Československá psychologie 2008 / ročník LII / číslo 2

Metodické studie

SOFTWARE PRO ADAPTIVNÍ TESTOVÁNÍ: CAT V PRAXI

PETR KVĚTON, MARTIN JELÍNEK, DENISA DENGLEROVÁ, DALIBOR VOBOŘIL
Psychologický ústav AV ČR, Brno

ABSTRACT

Software for adaptive testing: CAT in action

P. Květon, M. Jelínek, D. Denglerová, D. Vobořil

Nowadays a new approach to testing psychological and other characteristics is widely used. This approach is known as Computerized Adaptive Testing (CAT) and enables testing to become more effective and accurate. The basic idea of CAT is to administrate only those items that are most appropriate for a tested individual and providing maximum amount of information, speaking in terms of Item Response Theory (IRT) as the most common mathematical apparatus of CAT. This article presents original software developed on the ground of the Institute of Psychology, Academy of Sciences of the Czech Republic. The program is able to interactively administrate and choose appropriate items, estimate proficiency, and evaluate predefined condition for the end of the test. The current version is fully capable of processing dichotomous items test-banks. We call this little piece of software CATO™ - Computerized Adaptive Testing *Optimized.*

key words:
CAT,
IRT,
computerized adaptive testing,
item response theory,
software

klíčová slova:
CAT,
IRT,
počítačové adaptivní testování,
teorie odpovědi na položku,
software

Cílem adaptivního testování je zefektivnění testovací procedury při současném zachování přesnosti měření. Tohoto cíle je dosahováno záměrným výběrem takových položek, které pro danou osobu přinášejí maximální informaci o sledované charakteristice. Základní principy adaptivního testování byly již na stránkách Československé psychologie popsány (Květon, Jelínek, Denglerová, 2006). Pro účely předkládaného příspěvku tedy pouze shrneme tři základní kroky, které je potřeba v procesu adaptivního testování zajistit. Jedná se o zahájení testu, interaktivní výběr adekvátních položek a rozhodnutí o ukončení testu.

V odborné literatuře (Wise, Kingsburry, 2000; Linacre, 2000; Wainer et al., 2000) existuje shoda, že ideálním prostředkem pro uskutečnění těchto kroků je rodina matematických modelů souhrnně označovaná jako teorie odpovědi na položku (IRT – Item Response Theory). Jedná se však o poměrně sofistikovaný přístup

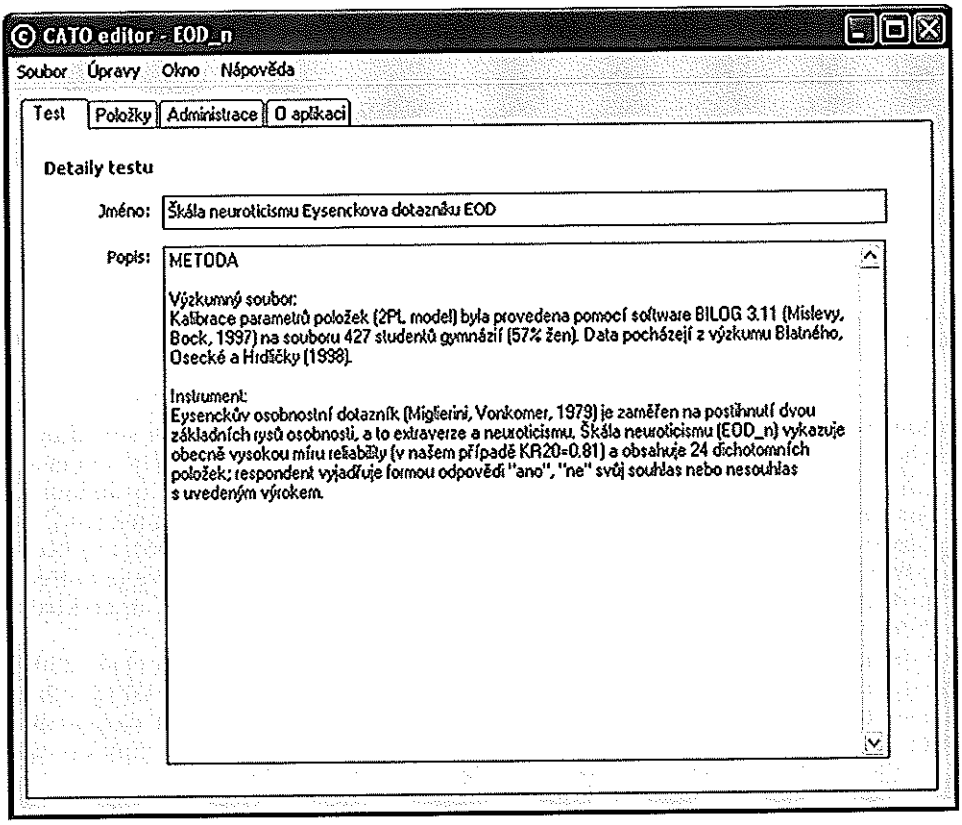
Došlo: 22. 2. 2007; P. K., M. J., D. D., D. V.; Psychologický ústav AV ČR, Veveří 97, 602 00 Brno; e-mail: kveton@psu.cas.cz

