K EDUKOMETRICKÝM PRINCIPŮM: ITEM RESPONSE THEORY

S. Komenda - J. Zapletalová

Ústav sociálního lékařství a zdravotní politiky LF UP Olomouc

Měření

Měření patří k významným doprovodným jevům lidské civilizace – přesněji řečeno, měření je vlivným faktorem jejího rozvoje. Úsilí řešit problémy mezilidského porozumění, komunikace v obchodu, přenosu technologií, dopravě a vzdělání bylo hybnou silou integrace lidských společenstev.

Pokud jde o vědu a techniku, průmyslové technologie a obchod, sotva kdo o potřebách a možnostech měření pochybuje. Daleko složitější a méně zřejmá je otázka měření v oblasti vzdělávacího procesu, psychologii a jiných humanitních oborech. Zkoušky a zkoušení jsou procedury užívané při výuce už tradičně; zkoušení má nepochybně svou historii jako součást zpětné vazby mezi učícím se subjektem a jeho učitelem, případně vlastní motivací naučit se. Přesto však teprve v posledních desetiletích je tato oblast chápána jako měření *per se*, které má být podrobováno všem kritériím jako měření fyzikální či technická. Jde vlastně o prosazování potřebnosti *metaměření*, tj. zjišťování a posuzování vlastností a kvality samotného procesu měření. Sem patří také zjišťování objektivnosti, reliability a validity měřicí metody.

Měření znalostí či schopností subjektů se uskutečňuje formou testů. Test může být považován za experiment (van der Linden, Hambleton 1997), jehož hlavním metodologickým problémem je kontrolovat chyby měření. Ty vznikají v důsledku toho, že na chování subjektu, jehož znalosti či schopnosti jsou předmětem měření, mají vliv nejenom veličiny v testu kontrolované, ale i faktory další, které už přímo neřídíme, a jež souhrnně označujeme jako pozadí, šum či chybové proměnné. Nejsou-li takové faktory pod přiměřenou kontrolou, nelze z experimentu, tj. z měření, činit platné závěry.

Kontrola vlivu přímo neřízených faktorů se provádí trojím způsobem:

- párováním či standardizací
- znáhodňováním
- statistickou úpravou

Jsou-li experimentální podmínky standardizovány, znamená to, že subjekty jednají při týchž úrovních neřízených (chybových) faktorů a vlivem těchto faktorů nemohou být vysvětlovány rozdíly ve výsledcích experimentu. Ačkoli je standardizace (matching) technika velice účinná, má i své nevýhody; k nim patří omezená zobecnitelnost. Výsledky experimentu se totiž váží (a jsou podmíněny) na to, jaká byla úroveň standardizovaných faktorů.

Znáhodnění se opírá o představu, že náhodná volba podmínek (neřízených faktorů) dává stejnou šanci všem experimentálním situacím – a v průměru žádnou z nich nezvýhodňuje. Vliv neřízených faktorů na chování subjektů by tedy měl být vyvážený.

Statistická úprava je technika řízení *post hoc, tj.* nepředpokládající zásah experimentátora do experimentu. Používá se, nebylo-li možno aplikovat standardizaci ani znáhodnění; vyžaduje, aby byly v průběhu pokusu měřeny všechny podstatné (relevantní) chybové proměnné. Těchto měření se pak dodatečně využívá k úpravě pozorovaných hodnot závislých veličin (chování subjektu), které se tak od vlivu chybových proměnných očišťuje.

Techniky statistického upravování se opírají o model – v němž je kvantitativně (matematickými formulemi) popsána souvislost chybových proměnných se závislými veličinami. Tohle se běžně dělá např. ve fyzice, kdy jsou vztahy mezi veličinami popsány spolehlivě prověřenými zákony, což umožňuje činit korekce při studiu chování složitých systémů v ekologii, meteorologii, geologii a jinde. Je-li závislou, sledovanou veličinou chování, bývá vliv doprovodných proměnných kontrolován např. v analýze kovariance, kde se vztah mezi doprovodnými a náhodnými vlivy postuluje jako lineární.

