíží k jeho opačnému významu. Pozor, jde často o odlišení velmi jemných rozdílů.

NAČERPAT

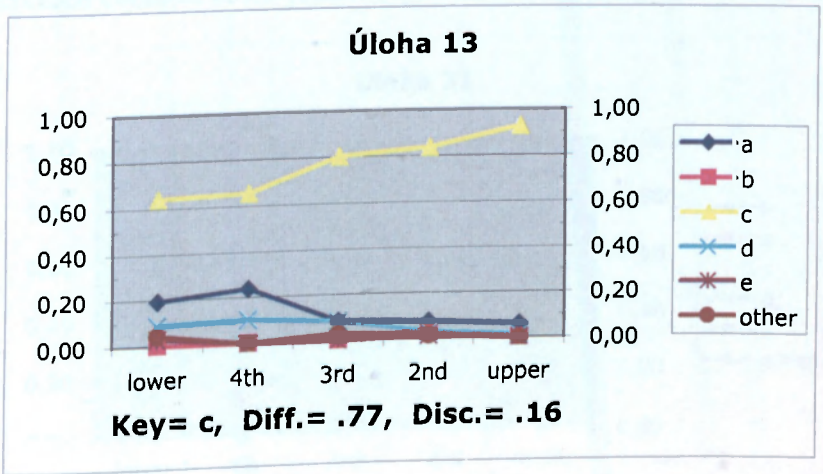
(A) odpustit
(B) dopustit
(C) vypustit
(D) uvolnit
(E) snížit

Výstup z programu ITEMAN (OSP, varianta D, úloha č. 13)

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
13	0-13	.77	.29	.14	A	.13	.18	.03	-.32	
					B	**.01**	**.02**	**.00**	**-.07**	
					C	.77	.64	.93	.14	*
					D	.06	.12	.04	-.25	
					E	**.01**	**.02**	**.00**	**-.12**	
					Other	.02	.00	.00	-.12	

Obdobná situace jako u úlohy 2 varianty C. Distraktory B a E zvolilo 1% z testovaných, neboť je lze snadno vyloučit jako nesmyslné. Odpověď D volilo též velmi málo testovaných (6%). Jediný skutečně funkční distraktor je tudíž A (13% odpovědí).

Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Úloha 31

Vaším úkolem je porovnat dvě hodnoty uvedené v rámečku.

A je počet různých prvočíselných dělitelů čísla 144,
B je počet různých prvočíselných dělitelů čísla 100.

A	*B*

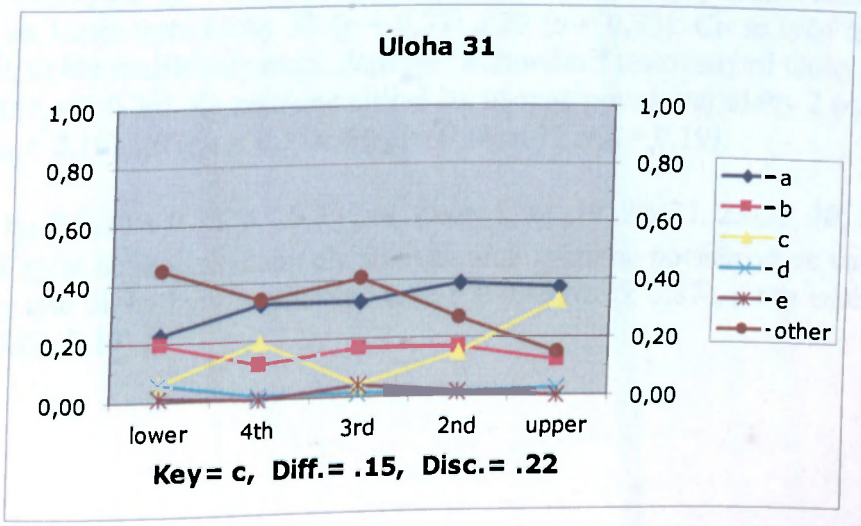
(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

Výstup z programu ITEMAN (OSP, varianta D, úloha č. 31)

		Item Statistics			Alternative Statistics				
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
31	2-1	.15	.24	.25	A	.33	.20	.40	-.04	
					B	.15	.16	.13	-.15	
					C	.15	.06	.30	.25	*
					D	.03	.01	.02	-.04	
					E	.01	.02	.01	-.07	
					Other	.32	.00	.00	-.49	

Správnou odpověď C zvolilo pouze 15% testovaných, a to častěji ti z lepší skupiny. Pro vzorek testovaných byla tato úloha poměrně obtížná. Distraktory A a B splnily svoji funkci. Opět relativně velká část testovaných neodpověděla (32%). Důvodem pro málo správných odpovědí a velký počet vynechaných odpovědí může být nízké povědomí testovaných o pojmu prvočíselný dělitel či způsobu, jak rozložit číslo na součin těchto dělitelů, je to však pouhá domněnka.

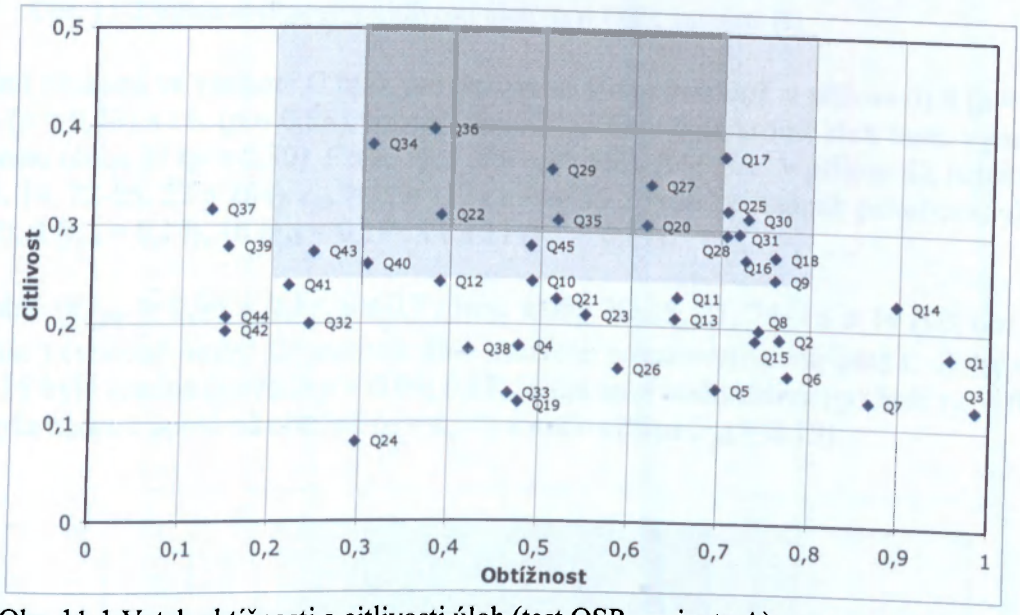
Ukázka grafu relativních četností voleb jednotlivých nabídek z programu LERTAP 5



Obtížnost a citlivost úloh podle KTT

Nejsnadnějšími úlohami ve variantě A byly pro testované úlohy (viz *diff.*[70] v příloze 6) 1 ($p = 0,96$), 3 ($p = 0,98$), 7 ($p = 0,87$) a 14 ($p = 0,90$), naopak nejtěžšími úlohami byly kromě těch vynechaných na konci testu úlohy 32 ($p = 0,24$) a 34 ($p = 0,31$). Co se týče citlivosti úloh (viz *disc.*[71] v příloze 6), dobře rozlišovaly mezi „lepšími" a „horšími" testovanými úlohy 17, 20, 22, 25, 27-31 a 34 (s $r_{pb} \geq 0,30$). Za nejméně citlivé lze naopak považovat úlohy 24 ($r_{pb} = 0,08$), 19 a 3 ($r_{pb} = 0,12$), 7 a 33 ($r_{pb} = 0,13$), 5 ($r_{pb} = 0,14$), 6 ($r_{pb} = 0,15$) a 6 ($r_{pb} = 0,15$).

Na základě obtížnosti a citlivosti úloh můžeme úlohy rozdělit na vhodné, podezřelé a zcela nevhodné (viz kap. 4 a obr. 11-1). Zcela vhodné ($r_{pb} \geq 0,30$ a $0,3 \leq p \leq 0,7$) jsou úlohy 17, 20, 22, 25, 28, 29 a 34 (viz šedý obdélník na obr. 11-1). Za naopak spíše nevhodné úlohy dle statistik úloh můžeme považovat ve variantě A úlohy 7 a 14. Úloha 7 byla snadná ($p = 0,87$), a tím také málo citlivá ($r_{pb} = 0,13$), úloha 14 byla ještě snadnější ($p = 0,90$), ale o trochu citlivější ($r_{pb} = 0,23$). Podezřelé úlohy jsme uvedly výše u funkčnosti distraktorů.

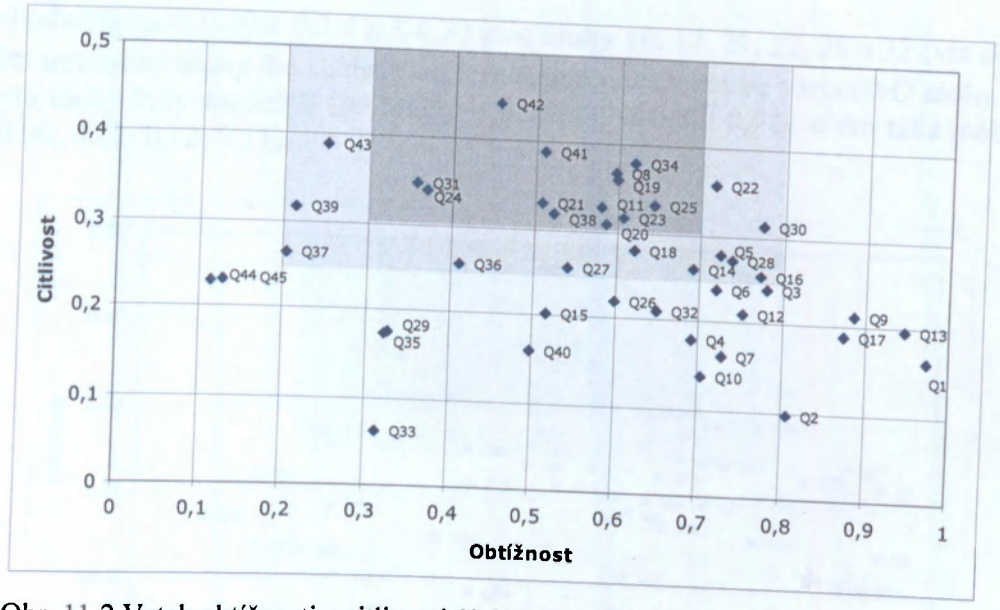


Obr. 11-1 Vztah obtížnosti a citlivosti úloh (test OSP, varianta A)

Nejsnazšími úlohami ve variantě B byly pro testované úlohy (viz *diff.* v příloze 6) 1 ($p = 0,97$), 13 ($p = 0,95$), 9 ($p = 0,89$) a 2 ($p = 0,81$), naopak nejtěžšími úlohami byly kromě těch posledních často vynechávaných na konci testu úlohy 33 ($p = 0,31$) a 29 ($p = 0,33$). Co se týče citlivosti úloh (viz *disc.* v příloze 6), dobře rozlišovaly mezi „lepšími" a „horšími" testovanými úlohy 8, 11, 19, 21, 22-25, 30, 31 a 34 (s $r_{pb} \geq 0,30$). Za nejméně citlivé lze naopak považovat úlohy 2 ($r_{pb} = 0,10$), 10 ($r_{pb} = 0,14$), 1 a 7 ($r_{pb} = 0,16$), 29 ($r_{pb} = 0,17$), 4 ($r_{pb} = 0,18$) a 17 ($r_{pb} = 0,19$).

Zcela vhodné (s $r_{pb} \geq 0,30$ a $0,3 \leq p \leq 0,7$) jsou úlohy 8, 11, 19, 20, 21, 23-25, 30, 31 (viz obr. 11-2). Za naopak spíše nevhodné úlohy dle statistik úloh můžeme považovat ve variantě B úlohy 9, 13 a 17. Všechny tyto úlohy byly snadné (po řadě $p = 0,89$; 0,95; 0,87), a tím také málo citlivé (po řadě $r_{pb} = 0,21$; 0,20; 0,19).

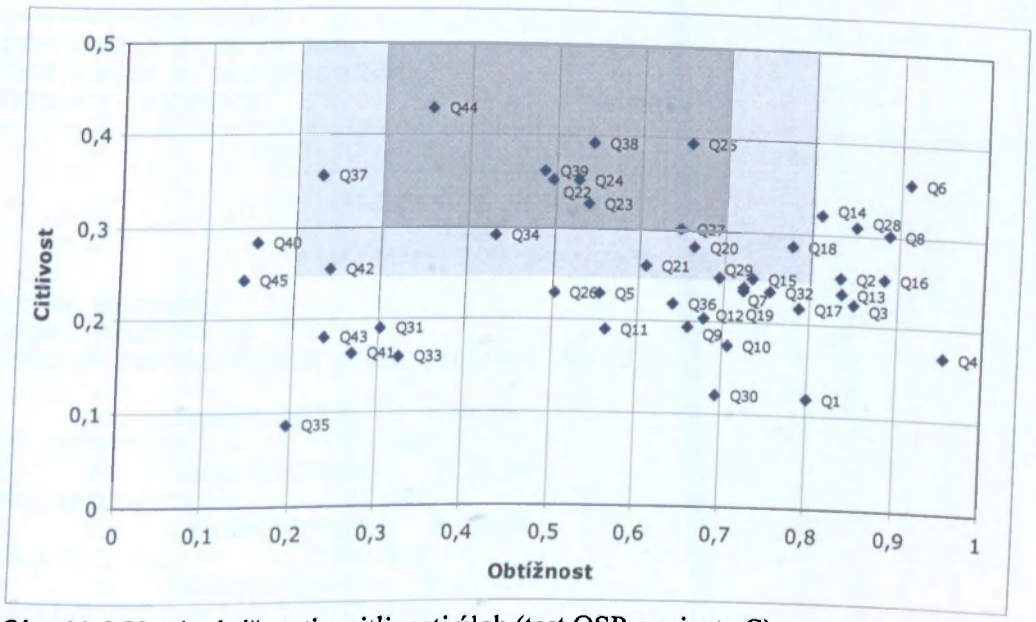
[70] Odpovídá ukazateli obtížnosti *p*.
[71] Odpovídá bodově biseriálnímu koeficientu.



Obr. 11-2 Vztah obtížnosti a citlivosti úloh (test OSP, varianta B)

Nejsnazšími úlohami ve variantě C byly pro testované úlohy (viz *diff.* v příloze 6) 4 ($p = 0,96$), 6 ($p = 0,91$), 8 ($p = 0,89$) a 16 ($p = 0,88$), naopak nejtěžší úlohou byla kromě těch často vynechávaných na konci testu úloha 31 ($p = 0,30$). Co se týče citlivosti úloh (viz *disc.* v příloze 6), nejcitlivější byly úlohy 6, 8, 14, 22-25, 27 a 28 (s $r_{pb} \geq 0,30$). Za nejméně citlivé lze naopak považovat úlohy 1 a 30 ($r_{pb} = 0,12$), 4 ($r_{pb} = 0,17$), 10 ($r_{pb} = 0,18$), 11 a 31 ($r_{pb} = 0,19$).

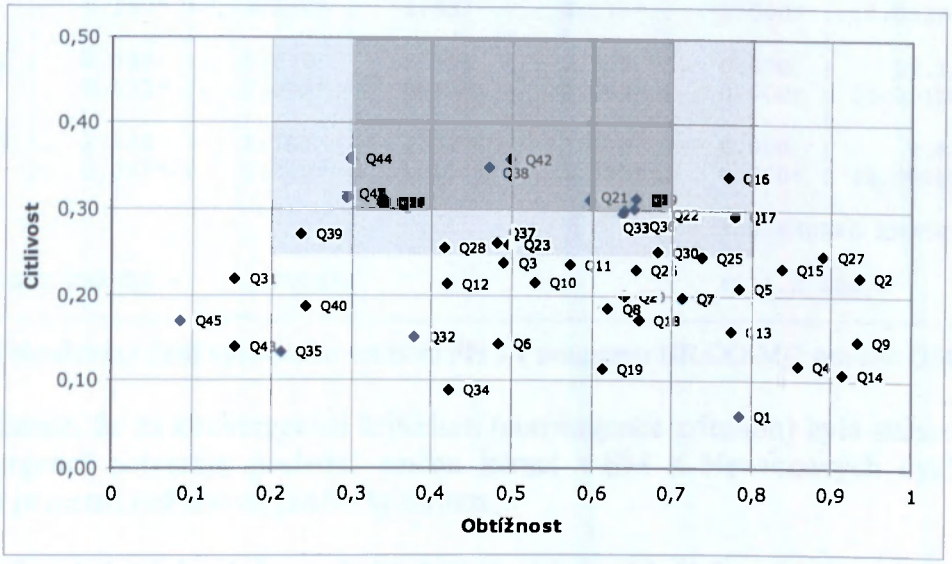
Zcela vhodné (s $r_{pb} \geq 0,30$ a $0,3 < p \leq 0,7$) jsou úlohy 22-25, 27, 34, 18 a 14 (viz obr. 11-3). Za naopak spíše nevhodné úlohy dle statistik úloh můžeme považovat ve variantě C úlohy 4, 16 a 31. Úlohy 4 a 16 byly snadné (po řadě $p = 0,96$; 0,88), a tím také málo citlivé (po řadě $r_{pb} = 0,17$; 0,25), úloha 31 byla naopak poměrně obtížná ($p = 0,30$) a málo citlivá ($r_{pb} = 0,19$).



Obr. 11-3 Vztah obtížnosti a citlivosti úloh (test OSP, varianta C)

Nejsnazšími úlohami ve variantě D byly pro testované úlohy (viz *diff.* v příloze 6) 2 ($p = 0,94$), 9 ($p = 0,93$), 14 ($p = 0,91$) a 27 ($p = 0,89$), naopak nejtěžší úlohou byla kromě opět těch často vynechávaných na konci testu úloha 24 ($p = 0,34$). Co se týče citlivosti úloh (viz *disc.* v příloze 5), nejcitlivější byly úlohy 16, 21, 22, 24, 29 a 33 (s $r_{pb} \geq 0,30$). Za nejméně citlivé lze naopak považovat úlohy 1 ($r_{pb} = 0,06$), 14 a 19 ($r_{pb} = 0,11$), 4 ($r_{pb} = 0,12$), 6 ($r_{pb} = 0,14$), 9 ($r_{pb} = 0,15$), 13 ($r_{pb} = 0,16$), 18 ($r_{pb} = 0,17$) a 8 ($r_{pb} = 0,18$).

Naprosto vhodné (s $r_{pb} \geq 0{,}30$ a $0{,}3 \leq p \leq 0{,}7$) jsou úlohy 16, 17, 21, 22, 29 a 33 (viz obr. 11-4). Za naopak spíše nevhodné úlohy dle statistik úloh můžeme považovat ve variantě D úlohy 1, 2, 4 a 14. Všechny tyto úlohy byly snadnější (po řadě p = 0,78; 0,94; 0,86; 0,91), a tím také málo citlivé (po řadě r_{pb} = 0,06; 0,22; 0,12; 0,11).



