

Metodické studie

TEORIE ODPOVĚDI NA POLOŽKU

TOMÁŠ URBÁNEK¹⁾

Psychologický ústav AV ČR, Brno

MICHAL ŠIMEČEK²⁾

Centrum výzkumu vývoje osobnosti a etnicity, Brno

ABSTRACT

Item response theory

T. Urbánek, M. Šimeček

The paper compares basic principles of the classical test theory (CTT) and item response theory (IRT). Main emphasis is laid on the presentation of the IRT models and their advantages for test construction. Comparison of CTT and IRT is focused on the issues of the item-test relationship, item characteristics, reliability and accuracy of measurement, and possibilities of test results interpretation.

key words:

item response theory,
classical test theory,
item characteristic curve,
reliability

klíčová slova:

teorie odpovědi na položku,
klasická teorie testů,
charakteristická křivka položky,
reliabilita

ÚVOD

Teorie odpovědi na položku (*IRT – item response theory*) není na scéně psychometrických přístupů žádným nováčkem, protože její kořeny sahají hluboko do minulého století. Souvislosti a podobnosti s ní lze také najít mj. v psychofyzice, v Thurstonových a Guttmanových postupech škálování a v raných úvahách o nelineární faktorové analýze. Samotná teorie odpovědi na položku je však datována od vydání knihy dánského matematika Rasche (1960) s názvem *Probabilistic models for some intelligence and attainment tests*, jejíž teze (od té doby mnohokrát zopakována a rozvinutá v dalších pracích) podstatně mění náhled na povahu měření u stále rostoucího počtu odborníků zabývajících se tvorbou psychologických nebo didaktických testů.

Prestože hlavní aplikační oblastí IRT jsou zatím didaktické testy všeho druhu a výkonové testy, u kterých lze jednoduše posoudit správnost řešení IRT proniká postupně i do oblasti měření osobnostních rysů. Ve světě již existuje množství metod založe-

Došlo: 26. 6. 2001; T. U., Psychologický ústav AV ČR, Veveří 97, 602 00 Brno; M. Š., Centrum výzkumu vývoje osobnosti a etnicity, Brno

¹⁾ Článek je za Psychologický ústav AV ČR podporován výzkumným záměrem s registračním číslem AV0Z7025918.

²⁾ Článek je za Centrum výzkumu vývoje osobnosti a etnicity podporován projektem číslo LN00A023 (MŠMT ČR) s názvem „Centrum výzkumu vývoje osobnosti a etnicity.“

ných na IRT, např. různé testy jazykových schopností, testy inteligence, ale jsou známy i adaptace osobnostních dotazníků. V České republice je však tento přístup dosud velmi málo známý. Pokud je autorům známo, neexistuje žádná česká adaptace metody založené na IRT.

Témata rozvíjená v knihách a článcích o IRT jsou početná a zcela jistě by zasloužila speciální knižní publikaci v češtině. Je pravděpodobné, že i na to časem dojde. Zatím jsme se v tomto článku pokusili představit aspoň základní rysy tohoto přístupu.

Rozhodli jsme se zcela záměrně představit IRT prostřednictvím jejího srovnání s klasickou teorií testů (*CTT – classical test theory*), protože na základě tohoto přístupu byla vytvořena velká většina testových a dotazníkových metod u nás používaných a její terminologie a systém jsou (nebo by měly být) obecně známy. Navíc, většina témat přítomných v CTT koresponduje s tématy řešenými v IRT, kde jsou ale řešena na konceptuálně i metodologicky vyšší úrovni.

Obě teorie, CTT i IRT, nelze podle našeho názoru plně pochopit bez použití matematického aparátu. Protože však víme o nechuti části psychologické veřejnosti k matematice, pokusili jsme se v článku omezit užití matematických výrazů na nutné minimum.

CELKOVÉ SROVNÁNÍ CTT A IRT

Hlavní rozdíl mezi CTT (klasickou teorií testů) a IRT (teorií odpovědí na položku) spočívá v postavení položek jako jednotek testu. Zatímco CTT se položkami zabývá v podstatě výhradně v kontextu konkrétního testu, jehož jsou součástí, základním stavebním kamenem IRT je *skutečně* položka a její vlastnosti (např. Hambleton, Swaminathan, Rogers, 1991; McDonald, 1999). Samozřejmě i v CTT je položka významným objektem zájmu, ale není oddělitelná od celku testu. Teprve IRT uvažuje o položce samostatně a o testu jako souboru takových samostatných položek. Toto „osvobození“ položek má několik praktických důsledků:

Za prvé, měření rysů (vlastností, dovedností, vědomostí atp.) již není nutné interpretovat výhradně v kontextu populace, pro kterou byla provedena standardizace testu (v IRT odpovídá této proceduře tzv. kalibrace položek). To vyplývá z faktu, že i parametry položek jsou nezávislé na souboru, na kterém byly kalibrovány (Hambleton, Swaminathan, Rogers, 1991 – viz konec tohoto oddílu). To má značné důsledky pro interpretaci výsledků testování – kromě postupu interpretace skrze tradiční normy byly vyvinuty nové typy indexů (viz dále).

Za druhé, z nezávislosti vlastností položek na složení celého testu vyplývá možnost variabilní délky a složení testu. Zatímco v CTT je složení a délka testu konstantní, aby byla garantována jistá minimální hodnota reliability a skutečnost, že test opravdu měří to, co se od něj očekává (tzn. jeho validitu), v IRT existuje možnost, že krátký test s několika málo vhodně zvolenými položkami bude mít vyšší reliabilitu než test dlouhý.