Studie byla vypracována za podpory grantu GAČR (č. 406/06/P396) a je součástí výzkumného záměru PsÚ AV ČR (reg. č. AV0Z70250504).

k měření, který vyžaduje jistou výpočetní kapacitu. Vzhledem k tomu, že proces adaptivního testování musí být interaktivní, je k administraci testů nutně využíváno počítačových technologií. Proto se také v této souvislosti hovoří o počítačovém adaptivním testování (v zavedené anglické zkratce CAT – Computerized Adaptive Testing).

Na mezinárodní scéně již existují nejrůznější implementace CAT, které fungují na základě IRT (např. FastTEST professional, ASC, 2000). Vzhledem k tomu, že tyto komerční produkty bývají poměrně drahé a ne vždy zcela vyhovují konkrétním potřebám, rozhodli jsme se vyvinout vlastní řešení, které by se svou funkčností jiným produktům vyrovnalo a zároveň lépe splňovalo naše požadavky. Nezanedbatelnou výhodou vlastního řešení je i dokonalejší porozumění problematice, získání užitečné praktické zkušenosti a transparentnost a kontrola výpočetních algoritmů.

Vzhledem k tomu, že v současné době jsme již dokončili verzi 0.9 software zvaného CATO™ (Computerized Adaptive Testing *optimized*), považujeme za přínosné popsat proces adaptivního testování na konkrétních příkladech realizovaných prostřednictvím tohoto programu[1]. CATO™ obsahuje rozhraní pro nastavení testovací procedury (CATO editor) a modul pro administraci testu klientovi (CATO-GO).