Obr. 11-4 Vztah obtížnosti a citlivosti úloh (test OSP, varianta C)

Obtížnost a citlivost úloh v IRT jsme zjišťovali pomocí parametrů b a a za použití 2-parametrového modelu. Hodnoty obou parametrů jsou uvedeny v příloze 7. IRT parametry pro všechny varianty testu OSP získané v kalibračním procesu (který konvergoval na hladině významnosti $\alpha = 0{,}001$, viz obr. 11-5) z programu BIOLOG-MG jsme dále porovnávali s charakteristikami úloh v KTT.

```
 CALIBRATION PARAMETERS
 ======================
 MAXIMUM NUMBER OF EM CYCLES:                25
 MAXIMUM NUMBER OF NEWTON CYCLES:            10
 CONVERGENCE CRITERION:                      0.001
 ------------------------------------------------------------------------------

                        ********************************
                          CALIBRATION OF MAINTEST
                                 TEST0001
                        ********************************

 METHOD OF SOLUTION:
 EM CYCLES (MAXIMUM OF    25)
 FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF 10)

CYCLE     19;    LARGEST CHANGE=    0.00196
 -2 LOG LIKELIHOOD =       23303.008
 CYCLE     20;    LARGEST CHANGE=    0.00168
 -2 LOG LIKELIHOOD =       23302.958
 CYCLE     21;    LARGEST CHANGE=    0.00144
 -2 LOG LIKELIHOOD =       23302.914
 CYCLE     22;    LARGEST CHANGE=    0.00124
 -2 LOG LIKELIHOOD =       23302.876
 CYCLE     23;    LARGEST CHANGE=    0.00106
 -2 LOG LIKELIHOOD =       23302.843
 CYCLE     24;    LARGEST CHANGE=    0.00091

[NEWTON CYCLES]
 -2 LOG LIKELIHOOD:          23302.8116
 CYCLE    25;    LARGEST CHANGE=    0.00083
```

Obr. 11-5 Ukázka části výstupního souboru PH 2 z programu BILOG-MG pro test OSP, variantu A

TEST0001; ITEM PARAMETERS AFTER CYCLE 25

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	LOADING S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF
ITEM0001	2.086	0.609	-3.422	0.520	0.000	5.0	2.0
	0.167*	0.130*	0.637*	0.111*	0.000*	(0.0839)	
ITEM0002	0.788	0.376	-2.098	0.352	0.000	12.5	5.0
	0.072*	0.073*	0.408*	0.069*	0.000*	(0.0515)	
ITEM0003	2.840	0.702	-4.047	0.574	0.000	2.6	0.0
	0.340*	0.245*	1.110*	0.201*	0.000*	(0.0000)	

* STANDARD ERROR

LARGEST CHANGE = 0.000831

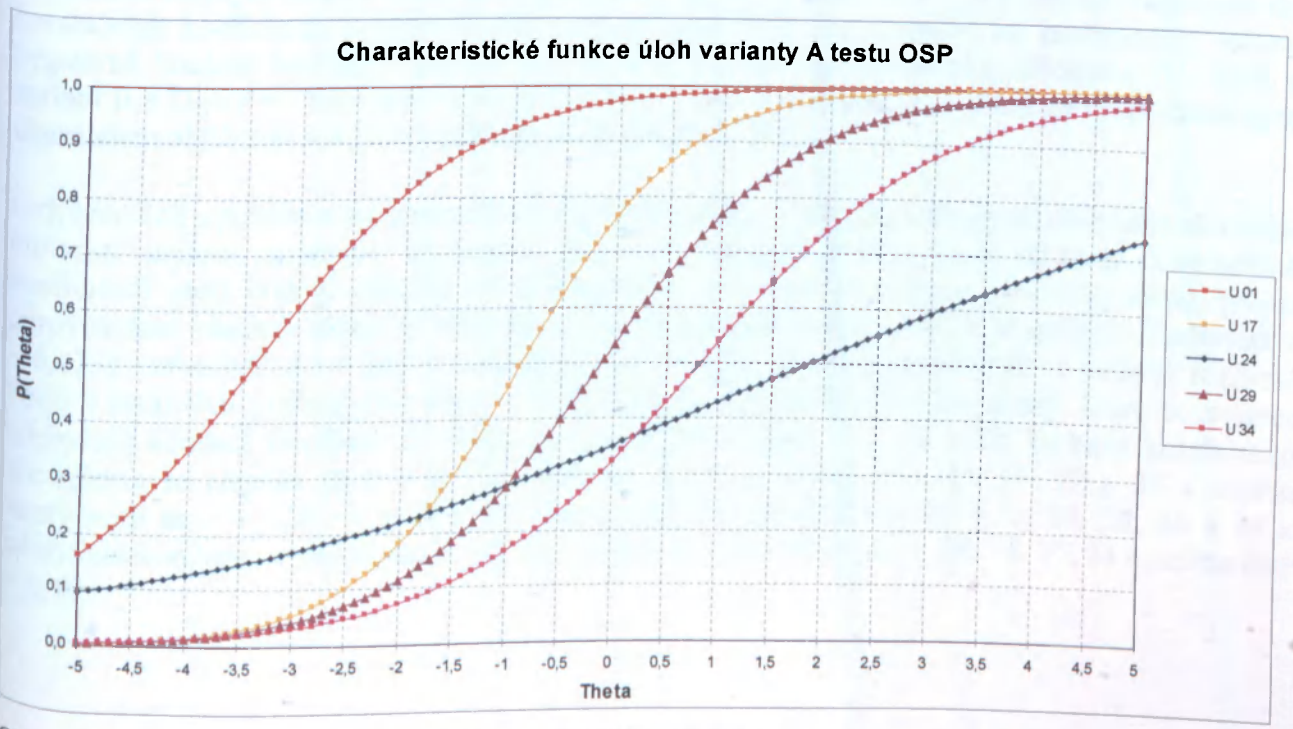
4211.2 198.0

Obr. 11-5 Pokračování části výstupního souboru PH 2 z programu BILOG-MG pro test OSP, variantu A

Z výstupu vyčteme, že za konvergenční kritérium (convergence criterion) byla stanovena hodnota 0,001. Konvergenci potvrzuje poslední změna iterací v EM a Newtonových cyklech (largest change), která je menší než konvergenční kritérium.

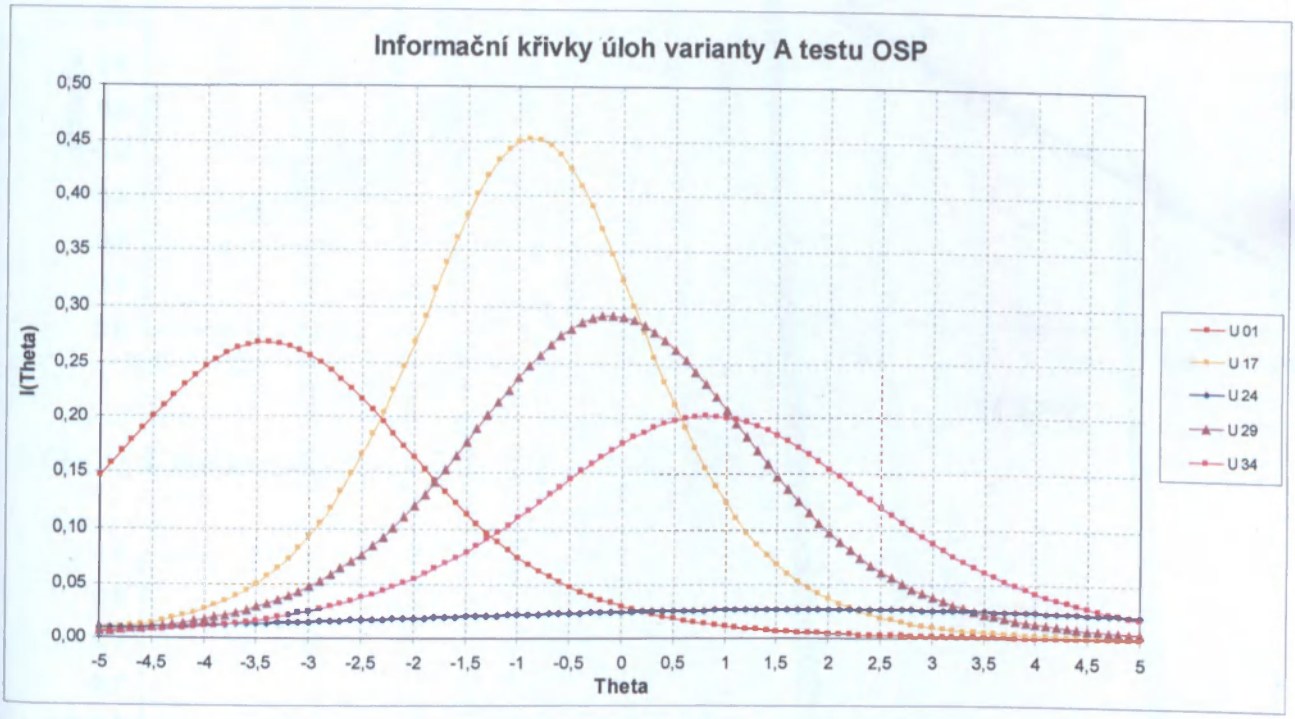
Ukázky charakteristických a informačních křivek vybraných úloh

Na obr. 11-6 jsou charakteristické funkce pěti úloh (1, 17, 24, 29 a 34) varianty A, které se na první pohled liší buď obtížností nebo/ a citlivostí. Z polohy bodu odpovídajícímu 50ti procentní pravděpodobnosti správné odpovědi (P = 0,5) u každé křivky vidíme, že úloha 24 byla nejtěžší (nadprůměrná obtížnost), potom úlohy 34, 29 a 17 a úloha 1 byla velmi snadná. Co se týče citlivosti úloh, nejméně strmá je charakteristická křivka úlohy 24, proto je tato úloha nejméně citlivá. Úlohy 1, 29 a 34 jsou téměř stejně obtížné (zhruba paralelní) a úloha 17 rozlišuje nejlépe ze všech. Samozřejmě jednotlivé úlohy nerozlišují stejně pro všechny úrovně schopnosti theta (vodorovná osa). K té stejné klasifikaci úloh podle obtížnosti a citlivosti dospějeme i při porovnání informačních funkcí těchto úloh (viz obr. 11-7).



Obr. 11-6 Charakteristické funkce vybraných pěti úloh z varianty A

Z polohy nejvyššího bodu informačních křivek úloh vzhledem k vodorovné (theta) i svislé ose (množství informace) je zřejmé, že úloha 17 výborně rozlišuje, především pro schopnost θ zhruba mezi -2 a 0,5. Velmi mizerná je úloha 24, která není vůbec citlivá. Průměrně obtížné je úloha 29.



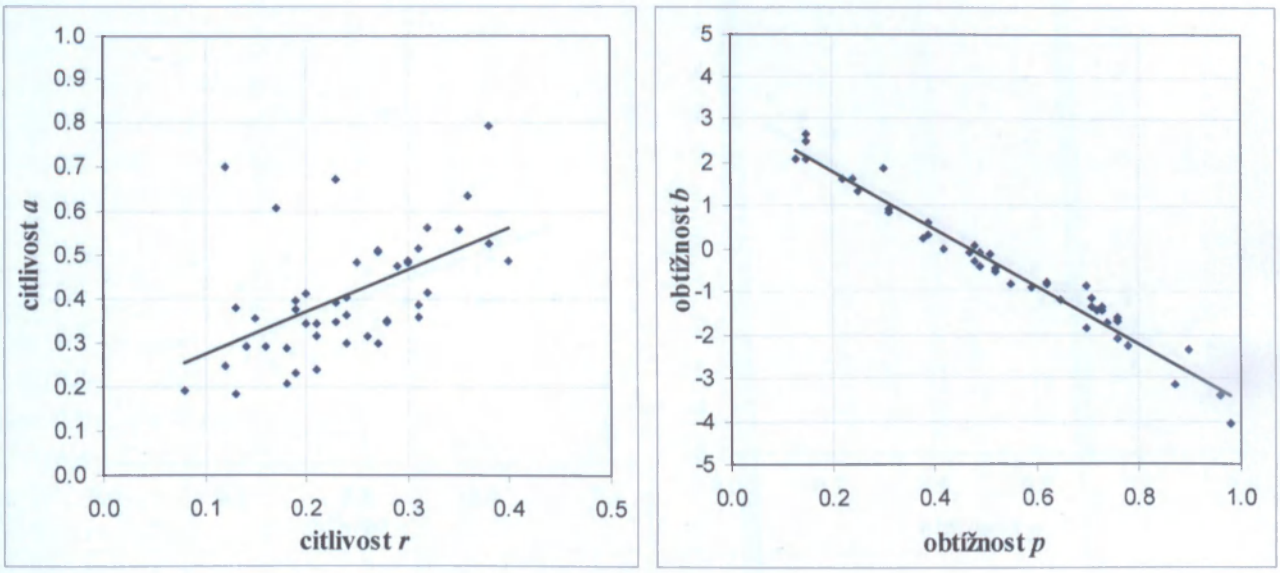
Obr. 11-7 Informační funkce vybraných pěti úloh z varianty A

Pro zajímavost uvádíme v přílohách 8 až 11 charakteristické a informační křivky všech úloh ve všech čtyřech variantách testu OSP.

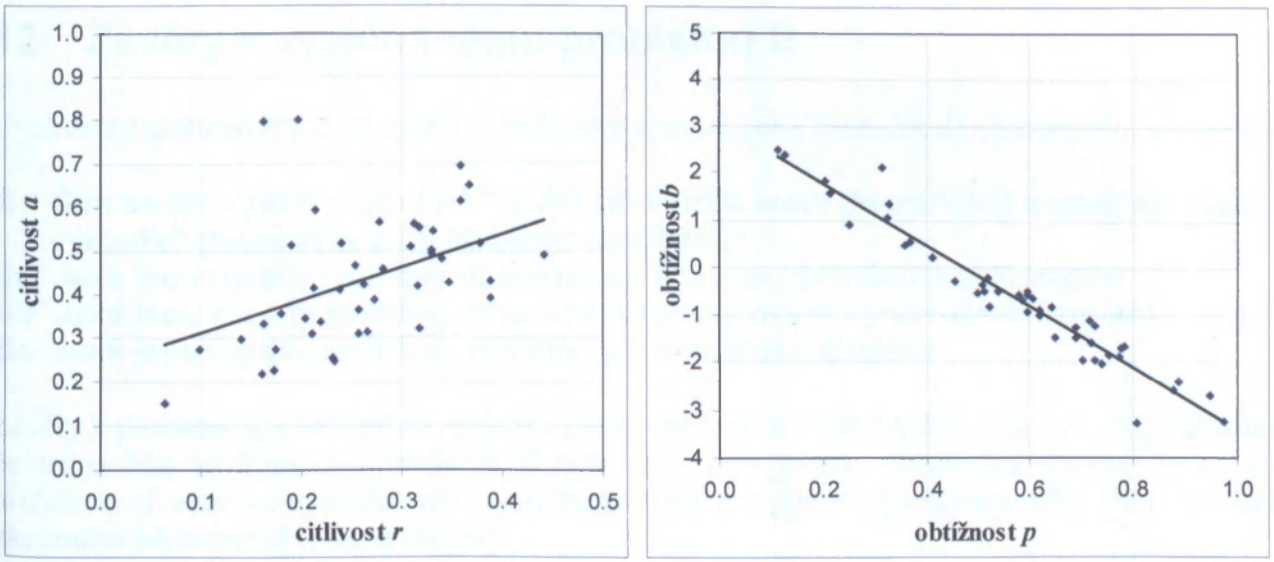
Porovnání obtížnosti a citlivosti úloh v KTT a IRT

Pro porovnání ukazatele obtížnosti *p* v KTT s parametrem obtížnosti *b* v IRT jsme použili korelační analýzu. Z rozptylů bodů v bodových grafech (viz obr. 11-8 až 11-11) a z hodnot Pearsonových korelačních koeficientů u jednotlivých variant testu OSP lze usoudit, že koeficient *p* vysoce negativně lineárně koreluje s parametrem *b*. U varianty A je korelační koeficient r = - 0,98, u variant B a D je r = - 0,97 a u varianty C je r = - 0,96. To znamená, že existuje těsný vztah mezi ukazatelem obtížnosti v KTT a parametrem obtížnosti v IRT.

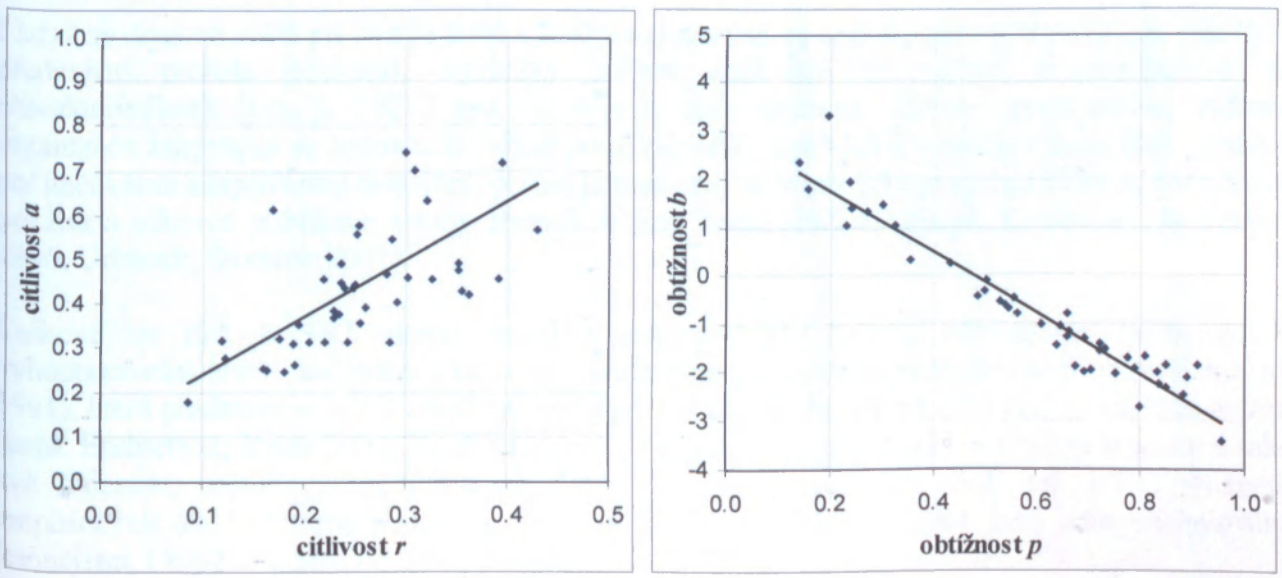
Citlivost úloh s použitím obou metod nebyla v tak těsném vztahu jako obtížnost úloh (oba ukazatele citlivosti nejsou navzájem významně pozitivně lineárně korelovány). Hodnoty korelačních koeficientů mezi bodově biseriálním koeficientem r_{pb} a parametrem *a* jsou dost nízké, pokud porovnáváme všechny úlohy. U varianty A je korelační koeficient r = 0,53, u varianty B dokonce r = 0,43, u varianty C je r = 0,61 a u varianty D je r = 0,57. To je názorně vidět i z velkých rozptylů bodů v bodových grafech (viz obr. 11-8 až 11-11). Vynecháme-li však úlohy, které se nejvíce odchylují, korelaci zlepšíme. U varianty A se jedná o úlohy 1, 3, 14 a 17, hodnota korelačního koeficientu se zlepšila na r = 0,77, u varianty B jde o úlohy 1, 9, 13, 17, 22 a 33 a změnu koeficientu na r = 0,70, u varianty C jsou se nejvíce odchylují úlohy 4, 8, 28, 38, 40 a 44 a koeficient korelace se zlepšil na r = 0,72 a u varianty D jde o úlohy 2, 9, 16, 17, 27, 44 a změnu na r = 0,74.



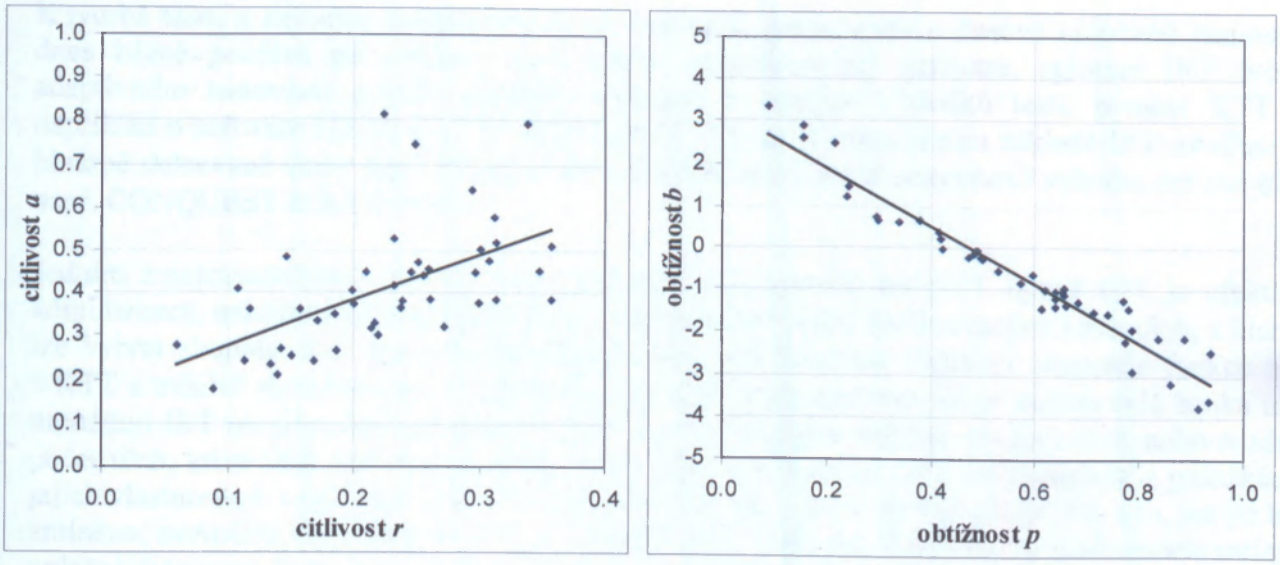
Obr. 11-8 Bodové grafy mezi a a r_{pb}, resp. b a p (varianta A)



Obr. 11-9 Bodové grafy mezi a a r_{pb}, resp. b a p (varianta B)



Obr. 11-10 Bodové grafy mezi a a r_{pb}, resp. b a p (varianta C)



Obr. 11-11 Bodové grafy mezi a a r_{pb}, resp. b a p (varianta D)

12 Závěry k výzkumnému problému B

Výzkumný problém B a další otázky, které s ním souvisí, jsme formulovali následujícím způsobem:

B Čím se liší využití metod KTT a IRT při tvorbě testu (teoreticky) a analýze jeho výsledků (teoreticky a při konkrétní aplikaci)?
Ba Jaké jsou rozdíly zmíněných metod při analýze výsledků testovaných
Bb Jaké jsou rozdíly zmíněných metod při posuzování vyrovnanosti variant
Bc Jaké jsou rozdíly zmíněných metod při položkové analýze

Závěry k problému B uvádíme na základě zjištění uvedených v kapitolách 4, 10 a 11. Připomínáme, že jedna část výzkumného problému B byla zkoumána pouze v teoretické rovině, druhá část problému, tj. analýza výsledků testu upřesněná třemi výzkumnými otázkami (Ba, Bb a Bc) byla zkoumána jak teoreticky, tak empiricky.

Závěry vyplývající z teoretického srovnání

I když se dnes ve světě při tvorbě testů a analýze testových výsledků využívá čím dál tím více IRT, především protože poskytuje výsledky měření nezávislé na vzorku testovaných a na charakteristikách úloh, je i KTT pro své výhody stále užitečná. Zatímco profesionální světové organizace zabývající se testováním běžně používají IRT například k vytváření bank úloh a také k počítačovému adaptivnímu testování, pokud je nám známo, u nás při vývoji testů IRT nebyla dosud použita a odborné publikace s touto tématikou jsou zcela ojedinělé (např. Komenda; Mazuchová 1995; Urbánek; Šimeček 2001).

Celkově lze říci, že IRT nabízí nesmírný potenciál pro tvorbu, administraci, skórování a vyhodnocování testu. Na jednu stranu je považována za moderní metodu (např. Hambleton aj. 1991), která představuje velké zlepšení oproti přístupům založeným na KTT, jak uvádí řada autorů (např. Embretson; Reise 2000; Hambleton aj. 1991, Sands aj. 1997), ale na druhou stranu má také svá omezení, protože je založena na silných předpokladech. Jak IRT, tak KTT používají empirických dat ve stejné podobě s tím rozdílem ale, že pomocí IRT jsou data analyzována jemnějším, i když mnohem náročnějším způsobem.

K tvorbě testu a zejména analýze testových výsledků, protože jde o časově náročnou činnost, se dnes běžně používá při aplikaci obou metod specializovaný software, aplikace IRT (včetně adaptivního testování) použití počítače vyžadují. K analýze výsledků testu pomocí KTT jde například o software ITEMAN, LERTAP 5 či TESTFACT, k analýze na základě IRT uveďme pro binárně skórované úlohy např. BILOG-MG a PARSCALE, pro vícestupňově vyhodnocované úlohy např. CONQUEST či MULTILOG.

Jedním z nejzásadnějších rozdílů mezi testováním založeném na KTT oproti IRT je efektivita administrace, nahlížení na test a zakládání bank úloh (vytváření kalibrovaných bank úloh, z kterých lze vybrat skupinu úloh pro každého testovaného individuálně). Zatímco ukazatelé funkce testu v KTT a zvláště skórování testu vycházejí z toho, že všem testovaným je zadána celá banka úloh, umožňují IRT metody, aby testovaným byly zadány naprosto odlišné soubory úloh nebo rozdílný počet úloh, které však měří stejný latentní rys nebo-li schopnost. IRT totiž uvažuje o položkách a jejich vlastnostech samostatně z hlediska latentního rysu, který mají položky měřit, a to, jak již bylo zmíněno, nezávisle na testu a na souboru testovaných. Test považuje IRT za soubor samostatných položek (viz např. Hambleton; Swaminathan; Rogers 1991, Baker 2001). Měřený rys je odhadován jako úroveň schopnosti (theta), je součástí IRT modelu popisujícího odpovídání testovaných na položky testu. Odhady schopnosti jsou nezávislé na použitých úlohách v testu. To umožňuje, aby jedinci se stejným skórem dosáhli různých úrovní schopnosti.

Oproti tomu KTT nahlíží na položky a jejich vlastnosti v kontextu konkrétního testu, je orientovaná na test, položky nejsou od celku testu oddělitelné, protože jsou korelovány s celkovým skórem. Jsou tedy závislé na souboru testovaných, kterým byly zadány. Odhad měřeného rysu (pravý skór, který je z praktických důvodů často transformován na z-skór, Baker 2001) testovaného vyplývá přímo ze skóru celého testu a nedovoluje žádné úvahy o odpovědích testovaných na položku. Nelze tedy předpokládat jako v případě IRT, jak testovaný v úloze odpoví.
Další předností IRT je to, že poskytuje účinné nástroje pro testování tzv. specifických chyb v úlohách (jejich zaujatost např. vůči demografickým rozdílům, pohlavím apod.; item bias) pomocí tzv. DIF funkce úlohy (differential item functioning)[72]. Oproti tomu metody založené na KTT jsou v hodnocení specifických typů chyb v úlohách zásadně limitovány (např. Drasgow 1987). V podstatě takové metody nemohou rozlišit mezi situací, kdy podskupiny mají různé průměry a test je zaujatý, a situací, kdy se průměry liší, ale test není zaujatý (takže jedna skupina má skutečně vyšší průměr v testu).

Pojetí reliability v KTT je zhruba analogické k pojetí informace v IRT ve smyslu, že vyšší hodnoty udávají lepší přesnost měření (nebo-li menší chybu měření). V KTT je koeficient reliability vysoký pro úlohy s dobrou citlivostí a obtížností (Hopkins 1998). V IRT jsou vyšší hodnoty informace asociovány s vyššími hodnotami parametrů citlivosti (a) a malými hodnotami parametru hádání (c). Obdobně jako je standardní chyba odhadu určitého skóru θ v IRT v inverzním vztahu k informaci, je standardní chyba měření v KTT v inverzním vztahu s reliabilitou testu. Čím více informace test na dané úrovni schopnosti při aplikaci IRT poskytuje, tím menší je chyba, s níž je úroveň schopnosti odhadována. Zásadním rozdílem mezi IRT a KTT je však to, že KTT předpokládá stejnou standardní chybu pro všechny skóry v testu a reliabilitu testových výsledků. Chyba je závislá na souboru testovaných, ale ne na úrovni měřeného rysu (vyplývá z její definice). Oproti tomu v IRT je standardní chyba proměnlivá pro všechny skóry a není závislá na souboru testovaných. Umožňuje zobecnění na různé populace.

Další výhodou IRT oproti KTT je to, že obtížnost položky a schopnost testovaných jsou nezávislé na sobě měřeny na téže škále, což umožňuje vybrat či odebrat úlohy odpovídající určité úrovni schopnosti testovaných, a tím vytvářet zcela paralelní varianty jednoho testu. IRT vypočítává podmíněné chyby měření založené na informační funkci testu pro všechny úrovně schopnosti, a tak

[72] Jak identifikovat zaujatost úloh uvádějí např. Hambleton aj. (1991), Baker (2001), Embretson; Reise (2000) aj.

můžeme z banky úloh vybrat či odebrat ty úlohy, které poskytují maximální či minimální množství informace pro konkrétní úroveň schopnosti.

Obecně lze říci, že IRT parametry jsou přesnější, protože nezávisí na charakteristikách skupiny testovaných, ale pokud použijeme velké a reprezentativní vzorky testovaných z reálné populace, jsou obě metody (CTT a IRT) rovnocenné, což dokládají i četné výzkumy v této oblasti (např. Stage 1997 a 2003, Bechger aj. 2003). Také náš malý výzkum na testu OSP ukázal velmi dobrou korelaci mezi ukazateli obtížnosti KTT a IRT, u citlivosti se korelace staly významnějšími až po vynechání nejvíce se vychylujících položek.

Nevýhodou IRT je to, že vyžaduje speciální počítačové programy a zkušenost práce s nimi, dále zapojení pracovníků, kteří mají rozsáhlejší zkušenosti se statistickou analýzou a jejím prováděním v počítačovém prostředí.

Dalším omezením IRT modelů je to, že vyžadují velké soubory testovaných a jsou vázány na určitý minimální počet úloh tvořících test k získání přesných a stabilních odhadů položkových parametrů. Suen (1990) uvádí, že zpravidla stačí 20-30 úloh a 150-200 testovaných u 1- parametrového modelu, 30 úloh a 500 testovaných u 2-parametrového modelu a 60 úloh a 1000 testovaných u 3-parametrového modelu. Oproti tomu KTT vyžaduje menší reprezentativní vzorky testovaných, což je částečně důležité při pilotáži testu.

Kromě toho stanovení položkových parametrů je v IRT poměrně komplikované a rozpoznání špatných položek je dokonce složitější než u KTT. Výhodou KTT je oproti tomu to, že používá relativně jednoduché matematické postupy.

Zásadním omezením IRT je silný předpoklad unidimenzionality banky úloh u IRT modelů (která způsobuje lokální nezávislost úloh), protože v praxi zpravidla nikdy není rozumný počet úloh perfektně unidimenzionální (Baker 2001). Kromě toho matematický popis většiny IRT modelů logistickými charakteristickými křivkami je velmi omezující. IRT předpokládá, že test bude odpovídat nějakému matematickému modelu, ale toho v praxi nelze vždy dosáhnout.

Test OSP nebyl obsahově homogenní, tudíž nebyla dodržena unidimenzionalita, a tím některé úlohy nevyhovovaly námi zvolenému 2-parametrovému modelu (zjišťuje se χ^2- testem). Ani po zvolení přesnějšího 3- parametrového modelu či navýšení intervalů schopnosti pro výpočet χ^2 statistik se situace nezlepšila. Protože ale proces kalibrace položkových parametrů konvergoval, použili jsme parametry k jejich porovnání s ukazateli v KTT.

Ba Jaké jsou rozdíly zmíněných metod při analýze výsledků testovaných
Bb Jaké jsou rozdíly zmíněných metod při posuzování vyrovnanosti variant

Při srovnání technických charakteristik testu získaných při analýze testových výsledků založené na KTT (pomocí programů ITEMAN a LERTAP 5) a analýze založené na IRT (pomocí programu BILOG-MG) jsme zjistili jejich téměř naprostou shodu. Údaje o tom jsou v tabulkách 10-2, 10-3 a na obrázcích 10-3 až 10-10. Srovnání jsme prováděli u výsledků čtyř variant testu OSP, z nichž každou bylo testováno 339 studentů.

Pokud jde o posouzení vyrovnanosti variant testu OSP, lze říci, že nehledě na použité metodě (KTT či IRT) jsme došli ke stejnému výsledku, že varianty testu OSP můžeme považovat za zhruba vyrovnané, i když se od sebe trochu liší. Varianty A a D jsou poněkud těžší a citlivější, ale přitom jejich realiabilita je trochu nižší než u variant B a C. Reliabilita ani jedné varianty není ale příliš vysoká, což je zřejmým nedostatkem všech analyzovaných variant testu OSP, především proto, že výsledky testu mají být podkladem pro závažné rozhodnutí přijetí či nepřijetí na vysokou školu.

V rámci výzkumné otázky Ba byla sledována ještě těsnost vztahu mezi korigovanými a nekorigovanými skóry. Těsnost vztahu byla velmi vysoká. U variant A, B, C a D bylo po řadě $r = 0{,}982$; 0,985; 0,985; 0,980. Je však nutné poznamenat, že pokud by výsledky testu nebyly ovlivněny faktorem času (část studentů test ve stanovené době nedokončila), byly by příslušné korelační koeficienty poněkud nižší. Použití korekce skórů je však sporné, neboť na jednu stranu se jí eliminuje náhodně dosažený výsledek v úlohách, na druhou stranu to však některé testované neoprávněně poškozuje. Kromě toho se s počtem nabízených odpovědí pravděpodobnost uhádnutí snižuje (v našem případě šlo o pět nabídek). To, že se v případě použití korigovaného skóru odebírá část bodu za chybnou odpověď, může studenty odrazovat od řešení (i hádání) úloh. Ti, co se úlohu alespoň pokusili řešit a udělali v ní malou chybu, která je stála ztrátu části bodu, je znevýhodňuje oproti těm, co se o odpověď vůbec nepokusili.

Bc Jaké jsou rozdíly zmíněných metod při položkové analýze

Obecně lze říci, že jsou IRT parametry přesnější, protože nezávisí na charakteristikách skupiny testovaných, ale pokud použijeme velké a reprezentativní vzorky testovaných z reálné populace, jsou obě metody (CTT a IRT) rovnocenné, což dokládají i četné výzkumy v této oblasti (např. Stage 1997 a 2003, Bechger aj. 2003).

Také náš malý výzkum na testu Obecné studijní předpoklady (OSP) ukázal velmi dobrou korelaci mezi ukazateli obtížnosti KTT a IRT, i když v případě citlivosti se korelace staly významnějšími až po vynechání nejvíce se vychylujících položek.

Jednou z nevýhod položkové analýzy založené na IRT je i to, co platí o celé IRT. Využívání IRT při položkové analýze, ať již při terénním ověřování testových úloh nebo při navrhování vyrovnaných testových variant, vyžaduje od těch, kteří testy konstruují, značné teoretické znalosti i zkušenost s využíváním vhodných IRT modelů a vhodných software.

DOPORUČENÍ

Bývá zvykem končit výzkumné studie i disertační práce dvěma krátkými kapitolami: „závěry" a „doporučeními k dalšímu výzkumu". Tuto tradici jsem nedodržela ze dvou důvodů. Téma práce bylo poměrně široké, a proto závěry byly již uvedeny na dvou jiných místech práce: v kap. 8 a kap. 12. „Doporučení k dalšímu výzkumu" jsem pak neuvedla z toho důvodu, že výzkum v této oblasti, tím mám na mysli především využívání item response theory i některé testové aplikace uváděné v práci, např. adaptivní testování, nejsou u nás téměř známy. Navíc se mi zdá, že tvorba testů, jejichž výsledky jsou podkladem pro závažná rozhodnutí (např. přijetí na vyšší druh školy), není založena na teorií a praxí ověřených postupech a mnohdy je ovlivněna komerčními zájmy některých institucí. Pokusím se alespoň v bodech uvést některá doporučení pro rozvoj v oblasti testování, a to nejen testování pomocí počítačů:

- vychovat mladé odborníky v oboru testování (mohlo by to být realizováno v magisterském a doktorském studiu na pedagogických i jiných fakultách VŠ)
- zprostředkovat kontakty se zahraničními odborníky v této oblasti (zejména USA, Velká Británie, Nizozemí a Švédsko)
- zajistit dostatek zahraniční odborné literatury
- umožnit těm, kteří v oblasti testování pracují, i mladým odborníkům studijní pobyty na významných institucích (ETS, ACT, CITO, CRESST) a účast na zahraničních konferencích
- rozhodnutí týkající se využití závažných testů by měla být založena na vyjádření odborníků v této oblasti a nikoli na čistě politických rozhodnutích

Byla bych ráda, kdyby tato práce byla alespoň malým podnětem ke zdokonalení stavu v oblasti testování u nás, zejména pak při seznámení s IRT, adaptivním testováním a jejich praktickým využitím.

Literatura:

ADELSBERGER, H. H.; COLLIS, B.; PAWLOWSKI, J. M. (Eds.). *Handbook on Information Technologies for Education and Training*. Berlin; Heidelberg; New York : Springer Verlag, 2002. ISBN 3-540-67803-4.

AERA; APA; NCME. *Standardy pro pedagogické a psychologické testování* (český překlad). Praha : Testcentrum, 2001. ISBN 80-86471-07-1.

ALESSI, S. M.; TROLLIP, S. R. *Multimedia for Learning: Methods and Development*. 3rd ed. Needham Heights, MA : Allyn & Bacon, 2001. ISBN 0-205-27691-1.

ANDERSON, L.W.; KRATHWOHL, D.R. (Eds.). A *Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York : Addison Wesley Longman, 2001a. ISBN 0-321-08405-5.

ANDRICH, D. A rating formulation for ordered response categories. *Psychometrika*, 1978, no. 43, s. 561-573.

ASSESSMENT SYSTEM CORPORATION. *User´s manual for the MicroCAT testing system*. (Version 3). St. Paul, MN : Author, 1988.

ATKINSON, R. C. Computerized instruction and the learning process. In: ATKINSON, R. C.; WILSON, H. A. Computer-assisted instruction: A book of readings. New York : Academic Press, 1969.

ATTALI, Y.; BURSTEIN, J. Automated Essay Scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment* [on-line]. Boston College : Technology and Assessment Study Collaborative, 2006, roč. 4, č.3 [cit. 1. 4. 2007]. ISSN 1540-2525. Dostupné na:<www.jtla.org>.

BAKER, F. *The Basics of Item response Theory*. Portsmouth, NH : Heinemann, 1985.

BAKER, F. *The Basics of Item response Theory*. 2nd ed. [on-line]. ERIC Clearinghouse on Assessment and Evaluation, 2001 [cit. 1. 4. 2007]. ISBN 1-886047-03-0. Dostupné na:<www.ericae.net/irt/baker>.

BAKER, F. B., KIM S. *Item Response Theory. Parameter Estimation Techniques*. New York : Marcel Dekker, Inc., 2004. ISBN: 0-8247-5825-0.

BAŽANTOVÁ, Z. Využití revidované Bloomovy taxonomie ve školní praxi. In *Sborník 14. konference ČAPV: Současné metodologické přístupy a strategie pedagogického výzkumu* [CD-ROM]. Plzeň: Západočeská univerzita v Plzni, 2006. ISBN 80-7043-483-X.

BAŽANTOVÁ, Z.; BYČKOVSKÝ, P. Tradiční a nové metody položkové analýzy. In *Sborník z mezinárodní konference Pedagogická evaluace ´06* [CD-ROM]. Ostrava: Pedagogická fakulta OU, 2006. ISBN 80-7368-272-9.

BECHGE, T.; GUNTER, M.; HUUB, H.; BÉGUIN, A. Using classical test theory in combination with item response theory. *Applied psychological measurement*, 2003, vol. 27, no. 5, s. 319-334.

BEUCHERT, A. K.; MENDOZA, J. L. A Monte Carlo comparison of ten item discrimination indices. *Journal of Educational Measurement*, 16, 1979, s. 109-117.

BINET, A.; SIMON, Th. A. Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, s. 191-244, 1905.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee´s ability. In LORD, F. M.; NOVICK, M. R. *Statistical theories of mental test scores*. Reading, MA : Addison-Wesley, 1968.

BLOOM, B.S. (Ed.). *Taxonomy of Educational Objectives, The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York: David McKay, 1956. ISBN 0-679-3029-3.

BOTZ, J.; LIENERT, G. A.; BOEHNKE, K. *Verteilungsfreie Methoden in der Biostatistik*. Berlin : Springer, 2000. ISBN 978-3540-67590-7.

BUGBEE, A.C.; BERNT, F.M. Testing By Computer: Findings in Six Years of Use 1982-1988. *Journal of Research on Computing in Education*, 1990, roč. 23, č. 1, s. 87-100.

BUNDERSON, C. V.; INOUYE, D. K.; OLSEN, J. B. The Four Generation of Computerized Educational Measurement. In LINN, R. L. (Ed.) *Educational Measurement*. 3rd ed. New York : Macmilan Publishing Company, 1989. ISBN 0-02-922400-4.

BURJAN, V. Tvorba a využívanie školských testov [online]. In *EXAM-info*, č. 1-7. Bratislava, 1999-2005 [cit. 1. 4. 2007]. Dostupné na: <www.exam.sk>.
BURSTEIN, J. The e-rater scoring engine: Automated Essay Scoring with natural language processing. In SHERMIS, M. D.; BURSTEIN, J. C (Eds.). *Automated Essay Scoring: A cross disciplinary approach*. Mahwah, NJ : Lawrence Erlbaum Associates, 2003.
BURSTEIN, J.; KUKICH, K.; WOLFF, S.; LU, C.; CHODOROW, M. *Computer analysis of essays*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA, 1998.
BYČKOVSKÝ, P. *Základy měření výsledků výuky: Tvorba didaktického textu*. Praha: ČVUT, 1983.
BYČKOVSKÝ, P.; BAŽANTOVÁ, Z. Tvorba tradičních a netradičních testových úloh. In *Sborník Národní konference o počítačích ve škole POŠKOLE 2005*, 2005. ISBN 80-239-4633-1
BYČKOVSKÝ, P.; MARKOVÁ, M. Využití software ITEMAN k položkové analýze. In *Sborník XI. Konference ČAPV. Sociální a kulturní souvislosti výchovy a vzdělávání* [CR-ROM]. Brno : PedF MU, 2003. ISBN 80-7315-046-8.
BYČKOVSKÝ, P.; KOTÁSEK, J. Revize Bloomovy taxonomie edukačních cílů. *Pedagogika*, 2004, roč. LIV, č. 3, s. 227-242.
CALFEE, R. To grade or not to grade. *IEEE Intelligent Systems*, 2000, vol. 15, no. 5, s. 35-37.
CHONG, H. Y. *A Simple Guide to the Item Response Theory (IRT)* [on-line]. 2006 [cit. 1. 4. 2007]. Dostupné na: <http://www.creative-wisdom.com>.
CLAMAN, C. (Ed.). *10 Real SATs*. New York : College Entrance Examination Board a ETS, 2003. ISBN 0-87447-705-0.
CLAUSER, B. E.; ROSS, L. P.; CLYMAN, S. G.; ROSE, K. M.; MARGOLIS, M. J., et al. Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement*, 1997b, no. 34, s. 141-161.
CLAUSER, B. E.; MARGOLIS, M. J.; CLYMAN, S. G.; ROSS, L. P. Development of a scoring algorithm to replace expert rating for a scoring of complex performance-based assessments. *Applied Measurement in Education*, 1997a, no. 10, s. 345-358.
COLEY, R. J. Technology´s Impact. A new study shows the effectiveness – and the limitations – of school technology [on-line]. Electronic School, September 1997 [cit. 1. 4. 2007]. Dostupné na: <www.electronic-school.com/0997f3>.
CONOLEY, J. C.; O´NEIL, H. F. (JR). A Primer for Developing Test Items. In O´NEIL, H. F. (JR). *Procedures for Instructional Systems development*. New York : Academic Press, Inc., 1979. ISBN 0-12-526660-X.
CASE, S.; SWANSON, B. *Constructing Written Test Questions For the Basic and Clinical Science* [on-line]. 3rd ed. Philadelphia, PA : National Board of Medical Examiners, 2002 [cit. 1. 4. 2007]. Dostupné na: <www.au.af.mil/au/awc/awcgate/documents/nbme_iwgindex.pdf>
ČERNOCHOVÁ, M. Příprava budoucích *eU*čitelů na *e*Instruction. Kladno : AISIS, 2003. ISBN 80-239-0938-X.
ČERNOCHOVÁ, M. O stavu a trendech využívání ICT v českých školách a v zahraničí. *Pedagogika. Nové technologie a nové formy ve vzdělávání*, 2006, roč. LVI, č. 4, s.316-334. Praha : UK PedF. ISSN 0031-3815.
DAVEY, T.; GODWIN, J.; MITTELHOLTZ, D. Developing and scoring innovative computerized writing assessment. *Journal of Educational Measurement*, 1997, no. 34, s. 21-41.
DENGLEROVÁ, D. Nové metody v diagnostice osobnosti aneb IRT a její přínos k testování osobnosti. In *Sborník konference CVVOE. Vývoj a utváření osobnosti v sociálních a etnických kontextech (víceoborový přístup)*. Brno : MU Brno, 2005. ISBN 80-210-3804-7.
DIKLI, S. An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment* [on-line]. Boston College : Technology and Assessment Study Collaborative, 2006, vol. 5, no.1. ISSN 1540-2525 [cit. 1. 4. 2007]. Dostupné na:<www.itla.org>.
DRASGOW, F.; OLSON-BUCHANAN, J. B. *Innovations in computerized assessment*. Mahwah, NJ: Lawrence Erlbaum, 1999.