Za třetí, z předchozího tvrzení vyplývá možnost inovací pro praxi konstrukce testů a samotného testování. Např. postup adaptivního testování, známý již téměř století (Kline, 1993; Hambleton, Swaminathan, Rogers, 1991), dostává díky IRT teoretické a metodické zakotvení. Jedna z možných podob adaptivního testování je taková, že se provede počáteční odhad např. inteligence probanda a poté se mu administrují další položky na základě určitého algoritmu (viz např. manuál k WAIS-R, Říčan, Šebek, Vágnerová, 1983). Podstatnou součástí tohoto postupu ale zpravidla bývá administrace položek v předem stanoveném pořadí. Součástí tohoto přístupu je předpoklad, že položky jsou v testu seřazeny podle rostoucí obtížnosti (v tradičním slova smyslu, tzn. v kontextu CTT). Jakmile proband dosáhne mezí svých schopností (což se projeví opakováním neúspěchem při řešení položek), procedura se „adaptivně“ ukončí.

Díky IRT získávají logika a postupy adaptivního testování další možnosti. V CTT je adaptivní procedura volena tak, aby nebyla adaptivní příliš, protože jinak by došlo k rozpadu konceptů reliability a validity testu jako celku. Protože v IRT je jednotkou testování položka, je adaptivní testovací procedura volena tak, aby se maximalizovala reliabilita (a tím i validita) konkrétního měření a aby přitom délka testu nebyla větší než je nezbytně nutné. V praxi to vypadá tak, že se v každém okamžiku testování volí položka, která v této chvíli přináší co největší množství informace o úrovni měřeného rysu probanda, čímž se zvyšuje přesnost měření (jeho reliabilita, viz dále). Pro každého probanda pak ad hoc vzniká test, který je mu doslova „ušit na míru“ (tailored – viz Kline, 1993; McDonald, 1999).

Základních rozdílů mezi CTT a IRT je mnoho a bylo by možné v jejich uvádění dálé pokračovat. Např. jeden z dalších závažných rozdílů spočívá v tom, že zatímco položková analýza u CTT je v zásadě spíše exploratorní postup analýzy dat, položková analýza u IRT je postup konfirmatorní (vždy musí být jasné, na základě jakého modelu se provádí odhad parametrů); teprve ve chvíli, kdy je doložena uspokojivá míra shody modelu s daty, se začínají interpretovat výsledky analýzy a činí se závěry o jednotlivých položkách.

Po tomto (spíše neformálním) uvedení hlavních rozdílů mezi CTT a IRT je na místě zmínit základní postuláty celé teorie (viz Hambleton, Swaminathan, Rogers, 1991):

- Odpověď probanda na položku lze předpovědět či vysvětlit pomocí množiny faktorů (latentních rysů) probanda.
- Vztah mezi pravděpodobností určité³⁾ odpovědi probanda na položku a latentními rysy⁴⁾ lze popsat pomocí monotónně rostoucí charakteristické funkce položky (*item characteristic curve – ICC*).

První předpoklad je v podstatě totožný s implicitním předpokladem CTT – i zde se prostřednictvím skupiny položek pokoušíme měřit nějaký rys, na jehož míru nelze usuzovat přímo. Ovšem druhý předpoklad je mnohem realističtější než v CTT, protože – na rozdíl od lineárního vztahu mezi pravděpodobností určité odpovědi na položku a mírou měřeného rysu uvažovaného v klasické teorii – je vztah mezi těmito hodnotami v IRT nelineární (viz dále). V tomto článku se budeme zabývat pouze jednoduššími modely, které pracují pouze s jedním latentním rysem, tzn. jsou jednorozměrné, protože jsou ilustrativnější a také mnohem lépe rozpracované.

ROLE POLOŽKY U CTT A IRT

Jedním z nejdůležitějších témat CTT jsou postupy položkové analýzy, kterých existuje několik desítek. Cílem těchto metod je vybrat takové položky, které v rámci právě konstruovaného (analyzovaného) testu tvoří spolu s ostatními položkami požadovaný konzistentní celek. S jistou mírou nadsázky je však možné tvrdit, že samotná položka vlastně podle CTT nic podstatného neměří. Bez kontextu celého testu nemá CTT nástroje, kterými by mohla studovat, zda položka skutečně měří, co měřit má. Měřicím nástrojem je až test, který je složen z vhodně vytvořených a vybraných položek.

V teorii odpovědi na položku (IRT), jak už sám název napovídá, je položka v centru naší pozornosti. O jednotlivé položce se sice také uvažuje v kontextu dalších položek, ale nikoli nutně z hlediska nějakého testu, ale zato vždy z hlediska latentního rysu,

³⁾ Rozhodli jsme se použít termín „určitá odpověď na položku“, který v tomto případě znamená: a) správnou odpověď, nebo b) odpověď svědčící pro nějakou míru latentního rysu (např. neuroticismu).

⁴⁾ Tímto vztahem se přesně rozumí pravděpodobnost určité odpovědi podmíněná úrovněmi latentních rysů.

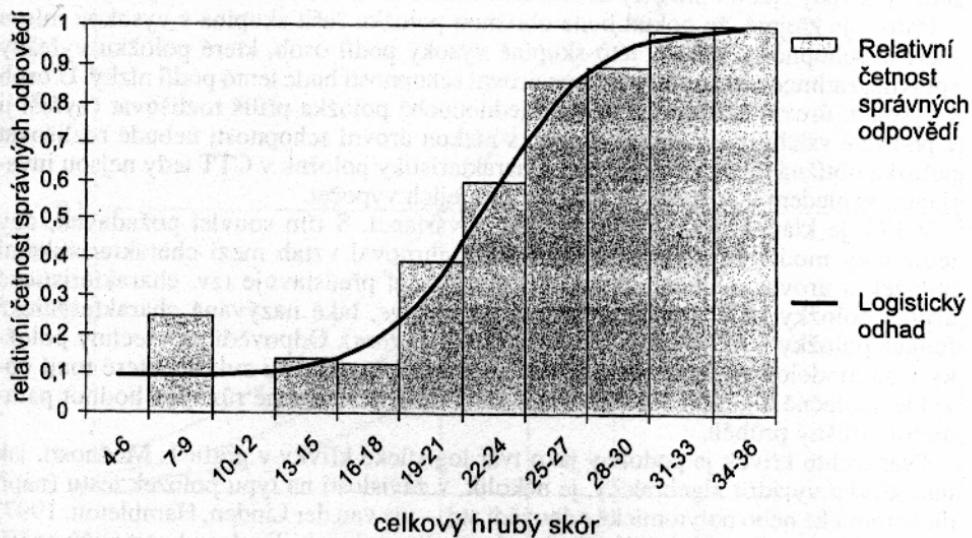
který má být položkou měřen. To vyplývá z výše uvedeného pojetí testu v IRT, který může být složen pro každého probanda z jiných položek, a přesto bude měřit tentýž rys.

Podívejme se teď na hlavní rozdíly mezi CTT a IRT z hlediska vlastností položek.

Vztah odpovědi na položku a měrené schopnosti

V CTT se postuluje lineární model vztahu mezi pravděpodobností určité odpovědi na položku a mírou rysu, který má položka (v kontextu daného testu) měřit. Tzn. pravděpodobnost toho, že konkrétní proband odpoví na položku správně nebo v diagnostickém směru, rovnoměrně roste v závislosti na tom, jak vysoký je celkový skóř tohoto probanda v testu. Obvyklým indexem vyjadřujícím míru vztahu mezi touto pravděpodobností a úrovní měřeného rysu je tedy v CTT korelace (v případě dichotomických odpovědí na položky bodově-biseriální), tzn. jedná se o lineárně regresní model vztahu.

Problém u tohoto modelu představují probandi s velmi vysokou, nebo naopak, s velmi nízkou úrovní měřeného rysu. U většiny položek dochází k absurdní situaci, kdy probandi s nízkou úrovní rysu mají zápornou pravděpodobnost správné (diagnostické) odpovědi a probandi s úrovní vysokou pravděpodobnost vyšší než jedna. Model je tedy zjevně nerealistický a může fungovat v nejlepším případě pouze jako hrubá aproximace.



Graf 1 Vztah relativní četnosti správné odpovědi na položku „ab9“ s celkovým hrubým skórem barevných progresivních matic

Ilustruje to příklad na grafu 1. Je zde znázorněn vztah mezi relativní četností správné odpovědi na položku „ab9“ Ravenových barevných progresivních matic (Raven, Court, Raven, 1991), v závislosti na celkovém hrubém skóru. Byly použity výsledky 608 dětí ze 2. a 3. třídy ZŠ. Na grafu je patrný nelineární tvar růstu relativní četnosti (kterou můžeme chápout jako odhad pravděpodobnosti správné odpovědi na položku). Tento nelineární vztah je approximován logistikou křivkou.

Důležitý rozdíl mezi CTT a IRT spočívá v tom, že zatímco u CTT je odhad měřeného rysu daného probanda založen přímo na hrubém skóru testu (který podle předpokladů CTT obsahuje určitý podíl chyby měření), u IRT se měřený rys odhaduje jako úroveň latentního rysu, který je součástí modelu vysvětlujícího odpovídání probandů na položky testu. Tento rozdíl se nemusí na první pohled zdát nijak závažný, ale ve skutečnosti je podstatný. Úroveň měřeného rysu vyplývá u CTT z hrubého skóru celého testu, takže proces jeho měření je s testem neoddělitelně spojen. V IRT je ale možné, aby osoby se stejným hrubým skórem dosáhly různých úrovní latentního rysu, protože záleží především na konkrétních položkách, ze kterých je tato úroveň odhadována.

Podívejme se nyní na to, jakým způsobem řeší IRT problém vztahu měřeného rysu a odpovědi na konkrétní položku.

Vlastnosti položek u CTT a IRT

Základními vlastnostmi položek v CTT jsou obtížnost (příp. tzv. popularita) a rozlišovací účinnost (nebo také diskriminační schopnost). Obtížnost je definována jako podíl (procento) osob, které položku zodpoví správně (jedná se tedy spíše o „jednoduchost“ – viz McDonald, 1999). Rozlišovací účinnost je definována jako korelace položky s celkovým skórem, popřípadě tzv. korigovaná korelace položky s celkovým skórem, což je korelace se skórem získaným ze zbylých položek testu (McDonald, 1999). Oba indexy se tedy zjišťují pro celý soubor bez ohledu na úroveň měřeného rysu probandů.

Přitom je zřejmé, že pokud bude obtížnou položku řešit skupina s vysokou mírou měřené schopnosti, bude v této skupině vysoký podíl osob, které položku vyřeší správně, zatímco ve skupině s nízkou úrovní schopnosti bude tento podíl nízký. U osob s vysokou úrovní schopnosti nebude jednoduchá položka příliš rozlišovat (vyřeší ji v podstatě všichni) a podobně u osob s nízkou úrovní schopnosti nebude rozlišovat položka obtížná (nevýřeší ji nikdo). Charakteristiky položek v CTT tedy nejsou invariantní vzhledem k souboru použitému pro jejich výpočet.

V IRT je kladen důraz právě na tuto invariantaci. S tím souvisí požadavek, aby teoretický model odpovídání na položku zahrnoval vztah mezi charakteristikami položek a úrovní měřeného rysu. Takový model představuje tzv. charakteristická křivka položky (*ICC – item characteristic curve*, také nazývaná charakteristická funkce položky – *ICF – item characteristic function*). Odpovědi na všechny položky jsou modelovány konkrétními ICC (pro každou položku zvlášť), které mají obvykle společné algebraické vyjádření, ale v důsledku obecně různých hodnot parametrů odlišný průběh.

Tvar těchto křivek je podobný jako tvar logistické křivky v grafu 1. Možnosti, jak tuto křivku vyjádřit algebraicky, je několik, v závislosti na typu položek testu (např. dichotomické nebo polytomické odpovědi atd. – viz van der Linden, Hambleton, 1997) a počtu parametrů představujících charakteristiky položek. Tři dosud nejpoužívanější parametry pro průběh ICC jsou obtížnost (*difficulty*), rozlišovací účinnost (*discrimination power*) a uhádnutelnost (*pseudo-guessing*).