Klasická teorie testování

Klasická teorie testování se opírá o představu, že experimentální podmínky jsou kontrolovány standardizací a znáhodněním a že všechny možné jiné zdroje variability v chování subjektů působí pouze nesystematicky, tj. některé jedním směrem, tak, že je možné chápat je jako náhodné vlivy.

V souladu s touto představou jsou pozorovaná (naměřená, zjištěná) testová skóre rozkládána na *skóre skutečné* a *skóre chybové.* Formálně zapsáno, označíme-li symbolem X_{ij} naměřené skóre subjektu i při zvládání testu j, a soudíme-li, že opakovaná měření se mohou lišit pouze v důsledku působení náhodných vlivů, pak klasický model předpokládá následující strukturu pozorování

$$X_{ij} = T_{ij} + E_{ij} \quad , \qquad (1)$$

kde T_{ij} je hodnota skutečného skóre a $E_{ij} = X_{ij} - T_{ij}$ je skóre chybové. T_{ij} je přitom zároveň střední (očekávaná) hodnota pozorovaného skóre subjektu i v testu j, protože střední hodnota chybového skóre E_{ij} se postuluje jako nula.

Model (1) je adekvátní pouze tehdy, jsou-li všechny podstatné faktory mající vliv na pozorované skóre, pod kontrolou. T_{ij} je určeno v závislosti na zvolených hladinách standardizovaných chybových faktorů. Skóre je plně určeno experimentálním plánem.

Považujeme-li subjekty za náhodně vybrané vzorky z nějaké širší populace, má T_{ij} charakter náhodné veličiny T_{j}. Model (1) tím přejde do tvaru

$$X_{j} = T_{j} + E_{j} \quad . \qquad (2)$$

(2) odpovídá modelu analýzy rozptylu (ANOVA) při jednoduchém třídění, kdy je faktor považován za náhodný.

IRT

Matematické modely zaměřené na statistickou úpravu testových skóre rozvíjí IRT. Známé IRT modely pro nula-jedničkové odpovědi např. upravují odpovědi s ohledem na takové vlastnosti testových položek jako jsou jejich obtížnost, diskriminační schopnost a přípustnost náhodného uhádnutí správné odpovědi. IRT se ovšem neomezuje jenom na nula-jedničkové odpovědi.

Modely jsou navrhovány v souladu se zásadou, že pro každý faktor, jehož vliv na odpověď se předpokládá, obsahuje model jediný parametr a předpokládá se specifický vliv faktoru na odpovědi při řešení položky. Je věcí „umění modelovat", aby byl model navržen tak, že vezme v úvahu interakci mezi těmito faktory a zároveň zajistí, aby byl model statisticky zvládnutelný.

IRT modely se týkají chování subjektu na úrovni položky testu, nikoli na úrovni skóre testu. Historicky první byl dichotomický formát, klasifikující odpovědi jako správné či nesprávné. Obvyklé kódování bylo přitom „1" (správná odpověď) a „0" (chybná odpověď). Předmětem modelování byla pravděpodobnost správné odpovědi, tj. možnosti subjektu se schopností θ, $-\infty < \theta < \infty$, vyřešit správně položku j, ovlivňující pravděpodobnost úspěchu svými dvěma parametry: svou „obtížností" a svou „diskriminační schopností", obvykle označenými jako b_j, $-\infty < b_j < \infty$ a a_j, $0 < a_j < \infty$.

Protože parametr schopnosti subjektu je strukturální parametr, který je předmětem měření, zatímco parametry a_j a b_j položky j se považují za šumové parametry, je obvykle pravděpodobnost úspěchu subjektu se schopností θ při řešení položky j označována symbolem $P_j(\theta)$, tj. jako funkce argumentu θ za specifických podmínek položky j. