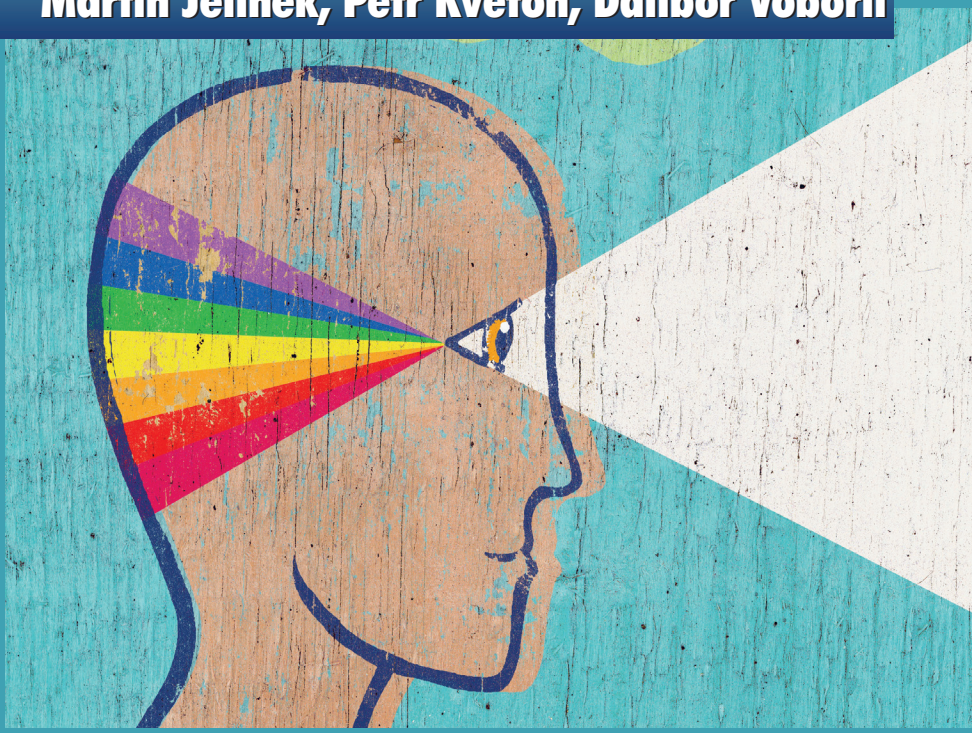
Martin Jelínek, Petr Květon, Dalibor Vobořil



# TESTOVÁNÍ V PSYCHOLOGII

## Teorie odpovědi na položku a počítačové adaptivní testování

PSYCHÉ

GRADA®

Martin Jelínek, Petr Květon, Dalibor Vobořil

# TESTOVÁNÍ V PSYCHOLOGII

## Teorie odpovědi na položku a počítačové adaptivní testování

Grada Publishing

Vznik publikace byl podpořen Grantovou agenturou České republiky v rámci grantového projektu GA ČR č. 406/09/P284 a výzkumným záměrem PSÚ AV ČR, v.v.i. reg. č. AV0Z70250504.

**PhDr. Martin Jelínek, Ph.D.**
**PhDr. Petr Květon, Ph.D.**
**PhDr. Dalibor Vobořil, Ph.D.**

# TESTOVÁNÍ V PSYCHOLOGII

## Teorie odpovědi na položku a počítačové adaptivní testování

Vydala Grada Publishing, a.s.
U Průhonu 22, 170 00 Praha 7
tel.: +420 234 264 401, fax: +420 234 264 400
www.grada.cz
jako svou 4580. publikaci

**Recenzoval:**
doc. PhDr. Tomáš Urbánek, Ph.D.

Odpovědná redaktorka Jana J. Kubínová
Sazba a zlom Antonín Plicka
Počet stran 160
Vydání 1., 2011