Obtížnost, rozlišovací schopnost, uhádnutelnost

V této části není z důvodů přesnosti možné vyhnout se matematickým vzorcům. Budou postupně představeny tři modely pro položky s dichotomickým formátem odpovídání. První model obsahuje pouze parametr obtížnosti položek, do dalších modelů je pak postupně přidán nejprve parametr rozlišovací účinnosti a posléze parametr uhádnutelnosti.

Logistický model s jedním parametrem

Model označovaný v originále zkratkou 1PL (*one-parameter logistic model*) je pokus o formalizaci vztahu mezi úrovní latentního rysu probanda a pravděpodobnosti určité odpovědi na položku, který je upravován pouze obtížností položky. Model 1PL má následující algebraické vyjádření:

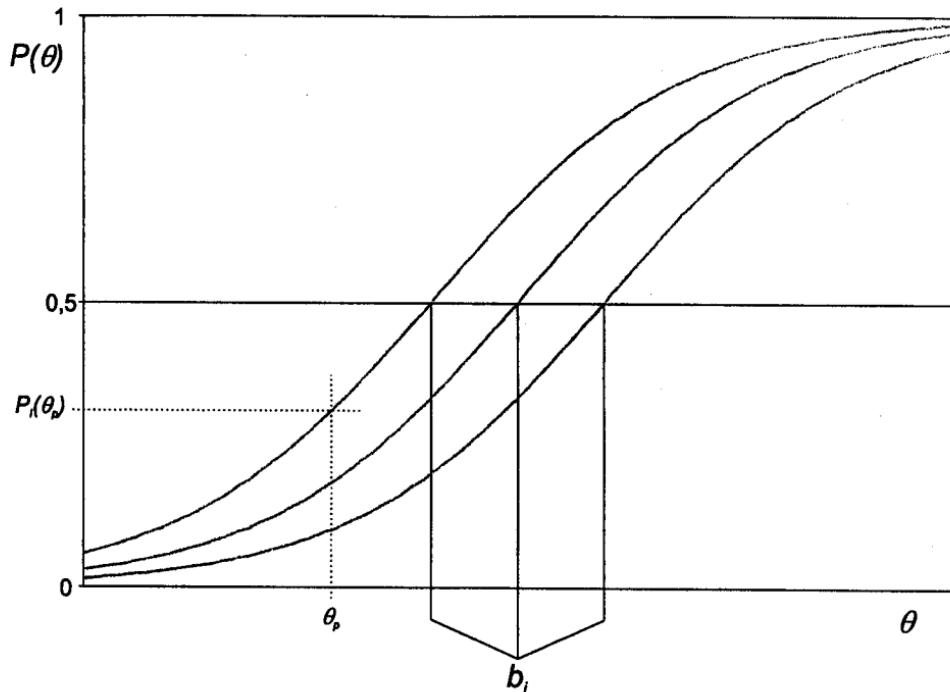
$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$$

kde $P_i(\theta)$ je pravděpodobnost určité odpovědi na položku i s obtížností b_i u probanda s úrovní latentního rysu θ ,

θ je úroveň latentního rysu probanda,

b_i je obtížnost položky i (viz níže),

D je konstanta ($D = 1,7$), která approximuje logistickou křivku do tvaru co nejpodobnějšího normálnímu kumulativnímu rozložení (které bylo používáno tradičně, ale má nevýhodné vlastnosti pro odhad modelu – viz Hambleton, Swaminathan, Rogers, 1991).



Graf 2 Tři charakteristické křivky položek pro model 1PL

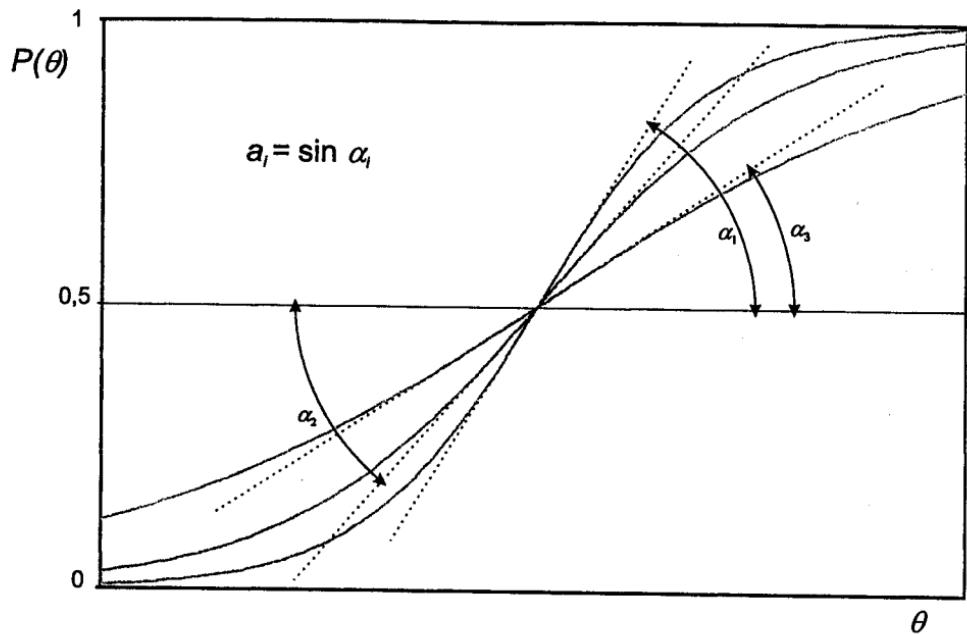
Tři křivky v grafu 2 jsou ICC tří různých položek, které se liší obtížností. Osa X představuje škálu latentního rysu θ a osa Y pravděpodobnost určité odpovědi na položku. Pokud zvolíme úroveň latentního rysu θ_p , a vedené kolmici z tohoto bodu na osu X k ICC zvolené položky I a z místa jejich průsečku vedené kolmici na osu Y, dostaneme zde hodnotu pravděpodobnosti určité odpovědi na danou položku $P_i(\theta_p)$ u probanda P se zvolenou úrovní latentního rysu θ_p (viz graf 2). Obtížnost dané

položky je definována jako bod na škále latentního rysu, kterému odpovídá pravděpodobnost 0,5 určité odpovědi; nejobtížnější je tedy položka, která je posunutá nejvíce vpravo.

Z tvaru křivky, který je monotónně rostoucí, vyplývá, že čím vyšší je úroveň latentního rysu probanda, tím vyšší je pravděpodobnost jeho určité odpovědi na položku. Ze sklonu ICC v určitém bodě jejího průběhu⁵⁾ se dá usuzovat na rozlišovací schopnost dané položky u probandů s touto úrovni latentního rysu. Čím strmější je tento sklon, tím větší je změna pravděpodobnosti určité odpovědi i při drobných změnách v úrovni latentního rysu – jinými slovy, tím citlivější je měření, prováděné touto položkou. Obecně platí, že nejcitlivěji položka rozlišuje mezi probandy, jejichž úroveň latentního rysu se nachází v blízkosti hodnoty obtížnosti dané položky (Hambleton, Swaminathan, Rogers, 1991).

Logistický model se dvěma parametry

Model 1PL je nevhodnější pro testy složené z položek, u kterých se předpokládá přibližně stejný průběh rozlišovací schopnosti. Obecně tomu však takto být nemusí, proto byl zaveden další parametr, který sumarizuje vlastnosti průběhu ICC, které jsme v předchozím odstavci trochu svévolně nazvali rozlišovací schopnost. Tento parametr



Graf 3 Tři charakteristické křivky položek pro model 2PL

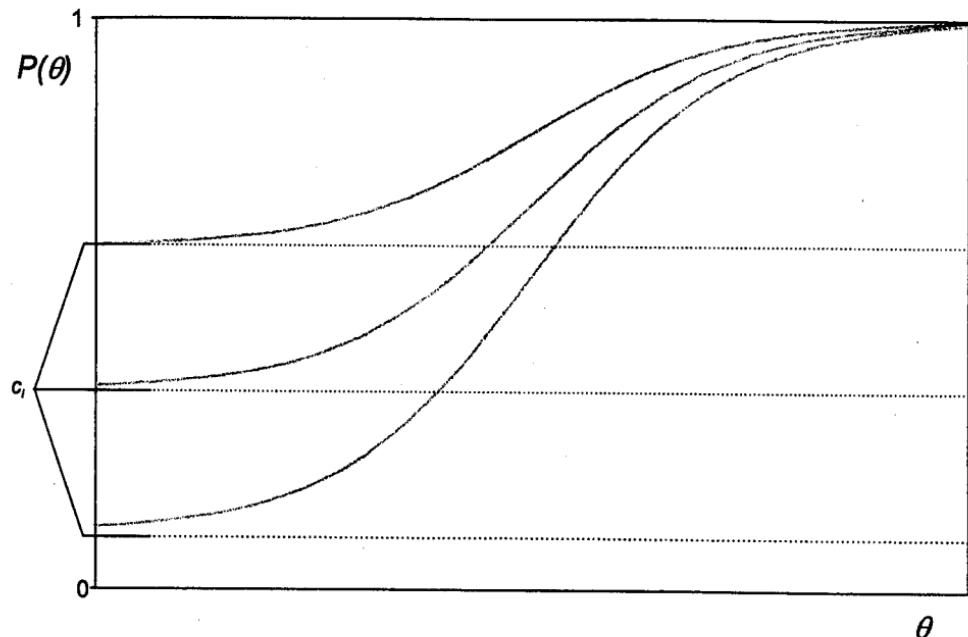
⁵⁾ Z matematického hlediska se vlastně jedná o derivaci charakteristické funkce položky v tomto bodě; u přímky bychom mluvili o její směrnici.

⁶⁾ Z nedostatku české terminologie nazýváme rozlišovací schopnost položky testu fakt, že odpovědi probandů na tuto položku se liší. Rozlišovací účinnost pak nazýváme konkrétní operacionální podobu této „schopnosti“ položky v IRT.

se nazývá rozlišovací účinnost⁶⁾ a model, ve kterém se poprvé objevuje, se zkracuje jako 2PL (*two-parameter logistic model*). Jeho vzorec následuje:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Významy všech parametrů tohoto modelu jsou totožné jako u modelu 1PL; kromě toho se zde zavádí další parametr a_i , který představuje diskriminační (rozlišovací) účinnost dané položky, jíž je konkrétně hodnota derivace dané ICC v bodě kolem hodnoty obtížnosti této položky. Čím je parametr a_i vyšší, tím lépe daná položka rozlišuje.



Graf 4 Tři charakteristické křivky položek pro model 3PL

V grafu 3 jsou tři ICC položek se stejnou obtížností, ale rozdílnou diskriminační účinností. Nejlépe rozlišuje položka, jejíž průběh je nejstrmější, protože tataž změna na škále latentního rysu vede u této položky k největší změně pravděpodobnosti určité odpovědi.

Logistický model se třemi parametry

U některých formátů odpovědí (např. nucené volby z několika málo možností) se může stát, že určitou (např. správnou) odpověď volí s poměrně vysokou pravděpodobností i probandí s extrémně nízkou úrovní latentního rysu. Nabízí se vysvětlení, že tito probandí jsou schopni správnou odpověď jednoduše uhádnout. U testu s volbou z osmi možností teoreticky nemůže četnost správných odpovědí klesnout pod 1/8 = 0,125. Jak to může vypadat v praxi, je vidět např. na grafu 1. Proto byl dvouparamet-

rový model obohacen o třetí parametr, nazývaný uhádnutelnost (*pseudo-guessing*). Vzorec tohoto modelu je následující:

$$P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-D_{ai}(\theta - b_i)}}$$

Opět jsou zde použity již zavedené parametry a navíc je zde parametr c_i , který představuje uhádnutelnost položky i .