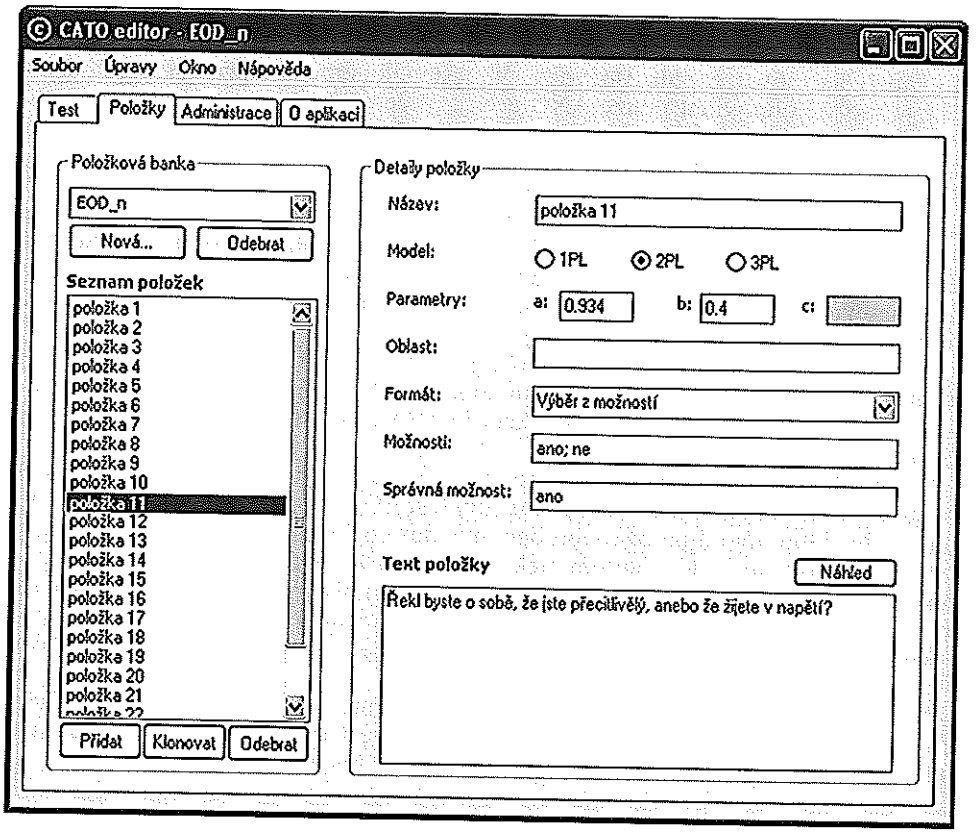
Obr. 1 Popis použitého testu

[1] Technické informace o programu viz příloha.

Příprava testu

Úvodním krokem při tvorbě adaptivního testu je výběr vhodných položek měřících zvolenou charakteristiku. V našem případě jsme zamýšleli zjišťovat úroveň neuroticismu a k tomuto účelu jsme zvolili soubor položek škály neuroticismu Eysenckova osobnostního dotazníku EOD.

Na otisku obrazovky je zobrazen úvodní dialog software CATO. Zde se nastavují základní charakteristiky vytvářeného testu, a to *jméno testu* a podrobnější *popis* testu. V popisu testu je vhodné poznamenat kromě věcného popisu instrumentu také např. detaily o procesu kalibrace (viz níže). Po nadefinování základních vlastností je třeba přikročit k vytvoření položkové banky.



Obr. 2 Vytváření položkové banky

Základním stavebním kamenem každého testu jsou položky, a proto nyní přikročíme k definování položkové banky. V našem případě, kdy pracujeme se škálou neuroticismu, postupně definujeme všech 24 položek této škály. Vytvoříme si tedy novou položkovou banku s názvem „EOD_n" a naplníme ji těmito položkami. Na obr. 2 vidíme, že u každé položky je nejprve třeba zadat jednoznačný identifikátor položky (*název*). Víme[2], že položky byly kalibrovány pomocí logistického IRT modelu 2PL, proto tento model zvolíme a zadáme parametr rozlišovací účinnosti[3] a parametr obtížnosti[4] *b*. Parametr *c* je v tomto případě automaticky pokládán za

nulový a my jej nezadáváme. Podrobnější popis jednotlivých modelů 1PL, 2PL a 3PL lze nalézt např. v Urbánek, Šimeček (2001).

V praxi se často setkáváme se situací, kdy je test považován za jednodimenzionální, třebaže v něm lze položky dle specifického zaměření rozdělit do několika skupin, resp. oblastí. Pro tyto případy je možné pro každou položku definovat pole *oblast*, což lze pak zohlednit v administraci testu střídáním položek z různých oblastí zatrhnutím atributu *Stratifikovat dle oblasti* na záložce *Administrace* (viz dále). Implementováním této funkce do CATO™ reagujeme na návrh Leunga, Changa a Haua (2003) ohledně problému balancování obsahu (content-balancing).

Pomocí pole *Formát* lze kromě nejužívanějšího „Výběr z možností" navolit také „Volná výpověď", kdy je volně zapsaná výpověď probanda porovnána s předem definovanou klíčovou (správnou, resp. diagnostickou) odpovědí v poli *Správná odpověď*. V případě, že je formát zvolen jako „Výběr z možností", je třeba tyto možnosti definovat v poli *Možnosti*. V našem ilustračním případě s položkou p11 je možné odpovídat pouze ANO a NE, přičemž jako diagnostická odpověď je nastaveno „ANO". Do pole *Text položky* je třeba vepsat znění položky v tom tvaru, jak je posléze během administrace probandovi prezentována. Text je možné formátovat standardními příkazy jazyka HTML, což dává poměrně široké možnosti ve vzhledu položek. I pro IT laika je pak jednoduché si položky připravit v kterémkoli HTML editoru (např. CoffeeCup HTML Editor 2007 nebo známější CuteHTML) a připravený HTML kód jednoduše vložit. Tento přístup pochopitelně také umožňuje jednoduché použití obrázků, zvuku a případně i dalších forem tzv. inovativních položek (Harmes, 1999).

Technologie CAT vychází v podstatě z jediné základní ideje (viz úvod článku), avšak jednotlivé fáze průběhu testování mohou být různě modifikovány. Software CATO základní druhy těchto modifikací implementuje a jejich nastavení je přístupné na záložce *Administrace*.

V horní části okna na obr. 3 je seznam nadefinovaných položkových bank. Pokud bychom chtěli najednou administrovat celý test EOD, včetně dimenze extraverze, museli bychom novou položkovou banku nadefinovat na záložce *Položky*, aby se v tomto seznamu posléze automaticky objevila. V procesu administrace budou položky obou bank promíchány, ačkoli výpočetní algoritmus je bude zpracovávat nezávisle a CATO ve výstupu předá odhady každého rysu zvlášť (extraverze i neuroticismu). Pro každou položkovou banku je nutné individuálně definovat i další parametry administrace.

V procesu administrace adaptivního testu můžeme ovlivňovat všechny tři fáze, které byly popsány v úvodu. Úvodní odhad rysu, kterým administrace musí začínat, umožňuje CATO provést buď na sledu několika náhodně[5] vybraných položek (viz

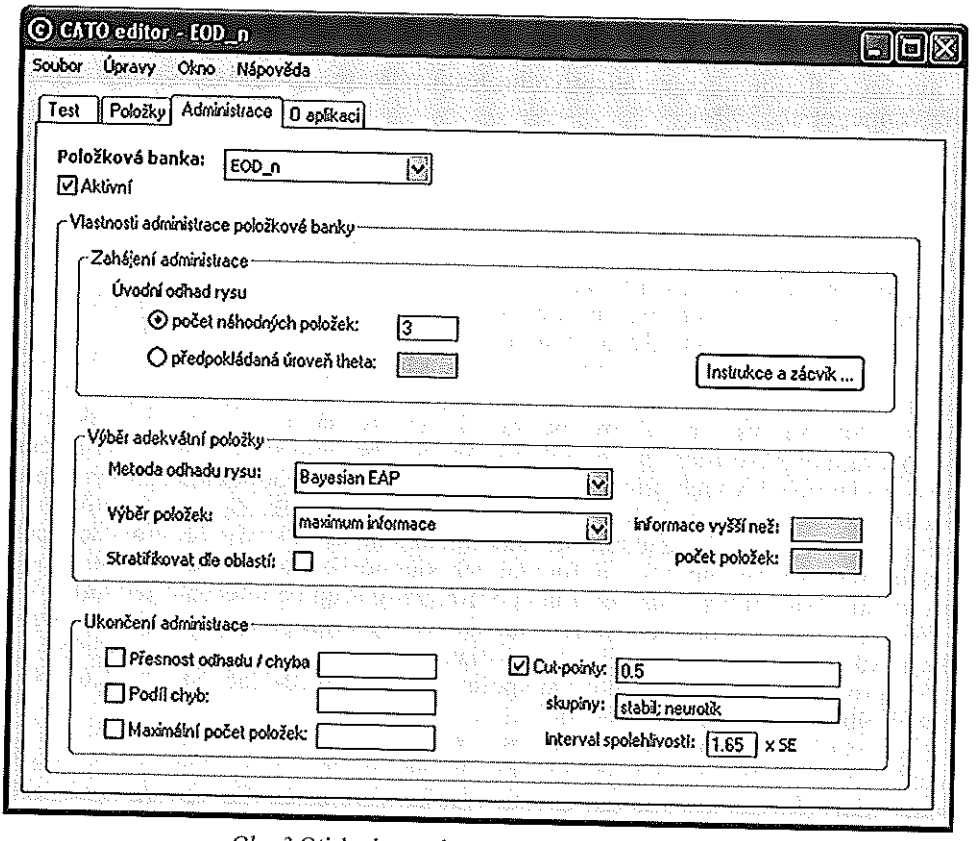
---

[2] Před tím, než definujeme položkovou banku v CAT software, je zapotřebí provést na dostatečném množství dat odhady parametrů položek dle zvoleného IRT modelu. My jsme pro tyto výpočty použili specializovaný software Bilog 3.11.

[3] Rozlišovací účinnost určuje sklon tzv. charakteristické křivky, resp. funkce položky. Vyšší číslo naznačuje strmější křivku, položka tedy lépe rozlišuje mezi osobami s různou úrovní měřené vlastnosti, a to zejména u osob, jejichž úroveň se přibližně pohybuje okolo obtížnosti dané položky.

[4] Obtížnost položky je definována jako taková úroveň latentního rysu, při které má proband 50% pravděpodobnost odpovědět správně, resp. v diagnostickém směru. Termín obtížnost zachováváme záměrně, i když se používá primárně u výkonově zaměřených položek.

[5] Náhodný výběr je zde jedinou možností, neboť ještě není stanoveno kritérium (odhad rysu) pro výběr adekvátních položek.



Obr. 3 Otisk obrazovky – nastavení administrace

pole *počet náhodných položek*) nebo lze zadat přesnou hodnotu rysu (*předpokládaná úroveň theta*) dle vlastního uvážení. Úvaha pro tuto hodnotu může vycházet buď z předběžných informací o probandovi, případně můžeme zvolit např. průměr dané populace. Zde je ovšem nutné mít na zřeteli, že stejný startovní bod pro všechny respondenty s sebou nese riziko tzv. řetězení položek v případě, že je test nadefinován tak, aby volil vždy nejadekvátnější položky.

Volba vhodných položek je vlastně druhou fází adaptivního testování. V rámci CATO je možné uplatnit tři různé přístupy. Buď jsou voleny vždy nejadekvátnější položky (tedy s maximálním informačním přínosem pro danou úroveň rysu), nebo náhodně volené položky s informačním přínosem vyšším než stanovená hodnota, a popřípadě i náhodně zvolené položky z množiny nejadekvátnějších.

Pro výpočet informačního přínosu[6] konkrétní položky je nutné znát úroveň latentního rysu, která se v procesu administrace stále upřesňuje. Software CATO umožňuje výběr mezi dvěma metodami odhadu, a to metodou maximální věrohodnosti (Maximum Likelihood) a Bayesovským odhadem (Expected a Posteriori). V našem ilustračním příkladu se škálou neuroticismu jsme zvolili metodu Bayesov-

[6] Algoritmy pro výpočet informačního přínosu položky v závislosti na úrovni latentního rysu lze najít ve většině pramenů zabývajících se IRT (např. Hambleton, Swaminathan, Rogers, 1991).

ského odhadu, neboť na rozdíl od Maximum Likelihood metody umožňuje odhad rysu i v případě, že jsou všechny položky zodpovězeny stejným směrem, což je v úvodní fázi administrace adaptivního testu poměrně obvyklý jev (podrobněji viz Wainer et al., 2000).

Pokud jsme již nadefinovali úvod testu a samotný průběh administrace položek, je třeba určit podmínky pro ukončení testu. Program CATO nabízí v současné verzi celkem pět variant ukončení, které je možné navzájem kombinovat. Pokud nastavíme více než jednu podmínku, test je ukončen v případě splnění kterékoli z nich. Podmínka *Přesnost chyby odhadu* umožňuje definovat úroveň chyby odhadu latentního rysu, při které považujeme odhad za natolik přesný, že další administrace by již nepřinášela diagnosticky relevantní údaje. *Podíl chyb* zohledňuje vzájemný vztah dvou po sobě následujících odhadů a s nimi spojených chyb. Jakmile podíl chyb dosáhne stanovené hodnoty, test je ukončen. Jakousi implicitní podmínkou ukončení testu je i vyčerpání všech položek z položkové banky. Pokud je však položková banka relativně rozsáhlá, je výhodné nadefinovat podmínku *Maximální počet položek*, která zabraňuje administraci přílišného množství položek v těch případech, kdy zkoumaná osoba odpovídá takovým způsobem, že nedojde ke splnění žádné z ostatních podmínek. Zajímavou možností pro potřeby screeningu je ukončení testu tím okamžikem, kdy je program schopen zařadit testovanou osobu do definovaných kategorií, a to s předem nastavenou přesností. Pomocí tzv. *Cut-point* jsou definovány hranice intervalů jednotlivých kategorií, které jsou pojmenovány v poli *Jména intervalů*. *Interval spolehlivosti* je definován násobkem standardní chyby. Platí tedy, že pokud zadáme např. hodnotu 1,65, program interpretuje interval spolehlivosti jako (θ-1,65*SE ; θ+1,65*SE) neboli 90% interval spolehlivosti.