ELIOT, C.; WOOLF, B. An Adaptive Student Centered Curriculum for an Intelligent Training system. *User Modeling and User-Adapted Interaction*, 1995, vol. 5, no. 1, s. 67-86.
EMBRETSON, S. E.; REISE, S. P. *Item response theory for Psychologists*. Mahwah, NJ : Lawrence Erlbaum Associates, 2000, ISBN 0-8058-2818-4.
GIALLUCA, K. A.; WEISS, D. J. *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement*. Research Report 79-6. Minneapolis : University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1979.
GOLDSTEIN, H.; WOOD, R. Five decades of item response modeling. *British Journal of Mathematical Statistical Psychology*, 1989, no. 42, s. 139-167.
GUILFORD, J. P. The Determination of Item Difficulty When Chance Success is a Factor. *Psychometrika*, 1936, vol. 1, no. 4, s. 259-264.
HALADYNA, T. M. *Writing Test Items to Evaluate Higher Order Thinking*. Needham Heights, MA : Allyn a Bacon, 1997. ISBN 0-205-17875-8.
HALADYNA, T. M. *Developing And Validating Multiple-Choice Test Items*. Mahwah, NJ : Lawrence Erlbaum 2004. ISBN 0-8058-4661-1.
HALADYNA, T. M.; DOWNING, S. M.; RODRIGUEZ, M. C. *Applied Measurement in Education* 2002, 15, č. 3, s. 309-333.
HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H.J. *Fundamentals of Item Response Theory*. Newbury Park, CA : Sage Publications, Inc., 1991. ISBN 0-8039-3647-8.
HARTKE, A. R. The use of latent partition analysis to identify homogeneity of an item population. *Journal of Educational Measurement*, 1978, vol. 15, no. 1, s. 43-47.
HENDRICKSON, A. B.; KOLEN, M. J. *IRT Equating of the MCAT* [online]. University of Iowa 2003 [cit. 1. 4. 2007]. Dostupné na: <www.aamc.org/students/mcat/research/monograph4.pdf>.
HENRISSEN, S. Gathering analyzing and using data on test items. In THORNDIKE, R. L. (Ed.), *Educational Measurement*. 2nd ed. Washington DC : American Council on Education, 1971.
HERMAN, J. L. Item Writing Techniques. In KEEVES, J. P. *Educational Research, Metholodogy and Measurement: AnInternational Handbook*. New York : Pergamon Press, 1988. ISBN 0-080365510-8.
HIRSCHMAN, L.; BRECK, E.; LIGHT, M.; BURGER, J. D.; FERRO, L. Automated grading of short answer tests. *IEEE Intelligent Systems*, 2000, vol. 15, no. 5, s. 31-35.
HIVELY, W.; PATTERSON, H. L.; PAGE, S. H. A „universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, vol. 5, no. 4, s. 275-289.
HNILIČKOVÁ, J.; JOSÍFKO, M.; TUČEK, M. *Didaktické testy a jejich statistické zpracování*. Praha : SPN, 1972.
HOI, K. S. *Principles of Test Theories*. Hillsdale, NJ : Lawrence Erlbaum Associates, Inc., 1990.
HOPKINS, K. D. *Educational and psychological measurement and evaluation* (8th ed.). Boston : Allyn and Bacon, 1998.
HORKAY, N.; BENNETT R. E.; ALLEN, N.; KAPLAN, B.; YAN, F. Does it Matter if I take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning, and Assessment* [online], 2006, vol. 5, no. 2 [cit. 1. 4. 2007]. Dostupné na <http://www.jtla.org>.
HRABAL, V.; LUSTIGOVÁ, Z.; VALENTOVÁ, L.: *Testy a testování ve škole*. Praha : Středisko vědeckých informací PedF UK, 1994.
HULIN, C. L.; DRASGOW, F.; PARSONS, C. K. *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin, 1983. ISBN: 0-87094284-0.
CHRÁSKA. M.: *Didaktické testy. Příručka pro učitele a studenty učitelství*. Brno : Paido, 1999.
IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood, IL : Scientific Software International, Inc. 2003. ISBN: 0-89498-053-X.
JELÍNEK, M., KVĚTON, P., DENGLEROVÁ, D. Adaptivní testování - základní pojmy a principy. *Československá psychologie*, roč. L, č. 2, s. 163-173. Praha : Academia, 2006. ISSN 0009-062X.
KESTLER, J., aj. *SPARKNOTES ACT*. New York : SparkNotes LLC, 2003. ISBN 1-58663-961-7.

KESTLER, J., aj. *SPARKNOTES 5 Practice Tests for the SAT II Math IIC*. New York : SparkNotes LLC, 2003. ISBN 1-58663-869-6.
Key Data on Information and Communication Technology in Schools in Europe. Brussel · Eurydice, 2004. ISBN 2-87116-370-7.
KIM, S. H. *A continuation ratio model for ordered category items*. Prezentováno na pravidelném setkání Psychometric Society. Chapel Hill, NC, 2002.
KINGSBURY, G. G.; ZARA, A. R. A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 1991, s. 241-261.
KOMENDA, S. Měření a metaměření znalostí. Olomouc : Nakladatelství Univerzity Palackého 2003. ISBN 80-244-0776-0.
KOMENDA, S. MAZUCHOVÁ, J. *Tvorba a testování testu*. Olomouc : Nakladatelství Univerzity Palackého, 1995. ISBN 80-7067-461-X.
KOMENDA, S. ZAPLETALOVÁ, J. *Analýza didaktického testu a její počítačová podpora*. Olomouc : Lékařská fakulta UP, 1996.
KUKICH, K. (2000). Beyond automated essay scoring. *IEEE Intelligent Systéme*, 2000, vol. 15, no. 5, s. 22-27.
KULIK, C. L.; KULIK, J. A. Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, 1991, vol. 7, no. 1 a 2, s. 75-94.
KULIK, J. A. Meta-Analytic Studies of Findings on Computer-Based Instruction. In BAKER, E. L.; O'NEIL, H. F. *Technology Assessment in Education and Training*. Hillsdale, NJ : Lawrence Erlbaum Associates, 1994. ISBN 0-8058-1247-4.
KVĚTON, P., KLIMUSOVÁ, H. Metodologické aspekty počítačové administrace psychodiagnostických metod. *Československá psychologie*, 2002, roč. 46, č. 3, s. 251-264.
LAWLEY, D. N. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 1943, no. 61, s. 273-287.
LANDAUER, T. K.; LAHAM, D.; FOLTZ, P.W. Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, 2003, vol. 10, no. 3, *s. 295-308.*
LANDAUER, T. K.; LAHAM, D.; FOLTZ, P. W. The Intelligent Essay Assessor. *IEEE Intelligent Systems*, 2000, vol. 15, no. 5, *s. 27-31.*
LEUNG, CH. K; CHANG, H. H.; HAU, K. Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods. *Journal of Technology, Learning, and Assessment*, 2003, vol. 2, no. 5, s. 2-15.
LIVINGSTON, S. A.; DORANS, N. J. *A Graphical Approach to Item Analysis*. Research Report RR 04-10. Princeton, NJ : Educational Testing Service, 2006.
LIVINGSTONE, S; BOBER, M. *UK Children Go Online* [online]. ECRV, 2004 [cit. 1. 4. 2007]. Dostupné na http://personal.lse.ac.uk/bober/UKCGOfinal-Report.pdf.
LORD, F. M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ : Lawrence Erlbaum, 1980. ISBN 0-89859-006-X.
LORD, F. M. Some test theory for tailored testing. In HOLTZMAN, W. H. (Ed) *Computer-assisted instruction, testing and guidance*. New York : Harper a Row, 1970.
LORD, F. M. The relation of the reliability of multiple-choice items to the distribution of item difficulty. *Psychometrika*, 1952, no. 17, s. 56-57.
LORD, F. M.; NOVICK, M. R. *Statistical Theories of Mental Test Scores*. Reading, MA : Addison-Wesley, 1968.
MAREŠ, J.: Psychodiagnostika podporovaná počítačem. Praha : Ústav pro informace ve vzdělávání, 1992.
MASTERS, G. N. A Rasch model for partial credit scoring. *Psychometrika*, 1982, no. 47, s. 149-174.
MAŠEK, J.; MICHALÍK, P.; VRBÍK, V. Otevřené technologie ve výuce. Plzeň : Západočeská univerzita v Plzni, 2004. ISBN 80-7043-254-3.
McDONALD, R. P. *Test theory. A unified treatment*. Mahwah, NJ : Lawrence Erlbaum Associates, Inc., 1999. ISBN 0-8058-3075-8.
Mediax Interactive Technologies. *MICROSCALE*. Black Rock, CT : Author, 1986.

MERRILL, M. D. Component Display Theory. In REIGELUTH, C. (Ed.). *Instructional Design Theories and Models*. Hillsdale, NJ : Lawrence Erlbaum, 1983, s. 279-333.
MILLMAN, J. Computer-based item generation. In *Criterion-Referenced Measurement. The State of the Art*. Baltimore, London : The John Hopkins University Press, 1980. ISBN 0-8018-2264-5.
MILLMAN, J.; GREENE, J. *The Specification and Development of Test of Achievement and Ability*. In LINN, R. L. (Ed.) *Educational Measurement*. 3rd ed. New York : Macmilan Publishing Company, 1989. ISBN 0-02-922400-4.
MOLNAR, A. R. *Computers in Education: A Brief History* [on-line]. June, 1997 [cit. 1. 4. 2007]. Dostupné na: <http://thejournal.com/the/printarticle/?id=13739>.
MURAKI, E. Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 1990, no.14, s. 59-71.
MURAKI, E.; BOCK, R. D. PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. Chicago : Scientific Software Int., 1993.
NELSON, L. R. *Item Analysis for Tests and Surveys Using Lertap 5*. Perth, Western Australia : Curtin University of Technology, 2001.
O´NEIL, H. F.; BAKER, E. L. *Technology assessment in software applications*. Hillsdale, NJ : Lawrence Erlbaum, 1994. ISBN 0-8058-1247-4.
O´NEIL, H. F.; BAKER, E. L. *Technology assessment in education and training*. Hillsdale, NJ : Lawrence Erlbaum, 1994. ISBN 0-8058-1249-0.
OOSTERHOF, A. C. Similarity of various item discrimination indices. *Journal of Educational Measurement*, 1969, no. 6, s. 1-9.
PELIKÁN, J. *Základy empirického výzkumu pedagogických jevů*. Praha : Karolinum, 1998. ISBN 80-7184-569-8.
PELIKÁN, J. Programovaná výuka v kombinaci s hypertextem. *Zpravodaj ÚVT MU*, 1998, roč. IX, č. 2, s. 9-13. ISSN 1212-0901.
PRŮCHA, J. Pedagogická věda a nové výzvy edukační praxe. *Pedagogika. Nové technologie a nové formy ve vzdělávání*, 2006, roč. LVI, č. 4, s.307-315. Praha : UK PedF. ISSN 0031-3815.
RASCH, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen : Danish Institute for Educational Research, 1960.
ROID, G. H.; HALADYNA, T. M. *A Technology For Test-Item Writing*. New York : Academic Press, 1982. ISBN 0-12-593250-2.
ROID, G. The Technology of Test-Item Writing. In O´NEIL, H. F. (JR). *Procedures for Instructional Systems development*. New York : Academic Press, 1979. ISBN 0-12-526660-X.
RUDNER, L. M. *An On-line, Interactive, Computer Adaptive Testing Tutoriál* [online].1998 [cit. 1. 4. 2007]. Dostupné na:<http://edres.org/scripts/cat>.
RUDNER, L. M. *An On-line, Interactive, Computer Adaptive Testing Mini-Tutorial* [online]. 1998 [cit. 1. 4. 2007]. Dostupné na:<http://edresearch.org/scripts/cat>.
RUDNER, L. M.; LIANG, T. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and* Assessment [online]. 2002, vol. 2, no. 1 [cit. 1. 4. 2007]. Dostupné na: <http://www.jtla.org>.
SANDS, W. A; WATERS, B. K.; McBRIDE, J. R. *Computerized adaptive testing: From inquiry to operations*. Washington, DC : American Psychological Association, 1997.
SHERIDAN, B.; ANDRICH, D.; LUO, G. *Welcome to RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models*. User´s Guide, 1996.
SHERMIS, M. D.; BURSTEIN, J. *Automated Essay Scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.
SCHINDLER, R. a kol. *Rukověť autora testových úloh*. Praha : Centrum pro zjišťování výsledků vzdělávání, 2006. ISBN 80-239-7111-5.
SIVIN-KACHALA, J.; BIALO, E. R. *Report on the Effectiveness of Technology in Schools 1990-1994*. Washington, D.C.: Software Publishers Association, 1994.

STAGE, C. *A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest READ* [online]. Umeå University: Department of Educational Measurement, 1997 [cit. 1. 4. 2007]. Dostupné na: <www.umu.se/edmeas/publikationer/pdf>

STARK, S.; CHERNYSHENKO, S.; CHUAH, D.; LEE, W.; WADLINGTON, P. *IRT Modeling Lab* [online]. University of Illinois, 2001 [cit. 1. 4. 2007]. Dostupné na: <http://work.psych.uiuc.edu/irt/>

SVATOŠ, T. Elektronická edukační média a cesty jejich evaluace. *Pedagogika. Nové technologie a nové formy ve vzdělávání*, 2006, roč. LVI, č. 4, s. 348-360. Praha : UK PedF. ISSN 0031-3815.

ŠKALOUDOVÁ, A. *Statistika v pedagogickém a psychologickém VÝZKUMU.* Praha : PedF UK, 1998. ISBN 80-86039-56-0.

TOMÁŠEK, V.; POTUŽNÍKOVÁ, E. *Netradiční úlohy. Problémové úlohy mezinárodního výzkumu PISA*. Praha : Ústav pro informace ve vzdělávání, 2004. ISBN 80-211-0484-8.

TROLLIP, S. R. The evaluation of a complex, computer-based flight procedures trainer. *Human Factors*, 1979, vol. 22, no. 1, s. 47-54.

URBÁNEK, T., ŠIMEČEK, M. Teorie odpovědi na položku. *Československá psychologie*, 2001, roč. 45, č. 5, s. 428-440.

URBINA, S. Essentials of Psychological Testing. New Jersey : John Wiley a Sons, Inc., 2004. ISBN 0-471-41978-8.

VAN DER LINDEN, W. J.; HAMBLETON, R.K. *Handbook of modern item response theory*, New York : Springer, 1997.

Věda, technika a průmysl v zemích OECD: výsledková tabulka 2003[online]. OECD[cit. 1. 4. 2007]. 2003 . Dostupné na <www.oecd.org/dataoecd/57/59/31429790.pdf>.

WAINER, H. *Computerized Adaptive Testing: A primer*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000. ISBN 0-8058-3511-3.

WAINER, H.; MISLEVY, R. J. Item Response Theory, Item Calibration, and Proficiency Estimation. In WAINER, H., et. al. *Computerized Adaptive Testing: A primer*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000. ISBN 0-8058-3511-3.

WEISS, D. J. *The Stratified adaptive computerized ability test*. Research report 73-3. Minneapolis : University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1973.

WEISS, D. J. Computerized Adaptive Achievement Testing. In O´NEIL, H. F. (JR). *Procedures for Instructional Systems Development*. New York : Academic Press, 1979. ISBN 0-12-526660-X.

WEISS, D. J. Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*. New York : Pergamon Press, 1988. ISBN 0-080365510-8

WEISS, D. J., Adaptive Testing. In KEEVES, J. P. *Educational Research, Metholodogy and Measurement: AnInternational Handbook.* New York : Pergamon Press, 1988. ISBN 0-080365510-8

WEISS, D. J. New horizons in testing: Latent trait test theory and computerized adaptive testing. New York : Academic Press, 1983.

WEISS, D. J. *Manual for the FastTEST Professional Testing System, Version 2*. St. Paul, Minnesora : Assessment Systems Corporation, 2006.

WIBERG, M. Classical test theory vs. item response theory. An evaluation of the theory test in the Swedish driving-licence test. *Educational Measurement*, 2004, no. 50. ISSN 1103-2685.

WIJEKUMAR, K. J.; MEYER, B. J. F.; WAGONES, D.; FERGUSON, L: Computer Technologies and Learning Outcomes. *British Journal of Educational Technology*, 2006, vol. 37, no. 2, s. 191-209.

WILSON, J. W. Evaluation of Learning in Secondary School Mathematics. In BLOOM, B. S.; HASTINGS, J. T.; MADAUS, G. F. *Handbook on Formative and Summative Evaluation of Student Learning*. New York : McGraw-Hill Book Company, 1971.

WOOLF, B. *AI in Education. Encyklopedia of Artificial Intelligence*. New York : John Wiley a Sons, Inc., 1992.

WRIGHT, B. D.; STONE, M. H. Best test design. Chicago : MESA Press, 1979.

YU, CH. H.; POPP, S. E. O. Test Equating by Common Items and Common Subjects: Concepts and Applications. *Practical Assessment Research and Evaluation* [online]. 2005, vol. 10, no. 4 [cit. 1. 4. 2007]. ISSN 1531-7714. Dostupné na: < http://pareonline.net/getvn.asp?v=10&n=4>.

ZIMOWSKI, M. F.; MURAKI, E.; MISLEVY, R. J.; BOCK, R.D. *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International, 1996.

ZIMOWSKI, M. F.; MURAKI, E.; MISLEVY, R. J.; BOCK, R.D. *BILOG-MG*. Lincolnwood, IL : Scientific Software International, 2003.

ZOUNEK, J. E-learning a vzdělávání. Několik pohledů na problematiku e-learningu. *Pedagogika. Nové technologie a nové formy ve vzdělávání*, 2006, roč. LVI, č. 4, s.335-347. Praha : UK PedF. ISSN 0031-3815.

User´s Manual for the ITEMAN Conventional Item Analysis Program. St.Paul, Minnesota : Assessment Systems Corporation,1989-1996

Software:

ITEMAN Classical Item Analysis Program [disketa], St. Paul, Minnesota : Assessment Systems Corporation 1996

LERTAP 5 Item and Test Analysis Using Excel, Windows Version [CD-ROM]. Test Analysis for Cognitive and Personality Tests, Mastery tests, Scales and Surveys. St. Paul, Minnesota : Assessment Systems Corporation, Curtin University of Technology, 2001-2002.

FastTEST Professional Testing System. Test Development Package [CD-ROM]. St. Paul, Minnesota : Assessment Systems Corporation

Zimowski, M., Muraki, E., Mislevy, R.; & Bock D. Bilog-MG [CD-ROM]. Mooresville : Scientific Software International, 2003.

Bibliografický záznam

BAŽANTOVÁ, Zuzana. *Využívání počítačů k testování.* Praha : Univerzita Karlova, Pedagogická fakulta, 2007. 160 s. Disertační práce (školitel: Petr Byčkovský).

Anotace

Disertační práce, která se zabývá problematikou využívání počítačů při testování, se skládá ze dvou částí: teoretické a výzkumné. Je zaměřená na řešení dvou problémů: charakterizování současného stavu využívání počítačů při testování (1) a odlišnostmi i shodami metody založené na klasické teorie testu (KTT) a metody vycházející z teorie odpovědi na položku (IRT) při tvorbě testu a analýze jeho výsledků (2). Druhý problém je řešen jak z teoretického hlediska, tak při aplikaci obou metod při analýze výsledků testu Obecné studijní předpoklady použitém v přijímacím řízení na Pedagogickou fakultu UK v roce 2006. V práci je popsáno také adaptivní testování a uvádí se zde využití speciálních software pro provádění klasické analýzy testových výsledků (ITEMAN a LERTAP 5) a pro jedno-, dvou- a tří-parametrovou IRT analýzu (BILOG-MG).

Klíčová slova

Využití informačních a komunikačních technologií při testování, klasická teorie testu (KTT), teorie odpovědi na položku (IRT), IRT parametry, charakteristická funkce položky/ testu, informační funkce položky/ testu, IRT modely, analýza testu, reliabilita, položková analýza, obtížnost položky, citlivost položky, analýza distraktorů, adaptivní testování.

SUMMARY

The thesis provided deals with the question of computer based testing. The study contains two main parts: theoretical and empirical. It addresses two problems: what is the state of the art of computers implementation in testing (1), and what is the difference between applying two theories – classical test theory (CTT), and item response theory (IRT) in test construction, and analysis of test scores (2). The second problem has been solved both theoretical and empirical.

No doubt, that computer implementation in test administration, construction, score analysis, test forms creating, and adaptive testing is very useful activity. It seems, as well, implementation of IRT in test construction and its score analysis has some advantages in comparison with CTT. On the other hand, mastering IRT is a quite difficult task, and there are strict demands to be fulfilled before IRT could be applied, too.

There are principles of adaptive testing described in the study. This topic has not been described in any educational journal till today. Special software for realizing classical item analysis (ITEMAN, LERTAP 5) and for realizing IRT analysis (BILOG-MG) are presented here. To my knowledge comparison of item analysis carried out by methods based on CTT and IRT has not been presented in the Czech Republic.

An analysis mentioned above has been performed on data of 1696 candidates during their entrance examinations to Faculty of Education, Charles University, Prague in 2006. Candidates have been tested by four forms of General Aptitude Test.

In general, it seems the item analysis made by IRT provides parameters that are more precise since they do not depend both on characteristics of testees and structure of test items. But when we apply both methods of item analysis on large and representative samples of testees, both methods can be considered to be equal. However, more research in this area is needed.

PŘÍLOHY

Příloha 1 : Přehled nejvýznamnějších světových organizací a firem činných v oblasti testování

zkratka	organizace	stát	popis	URL odkaz
ETS	Educational Testing Service	USA	Největší světová organizace zabývající se tvorbou didaktických testů (i počítačových a počítačových adaptivních) a problematikou pedagogických měření, má téměř 3000 zaměstnanců!	www.ets.org
CITO	Národní institut pro tvorbu testů	Holandsko	Největší evropská a pátá největší světová organizace specializující se na tvorbu školských testů (i počítačových a počítačových adaptivních) a problematikou pedagogických měření (500 stálých a 2000 externích zaměstnanců). Cito má svá zastoupení i mimo Evropu v USA, Turecku a Japonsku. Podílí se na zahraničních vzdělávacích projektech nejen v Evropě, ale téměř po celém světě.	www.cito.nl
NFER	National Foundation for Educational Research	Velká Británie	Největší anglická organizace zabývající se pedagogickým výzkumem, tvorbou školských testů a problematikou pedagogických měření.	www.nfer.ac.uk
ACER	Australian Council for Educational Research	Austrálie	Největší australská organizace zabývající se pedagogickým výzkumem, tvorbou školských testů a problematikou pedagogických měření	www.acer.edu.au
CRESST	National Center for Research on Evaluation, Standards and Student Testing	USA	Velké vědecké pracoviště zabývající se výzkumem v oblasti pedagogického měření. Umožňuje bezplatné stáhnutí stovek výzkumných zpráv (v PDF formátu) o různých aspektech testování.	www.cresst96.cse.ucla.edu
	The College Board	USA	Národní nezisková členská asociace, založena roku 1900, se skládá z více než 4300 středních a vysokých škol, univerzit a jiných vzdělávacích organizací. Zabývá se tvorbou školských testů a problematikou pedagogických měření. Každý rok připravuje studenty na přijímací testy, nabízí jim poradenství, finanční pomoc, registraci ke zkouškám. Mezi její nejznámější programy patří testy SAT, PSAT/NMSQT a the Advanced Placement program.	www.collegeboard.com
NCME	National Council on Measurement in Education	USA	Organizace sdružující odborníky z oblasti pedagogických měření, vydává dva odborné časopisy věnované problematice testů: teoreticky laděný Journal of Educational Measurement a prakticky orientovaný Educational Measurement: Issues and Praktice.	http://ncme.ed.uiuc.edu
AERA	American Educational Research Association	USA	Americká asociace pedagogického výzkumu zabývající se problematikou měření a evaluace vzdělávání.	http://aera.net
	Assessment Systems Corporation	USA	Firma specializující se na vývoj software pro tvorbu testů, bank úloh, elektronické testování, adaptivní testování a software pro analýzu výsledků testů.	www.assess.com
AEA-Europe	Association for educational assessment	Evropa (Velká Británie, Švédsko, Nizozemí)	AEA je evropská asociace podporující diskuzi a vzájemnou spolupráci (na výzkumu, projektech apod.) mezi evropskými organizacemi, institucemi i jednotlivci aktivními v oblasti hodnocení a měření výsledků vzdělávání. Pořádá konference a vyvíjí publikační činnost.	www.aea-europe.net
CKE	Centralna Komisja Ekzaminacyjna	Polsko	Státní organizace, jejímž cílem je příprava a organizování systému pedagogické evaluace v Polsku.	http://www.cke.edu.pl/
EXAM		SR	Soukromá firma zabývající se systematicky a profesionálně problematikou školských testů.	www.exam.sk
SCIO		CR	Soukromá firma zabývající se tvorbou a vyhodnocováním školských testů, nabízí také on-line testy (jak předmětové, tak obecných studijních předpokladů) pro žáky ZŠ a SŠ jako přípravu k přijímacím zkouškám (jsou ihned vyhodnocované).	www.scio.cz
	Centrum moderního vzdělávání	CR	Soukromá firma nabízející e-learning pro školy. Věnuje se současně i firemnímu e-learningu a rozšířila své aktivity do dalších 6 evropských zemí.	http://www.modern-education.net/
CERMAT	Centrum pro zjišťování výsledků vzdělávání	CR	Státní organizace pověřená koordinací příprav reformy maturitní zkoušky v ČR, zabývá se evaluací výsledků vzdělávání, výzkumem a vývojem v oblasti testování.	www.cermat.cz

Příloha 2 : Přehled současných počítačových testů, včetně některých adaptivních[1] (označené *)

název testu	zkratka	tvůrce testu	popis testu	URL odkaz
Graduate Record Examination	**GRE***	ETS (USA)	Test studijních předpokladů používaný v USA při přijímacím řízení na postgraduální studium.	www.ets.org/portal/site/ets
Graduate Management Admission Test	**GMAT***	ETS (USA) pro potřeby Graduate Management Admission Council	Test je obvykle vyžadován při přijímacím řízení do programů Master of Business Administration (MBA) nejen ve Spojených státech, ale ve školách po celém světě, které studium tohoto manažerského programu nabízejí.	www.mba.com/mba
Multistate Pharmacy Jurisprudence Examination	**MPJE***	The National Association of Boards of Pharmacy (NABP) (USA)	Test právní způsobilosti farmaceutů vyžadovaný jednotlivými státy v USA.	http://www.nabp.net
North American Pharmacist Licensure Examination	**NAPLEX***	The National Association of Boards of Pharmacy (NABP) (USA)	Test odborné způsobilosti farmaceutů vyžadovaný jednotlivými státy v USA.	http://www.nabp.net
National Board for Professional Teaching Standards	**NBPTS**	National Board for Professional Teaching Standards (NBPTS, USA)	Slouží k dalšímu vzdělávání učitelů a certifikaci dosažené úrovně. Pro každý předmět a stupeň školy lze získat příslušný certifikát na základě až tříletého studa a zkoušky vykonávané vzdáleně. Uchazeč musí vypracovat materiály na určená témata doplněné videonahrávkami a ukázkami práce studentů, tuto část skóruje 12 nezávislých vyškolených hodnotitelů. Poté v test centru řeší 6 praktických úloh na počítači.	http://www.nbpts.org/
Professional Assessments for Beginning Teachers	**PRAXIS I**	ETS (USA)	Test využívaný některými univerzitami pro výběr uchazečů o učitelské obory. Testuje psaní, porozumnění textu a matematiku. Je k dispozici v paírové i počítačové podobě. Některé státy USA jej vyžadují jako součást odborné učitelské kvalifikace pro získání certifikátu PRAXIS II, III.	http://www.ets.org/
Test of English as a Foreign Language	**TOEFL***	ETS (USA)	Mezinárodně uznávaná americká standardizovaná jazyková zkouška. Má ohodnotit úroveň znalostí anglického jazyka u uchazečů o studium v angličtině, u nichž není angličtina rodným jazykem. Je určena hlavně studentům se střední a vyšší pokročilostí. V České republice se testuje již pouze počítačově.	http://www.ets.org/toefl
International English Language Testing System	**IELTS**	IELTS International	Test používaný k ohodnocení jazykových znalostí u cizinců. Od roku 2005 je v dispozici i v počítačové verzi.	http://www.ielts.org/
College Level Examination Program Tests	**CLEP**	College Board (USA)	Sada 90 min testů používaných jako semestrální hodnocení znalostí univerzitních studentů. Většina z nich je MC, některé obsahují otevřené otázky.	http://www.colleoeboard.com/student/testino/clep/exams.html

[1] Úplný přehled současných významných počítačových adaptivních testů je uveden v kap. 7.

OBECNÉ STUDIJNÍ PŘEDPOKLADY

VARIANTA

A

Zadání neotvírejte, počkejte na pokyn!

Zopakujte si základní informace a pokyny ke zkoušce:

- U každé z úloh je vždy právě jedna odpověď správná.
- Za každou správně vyřešenou úlohu získáváte bod, za každou špatně vyřešenou úlohu se vám odečítá část bodu.
- Úlohy můžete řešit v libovolném pořadí.
- Test obsahuje 45 úloh, na řešení máte 45 minut.

Červen 2006

OBECNÉ STUDIJNÍ PŘEDPOKLADY

VARIANTA A

VERBÁLNÍ ODDÍL

V každé z následujících vět jsou jedno nebo dvě prázdná místa, která značí, že ve větě bylo něco vynecháno. Za každou větou najdete několik možností – slov nebo dvojic slov. Vyberte slovo nebo dvojici slov, která se **nejlépe** hodí do **příslušné věty** jako celku.

1.

Povrch Evropy je výškově i tvarově _____ , neboť vznikl _____ horotvornými pochody v různých geologických dobách.

(A) různý – stejnými
(B) podobný – dávnými
(C) rozmanitý – různými
(D) zanedbaný – zbytečnými
(E) zvláštní – rychlými

2.

Šéf disciplinární komise řekl, že případ nelze _____ jako úmyslnou a neomluvenou _____ u dopingového testu.

(A) hodnotit – absenci
(B) ocenit – neochotu
(C) posuzovat – sabotáž
(D) vyšetřit – přítomnost
(E) prokázat – zaujatost

3.

Vyšetřovatelé tak pomocí nastrčeného vězně dostávali _____ informace o jeho trestné _____ .

(A) neúplné – aktivitě
(B) podrobné – činnosti
(C) nedůležité – minulosti
(D) ucelené – charakteristice
(E) cenné – sazbě

4.

V poslední době se již množí _____ také od zákazníků, kteří na závažné _____ v prodejnách s potravinami na ně upozorňují.

(A) návrhy – podvody
(B) zprávy – obtížnosti
(C) dopisy – inovace
(D) stížnosti – obezličky
(E) podněty – nesrovnalosti

5.

Viz. „staré _____" jsou stále přítomny ve státní _____.

(A) aparáty – televizi
(B) struktury – správě
(C) modely – policii
(D) funkce – struktuře
(E) gardy – politice

Každá z následujících úloh se skládá z dvojice slov nebo slovních spojení, za kterými následuje pět možností – pět dvojic slov nebo slovních spojení. Z těchto pěti možností vyberte dvojici, v níž se vztah mezi členy **nejvíce** blíží vztahu v zadané dvojici.

6.

LOUČ : LAMPA

(A) arch : kniha
(B) brk : pero
(C) vrata : dveře
(D) lomoz : hluk
(E) sazenice : rostlina

7.

STANDARD : NADSTANDARD

(A) starý : moderní
(B) těžký : lehčí
(C) balík : zásilka
(D) večeře : hostina
(E) čištění : odstranění

8.

LOĎ : PLACHETNICE

(A) kopírování : kopírka
(B) stůl : židle
(C) porcelán : keramika
(D) auto : motorka
(E) postel : palanda

9.

ROSTLINA : HNOJIVO

(A) baterie : energie
(B) hodinky : ručičky
(C) pes : bouda
(D) dítě : kaše
(E) člověk : vitaminy

Text k úlohám 10 až 12

Neuvěřitelné úspěchy nepočetných evropských dobyvatelů nad mnohonásobnou vojenskou přesilou a rychlé rozvrácení vysoce organizovaných kulturně-civilizačních útvarů Aztéků, Mayů i Inků jsou zpravidla vysvětlovány odkazy na počáteční víru, že příchozí jsou bohové (Cortésův vstoupil podle křesťanského kalendáře na pobřeží v pátek 22. dubna 1519, tedy podle kalendáře Aztéků v roce, měsíci a dokonce neuvěřitelnou náhodou i ve dnech, kdy se podle legendy měl navrátit bůh Quetzalcoatl a ujmout se vlády), převahou evropské výzbroje (palné zbraně, ocelové meče) a způsobem válčení (Aztékům šlo o počty zajatců, kteří pak mohli být obětováni, ne o počet zabitých nepřátel v boji), koňmi (ti byli s jezdcem považováni Indiány za jednu bytost, která se může rozdělit), bojovými psy, nemocemi (druhou fázi Cortésova tažení ovlivnila masová úmrtnost domorodců v důsledku zavlečených neštovic) a povahovými vlastnostmi Evropanů!

Právě rozvinutá subjektivita Evropana je v konečné instanci rozhodujícím elementem, který vedl k tomu, že střetnutí s kulturně-civilizačními útvary amerického kontinentu vždy bylo rozhodnuto v jeho prospěch. Rozhodovala nesmírná, takřka sebezničující aktivita (již sama výprava za moře znamenala prodat veškerý majetek, zadlužit se, riskovat mnohdy vše ve jménu nejistoty); – „honra y provecho" – čest a zisk, touha po bohatství, půdě, otrocích spolu s vírou v zázrak, živenou legendami, to vše překrylo rozsah nebezpečí.

Znamenalo to zkrátka svobodně disponovat možnostmi, které si otevřel člověk – dobyvatel, který si svou činností sám dává hodnotu na straně jedné a ochromení až kulturně-civilizační trauma – šok (následná neschopnost vyrovnat se s běžnými životními situacemi v novém kontextu, snížení porodnosti) spolu s násilným přesídlováním z předkolumbovských sídel do měst a vesnic, zajišťující sociální a vojenskou kontrolu, vyřazením ze známých sociálních a ekonomických kontextů (rozsah centrální mocí garantovaných sociálních podpor a jistot v říši Inků před příchodem dobyvatelů vyvolával obdivný úžas předáků socialistických stran na počátku 20. stol.) na straně druhé.

(Oldřich Ševčík: Architektura – historie – umění, Praha 2002)

10.

Které z následujících tvrzení je v souladu s informacemi v uvedeném textu?

(A) Aztékové v boji nebrali zajatce.
(B) Evropští dobyvatelé měli vojenskou, morální a početní převahu.
(C) V říši Inků fungoval systém sociálních podpor obyvatelstva.
(D) Cortés si svůj příjezd schválně načasoval podle legendy o příchodu bohů.
(E) Mayský bůh Quetzalcoatl vypadal na dobových vyobrazeních přesně jako Cortés.

11.

Které z následujících tvrzení vystihuje jeden z hlavních rozdílů ve způsobu válčení mezi Aztéky a dobyvateli?

(A) Dobyvatelé nemilosrdně využívali nižší odolnosti Aztéků vůči neštovicím.
(B) Aztékům nešlo o počet zajatců.
(C) Dobyvatelé používali moderní manévrovací techniky a byli lépe organizovaní.
(D) Pro Aztéky nebyl důležitý počet nepřátel zabitých v boji.
(E) Žádná z možností (A) až (D) není správná.

12.

Které z následujících tvrzení **nevyplývá** z informací v uvedeném textu?

(A) Kulturně-civilizační útvary Aztéků, Mayů a Inků byly před příchodem dobyvatelů vysoce organizované.
(B) Aztékové usilovali v boji o co největší počet zajatců.
(C) Kulturně-civilizační šok se projevil mimo jiné i snížením porodnosti.
(D) Ve druhé fázi Cortésova tažení zemřelo mnoho domorodců na neštovice.
(E) Systém centrální moci dobyvatelů vyvolal později obdivný úžas předáků socialistických stran.

Každá z následujících úloh obsahuje slovo nebo slovní spojení, za kterým je uvedeno pět možností. K danému slovu vyberte to, které se **nejvíce** blíží k jeho **opačnému významu**. Pozor, v úlohách jde často o odlišení velmi **jemných** rozdílů.

13.

ZAČÍNAJÍCÍ

(A) zkušený
(B) poslední
(C) trvanlivý
(D) postupující
(E) zkoušející

14.

TĚSNĚ

(A) s potížemi
(B) s radostí
(C) s pomocí
(D) s rezervou
(E) se štěstím

15.

PŘIBLIŽNÝ

(A) celkový
(B) poměrný
(C) přesný
(D) vzdálený
(E) kompletní

16.

VSTOUPIT

(A) odejít
(B) vypadnout
(C) odstoupit
(D) zanechat
(E) odjet

17.

DISKUTABILNÍ

(A) neřešitelný
(B) monologový
(C) zřejmý
(D) sporný
(E) vyřešitelný

Analytický oddíl

Každá úloha nebo skupina úloh je založena na textu nebo souboru podmínek. Před vlastním řešením si pečlivě přečtěte zadání. Rozlišujte, které podmínky se týkají celé série úloh a které podmínky jsou uvedeny pouze pro jednu jedinou úlohu. Jsou-li úlohy založené na textu, vycházejte **pouze** z informací, které jsou v tomto textu obsažené. U některých úloh bude užitečné, když si pomůžete hrubým náčrtkem. Ke každé otázce vyberte tu **nejlepší** z nabízených odpovědí. Pouze jedna odpověď je správná.

Text k úlohám 18 až 21

Na parkovišti stojí ve třech řadách celkem osm aut. Každé auto je jiné barvy (bílé, černé, červené, fialové, modré, stříbrné, zelené, žluté). Víme, že:

- V první řadě stojí čtyři auta.
- Ve druhé řadě parkuje více aut než ve třetí řadě.
- V každé řadě parkuje alespoň jedno auto.
- Červené auto a žluté auto parkují ve stejné řadě.
- Stříbrné auto stojí ve druhé řadě.
- Modré auto stojí v jiné řadě než stříbrné auto, ale ve stejné jako zelené auto.
- Fialové auto ani černé auto nestojí ve stejné řadě jako bílé auto.

18.

Které z následujících tvrzení je určitě pravdivé?

(A) V první řadě stojí žluté auto.
(B) Ve druhé řadě stojí černé auto.
(C) Ve druhé řadě parkují právě dvě auta.
(D) Ve druhé řadě parkují právě tři auta.
(E) Ve třetí řadě parkují právě dvě auta.

19.

Které z následujících tvrzení je určitě **nepravdivé**?

(A) Bílé auto stojí v první řadě.
(B) Fialové auto stojí v první řadě.
(C) Bílé auto stojí ve třetí řadě.
(D) Černé auto stojí ve stejné řadě jako stříbrné auto.
(E) Modré auto stojí ve stejné řadě jako žluté auto.

20.

Ve které řadě může stát žluté auto?

(A) pouze v první řadě
(B) pouze ve třetí řadě
(C) pouze v první nebo ve druhé řadě
(D) pouze ve druhé nebo ve třetí řadě
(E) ve kterékoliv řadě

21.

Pokud stojí žluté auto ve stejné řadě jako zelené auto, které z následujících tvrzení je určitě pravdivé?

(A) Černé auto stojí ve stejné řadě jako modré auto.
(B) Bílé auto stojí ve stejné řadě jako stříbrné auto.
(C) Modré auto nestojí ve stejné řadě jako červené auto.
(D) Fialové auto nestojí ve stejné řadě jako stříbrné auto.
(E) Fialové auto stojí ve druhé řadě.

TEXT K ÚLOHÁM 22 AŽ 24

Čtyři kamarádi (Franta, Jirka, Pavel, Tomáš) maturovali ze stejných předmětů (angličtina, čeština, dějepis, zeměpis) V každém předmětu dostal jeden z nich jedničku, další dvojku, jiný trojku a jiný čtyřku. Přitom víme, že:

- Franta měl jedničku z češtiny a čtyřku ze zeměpisu.
- Tomáš má nejlepší průměr – dvě celé – a neměl žádnou čtyřku, ale ani dvě jedničky.
- Jirka dostal z češtiny a dějepisu stejnou známku.
- Pavel dostal z angličtiny stejnou známku jako Jirka z češtiny.
- Jirka měl čtyřku z dějepisu, zato Tomáš z něj měl jedničku.
- Franta dostal ze dvou předmětů trojku.
- Franta a Pavel měli stejný průměr.

22.

Které z následujících tvrzení je určitě **nepravdivé**?

(A) Tomáš měl ze zeměpisu trojku.
(B) Tomáš měl ze zeměpisu dvojku.
(C) Pavel měl tři dvojky.
(D) Jirka měl průměr 2,5.
(E) Jirka měl dvě jedničky.

23.

Co všechno mohl Tomáš dostat z češtiny?

(A) jen jedničku
(B) jen dvojku
(C) jen jedničku nebo dvojku
(D) jen jedničku nebo trojku
(E) jen dvojku nebo trojku

24.

Kdo všechno mohl dostat z některého předmětu trojku?

(A) jen Franta
(B) jen Pavel a Franta
(C) jen Pavel a Tomáš
(D) jen Pavel, Franta a Tomáš
(E) všichni čtyři

25.

Příliš intenzivní nebo nehospodárné využívání nerostných zdrojů je v protikladu s principy udržitelného rozvoje.

Které z následujících tvrzení vyplývá z uvedeného textu?

(A) Intenzivní používání výrobků z nerostných zdrojů není udržitelné.
(B) Principy udržitelného rozvoje se týkají pouze nerostných surovin.
(C) Principy udržitelného rozvoje nabádají k nevyužívání nerostných zdrojů.
(D) Příliš intenzivní využívání nerostných zdrojů nepovede k rozvoji.
(E) Nerostné zdroje je podle principů trvale udržitelného rozvoje třeba využívat hospodárně.

26.

Při šití se spotřeba látky řídí šířkou látky, délkou oděvu a objemem postavy. Přibližné množství určíme z plných délek, tedy délek s přídavky na švy.

Které z následujících tvrzení vyplývá z uvedeného textu?

(A) Pro určení spotřeby látky na ušití oděvu záleží hlavně na výšce postavy a jejím objemu.
(B) Spotřeba látky se přibližně určuje podle délek s přídavky na švy.
(C) Prázdné délky jsou délky bez přídavků na švy.
(D) Při šití vybíráme šíři látky podle šíře postavy.
(E) Šíři látky si uzpůsobíme podle délky odčvu a šíře postavy.

27.

Zůstavitel může vydědit potomka, jestliže v rozporu s dobrými mravy neposkytl zůstaviteli potřebnou pomoc v nemoci, ve stáří nebo jiných závažných případech; jestliže o zůstavitele trvale neprojevuje opravdový zájem, který by jako potomek projevovat měl, byl odsouzen pro úmyslný trestný čin k trestu odnětí svobody v trvání nejméně jednoho roku, nebo jestliže trvale vede nezřízený život.

Které z následujících tvrzení vyplývá z uvedeného textu?

(A) Potomek může být vydědčn, pokud přechodně vede nezřízený život.
(B) Zůstavitel může vydědit potomka, pokud by zůstavitel byl odsouzen pro úmyslný trestný čin k trestu odnětí svobody v trvání nejméně jednoho roku.
(C) Zůstavitel nesmí vydědit potomka, pokud se o něj stará v nemoci.
(D) Potomek může být vydčdčn, pokud o zůstavitele trvale neprojevuje opravdový zájem, který by jako potomek projevovat měl.
(E) Pokud potomek trvale jedná v rozporu s dobrými mravy, může ho zůstavitel vydědit.

28.

Spaluje tuky a především odbourává stres. Udržuje svalové napětí, posiluje srdce a cévy. Ve světě dnes vítězí nad joggingem, který nepříznivě působí především na klouby nohou a namáhá nadměrně páteř. Jde o chůzi ve velmi rychlém tempu. Jedině taková má všechny vyjmenované kladné účinky.

Které z následujících tvrzení vyplývá z uvedeného textu?

(A) Chůze ve velmi rychlém tempu odbourává stres.
(B) Chůze ve velmi rychlém tempu nemá žádné nepříznivé účinky.
(C) Pomalá chůze nespaluje tuky.
(D) Jogging je zdraví nebezpečný.
(E) Žádné z tvrzení (A) až (D) nevyplývá z textu.

29.

Koala žije v Austrálii. Vypadá jako malý, roztomilý medvídek, ale je to vačnatec. Žije v nejvyšších větvích blahovičníků – eukalyptů. Samice rodí pouze jediné mládě.

Jsou dána tři tvrzení:

I. Koala je malý roztomilý medvídek.
II. V Austrálii rostou blahovičníky.
III. Vačnatci jsou charakterističtí tím, že jejich samice rodí pouze jediné mládě.

Které z těchto tvrzení vyplývá z textu?

(A) jen tvrzení I
(B) jen tvrzení II
(C) jen tvrzení III
(D) jen tvrzení I a II
(E) všechna tři tvrzení

30.

Nejvýznamnějším celosvětovým problémem je globální oteplování, které souvisí s celou řadou antropogenních procesů, jako například spalováním fosilních paliv, odlesňováním, intenzifikací zemědělství či průmyslovými emisemi.

Které z následujících tvrzení vyplývá z uvedeného textu?

(A) Významným problémem je odlesňování a následná intenzifikace zemědělství.
(B) Globální oteplování nijak nesouvisí s přirozenými procesy v přírodě, je způsobeno jen lidskou činností.
(C) Globální oteplování přinese celou řadu problémů pro antropogenní procesy.
(D) Spalování fosilních paliv produkuje průmyslové emise, které jsou velkým celosvětovým problémem.
(E) Na globálním oteplování se podílejí odlesňování, intenzifikace zemědělství, spalování fosilních paliv, průmyslové emise a další procesy.

KVANTITATIVNÍ ODDÍL

Není dovoleno používat kalkulačky!

Není-li uvedeno jinak, jsou všechna použitá čísla reálná. Čáry, které se jeví jako přímé, považujte za přímky. ***O velikosti neoznačených částí obrazců nelze dělat žádné předpoklady.*** Geometrické úlohy řešte pomocí matematických znalostí, nikoli odhadem či měřením z obrázku.

V úlohách 31 až 36 je vaším úkolem porovnat dvě hodnoty.

Informace týkající se jedné nebo obou hodnot jsou uvedeny vždy uprostřed nad rámečkem s hodnotami.

31.

počet hran krychle	12

(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

32.

Ivan vyrobí za jednu hodinu dvakrát tolik vlaštovek než Jakub, ale o 10 vlaštovek méně než Matouš.

Počet vlaštovek, které vyrobí za hodinu Matouš.	Trojnásobek počtu vlaštovek, které vyrobí za hodinu Jakub.

(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

33.

V loteriích *Superšance* i *Ultrašance* vyhrává pouze ten, kdo uhodne všechna tažená čísla.
V loterii *Superšance* se losuje 6 čísel ze 40,
v loterii *Ultrašance* se losuje 37 čísel ze 40.

pravděpodobnost výhry v *Superšanci*	pravděpodobnost výhry v *Ultrašanci*

(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

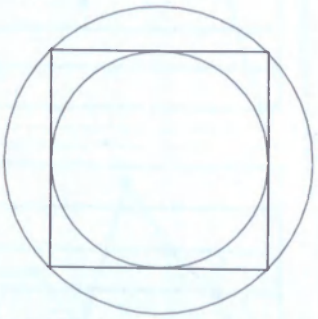
34.

$x > y$

$x + 3y$	$2x + y$

(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

35.



obsah malého kruhu	polovina obsahu čtverce

(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

36.

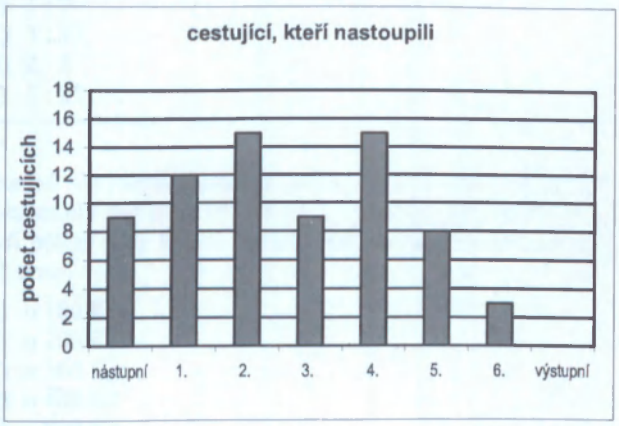
Máme číslo. Pokud ho postupně odmocníme, vynásobíme osmi, přičteme dvě a vynásobíme pěti, dostaneme číslo o 10 menší než sedminásobek čísla 20.

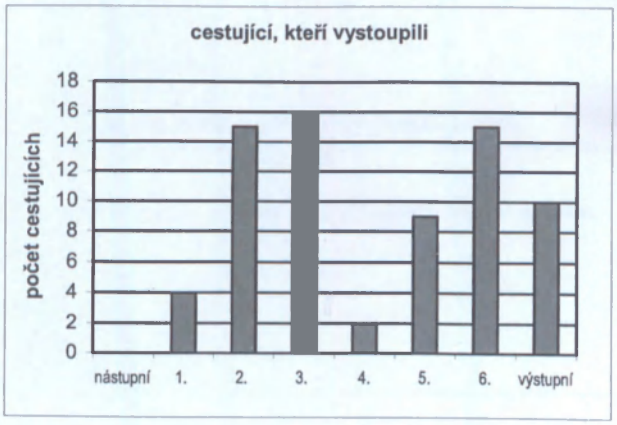
původní číslo	10

(A) Větší je hodnota vlevo.
(B) Větší je hodnota vpravo.
(C) Obě hodnoty jsou stejně velké.
(D) Nelze určit, která hodnota je větší.
(E) Žádná z možností (A) až (D) není správná.

GRAF K ÚLOHÁM 37 AŽ 39

Lanová dráha má nástupní a výstupní zastávku a šest zastávek mezi nimi. V grafech je znázorněno, jak se měnil počet cestujících během jedné cesty.





37.

O kolik více nebo méně cestujících jelo mezi druhou a třetí zastávkou v porovnání s cestou od první ke druhé zastávce?

(A) o 3 cestující méně
(B) o 3 cestující více
(C) o 8 cestujících více
(D) o 17 cestujících více
(E) Počet cestujících byl stejný.

38.

Které z následujících tvrzení je v souladu s údaji v grafech?

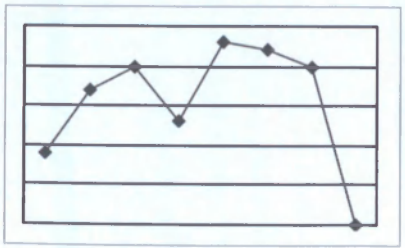
(A) Na páté zastávce nastoupilo více lidí, než vystoupilo.
(B) Od čtvrté zastávky se počet přepravovaných cestujících snižoval.
(C) Počet přepravovaných cestujících mezi zastávkami nikdy nebyl nižší než 10.
(D) Mezi druhou a třetí zastávkou se nepřepravoval nikdo.
(E) Žádné z tvrzení (A) až (D) není v souladu s údaji v grafech.

39.

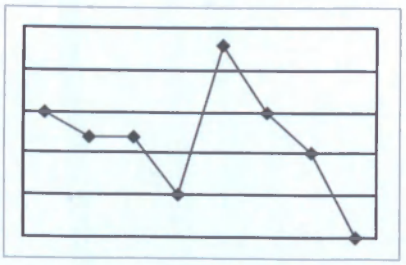
Z následujících grafu vyberte ten, který nejlépe vystihuje počty cestujících na lanové dráze na jednotlivých zastávkách.

(A)

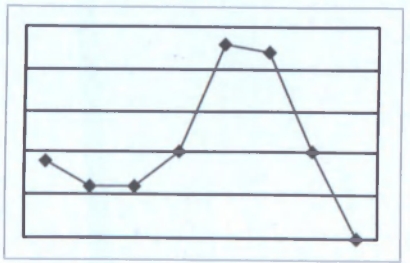
(B)



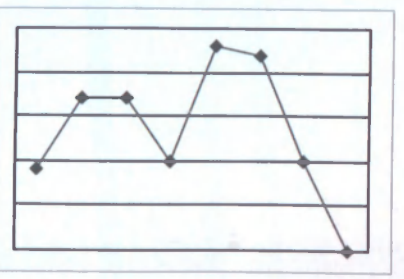
(C)



(D)



(E)



40.

Objemy malé a velké krychle jsou v poměru 1 : 27. V jakém poměru jsou délky jejich hran?

(A) 1 : 3
(B) 1 : 9
(C) 3 : 27
(D) 2 : 3
(E) 1 : 27

41.

Běta si v obchodě koupila jeden svetr a jedny kalhoty. Dohromady zaplatila 1800 Kč. Cena svetru byla stejná jako čtyři pětiny ceny kalhot. O kolik korun byly kalhoty dražší než svetr?

(A) o 180 Kč
(B) o 200 Kč
(C) o 360 Kč
(D) o 400 Kč
(E) o 800 Kč

42.

Desetipatrový dům má výšku 36 m. Mezi jednotlivými patry je 24 schodů. Kolik schodů by bylo mezi jednotlivými patry, kdyby byl každý schod o 5 cm vyšší?

(A) 24
(B) 21
(C) 20
(D) 18
(E) 15

43.

Noviny, vycházející denně mimo neděli, stojí ve všední den 8 Kč a v sobotu o čtvrtinu více. Jaká je výše předplatného na 30 týdnů, mají-li předplatitelé nárok na slevu ve výši 15 %?

(A) 206 Kč
(B) 1275 Kč
(C) 1375 Kč
(D) 1500 Kč
(E) Žádná z možností (A) až (D) není správná.

44.

Akvárium o rozměrech 40 × 50 × 60 cm se plní rychlostí 2 litry za minutu. Jak dlouho bude trvat naplnění akvária z jedné poloviny?

(A) 3 minuty
(B) 6 minut
(C) 0,5 hodiny
(D) 1 hodinu
(E) Žádná z možností (A) až (D) není správná.

45.

Určete součin součtu sedmi a tří a podílu jedné a dvou.

(A) 1
(B) 2
(C) 4
(D) 5
(E) 10

Konec zkoušky!

Příloha 4 : Statistické charakteristiky variant A, B, C, D testu OSP získané z programů ITEMAN a LERTAP

	Varianta A								Varianta B							
	celkem		verbal		analyt		kvantit		celkem		verbal		analyt		kvantit	
	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor
počet úloh	45		17		13		15		45		17		13		15	
Počet testovaných	451								439							
Průměrný hrubý skór	24,62	21,74	12,27	11,28	7,46	6,64	4,89	3,82	25,92	23,26	12,74	11,84	7,62	6,75	5,57	4,66
Rozptyl skórů	37,32	48,79	6,89	10,06	7,29	9,38	8,71	9,26	42,74	55,14	7,65	11,05	8,43	10,96	10,37	10,68
Směrodatná odchylka skórů	6,11	6,98	2,63	3,17	2,70	3,06	2,95	3,04	6,54	7,43	2,77	3,32	2,90	3,31	3,22	3,27
Sikmost rozložení	0,06	0,03	-0,63	-0,68	-0,13	-0,14	0,62	0,86	-0,04	-0,08	-0,70	-0,67	-0,20	-0,20	0,21	0,41
Spičatost rozložení	0,33	0,31	0,46	0,67	-0,50	-0,43	0,37	0,87	-0,16	-0,15	0,49	0,37	-0,65	-0,62	-0,72	-0,59
Nejnižší dosažený skór	5	-2,5	3	-0,5	0	-1,75	0	-3,75	2	0	2	-0,5	0	-2	0	-1,5
Nejvyšší dosažený skór	45		17		13		15		42	41,5	17		13		14	13,75
Medián	25	21,5	12	11,5	8	6,75	5	3,5	26	23,25	13	12	8	6,75	5	4,25
Koef. reliability (Cronb. alfa)	0,78	0,77	0,60	0,60	0,64	0,62	0,72	0,64	0,81	0,80	0,65	0,64	0,70	0,69	0,75	0,68
Index reliability	0,88	0,88	0,77	0,77	0,80	0,79	0,85	0,80	0,90	0,89	0,81	0,80	0,84	0,83	0,87	0,83
Směrodatná chyba měření	2,86	3,36	1,66	2,02	1,61	1,88	1,57	1,82	2,87	3,36	1,63	1,98	1,58	1,85	1,61	1,84
Relativní průměrný skór	0,55	0,48	0,72	0,66	0,57	0,51	0,33	0,25	0,58	0,52	0,75	0,70	0,59	0,52	0,37	0,31
r_{pb} (průměrná citlivost úloh)	0,24	0,23	0,22	0,22	0,28	0,27	0,32	0,30	0,26	0,25	0,26	0,25	0,33	0,31	0,36	0,30
r_b (průměrná citlivost úloh)	0,33								0,36							

	Varianta C								Varianta D							
	celkem		verbal		analyt		kvantit		celkem		verbal		analyt		kvantit	
	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor	nekor	kor
počet úloh	45		17		13		15		45		17		13		15	
Počet testovaných	467								339							
Průměrný hrubý skór	26,86	24,33	13,17	12,42	8,38	7,70	5,31	4,21	25,42	22,63	12,19	11,28	8,03	7,16	5,20	4,19
Rozptyl skórů	39	52,05	6,82	9,73	7,70	10,00	8,90	9,92	34,74	46,49	6,40	9,06	7,34	9,75	8,82	9,43
Směrodatná odchylka skórů	6,24	7,21	2,61	3,12	2,77	3,16	2,98	3,15	5,89	6,82	2,53	3,01	2,71	3,12	2,97	3,07
Sikmost rozložení	-0,24	-0,41	-1,07	-1,09	-0,37	-0,44	0,37	0,46	0,10	0,04	-0,27	-0,27	-0,26	-0,21	0,41	0,64
Spičatost rozložení	0,41	0,71	1,49	1,63	-0,33	-0,19	-0,32	-0,17	0,05	0,19	-0,44	-0,46	-0,36	-0,40	-0,26	0,06
Nejnižší dosažený skór	6	-2,5	3	-0,5	0	-2	0	-1,75	8	2,5	6	3,25	0	-3,25	0	-1,25
Nejvyšší dosažený skór	44	43,75	17		13		14	13,75	44	43,75	17		13		15	
Medián	27	24,50	14	13,25	8	7,75	5	3,75	25	22,50	12	11,50	8	7,25	5	3,75
Koef. reliability (Cronb. alfa)	0,80	0,79	0,63	0,63	0,68	0,67	0,71	0,66	0,76	0,76	0,56	0,55	0,66	0,64	0,71	0,64
Index reliability	0,89	0,89	0,80	0,79	0,83	0,82	0,84	0,81	0,87	0,87	0,75	0,74	0,81	0,80	0,84	0,80
Směrodatná chyba měření	2,81	3,27	1,58	1,90	1,56	1,82	1,60	1,84	2,87	3,37	1,67	2,02	1,57	1,87	1,60	1,84
Relativní průměrný skór	0,60	0,54	0,78	0,73	0,65	0,59	0,35	0,28	0,57	0,50	0,72	0,66	0,62	0,55	0,35	0,28
r_{pb} (průměrná citlivost úloh)	0.26	0,25	0,25	0,24	0,31	0,30	0,32	0,27	0,23	0,22	0,19	0,19	0,30	0,28	0,31	0,26
r_b (průměrná citlivost úloh)	0.35								0,31							