Vytiskla Tiskárna PROTISK, s. r. o., České Budějovice



**ISBN 978-80-247-3515-3 (tištěná verze)**
**ISBN 978-80-247-7198-4 (elektronická verze ve formátu PDF)**

# OBSAH

# PŘEDMLUVA

Pojmy teorie odpovědi na položku a počítačové adaptivní testování jsou v posledních letech velmi často zmiňovány v souvislosti s moderními trendy v psychologické (i jiné) diagnostice. Ambicí této knihy je poskytnout českému čtenáři v ucelené formě popis základních principů teorie odpovědi na položku (*Item Response Theory – IRT*) a počítačového adaptivního testování (*Computerized Adaptive Testing – CAT*), které je na ní založené.

Teorie odpovědi na položku představuje komplexní matematický aparát, který se snaží postihnout situaci, kdy jedinec odpovídá na testovou položku. Díky tomuto aparátu dokážeme předvídat, jak jedinec s určitou úrovní schopností odpoví na konkrétní položku. Základem IRT je matematický model, který udává pravděpodobnost určité odpovědi v závislosti na úrovni latentního rysu jedince a charakteristikách konkrétní položky. Pojmem latentního rysu zde rozumíme uvažované, přímo nepozorovatelné charakteristiky respondentů, k jejichž odhadu chceme v procesu testování dospět. Každá položka je tedy popsána pomocí tzv. charakteristické funkce položky *Item Characteristic Function* (*ICF*), která vyjadřuje vztah mezi latentním rysem a pravděpodobností dané odpovědi. Na základě definovaného vztahu mezi latentním rysem a pravděpodobností odpovědí jsou odvozeny koncepty podmíněné chyby měření a informačního přínosu, které se uplatňují v celé řadě psychometrických aplikací.

Teorie odpovědi na položku je alternativou ke klasické teorii testů (*Clasical Test Theory – CTT*). Modelově zakotvená teorie odpovědi na položku přináší do procesu měření řadu výhod. IRT považuje test za množinu samostatných položek a základní jednotkou, pro kterou jsou různé modely vytvořeny, je jednotlivá položka. Naproti tomu v CTT není položka konceptuálně oddělitelná od celku konkrétního testu. Zejména díky osamostatnění testové položky IRT nabízí vhodný aparát pro uplatnění obecných principů tzv. adaptivního testování. Základní idea adaptivního testování je přitom jednoduchá: zadávejte testované osobě k řešení pouze takové položky, které umožňují ideálně posoudit úroveň měřeného rysu. V praxi její uplatnění vyžaduje buď zkušeného administrátora, který je schopen odhadnout úroveň rysu testované osoby, vybrat přiměřeně

obtížnou položku a po jejím zodpovězení opět celý proces opakovat, anebo využití matematického aparátu, který by v reálném čase provedl totéž – obvykle IRT v kombinaci s počítačovou technikou. Adaptivní testování je tak v dnešní době převážně chápáno jako počítačové adaptivní testování založené na principech IRT.

Čtenář bude postupně seznámen se základními IRT modely pro položky dichotomního a polytomního formátu, předpoklady těchto modelů, dále se způsoby, jakými lze dospět k odhadům parametrů položek a osob na škále latentního rysu, s možnostmi jejich transformace a s posuzováním vhodnosti IRT modelů vzhledem k datům. Pozornost je dále věnována praktickým aplikacím IRT, jako je rozpoznání odlišného fungování položek u různých skupin respondentů a využití konceptu informačního přínosu pro konstrukci testových metod. Teoretický popis aspektů IRT je průběžně doplňován o praktické příklady na reálných datech, a to včetně konkrétních tipů pro práci s renomovanými IRT software Bilog-MG, Parscale a Multilog. Tuto část knihy navíc doprovází on-line materiál, který obsahuje výpočetní šablony implementované v tabulkovém procesoru, které pomohou lepšímu pochopení výpočetních problémů a představují pro čtenáře užitečnou pomůcku k provádění vlastních analýz. V textu jsou tyto šablony označeny pořadovým číslem a QR kódem, v příloze uvádíme příslušný seznam internetových adres.