Nastavením podmínek ukončení testu jsme se dostali do situace, kdy máme test kompletně připravený k administraci klientům. Pro účely administrace slouží samostatný programový modul CATO-GO, který je na CATO editoru nezávislý. Výhodou samostatného administračního modulu je efektivní testování bez supervize, neboť klient nemá žádnou možnost ovlivňovat (ať již omylem či záměrně) nastavení testu.

Administrace testu

V další části textu budou prezentovány dva průběhy adaptivní administrace testu pomocí programu CATO, a to jeden typický s ukončením na základě přesnosti chyby odhadu a druhý více specifický případ, kdy máme v úmyslu testovanou osobu zařadit do předem definované kategorie (Cut-point ukončení). Na závěr budou prezentovány základní ekvivalenční údaje pro adaptivní a klasický způsob administrace.

Typický průběh administrace ukončené minimální chybou odhadu

Následující tab. 1 uvádí typický průběh postupného upřesňování odhadu latentního rysu[7]. Po prvních třech administrovaných položkách, které byly vybrány náhodně, jsou vždy vybírány takové položky, které disponují maximálním informačním přínosem pro aktuálně odhadovanou úroveň theta (nastavení postupu administrace v SW CATO viz obr. 3). S každou odpovědí probanda získáváme další informaci,

[7] Škála latentního rysu je nastavena tak, aby přibližně odpovídala z-rozložení (průměr 0; standardní odchylka 1).

díky níž dochází k postupnému snižování chyby odhadu. Po administraci 14. položky v našem příkladu klesá standardní chyba odhadu pod předem definovanou hodnotu 0,5. Konečný odhad neuroticismu je považován za dostatečně přesný (dle nastavení programu) a administrace je ukončena.

Tab. 1 Detaily administrace s ukončením minimální chybou

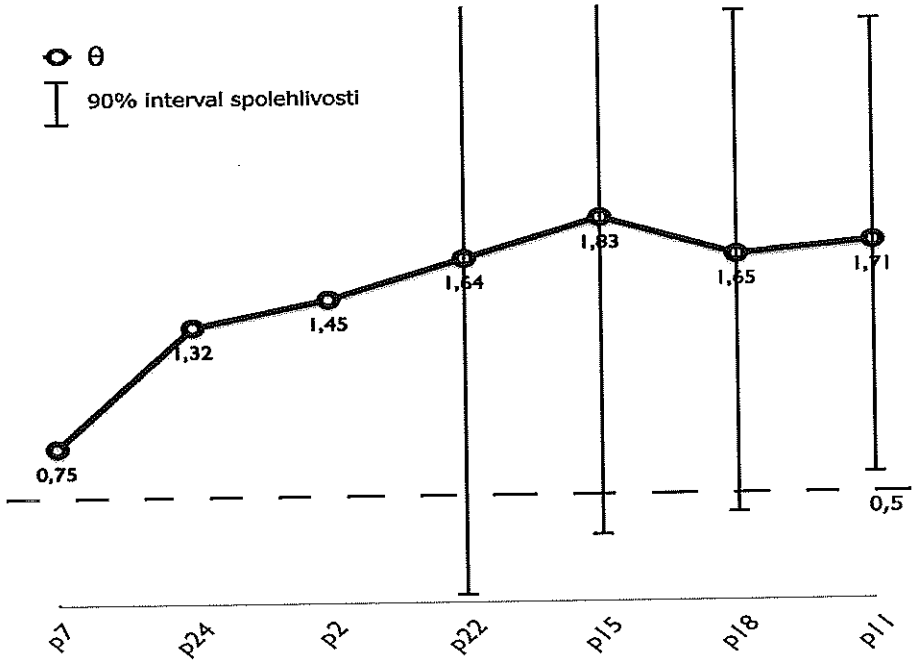
pořadí položky	položka	odpověď	rozlišovací účinnost (a)	obtížnost (b)	odhad rysu – neuroticismus (θ)	SE_θ
1	p20	1	0,79	0,27	0,55	1,51
2	p3	1	0,59	-0,59	0,70	1,31
3	p17	1	0,80	-0,01	0,90	1,08
4	p22	1	0,92	0,74	1,19	0,90
5	p11	0	0,93	0,40	0,76	0,69
6	p10	0	0,86	0,20	0,49	0,60
7	p15	1	0,72	1,20	0,70	0,58
8	p19	1	0,63	1,06	0,86	0,57
9	p7	1	0,62	1,60	1,02	0,56
10	p18	0	0,63	2,01	0,96	0,54
11	p24	0	0,61	2,08	0,91	0,52
12	p14	1	0,70	-0,49	0,95	0,51
13	p2	1	0,48	0,10	1,00	0,50
14	p16	0	0,42	0,73	**0,93**	**0,49**