Příloha 5 : Změna reliability po vynechání příslušné úlohy

	Varianta A		Varianta B		Varianta C		Varianta D	
	alfa = 0,780		alfa = 0,807		alfa = 0,798		alfa = 0,763	
bez úlohy	alfa	změna	alfa	změna	alfa	změna	alfa	změna
Q1	0,779	-0,002	0,807	-0,001	0,798	**0,000**	0,765	**0,002**
Q2	0,778	-0,002	0,808	**0,001**	0,795	-0,004	0,760	-0,003
Q3	0,780	**0,000**	0,805	-0,003	0,795	-0,003	0,759	-0,005
Q4	0,779	-0,001	0,807	-0,001	0,797	-0,001	0,763	-0,001
Q5	0,780	**0,000**	0,803	-0,004	0,796	-0,003	0,760	-0,003
Q6	0,780	-0,001	0,805	-0,003	0,793	-0,005	0,763	0,000
Q7	0,780	-0,001	0,807	-0,001	0,795	-0,003	0,760	-0,003
Q8	0,778	-0,002	0,800	-0,007	0,794	-0,004	0,761	-0,002
Q9	0,776	-0,004	0,805	-0,002	0,797	-0,002	0,762	-0,001
Q10	0,776	-0,004	0,808	**0,000**	0,797	-0,001	0,760	-0,004
Q11	0,777	-0,004	0,802	-0,006	0,797	-0,001	0,759	-0,005
Q12	0,776	-0,004	0,805	-0,002	0,796	-0,002	0,760	-0,004
Q13	0,778	-0,003	0,806	-0,002	0,795	-0,003	0,762	-0,002
Q14	0,777	-0,003	0,804	-0,004	0,793	-0,006	0,763	-0,001
Q15	0,778	-0,002	0,806	-0,001	0,795	-0,004	0,759	-0,004
Q16	0,776	-0,005	0,804	-0,003	0,795	-0,003	0,755	-0,008
Q17	0,771	-0,009	0,806	-0,002	0,796	-0,003	0,757	-0,007
Q18	0,775	-0,005	0,803	-0,004	0,794	-0,005	0,761	-0,002
Q19	0,781	**0,001**	0,801	-0,007	0,795	-0,003	0,764	**0,001**
Q20	0,774	-0,006	0,802	-0,005	0,794	-0,005	0,760	-0,003
Q21	0,777	-0,003	0,802	-0,006	0,794	-0,004	0,755	-0,008
Q22	0,774	-0,007	0,801	-0,006	0,791	-0,007	0,756	-0,007
Q23	0,778	-0,003	0,802	-0,005	0,792	-0,006	0,758	-0,006
Q24	0,782	**0,002**	0,801	-0,006	0,791	-0,007	0,756	-0,008
Q25	0,774	-0,007	0,802	-0,006	0,790	-0,009	0,758	-0,005
Q26	0,780	-0,001	0,805	-0,002	0,796	-0,003	0,759	-0,004
Q27	0,772	-0,008	0,804	-0,003	0,793	-0,005	0,759	-0,004
Q28	0,775	-0,006	0,804	-0,004	0,793	-0,005	0,758	-0,005
Q29	0,772	-0,009	0,807	-0,001	0,795	-0,003	0,755	-0,008
Q30	0,774	-0,007	0,803	-0,005	0,799	**0,001**	0,758	-0,005
Q31	0,774	-0,006	0,801	-0,006	0,797	-0,002	0,759	-0,004
Q32	0,778	-0,003	0,806	-0,002	0,795	-0,003	0,762	-0,001
Q33	0,781	**0,001**	0,810	**0,003**	0,798	0,000	0,756	-0,007
Q34	0,771	-0,009	0,800	-0,007	0,793	-0,005	0,765	**0,002**
Q35	0,774	-0,007	0,807	-0,001	0,799	**0,001**	0,762	-0,001
Q36	0,770	-0,010	0,804	-0,003	0,796	-0,002	0,756	-0,007
Q37	0,775	-0,006	0,804	-0,004	0,791	-0,007	0,758	-0,006
Q38	0,779	-0,001	0,802	-0,005	0,790	-0,009	0,754	-0,010
Q39	0,776	-0,005	0,802	-0,005	0,791	-0,008	0,757	-0,006
Q40	0,776	-0,005	0,807	0,000	0,794	-0,004	0,761	-0,003
Q41	0,777	-0,004	0,800	-0,008	0,797	-0,001	0,755	-0,008
Q42	0,778	-0,002	0,798	-0,010	0,795	-0,004	0,753	-0,010
Q43	0,775	-0,005	0,800	-0,007	0,797	-0,001	0,762	-0,001
Q44	0,778	-0,003	0,805	-0,003	0,788	-0,010	0,754	-0,010
Q45	0,775	-0,006	0,805	-0,003	0,795	-0,003	0,761	-0,002

	A	B	C	D	E	other	diff.	disc.	?
Q1	1%	2%	96%		1%	0%	0,96	0,17	D
Q2	76%		11%	4%	4%	5%	0,76	0,19	B
Q3	1%	98%			0%	0%	0,98	0,12	CD
Q4	8%	8%	2%	27%	48%	7%	0,48	0,18	
Q5	2%	70%	6%	4%	14%	5%	0,70	0,14	
Q6	6%	78%	7%	3%	4%	3%	0,78	0,15	D
Q7	6%	1%	1%	87%	4%	1%	0,87	0,13	
Q8	1%	2%	6%	15%	74%	2%	0,74	0,20	
Q9	13%	2%	1%	6%	76%	2%	0,76	0,25	
Q10	2%	26%	49%	7%	3%	14%	0,49	0,24	E
Q11	3%	2%	10%	65%	9%	12%	0,65	0,23	
Q12	13%	12%	17%	7%	39%	13%	0,39	0,24	
Q13	65%	20%	1%	8%	1%	4%	0,65	0,21	
Q14	2%	2%	0%	90%	3%	3%	0,90	0,23	
Q15	2%	3%	74%	16%	4%	2%	0,74	0,19	
Q16	73%	2%	21%	1%	1%	4%	0,73	0,27	
Q17	10%	6%	70%	4%	7%	3%	0,70	0,38	
Q18	4%	2%	5%	76%	4%	10%	0,76	0,27	
Q19	48%	6%	6%	6%	8%	26%	0,48	0,12	
Q20	12%	1%	62%	3%	7%	16%	0,62	0,30	
Q21	5%	5%	9%	5%	52%	24%	0,52	0,23	D
Q22	6%	8%	39%	6%	10%	32%	0,39	0,31	
Q23	1%	10%	5%	2%	55%	27%	0,55	0,21	
Q24	2%	7%	2%	30%	31%	29%	0,30	0,08	E
Q25	5%	2%	1%	14%	71%	7%	0,71	0,32	
Q26	13%	59%	13%	2%	6%	8%	0,59	0,16	C
Q27	1%	7%	3%	62%	12%	16%	0,62	0,35	
Q28	71%	4%	1%	5%	9%	10%	0,71	0,29	
Q29	4%	51%	8%	17%	14%	7%	0,51	0,36	
Q30	2%	7%	5%	2%	73%	12%	0,73	0,31	
Q31	6%	11%	72%	1%	0%	10%	0,72	0,30	
Q32	17%	22%	11%	24%	1%	26%	0,24	0,20	CE
Q33	29%	47%	2%	1%	1%	21%	0,47	0,13	A
Q34	20%	16%	14%	31%	1%	18%	0,31	0,38	E
Q35	52%	7%	7%	3%	0%	31%	0,52	0,31	
Q36	15%	38%	2%	4%	1%	41%	0,38	0,40	
Q37	16%	18%	9%	2%	13%	41%	0,13	0,32	D
Q38	2%	42%	3%	2%	10%	41%	0,42	0,18	A
Q39	11%	14%	8%	5%	15%	47%	0,15	0,28	D
Q40	31%	9%	1%	1%	17%	42%	0,31	0,26	B
Q41	5%	22%	16%	2%	5%	50%	0,22	0,24	AE
Q42	5%	5%	6%	15%	2%	67%	0,15	0,19	BC
Q43	4%	25%	8%	3%	10%	51%	0,25	0,27	CD
Q44	7%	8%	15%	5%	6%	59%	0,15	0,21	ADE
Q45	1%	2%	2%	10%	2%				

Res =	A	B	C	D	E	other	diff.	disc.	?
Q1	97%	0%		1%	1%	0%	0,97	0,16	C
Q2	10%	6%	2%	81%		1%	0,81	0,10	E
Q3	0%	13%	7%	0%	78%	1%	0,78	0,24	
Q4	69%	3%	7%	6%	13%	3%	0,69	0,18	B
Q5	14%	3%	72%	2%	7%	1%	0,72	0,28	
Q6	3%	72%	9%	11%	0%	5%	0,72	0,24	E
Q7	5%	12%	73%	2%	3%	5%	0,73	0,16	E
Q8	11%	8%	9%	7%	59%	5%	0,59	0,37	
Q9	6%	1%	1%	89%	2%	2%	0,89	0,21	
Q10	1%	4%	2%	70%	15%	8%	0,70	0,14	
Q11	58%	7%	8%	8%	7%	12%	0,58	0,33	
Q12	1%	12%	75%	3%	2%	7%	0,75	0,21	
Q13	95%	1%	2%	1%		1%	0,95	0,20	E
Q14	24%	69%	1%	1%	3%	2%	0,69	0,26	
Q15	4%		24%	15%	52%	6%	0,52	0,20	B
Q16	3%	8%	2%	1%	77%	8%	0,77	0,25	
Q17	0%	4%	0%	87%	6%	2%	0,87	0,19	
Q18	6%	62%	3%	3%	3%	22%	0,62	0,28	
Q19	3%	2%	4%	60%	21%	11%	0,60	0,36	
Q20	3%	12%	6%	2%	59%	19%	0,59	0,31	
Q21	7%	7%	5%	51%	3%	28%	0,51	0,33	
Q22	72%	3%	4%	5%	6%	10%	0,72	0,36	
Q23	4%	7%	17%	4%	61%	8%	0,61	0,32	
Q24	9%	2%	37%	10%	25%	17%	0,37	0,34	E
Q25	11%	64%	4%	11%	4%	6%	0,64	0,33	
Q26	4%	60%	13%	5%	7%	12%	0,60	0,22	
Q27	6%	13%	8%	54%	3%	16%	0,54	0,26	
Q28	2%	1%	74%	4%	2%	18%	0,74	0,27	
Q29	4%	2%	47%	33%	2%	12%	0,33	0,17	C
Q30	3%	1%	0%	4%	77%	13%	0,77	0,31	
Q31	36%	21%	2%	9%	1%	31%	0,36	0,35	
Q32	18%	65%	5%	0%		12%	0,65	0,21	E
Q33	2%	1%	31%	34%	3%	28%	0,31	0,06	D
Q34	5%	2%	62%	8%	2%	21%	0,62	0,38	
Q35	26%	32%	1%	1%	0%	39%	0,32	0,17	C
Q36	18%	41%	8%	3%	1%	30%	0,41	0,26	A
Q37	10%	2%	20%	3%	23%	40%	0,20	0,27	A
Q38	52%	2%	1%	1%	2%	42%	0,52	0,32	
Q39	0%	51%	1%	0%	21%	26%	0,21	0,32	
Q40	3%	4%	50%	6%	13%	25%	0,50	0,16	DE
Q41	5%	3%	1%	5%	51%	34%	0,51	0,39	
Q42	1%	2%	3%	4%	46%	45%	0,46	0,44	
Q43	6%	3%	25%	4%	4%	57%	0,25	0,39	A
Q44	3%	5%	6%	11%	1%	74%	0,11	0,22	

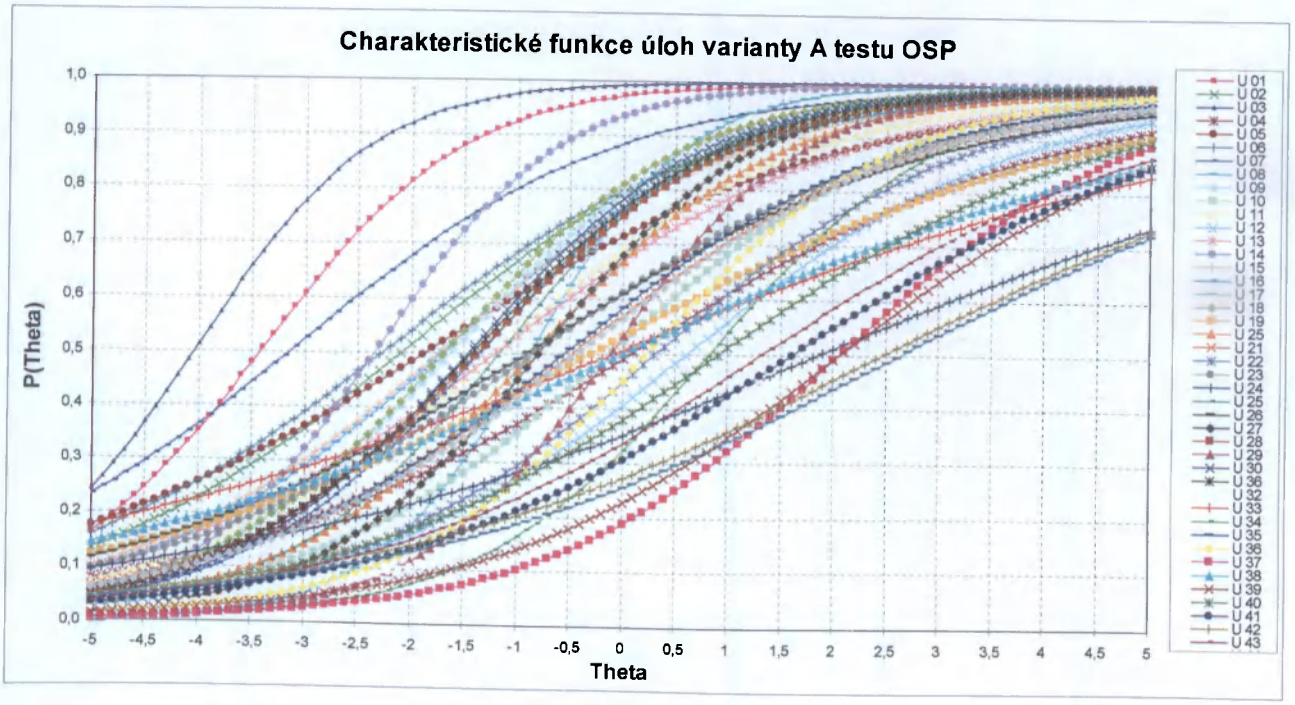
Res =	A	B	C	D	E	other	diff.	disc.	?
Q1	15%	2%	80%	1%	1%	1%	0,80	0,12	
Q2	0%	83%	13%	1%	1%	1%	0,83	0,25	
Q3	85%	3%	6%	3%	2%	2%	0,85	0,22	
Q4	0%	1%	1%	96%	1%	1%	0,96	0,17	
Q5	4%	30%	3%	3%	56%	5%	0,56	0,23	
Q6	1%	3%	2%	1%	91%	1%	0,91	0,36	
Q7	22%	2%	72%	1%	1%	1%	0,72	0,24	BE
Q8	6%	2%	0%	89%	2%	2%	0,89	0,30	
Q9	3%	3%	66%	13%	10%	5%	0,66	0,20	B
Q10	4%	6%	5%	1%	71%	12%	0,71	0,18	
Q11	12%	10%	1%	56%	4%	17%	0,56	0,19	
Q12	68%	0%	5%	1%	3%	22%	0,68	0,20	B
Q13	5%	84%	2%	4%	3%	3%	0,84	0,24	
Q14	1%	4%	9%	81%	3%	3%	0,81	0,32	
Q15	3%	16%	1%	4%	73%	3%	0,73	0,25	
Q16	1%	3%	88%	3%	1%	2%	0,88	0,25	
Q17	15%	3%	79%	1%	0%	2%	0,79	0,22	
Q18	1%	6%	4%	78%	2%	8%	0,78	0,29	
Q19	2%	1%	13%	3%	72%	9%	0,72	0,24	
Q20	66%	6%	5%	7%	2%	14%	0,66	0,28	
Q21	2%	0%	2%	61%	15%	20%	0,61	0,26	
Q22	2%	5%	50%	7%	7%	29%	0,50	0,35	
Q23	7%	54%	7%	2%	0%	30%	0,54	0,33	
Q24	1%	5%	6%	53%	1%	34%	0,53	0,35	
Q25	66%	2%	11%	2%	7%	12%	0,66	0,39	
Q26	4%	31%	5%	4%	50%	5%	0,50	0,23	
Q27	2%	9%	65%	8%	4%	13%	0,65	0,30	
Q28	1%	8%	3%	85%		3%	0,85	0,31	E
Q29	12%	3%	69%	5%	4%	6%	0,69	0,25	
Q30	1%	1%	11%	9%	69%	8%	0,69	0,12	C
Q31	30%	23%	17%	5%	1%	23%	0,30	0,19	B
Q32	75%	8%	4%	4%		9%	0,75	0,23	E
Q33	33%	15%	16%	3%	0%	33%	0,33	0,16	
Q34	37%	43%	4%	1%	1%	13%	0,43	0,29	
Q35	20%	17%	7%	13%	6%	37%	0,20	0,09	BD
Q36	6%	64%	2%	10%	0%	18%	0,64	0,22	D
Q37	3%	5%	9%	8%	23%	52%	0,23	0,36	
Q38	1%	3%	3%	54%	1%	38%	0,54	0,39	E
Q39	5%	49%	4%	1%	1%	41%	0,49	0,36	
Q40	4%	5%	6%	16%	4%	65%	0,16	0,28	B
Q41	6%	27%	10%	8%	2%	47%	0,27	0,16	D
Q42	2%	24%	5%	34%	1%	34%	0,24	0,26	
Q43	4%	24%	4%	27%	5%	36%	0,24	0,18	D
Q44	6%	2%	6%	13%	36%	38%	0,36	0,43	
Q45	9%	19%	10%	14%	2%	46%	0,14	0,24	BE

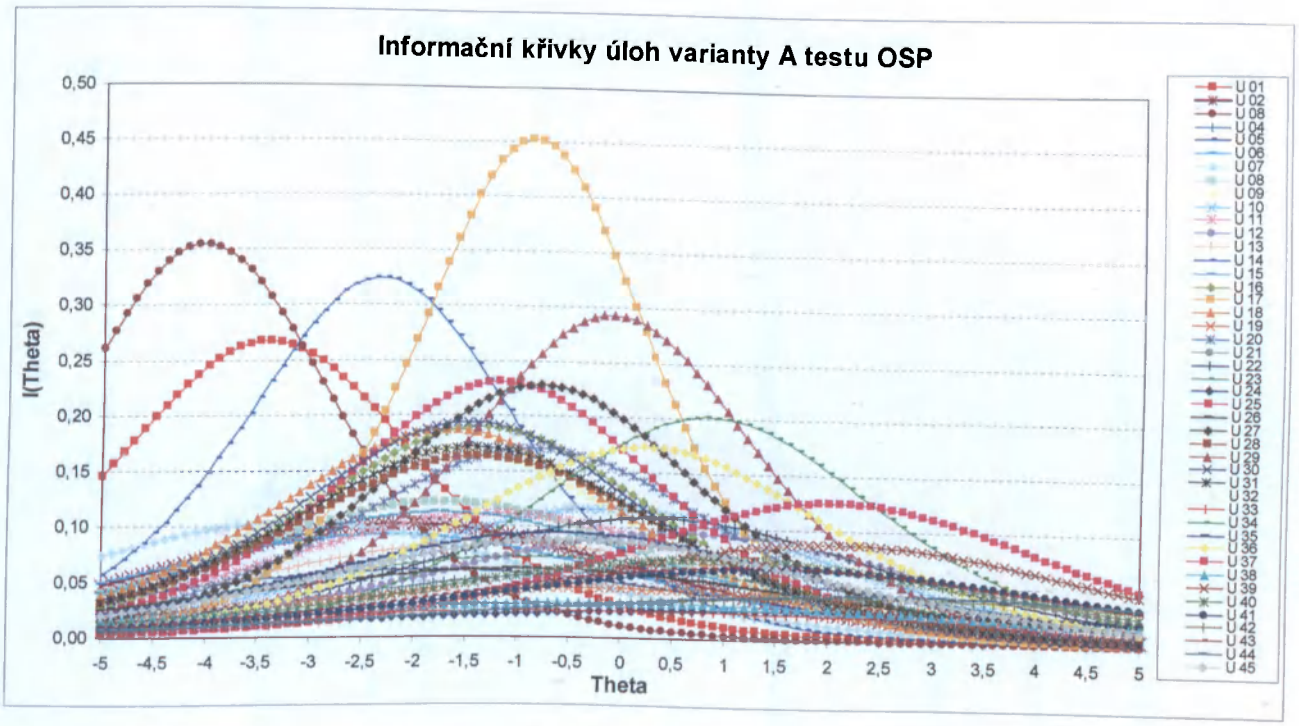
Res =	A	B	C	D	E	other	diff.	disc.	?
Q1	1%	1%	19%	1%	78%	1%	0,78	0,06	
Q2	1%	4%	94%	1%	0%	1%	0,94	0,22	
Q3	23%	5%	4%	49%	9%	10%	0,49	0,24	
Q4	86%	8%	5%	0%	0%	0%	0,86	0,12	E
Q5	1%	1%	4%	14%	78%	2%	0,78	0,21	A
Q6	8%	48%	7%	15%	10%	12%	0,48	0,14	
Q7	5%	8%	11%	71%	1%	4%	0,71	0,20	
Q8	3%	15%	7%	8%	62%	6%	0,62	0,18	
Q9	1%	93%	3%	1%	0%	2%	0,93	0,15	AE
Q10	53%	2%	10%	12%	2%	22%	0,53	0,22	
Q11	6%	7%	57%	5%	3%	22%	0,57	0,24	
Q12	4%	4%	2%	42%	22%	25%	0,42	0,21	C
Q13	13%	1%	77%	6%	1%	2%	0,77	0,16	B
Q14	0%	0%	1%	5%	91%	2%	0,91	0,11	
Q15	1%	84%	4%	2%	9%	1%	0,84	0,23	A
Q16	77%	2%	4%	13%	2%	2%	0,77	0,34	
Q17	2%	9%	1%	7%	78%	2%	0,78	0,29	
Q18	12%	2%	66%	6%	9%	5%	0,66	0,17	
Q19	61%	4%	5%	4%	12%	13%	0,61	0,11	C
Q20	1%	11%	6%	5%	64%	13%	0,64	0,20	A
Q21	17%	3%	4%	60%	2%	15%	0,60	0,31	BE
Q22	65%	1%	3%	0%	15%	16%	0,65	0,30	
Q23	19%	4%	50%	4%	2%	21%	0,50	0,26	
Q24	6%	4%	12%	19%	34%	25%	0,34	0,31	B
Q25	9%	74%	4%	3%	8%	3%	0,74	0,25	
Q26	65%	1%	11%	6%	9%	8%	0,65	0,23	
Q27	1%	2%	4%	0%	89%	4%	0,89	0,25	
Q28	2%	42%	3%	19%	26%	8%	0,42	0,26	
Q29	65%	13%	4%	9%	1%	8%	0,65	0,31	
Q30	3%	9%	4%	68%	2%	14%	0,68	0,25	
Q31	33%	15%	15%	3%	1%	32%	0,15	0,22	A
Q32	38%	25%	4%	6%	1%	26%	0,38	0,15	C
Q33	20%	64%	1%		1%	14%	0,64	0,30	D
Q34	4%	42%	5%	10%		40%	0,42	0,09	DE
Q35	15%	18%	13%	21%	0%	33%	0,21	0,13	AE
Q36	1%	3%	64%	3%	1%	27%	0,64	0,30	
Q37	7%	48%	13%	11%	1%	20%	0,48	0,26	
Q38	47%	9%	5%	4%	3%	32%	0,47	0,35	
Q39	3%	6%	4%	24%	8%	56%	0,24	0,27	
Q40	3%	8%	24%	9%	4%	53%	0,24	0,19	A
Q41	5%	3%	13%	3%	29%	47%	0,29	0,31	B
Q42	1%	17%	50%	7%	1%	24%	0,50	0,36	AE
Q43	4%	15%	8%	7%	8%	58%	0,15	0,14	ACDE
Q44	1%	7%	9%	30%	3%	50%	0,30	0,36	ABE
Q45	8%	12%	5%	9%	3%	63%	0,09	0,17	BC