V grafu 4 jsou tři ICC položek, jejichž parametry obtížnosti a diskriminační účinnosti jsou stejné, ale liší se v parametrech uhádnutelnosti. Nejvyšší hodnotu uhádnutelnosti má položka, jejíž levá asymptota leží nejvýš.

Model 3PL není zcela bezproblémový a mnoho autorů ho z různých důvodů kritizuje (např. Wright, 1999). Jedním z důvodů, proč se tento model příliš nedoporučuje, je problém s identifikovatelností parametrů v průběhu odhadu jejich hodnot. Navíc parametr uhádnutelnosti fakticky posouvá hodnoty předchozích dvou parametrů, čímž mění jejich definice a vlastnosti, které byly uvedeny výše. Uvedené tři modely nejsou jediné, naopak patří spíše k těm jednodušším. Existují modely např. pro polytomické formáty odpovědí na položky nebo pro škály (viz van der Linden, Hambleton, 1997).

V celém článku chybí podstatná součást všech knih zabývajících se IRT – otázky odhadu úrovně latentního rysu probanda a parametrů jednotlivých položek v závislosti na konkrétním zvoleném modelu; jinými slovy: konkrétní postupy položkové analýzy a samotného měření. Rozhodli jsme se tato téma z článku určeného široké psychologické veřejnosti vyloučit, protože se jedná o poměrně složitý matematický problém, který je nutné řešit za pomocí specializovaného software (viz např. Hambleton, Swaminathan, Rogers, 1991)⁷⁾. Zájemce o tuto problematiku můžeme odkázat na internet, kde lze nalézt značné množství konkrétních (ovšem často značně technických) návodů, jak postupovat⁸⁾.

RELIABILITA A PŘESNOST MĚŘENÍ V CTT A IRT

V CTT existuje jednoduchý vztah mezi reliabilitou testu a standardní chybou měření rysu tímto testem. Reliabilita je obecně definována jako podíl rozptylu pravého skóru a rozptylu hodnot naměřených testem:

$$\rho_{xx} = \frac{\sigma_T^2}{\sigma_x^2}$$

kde ρ_{xx} je reliabilita daného testu, σ_T^2 je rozptyl pravého skóru a σ_x^2 je rozptyl hodnot testových skóru. Tato definice je spíše teoretická a v praxi se používají různé metody odhadu reliability (např. McDonald, 1999). Standardní chyba měření se pak definuje jako

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx}}$$

⁷⁾ Velmi obsáhlý zdroj informací o knihách a speciálním software pro IRT (ale i CTT) představuje URL <http://www.assess.com>.

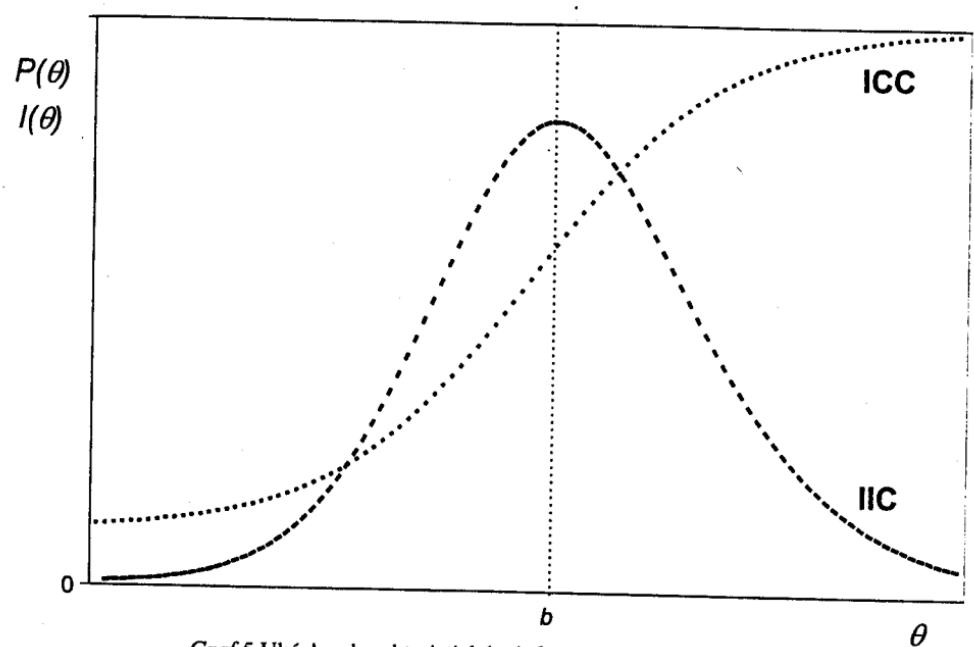
⁸⁾ Nejužitečnější se nám jeví zejména tyto dvě adresy: <http://www.b-a-h.com/papers/paper9701.html> a <http://www.b-a-h.com/papers/note9801.html>, které obsahují podrobný popis EM algoritmu pro odhad parametrů modelů IRT.

kde σ_{ϵ} je standardní chyba měření, σ_x je směrodatná odchylka hodnot testu a ρ_{xx} je jeho reliabilita. Na základě standardní chyby měření je možné sestrojit intervaly spolehlivosti naměřené úrovňě rysu pro všechny probandy společně (všichni probandí mají tedy totičný rozsah intervalu spolehlivosti). Z výše uvedených vzorců totiž vyplývá, že standardní chyba měření není závislá na úrovni měřeného rysu. Jako všude v CTT reliabilita, a tím i standardní chyba měření jsou tedy vlastnosti testu jako celku získaného pro určitý konkrétní soubor. Délka testu v CTT je volena tak, aby byla zajištěna dostatečná přesnost při testování všech probandů, kterým je test určen. Testy jsou tedy pro většinu probandů z hlediska délky naddimenzované.