Druhá část knihy je zaměřena na důkladné seznámení s principy adaptivního testování v historické perspektivě a na současné pojetí adaptivního testování, kdy je synonymem adaptivního testování počítačové adaptivní testování založené na IRT. Kapitoly o počítačovém adaptivním testování jsou založeny na zkušenostech autorů s vytvářením vlastního CAT software. Tento software je také využit pro demonstraci výhod i problémů, které adaptivní testování přináší do diagnostické praxe.

# I.
# TEORIE
# ODPOVĚDI NA POLOŽKU

# 1. HISTORICKÉ SOUVISLOSTI TEORIE ODPOVĚDI NA POLOŽKU

Podobně jako pro celý obor psychologie také v oblasti diagnostiky individuálních charakteristik platí výrok o dlouhé historii a krátké minulosti. Například Allenová a Yenová (2002) zmiňují první úřednické zkoušky způsobem formalizovaného testování již před třemi tisíci lety ve starověké Číně. Formální testovací postupy byly využívány u kandidátů na různé úřady, přičemž byla uplatňována pravidla, která ve své podstatě přetrvala až do dnešní doby, například anonymita testování nebo hodnocení dvěma nezávislými examinátory zajišťující objektivitu testování.

Další impulsy do vývoje testování přicházely zejména od vzdělávacích a vojenských institucí (potřeba vstupního, výstupního i průběžného testování znalostí a schopností, hodnocení rekrutů apod.). Období první světové války, konkrétně rok 1917, představovalo důležitý mezník ve vývoji psychologického testování. Skupina psychologů v čele s tehdejším prezidentem Americké psychologické asociace Robertem M. Yerkesem, která se inspirovala u velkých autorit – jako Galton, Binet, Pearson a dalších (blíže viz např. Hunt, 2000) – se rozhodla pro vytvoření skupinově administrovaného testu inteligence sloužícího k objektivnímu testování rekrutů v americké armádě (DuBois, 1970). Jejich snaha nakonec vyústila v tzv. Army Alpha Test, který se stal vzorem pro pozdější výkonové testy charakteristické požadavkem na objektivitu skórování, zajištěnou většinou použitím uzavřených položek s volbou z více možností. Takto byly vlastně položeny pevné základy pro klasickou teorii testů, která přetrvává a rozvíjí se až do dnešní doby.

První předzvěsti principů IRT lze zpětně vysledovat již u Luise Thurstonea v jeho analýze *Binetova a Simonova testu dětského mentálního vývoje* (Bock, 1997). V této studii Thurstone pro každou položku spočítal procento úspěšných dětí a graficky znázornil vztah věku a úspěšnosti, přičemž výsledný graf svým esovitým tvarem nápadně připomíná tzv. charakteristické křivky položek, jak jsou definovány v rámci IRT. Další zajímavé historické propojení teorie odpovědi na položku nalezl Bock v oboru toxikologie, ve kterém na začátku 19. století existovaly snahy modelovat reakce mikroorganismů na zvyšující se

dávky toxinu. Výsledný model měl v zásadě podobu kumulativního normálního rozložení vyjadřujícího úmrtnost organismů. Hlavním cílem bylo nalézt takové množství toxinu, při němž umírá 50 % organismů. S notnou dávkou nadsázky lze pravděpodobnost úmrtí organismů chápat jako paralelu k pravděpodobnosti správné odpovědi na položku a tzv. mediánovou efektivní dávku jako obdobu její obtížnosti. Propojení toxikologických výzkumů a pozdějšího IRT je ještě zřejmější na řešení problému s přirozenou úmrtností organismů, která tehdejším výzkumníkům systematicky nadsazovala účinnost testovaného toxinu. Výše popsaný model úmrtnosti tedy obohatili o parametr přirozené úmrtnosti, vysledovaný na kontrolní skupině. Výsledný matematický vzorec je v současné době prakticky ve stejné podobě využíván pro vyjádření pravděpodobnosti správné odpovědi v rámci tříparametrového logistického modelu s tím rozdílem, že přirozená úmrtnost je zde nahrazena uhádnutelností správné odpovědi (Bock, 1997).