Zařazení osoby do předem definované diagnostické kategorie

Kolem každého průběžného odhadu latentního rysu (neuroticismu) je konstruován interval spolehlivosti, který je neustále porovnáván s předem definovanou hranicí – „cut-point" (v tomto případě 90% interval spolehlivosti, tedy $1.65 \times SE_\theta$, a cut-point 0.5, tedy polovina směrodatné odchylky nad průměrem uvažované populace – viz obr. 3). V případě, že interval nepřekrývá žádný cut-point, dochází k ukončení administrace a výsledkem testu je určení kategorie, do které osoba spadá. V našem případě jsme definovali jeden cut-point, tzn. dvě kategorie charakteristické vyšší (neurotická osobnost) a nižší (emočně stabilní osobnost) úrovní neuroticismu. Na grafu 1 je patrné, že po administraci položky p11 je interval spolehlivosti celý nad úrovní cut-point a osoba je zařazena do skupiny „neurotiků". K dostatečně spolehlivému screeningovému odhadu tak došlo po sedmi položkách (včetně tří úvodních náhodně zvolených položek, viz nastavení administrace CATO) a nebylo tedy nutné pokračovat prezentací dalších položek.

Zhodnocení ekvivalence

Pro účely jednoduchého ověření ekvivalence adaptivně a klasicky administrované škály neuroticismu jsme provedli simulaci na reálných datech. Nejprve byl zjištěn hrubý skór neuroticismu sečtením diagnostických odpovědí všech 427 osob



Graf 1 Administrace s ukončením po zařazení do definované kategorie

(z klasické administrace). Následně byly odpovědi na položky využity pro simulaci adaptivně administrovaného testu. Program provedl pro každou jednotlivou osobu virtuální administrační sezení, přičemž místo dotazování reálných osob byla příslušná odpověď automaticky dohledána v databázi odpovědí zjištěných v rámci klasické (reálné) administrace.

Korelační analýzou byla zjištěna vysoká úroveň ekvivalence výsledků ($r^2 = 0{,}916$). Efektivita adaptivního testování je patrná z tab. 1, která uvádí počet administrovaných položek potřebný k dosažení předem definované úrovně přesnosti měření (SE = 0,5).

U většiny osob (86,7 %) bylo nutné administrovat méně než 24 položek tvořících kompletní škálu neuroticismu. U 57 osob nebyl odhad rysu dostatečně přesný, a proto byla administrace ukončena až vyčerpáním položkové banky. Tyto osoby jsou charakteristické extrémní úrovní rysu (na obou pólech). Položková banka obsahuje samozřejmě zejména položky adekvátní většině osob běžné populace, a proto v extrémních úrovních rysu tyto položky nepodávají dostatečnou informaci nutnou k dosažení požadované přesnosti. Pokud bychom tvořili původní adaptivní test neuroticismu, bylo by v tomto smyslu jedním z hlavních úkolů zajistit dostatečné množství položek tak, aby svou obtížností pokryly celou šíři teoreticky uvažovaného rozsahu latentního rysu.