Informace a chyba měření

Je všeobecně známo, že přesnost měření závisí na citlivosti měřícího nástroje pro určitou úroveň měřené veličiny. (Asi nikdo by na obyčejných kuchyňských vahách nevážil svoje zavazadlo před cestou letadlem, ale ani miligramovou hmotnost pro nějaký chemický pokus.) Pokud se testem pokoušíme měřit úroveň schopnosti probanda, který stěží zvládá ty nejjednodušší položky na začátku testu, dopouštíme se zkreslení, o jehož velikosti nic nevíme. Toto zkreslení je ale nižší, pokud zvolíme test, který je svou celkovou obtížností pro probanda vhodnější. Částečně se tento problém řeší i v rámci CTT – v praxi se klade důraz na adekvátnost metody možnostem probanda, ale tento požadavek nemá formální zakotvení v teorii.

IRT řeší otázku chyby měření zcela odlišným způsobem. Intuitivně přijatelný postulát říká, že velikost chyby klesá s přírůstkem informace (s redukcí neurčitosti). Každá administrovaná položka přináší o probandovi jistou informaci, se kterou je IRT schopna pracovat. Tato informace závisí nejenom na odhadu úrovni měřeného latentyho rysu, ale i na vlastnostech položky. Jedná se tedy opět o křivku, nazývanou informační



křivka položky (IIC – *item information curve, item information function*), jejíž vzorec je následující:

$$I_i(\theta) = \frac{[P_i(\theta)]^2}{[P_i(\theta)] Q_i(\theta)}$$

kde $P'_i(\theta)$ je derivace pravděpodobnosti určité odpovědi na položku i pro osoby s úrovní latentního rysu θ , $P_i(\theta)$ je opět tato pravděpodobnost a $Q_i(\theta)$ je pravděpodobnost jiné než určité odpovědi, tzn. $Q_i(\theta) = 1 - P_i(\theta)$. Tato funkce má zvonovitý tvar s maximem v blízkosti hodnoty parametru obtížnosti dané položky (graf 5). Z toho vyplývá již zmíněný fakt, že položka nejlépe rozlišuje (přináší nejvíce informace) u probandů s úrovní latentního rysu blízko hodnoty obtížnosti položky.

Celkový informační přínos testu (tzn. dosud administrovaných položek) vyplývá z informační funkce testu, která je prostým součtem informačních funkcí (křivek) administrovaných položek. Vzorec je tedy:

$$I(\theta) = \sum_{i=1}^n I_i(\theta)$$

kde $I(\theta)$ je informační funkce testu, $I_i(\theta)$ je informační funkce administrované položky i a n je celkový počet administrovaných položek. Ze všech těchto vzorců vyplývá, že informační přínos testu je obecně různý pro různé úrovně latentního rysu.

Informační přínos testu lze jednoduše převést na standardní chybu měření podmíněnou danou úrovní latentního rysu. Rovná se odmocnině pětadvacáté hodnoty informace testu:

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

kde θ je konečný odhad úrovně latentního rysu konkrétního probanda testem s informačním přínosem $I(\theta)$. Nyní lze na základě známé chyby měření určit intervaly spolehlivosti pro odhad úrovně latentního rysu daného probanda.

MOŽNOSTI INTERPRETACE VÝSLEDKŮ U CTT A IRT

Jak vyplývá z předchozího textu, představují modely IRT mnohem propracovanější řešení problému, jak měřit psychické jevy než CTT. Tato větší sofistikovanost má další důsledky pro práci s takto získanými výsledky testování.

Všechny tyto nové možnosti vyplývají z povahy odhadu úrovně latentního rysu θ , který je vlastně jádrem měření v IRT (na rozdíl od CTT, kde se vše odvíjí od hrubého skóru). V CTT je nutné hrubé skóry ještě nějakým způsobem zakotvit – bud' vzhledem ke kritériu, nebo vzhledem k normám (viz např. standardy pro pedagogické a psychologické testování, 2001), aby bylo možné věcně interpretovat výsledky konkrétní osoby. Nedostatky těchto postupů jsou všeobecně známy. V IRT vystupuje jako výsledek měření odhad úrovně latentního rysu θ , který je možné transformovat různým způsobem, a to jak na tradičně (v CTT) používané normy nebo kriteriálně zakotvené indexy, tak na tzv. úroveň zběhlosti (*proficiency level*), relativní index zběhlosti (*relative proficiency index*) nebo vývojové pásmo (*developmental zone*). Tyto nové typy interpretacích pomůcek by se slušelo krátce představit.

Úroveň zběhlosti (*proficiency level*) je kriteriálně interpretovaná informace (viz standardy pro pedagogické a psychologické testování, 2001) o tom, jakou úroveň obtížnosti úkolů je určitá osoba schopna vyřešit s danou pravděpodobností.

Index relativní záběhlosti (RPI – relative proficiency index) vychází z úrovně záběhlosti, kterou dává do vztahu s výkonem, jaký podávají jedinci stejného věku jako testovaná osoba. Jedná se o míru úspěšnosti této osoby při řešení úkolů, které jeho vrstevníci zvládají s 90% úspěšností (což je kritérium často používané pro posouzení mistrovství v nějakém typu úkolů). RPI má podobu zlomku, ze kterého je na první pohled patrné, jestli je výkon probanda nad, nebo pod úrovní výkonu jeho vrstevníků (Woodcock, 1999).

Vývojové pásmo (developmental zone) představuje další rozpracování RPI. Jedná se o úsek na škále latenciho rysu, pro kterou byly vytvořeny věkové nebo ročníkové normy. Úkoly v tomto úseku budou pro daného probanda přiměřené jeho schopnostem, což může znamenat velký přínos např. pro jeho učitele.

Závěr vyplývající z testování pak může znít např. takto (jedná se o 10-letého Carlose, chlapce hispánského původu, žáka 4. ročníku Kennedy Middle School, testovaného jedním z jazykových testů používaných v USA u bilingvních dětí): „Carlos vykazuje velmi omezenou až omezenou úroveň záběhlosti v angličtině (úroveň 2 – 3). Jeho výkon v tomto testu je srovnatelný s průměrem anglicky mluvících studentů s ročníkovým ekvivalentem 1,8 (tj. konec 2. třídy). Úkoly testu vyžadující úroveň angličtiny pod úrovní ročníkového ekvivalentu 1,3 budou pro něj spíše snadné; úkoly nad úrovní 2,4 budou pro něj spíše obtížné.“ (Woodcock, 1999, s. 122.) Index relativní záběhlosti (RPI) tohoto žáka byl 27/90. Úroveň jazykových schopností ve španělštině byla přitom vysoce nadprůměrná, takže je možné s jistotou vyloučit kognitivní deficit. Slezna učitelka Stricklandová nyní přesně ví, které kapitoly v učebnicích angličtiny pro 1. a 2. ročník musí s Carlosem procvičit, než bude schopen pokračovat dále.

Tyto příklady a indexy byly vytvořeny převážně pro výkonové a didaktické testy, ale lze si představit jejich obdobu pro dispozice, postoje a jiné psychické charakteristiky.

ZÁVĚREČNÉ POZNÁMKY

Po přečtení předchozích odstavců může vyvstat otázka, jaké jsou výhody metod vytvořených na základě IRT oproti metodám založeným na CTT. Na prvním místě je třeba zmínit, že modely IRT jsou propracovanější a obecně více odpovídají realitě testování (v současné době testování výkonu a schopností, ale vývoj zdaleka není ukončen). Další výhoda spočívá ve faktu, že měřený rys se odhaduje jako latentní rys na základě modelu a není nutné na něj usuzovat prostřednictvím nějakého typu norem, i když IRT tvorbu norem nevyulučuje, spíše umožňuje tvorbu nových indexů užitečných pro interpretaci výsledků testování. Naprostě zásadní výhodou všech metod založených na IRT je jejich nezávislost na populaci, ze které byl vybrán kalibrační soubor, což vyplývá z předchozího bodu. S tím také souvisí možnost vytvářet počítačové adaptivní testy, jejichž délka není konstantní a závisí na zvoleném požadavku přesnosti měření. Již dnes je téměř jisté, že počítačové adaptivní testování bude v budoucnu představovat značnou část testové psychodiagnostiky a didaktického testování.

Jedna z reprezentativních učebnic CTT – *The Handbook of Psychological Testing* (Kline, 1993) – podává o IRT mírně zkreslený obraz. Vnímá ji pouze jako nástroj k vytváření „na míru šitych“ (*tailored*) testů (a to převážně výkonových) a uzavírá konstatováním, že většina kvalitních testů byla stejně vytvořena pomocí CTT. Na to je třeba namítat, že IRT je mladší teorií, a proto je zatím aplikací nutně méně. Dále Kline (1993) varuje před přílišnou fascinací modely IRT, která může podle jeho názoru odvádět pozornost od pečlivé přípravy znění položek. I když tomuto argumentu nelze upřít racionální jádro, na okouzlení elegancí modelu není v psychometrice

(a koneckonců ani v životě) nic špatného. Teprve dlouhodobá zkušenosť a práce s konkrétními aplikacemi může přinést údaje nutné pro vyhodnocení užitečnosti těchto přístupů.

Tento článek se snaží pokrýt hlavní téma IRT, a to pomocí srovnání s klasickou teorií testů (CTT). Jistě by byly možné jiné přístupy k této problematice a akcentace dalších témat. V české psychologické literatuře nenajdeme o IRT téměř nic. Každá práce tohoto rozsahu je tedy nutně neúplná a zjednodušující. Věříme však, že se psychodiagnostické metody založené na IRT stanou brzy součástí repertoáru českých psychologů (aspoň pokud se chce česká testová psychodiagnostika srovnávat se světem, ve kterém se používání IRT stalo za poslední desetiletí běžným standardem). V souvislosti se začleňováním do celoevropských a světových institucí tu hrozí reálné nebezpečí, že se testy založené na IRT u nás sice stanou standardem, ale bez toho, aby byla široká psychologická veřejnost dostatečně seznámená s teoretickými východisky jejich konstrukce.

LITERATURA

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2001): Standardy pro pedagogické a psychologické testování. Praha, Testcentrum.
- Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991): Fundamentals of Item Response Theory. Newbury Park, CA, SAGE Publications, Inc.
- Kline, P. (1993): The handbook of psychological testing. London, Routledge.
- McDonald, R. P. (1999): Test theory. A unified treatment. Mahwah, NJ, Lawrence Erlbaum Associates, Inc., Publishers.
- Rasch, G. (1960): Probabilistic models for some intelligence and attainment tests. Copenhagen, Danish Institute for Educational Research.
- Raven, J. C., Court, J. H., Raven, J. (1991): Farebné progresívne matice (CPM). Bratislava, Psychodiagnostika, spoločnosť s. r. o.
- Říčan, P., Šebek, M., Vágnerová, M. (1983): WAIS-R. Wechslerův inteligenční test pro dospělé. Bratislava, Psychodiagnostické a didaktické testy, n. p.
- van der Linden, W. J., Hambleton, R. K. (1997): Handbook of modern Item Response Theory. New York, Springer-Verlag, Inc.
- Woodcock, R. W. (1999): What can Rasch-based scores convey about a person's test performance? In: Embretson, S. E., Hershberger, S. L. (Eds.): The new rules of measurement. Mahwah, NJ, Lawrence Erlbaum Associates, Inc., Publishers, 105 – 127.
- Wright, B. D. (1999): Fundamental measurement for psychology. In: Embretson, S. E., Hershberger, S. L. (Eds.): The new rules of measurement. Mahwah, NJ, Lawrence Erlbaum Associates, Inc., Publishers, 65 – 104.

SOUHRN

Článek srovnává základní principy klasické teorie testů (CTT) a teorie odpovědi na položku (IRT). Hlavní důraz je kláden na představení modelů IRT a jejich výhod při tvorbě testových metod. Srovnání CTT a IRT se soustředuje na otázky vztahu položky a celého testu, vlastností položek, reliability a přesnosti měření a možností interpretace výsledků testování.