Přibližně od šedesátých let minulého století se paralelně s klasickou testovou teorií začala rozvíjet teorie odpovědi na položku jako formálně odlišený přístup k testování individuálních charakteristik. V rámci vývoje IRT lze vysledovat dvě relativně samostatné linie uvažování (Embretson, Reise, 2000), které lze zjednodušeně označit jako americkou a evropskou. V USA je za formální počátek IRT přístupu považováno vydání knihy Lorda a Novicka (1968) s názvem *Statistické teorie mentálních testových skórů* (*Statistical Theories of Mental Test Scores*). Ačkoli kniha byla primárně zaměřena na klasickou teorii testů, obsahovala také několik kapitol Allana Birnbauma, které shrnují předchozí technické zprávy pro americké vzdušné síly a představují základy moderní IRT (Wainer, 2000). Další vývoj teorie odpovědi na položku je v USA spojen zejména se jmény Fumiko Samejima, David Thissen, Darrell Bock nebo Robert J. Mislevy. Poslední dva jmenovaní autoři stojí také v pozadí vzniku pravděpodobně nejužívanějšího softwaru pro praktickou aplikaci IRT s názvem Bilog, která je momentálně k dispozici ve verzi Bilog-MG 3 (Zimowski, Muraki, Mislevy, Bock, 2003). Evropská linie IRT je neodmyslitelně spjata se jménem dánského matematika Georga Rasche a jeho knihou *Pravděpodobnostní modely pro některé inteligenční a výkonové testy*. Další vývoj evropského uvažování v oblasti IRT posouvali zejména Erling B. Andersen, Wim J. van der Linden, Cees A. W. Glas a mnozí další.

Přes veškeré výhody přístupu IRT oproti CTT je poměrně zarážející relativně malé povědomí o IRT mezi širší odbornou psychologickou veřejností. Důvodů pro tento fakt lze najít hned několik. Jak poznamenává Embretson a Reise

(2000), typická kariéra Ph.D. studenta se specializací v oblasti IRT vrcholí neodmítnutelnou nabídkou zaměstnání ve společnostech zabývajících se vývojem testů, respektive v armádní testovací laboratoři, kde se posléze zabývá implementací IRT do vlastních testových baterií. Ačkoli taková kariéra je výhodná jak pro studenta, tak pro jeho zaměstnavatele, psychologická odborná veřejnost zůstává v tak progresivní oblasti, jakou IRT bezesporu je, v podstatě mimo hlavní dění. Dalším důvodem malého povědomí o IRT je určitě i slabá uživatelská přívětivost dostupných IRT software. Třebaže jsou matematicky na špičkové úrovni, jejich dokumentaci a ovládací rozhraní zvládne pouze uživatel s neobvykle silnou motivací a nadprůměrnou orientací v problematice. V neposlední řadě hraje velkou roli obecně vysoká složitost matematického pozadí celé teorie.

# 2. ZÁKLADNÍ IRT MODELY

Modely, se kterými se v rámci teorie odpovědi na položku pracuje, lze na obecné úrovni rozdělit podle počtu uvažovaných dimenzí měřené charakteristiky na jedno- a vícedimenzionální. Příkladem jednodimenzionálního uvažování je soubor položek měřících obecný faktor inteligence, u kterého se předpokládá, že odpověď na každou položku je téměř výhradně ovlivněna tímto obecným faktorem. Vícedimenzionální uvažování je v psychologické diagnostice také poměrně časté (např. Eysenckovy osobnostní dotazníky, dotazníky založené na teorii Big Five apod.). V praxi však jednotlivé dimenze bývají měřeny pomocí navzájem se nepřekrývajících souborů položek, mluvíme zde o tzv. mezipoložkové multidimenzionalitě (de Ayala, 2009). V rámci IRT je pak v rámci jednoho testu jednoduše aplikován jednodimenzionální model na každý samostatný soubor položek. Méně častá je tzv. vnitropoložková multidimenzionalita, kdy se při odpovědi na konkrétní položku výrazně uplatňuje více než jeden latentní rys. Příkladem může být metoda složená z několika položek zaměřených na koncept rodičovského zájmu. Lze předpokládat, že v pozadí odpovědí rodičů můžeme nalézt hned několik rysů více či méně ovlivňujících tyto odpovědi (např. zodpovědnost, láska, dominance a další). Zde je již nutné uplatnit speciálně navržené multidimenzionální IRT modely[1]. V této knize se budeme zabývat výhradně jednodimenzionálními modely, neboť většina psychologických testů je konstruována pro měření jedné dimenze či souboru několika navzájem konceptuálně nezávislých dimenzí.

Podle jiného kritéria lze IRT modely dělit na dichotomní a polytomní. Dichotomní modely jsou určeny pro binárně skórované položky používané zejména ve výkonových testech. Ačkoli u nich může být počet kategorií odpovědí různý, ve výsledku se pracuje pouze se dvěma hodnotami (0 a 1). Vzhledem k tomu, že se IRT dobře osvědčila v oblasti výkonových testů, byly posléze intenzivně vyvíjeny i polytomní modely určené pro položky s více než dvěma skórovacími

[1] Popis IRT modelů pracujících s multidimenzionálními konstrukty viz např. van der Linden, Hambleton (1997).

kategoriemi (např. posuzovací škály Likertova typu typicky užívané v diagnostice osobnosti, postojů apod.).

Při obecném představení jednotlivých modelů a výkladu jejich parametrů vycházíme z několika základních pramenů. Pro dichotomní IRT modely jsou to Wainer, Mislevy (2000), Hambleton, Swaminathan, Rogers (1991), Embretson, Reise (2000), Barton, Lord (1981) a de Ayala (2009). U polytomních modelů jsou to pak Nering, Ostini (2010), van der Linden, Hambleton (1997) a Embretson, Reise (2000). V textu jsou teoretické popisy problematiky doplněny o ukázky na reálných datech. Popis použitých datových souborů uvádíme v příloze 1.

## DICHOTOMNÍ IRT MODELY

Základní typy IRT modelů pro dichotomní položky jsou tzv. jednoparametrový logistický (1PL) model (zvaný též Raschův model), ve kterém jsou položky charakterizovány pouze jedním parametrem – obtížností –, dvouparametrový logistický (2PL) model s parametry obtížnosti a rozlišovací účinnosti a tříparametrový logistický (3PL) model s parametry obtížnosti, rozlišovací účinnosti a pseudouhádnutelnosti. Pro úplnost zmiňujeme také čtyřparametrový logistický (4PL) model, který 3PL model doplňuje o parametr ledabylosti.

U jednotlivých modelů vždy nejprve popíšeme jejich logiku (charakteristickou funkci položky), odvození informační funkce a způsob odhadu latentního rysu. V příloze knihy uvádíme způsob nastavení software Bilog-MG 3 s popisem základních proměnných ovlivňujících jednotlivé výpočty.

### JEDNOPARAMETROVÝ LOGISTICKÝ MODEL (1PL)

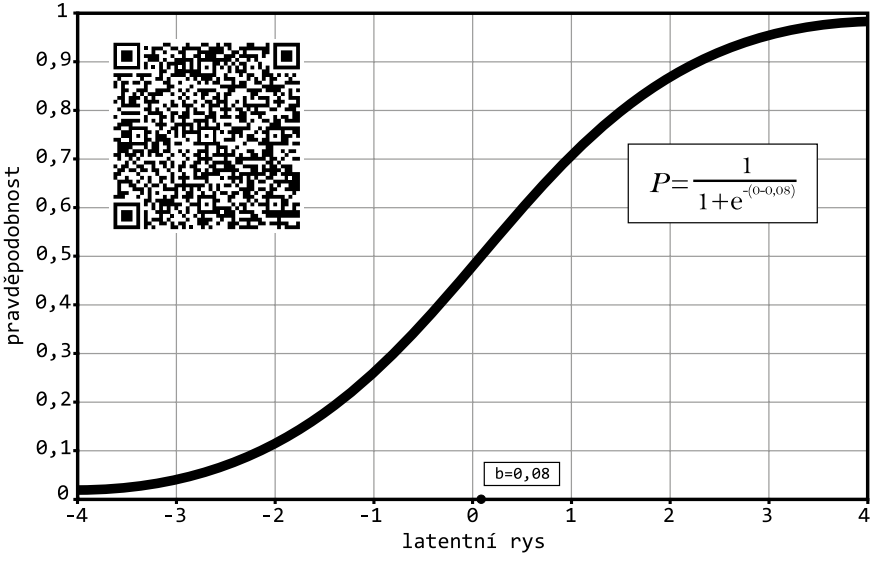
Model 1PL představuje nejjednodušší variantu ze skupiny IRT modelů. Vztah mezi úrovní latentního rysu probanda a pravděpodobností klíčové[2] odpovědi na určitou položku je dán pouze obtížností položky (*difficulty*). Charakteristickou funkci položky lze formálně vyjádřit jako

[2] V dalším textu budeme na místech, kde máme na mysli správnou odpověď nebo odpověď v diagnostickém směru, používat termín klíčová odpověď jako termín nadřazený.

$$P\left(U_{si}=1|\theta_s,b_i\right)=\frac{1}{1+e^{-(\theta_s-b_i)}}$$

Pravděpodobnost klíčové odpovědi ($U_{si}$ = 1) je tedy predikována z interakce mezi individuální úrovní latentního rysu $\theta$ probanda *s* a obtížností *b* položky *i*. Tyto dva parametry jsou ve výrazu použity v exponentu Eulerova čísla *e*, které tvoří základ přirozených logaritmů a má hodnotu přibližně 2,718.

Charakteristická křivka položky definovaná ve výše uvedeném vzorci má monotónně rostoucí průběh, a platí tedy intuitivní předpoklad, že s vyšší úrovní latentního rysu probanda vzrůstá pravděpodobnost klíčové odpovědi na položku. V grafu 1 je znázorněna ICF ukázkové položky *p*2 ze škály neuroticismu testu EOD (znění položky *Dělá Vám značné těžkosti říci někomu ne?*).




***Graf 1*** *ICF položky p2 dle 1PL modelu*

Obtížnost položky je definována jako úroveň latentního rysu, při které má proband 50% pravděpodobnost odpovědět diagnosticky. Parametr obtížnosti *b* je tedy vyjádřen na stejné škále jako latentní rys (škála odpovídá normálnímu z-rozdělení s průměrem 0 a standardní odchylkou 1). Je výhodou IRT, že na základě charakteristické funkce jsme schopni při znalosti úrovně rysu probanda odhadnout jeho pravděpodobnou odpověď na položku. Pokud bychom vedle sebe zobrazili ICF položek stejného testu odhadnuté na základě 1PL modelu, jednotlivé křivky by se od sebe lišily pouhým posunutím doleva nebo doprava dle různých obtížností, neboť parametr rozlišovací účinnosti udávající sklon křivky je fixován na stejnou hodnotu. Tento parametr bude podrobněji představen v následující kapitole o 2PL modelu.

V tabulce 1 jsou uvedeny odhady parametrů všech položek škály neuroticismu. Součástí tabulky je pro srovnání také obtížnost položek odhadnutá na základě klasické testové teorie, tedy jako procento správných odpovědí.