Příloha 7 : Hodnoty parametrů obtížnosti a citlivosti úloh testu OSP

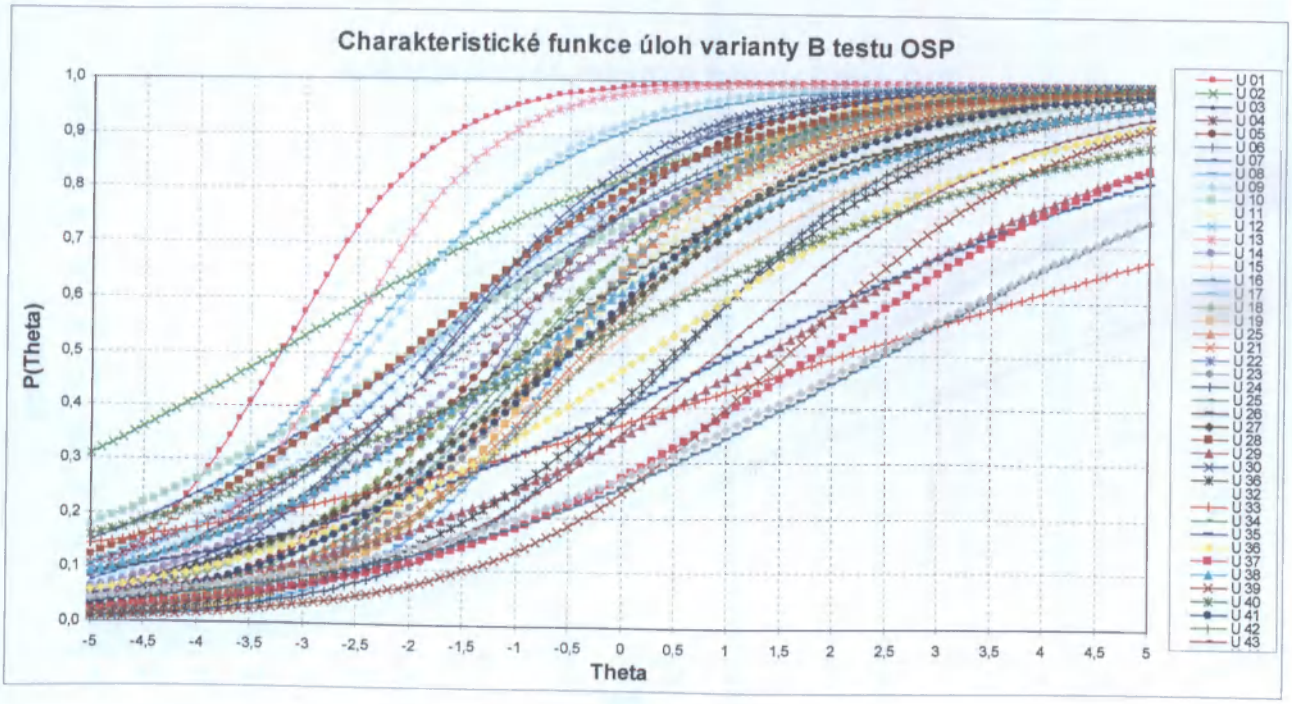
	varianta A				varianta B				varianta C				varianta D			
	obtížnost		citlivost		obtížnost		citlivost		obtížnost		citlivost		obtížnost		citlivost	
	KTT	IRT	KTT	IRT	KTT	IRT	KTT	IRT	KTT	IRT	KTT	IRT	KTT	IRT	KTT	IRT
úloha	*p*	*b*	r_{pb}	*a*	*p*	*b*	r_{pb}	*a*	*p*	*b*	r_{pb}	*a*	*p*	*b*	r_{pb}	*a*
1	0,96	-3,42	0,17	0,609	0,97	-3,22	0,16	0,798	0,80	-2,74	0,12	0,317	0,78	-2,88	0,06	0,279
2	0,76	-2,1	0,19	0,376	0,81	-3,26	0,10	0,272	0,83	-2	0,25	0,558	0,94	-2,45	0,22	0,814
3	0,98	-4,05	0,12	0,702	0,78	-1,68	0,24	0,52	0,85	-2,33	0,22	0,508	0,49	-0,08	0,24	0,382
4	0,48	0,074	0,18	0,287	0,69	-1,48	0,18	0,356	0,96	-3,41	0,17	0,615	0,86	-3,16	0,12	0,36
5	0,7	-1,87	0,14	0,291	0,72	-1,21	0,28	0,566	0,56	-0,46	0,23	0,384	0,78	-1,93	0,21	0,45
6	0,78	-2,27	0,15	0,358	0,72	-1,58	0,24	0,412	0,91	-1,85	0,36	1,108	0,48	-0,07	0,14	0,268
7	0,87	-3,15	0,13	0,38	0,73	-1,95	0,16	0,334	0,72	-1,43	0,24	0,447	0,71	-1,51	0,20	0,405
8	0,74	-1,72	0,2	0,411	0,59	-0,5	0,37	0,65	0,89	-2,1	0,30	0,743	0,62	-0,98	0,18	0,349
9	0,76	-1,63	0,25	0,485	0,89	-2,41	0,21	0,596	0,66	-1,28	0,20	0,352	0,93	-3,58	0,15	0,485
10	0,49	-0,12	0,24	0,403	0,70	-1,97	0,14	0,297	0,71	-1,96	0,18	0,321	0,53	-0,56	0,22	0,331
11	0,65	-1,21	0,23	0,393	0,58	-0,58	0,33	0,497	0,56	-0,79	0,19	0,311	0,57	-0,84	0,24	0,364
12	0,39	0,603	0,24	0,364	0,75	-1,85	0,21	0,417	0,68	-1,9	0,20	0,313	0,42	0,237	0,21	0,319
13	0,65	-1,21	0,21	0,346	0,95	-2,68	0,20	0,803	0,84	-2,26	0,24	0,493	0,77	-2,19	0,16	0,361
14	0,9	-2,35	0,23	0,671	0,69	-1,27	0,26	0,426	0,81	-1,67	0,32	0,635	0,91	-3,76	0,11	0,407
15	0,74	-1,74	0,19	0,395	0,52	-0,22	0,20	0,346	0,73	-1,53	0,25	0,445	0,84	-2,12	0,23	0,527
16	0,73	-1,36	0,27	0,512	0,77	-1,87	0,25	0,468	0,88	-2,46	0,25	0,58	0,77	-1,22	0,34	0,792
17	0,7	-0,88	0,38	0,792	0,87	-2,57	0,19	0,516	0,79	-1,93	0,22	0,457	0,78	-1,44	0,29	0,642
18	0,76	-1,69	0,27	0,509	0,62	-1	0,28	0,459	0,78	-1,69	0,29	0,549	0,66	-1,34	0,17	0,332
19	0,48	-0,29	0,12	0,249	0,60	-0,59	0,36	0,616	0,72	-1,56	0,24	0,437	0,61	-1,46	0,11	0,241
20	0,62	-0,85	0,3	0,486	0,59	-0,68	0,31	0,511	0,66	-1,16	0,28	0,468	0,64	-1,2	0,20	0,373
21	0,52	-0,48	0,23	0,349	0,51	-0,36	0,33	0,504	0,61	-1,01	0,26	0,393	0,60	-0,64	0,31	0,581
22	0,39	0,313	0,31	0,389	0,72	-1,11	0,36	0,693	0,50	-0,31	0,35	0,491	0,65	-1,05	0,30	0,504
23	0,55	-0,84	0,21	0,318	0,61	-0,64	0,32	0,555	0,54	-0,59	0,33	0,455	0,50	-0,24	0,26	0,456
24	0,3	1,853	0,08	0,193	0,37	0,543	0,34	0,484	0,53	-0,55	0,35	0,477	0,34	0,618	0,31	0,476
25	0,71	-1,17	0,32	0,565	0,64	-0,82	0,33	0,546	0,66	-0,81	0,39	0,718	0,74	-1,57	0,25	0,449
26	0,59	-0,9	0,16	0,292	0,60	-0,93	0,22	0,337	0,50	-0,09	0,23	0,369	0,65	-1,12	0,23	0,417
27	0,62	-0,79	0,35	0,562	0,54	-0,51	0,26	0,366	0,65	-1,05	0,30	0,459	0,89	-2,11	0,25	0,746
28	0,71	-1,37	0,29	0,475	0,74	-2,02	0,27	0,39	0,85	-1,86	0,31	0,703	0,42	0,394	0,26	0,442
29	0,51	-0,12	0,36	0,637	0,33	1,362	0,17	0,275	0,69	-1,32	0,25	0,436	0,65	-0,95	0,31	0,52
30	0,73	-1,47	0,31	0,518	0,77	-1,72	0,31	0,562	0,69	-2	0,12	0,278	0,68	-1,26	0,25	0,473
31	0,72	-1,43	0,3	0,487	0,36	0,485	0,35	0,43	0,30	1,488	0,19	0,26	0,15	2,616	0,22	0,303
32	0,24	1,613	0,2	0,346	0,65	-1,47	0,21	0,308	0,75	-2,04	0,23	0,378	0,38	0,682	0,15	0,252
33	0,47	-0,09	0,13	0,185	0,31	2,118	0,06	0,149	0,33	1,065	0,16	0,254	0,64	-1,03	0,30	0,45
34	0,31	0,817	0,38	0,528	0,62	-0,89	0,38	0,518	0,43	0,256	0,29	0,403	0,42	-0,02	0,09	0,177
35	0,52	-0,56	0,31	0,362	0,32	1,06	0,17	0,227	0,20	3,306	0,09	0,179	0,21	2,474	0,13	0,234
36	0,38	0,225	0,4	0,49	0,41	0,24	0,26	0,311	0,64	-1,47	0,22	0,313	0,64	-1,4	0,30	0,377
37	0,13	2,051	0,32	0,417	0,20	1,843	0,27	0,315	0,23	1,055	0,36	0,428	0,48	-0,15	0,26	0,386
38	0,42	-0,02	0,18	0,21	0,52	-0,84	0,32	0,323	0,54	-0,7	0,39	0,457	0,47	-0,22	0,35	0,452
39	0,15	2,055	0,28	0,349	0,21	1,57	0,32	0,422	0,49	-0,43	0,36	0,422	0,24	1,228	0,27	0,32
40	0,31	0,907	0,26	0,316	0,50	-0,54	0,16	0,219	0,16	1,821	0,28	0,315	0,24	1,467	0,19	0,265
41	0,22	1,601	0,24	0,299	0,51	-0,5	0,39	0,432	0,27	1,357	0,16	0,251	0,29	0,755	0,31	0,384
42	0,15	2,469	0,19	0,233	0,46	-0,27	0,44	0,492	0,24	1,431	0,26	0,357	0,50	-0,26	0,36	0,511
43	0,25	1,318	0,27	0,301	0,25	0,903	0,39	0,393	0,24	1,974	0,18	0,248	0,15	2,875	0,14	0,21
44	0,15	2,625	0,21	0,239	0,11	2,489	0,23	0,254	0,36	0,335	0,43	0,567	0,30	0,678	0,36	0,385
45	0,49	-0,44	0,28	0,352	0,13	2,373	0,23	0,245	0,14	2,363	0,24	0,315	0,09	3,334	0,17	0,242
r	-0,98		0,53		-0,97		0,43		-0,96		0,61		-0,97		0,57	

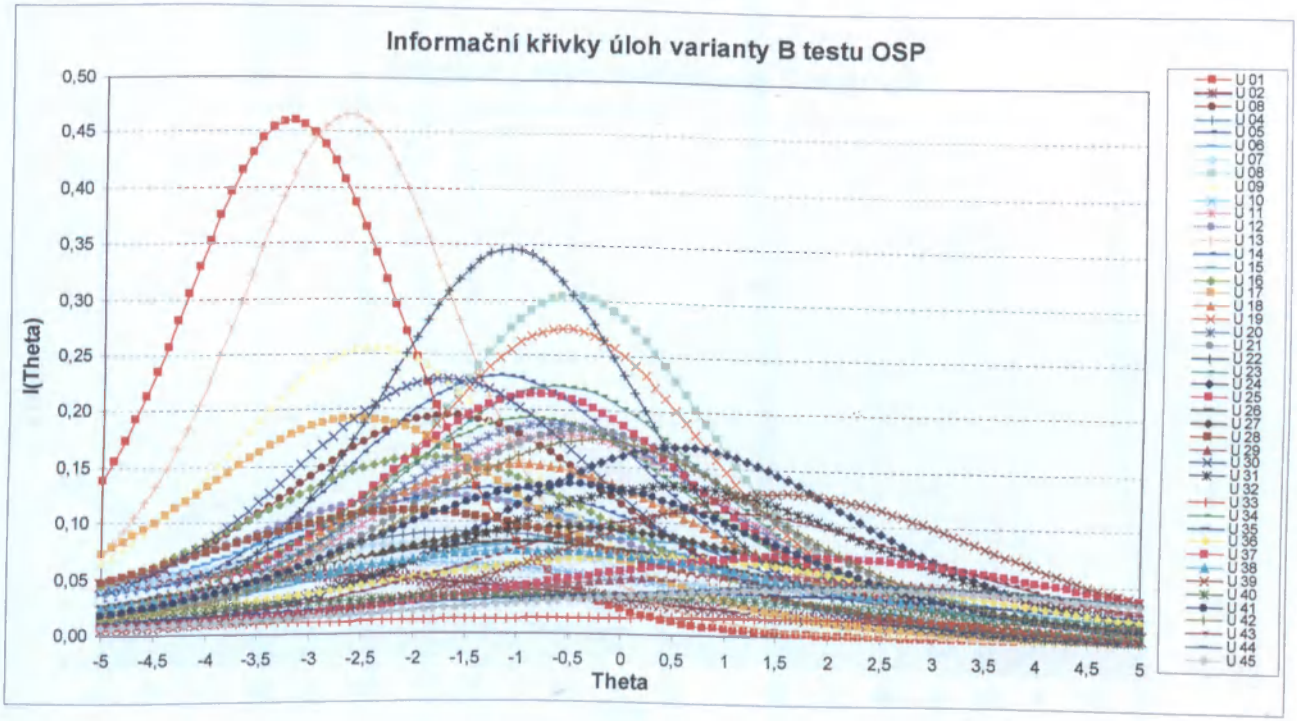
Příloha 8 : Charakteristické a informační křivky všech úloh (test OSP, varianta A)



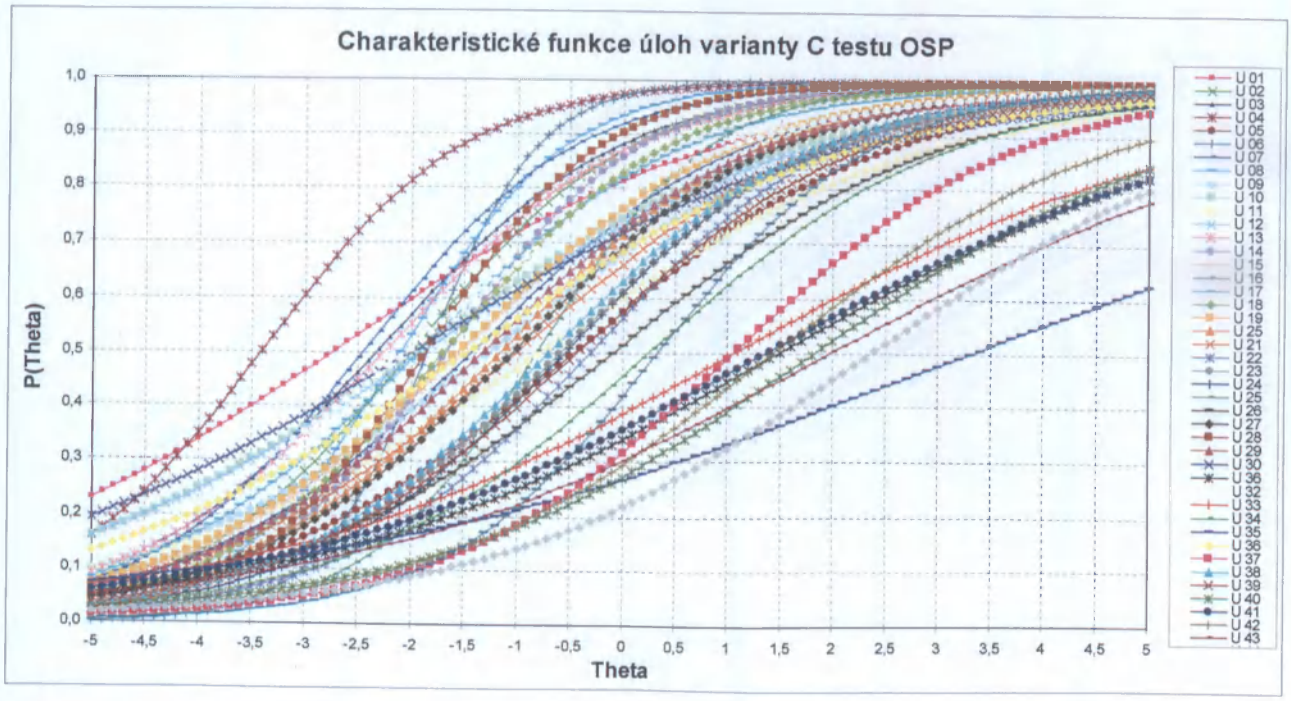


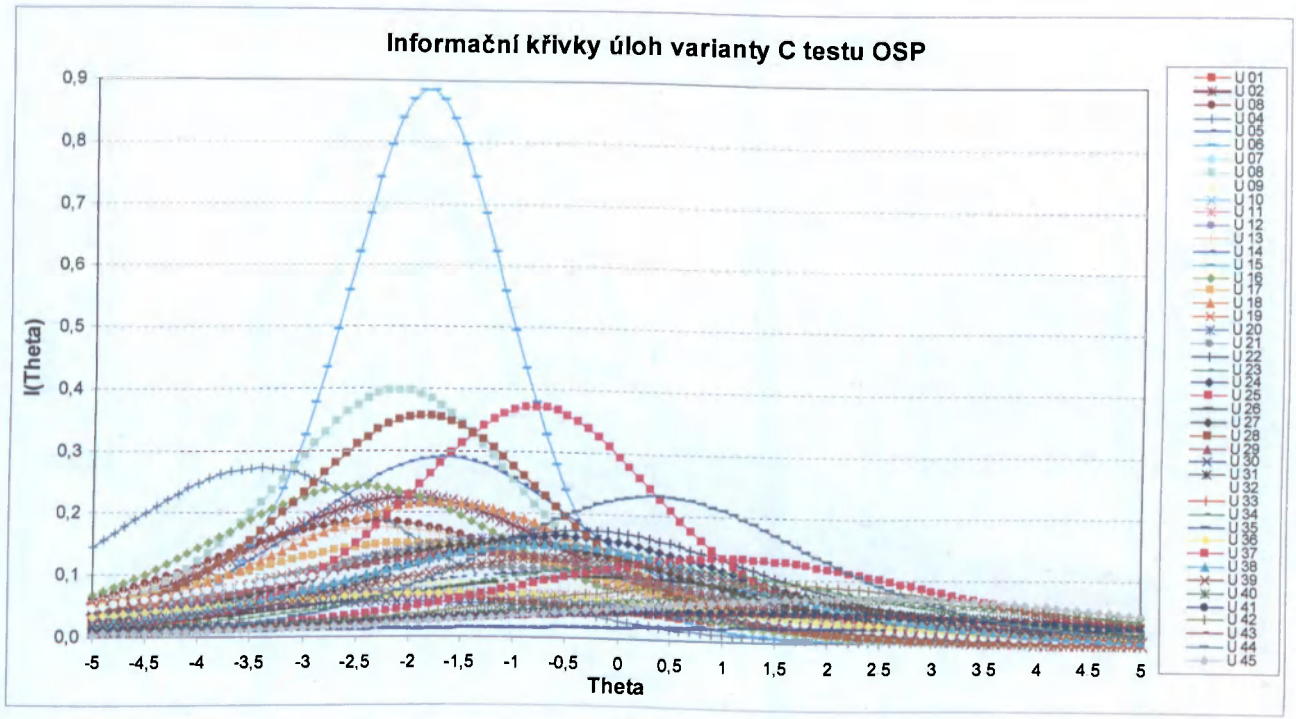
Příloha 9 : Charakteristické a informační křivky všech úloh (test OSP, varianta B)





Příloha 10 : Charakteristické a informační křivky všech úloh (test OSP, varianta C)





Příloha 11 : Charakteristické a informační křivky všech úloh (test OSP, varianta D)