ZÁVĚR

V příspěvku jsme se pokusili na původním software ilustrovat možnosti, které počítačové adaptivní testování přináší do psychodiagnostické praxe. Software CATO (Computerized Adaptive Testing *optimized*) byl vytvářen s důrazem na uživatelskou

Tab. 2 Počet položek nutných k požadované přesnosti měření v adaptivní administraci

Počet administrovaných položek	Počet osob (N = 427)	%	Kumulativní %
11	9	2,1	2,1
12	83	19,4	21,5
13	107	25,1	46,6
14	56	13,1	59,7
15	30	7,0	66,7
16	20	4,7	71,4
17	15	3,5	74,9
18	21	4,9	79,9
19	12	2,8	82,7
20	4	0,9	83,6
21	5	1,2	84,8
22	5	1,2	85,9
23	3	0,7	86,7
24	57	13,3	100,0

přívětivost a jednoduchost použití. Aktuální verze 0.9 je schopna administrace a vyhodnocení adaptivních testů tvořených dichotomně skórovanými položkami. V rámci budoucího rozšíření software bychom chtěli ještě pracovat na implementaci nových funkcí, avšak současně zachovat původní ideu, kterou je zpřístupnění CAT širší odborné veřejnosti. Nejbližším cílem je zahrnutí podpory IRT modelů, umožňujících práci s položkami polytomního formátu, které bývají zejména v psychologii osobnosti součástí používaných testů a škál.

LITERATURA

ASC (2000): The FastTEST Professional Testing System. St. Paul, MN: Assessment Systems Corporation.

Blatný, M., Osecká, L., Hrdlička, M. (1998): Zdroje sebehodnocení u temperamentových typů. Československá psychologie, 42, 297-305.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991): Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Harmes, J. C. (1999): Computer – based testing: Toward the design and use of innovative items, University of South Florida. Vyhledáno na http://www.coedu.usf.edu/itphdsem/eme7938/ch899.pdf

Jelínek, M., Květon, P., Denglerová, D. (2006): Adaptivní testování – základní pojmy a principy. Československá psychologie, 50, 2, 163-173.

Leung, C. K., Chang, H. H., Hau, K. T. (2003): Computerized adaptive testing: A comparison of three content balancing methods. Journal of Technology, Learning, and Assessment, 5, 2-15.

Linacre, J. M. (2000): Computer-adaptive testing: A methodology whose time has come. MESA Memorandum No. 69. Vyhledáno 14.1.2005 na http://www.rasch.org/memo69.pdf.

Miglierini, B., Vonkomer, J. (1979): Eysenckov osobnostný dotazník – EOD. Bratislava: Psychodiagnostické a didaktické testy.

Mislevy, R. J., Bock, R. D. (1997): BILOG 3.11: Item Analysis and Test Scoring with Binary Logistic Models. Scientific Software, Inc.

Urbánek, T. Šimeček, M. (2001): Teorie odpovědi na položku. Československá psychologie, 5, 428-440.

Wainer, H., Dorans, N. J., Eignor, D.,

Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., Thissen, D. (2000): Computerized adaptive testing: A primer (2nd edition). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Wise, S. L., Kingsbury, G. G. (2000): Practical issues in developing and maintaining a computerized adaptive testing program. Psicológica, 21, 135-155.

SOUHRN

Počítačové adaptivní testování představuje nový přístup k testování psychologických (i jiných) charakteristik, který umožňuje proces testování zefektivnit a zpřesnit. Základní ideou je administrace pouze takových položek, které jsou pro danou testovanou osobu adekvátní a poskytují tedy v terminologii Teorie odpovědi na položku (IRT – Item Response Theory), která je pro adaptivní testování základním matematickým aparátem, maximum informace. Cílem příspěvku je představení původního software vzniklého na půdě Psychologického ústavu Akademie věd ČR, který implementuje funkce pro interaktivní administraci a výběr adekvátních položek, odhad měřené charakteristiky, a vyhodnocení definované podmínky ukončení testu. V současné době je program schopen bezproblémově pracovat s testy tvořenými dichotomně skórovanými položkami. Software byl pojmenován Computerized Adaptive Testing *optimized*, ve zkratce CATO™.

PŘÍLOHA

O aplikaci CATO