***Tab. 1*** *Obtížnosti položek škály neuroticismu dle CTT a IRT*

| Položka | *p* | *b* | $se_b$ |
|---|---|---|---|
| 1 | 0,71 | -1,05 | 0,12 |
| 2 | 0,48 | 0,08 | 0,11 |
| 3 | 0,62 | -0,59 | 0,11 |
| 4 | 0,70 | -1,00 | 0,11 |
| 5 | 0,75 | -1,30 | 0,12 |
| 6 | 0,75 | -1,32 | 0,13 |
| 7 | 0,20 | 1,64 | 0,13 |
| 8 | 0,77 | -1,45 | 0,13 |
| 9 | 0,80 | -1,66 | 0,13 |
| 10 | 0,45 | 0,24 | 0,11 |
| 11 | 0,39 | 0,52 | 0,12 |
| 12 | 0,76 | -1,36 | 0,12 |
| 13 | 0,65 | -0,72 | 0,11 |
| 14 | 0,61 | -0,54 | 0,11 |
| 15 | 0,24 | 1,36 | 0,13 |
| 16 | 0,38 | 0,57 | 0,11 |
| 17 | 0,50 | -0,02 | 0,11 |
| 18 | 0,14 | 2,10 | 0,15 |
| 19 | 0,28 | 1,11 | 0,12 |

| Položka | *p* | *b* | $se_b$ |
|---|---|---|---|
| 20 | 0,43 | 0,32 | 0,11 |
| 21 | 0,44 | 0,29 | 0,10 |
| 22 | 0,31 | 0,96 | 0,12 |
| 23 | 0,48 | 0,09 | 0,10 |
| 24 | 0,14 | 2,12 | 0,15 |

*p* – obtížnost položky dle CTT; *b* – obtížnost položky dle IRT; $se_b$ – standardní chyba odhadu obtížnosti dle IRT

Obtížnosti položek odhadnuté na základě 1PL modelu se pohybují mezi -1,660 a 2,118. Čím je hodnota parametru vyšší, tím je položka obtížnější. V CTT se místo obtížnosti jedná spíše o tzv. jednoduchost, a platí tedy, že vyšší hodnota indikuje snadnější položku. Kromě samotných odhadů parametrů obtížnosti dle IRT je vypočtena také jejich standardní chyba, což CTT neumožňuje. Je možné si povšimnout, že se standardní chyba odhadu zvyšuje směrem k extrémním položkám. Toto zjištění lze jednoduše vysvětlit. Parametr položky je nejpřesněji odhadnut v případě, kdy máme k dispozici velké množství odpovědí od osob, pro které je daná položka adekvátně obtížná. Rozložení latentního rysu neuroticismu v našem vzorku bylo přibližně normální, a proto byly parametry extrémně obtížných položek odhadnuty s menší jistotou než položky průměrně obtížné.

Při popisu 1PL modelu jsme zmínili, že výhodou IRT je možnost odhadnutí pravděpodobnosti klíčové odpovědi na základě úrovně latentního rysu. Pokud uvažujeme dále, lze na základě tohoto údaje určit míru informačního potenciálu dané položky pro konkrétního jedince, která je formálně vyjádřena tzv. informační funkcí položky (*Item Information Function – IIF*). V případě 1PL modelu ji lze algebraicky vyjádřit jako

$$I_i(\theta) = P_i(\theta)Q_i(\theta)$$

kde $P_i(\theta)$ je pravděpodobnost klíčové odpovědi na položku *i* podmíněná úrovní latentního rysu a $Q_i(\theta) = 1 - P_i(\theta)$ je pravděpodobnost odpovědi opačné. Z rovnice vyplývá, že maximální hodnota informačního přínosu je rovna 0,25. Maximum se nachází v bodě obtížnosti dané položky, neboť při této úrovni latentního rysu je pravděpodobnost diagnostické odpovědi rovna pravděpodobnosti odpovědi opačné, tedy 0,5. Logika výpočtu odpovídá intuitivnímu předpokladu, že nemá smysl probandovi předkládat položky, u kterých prak-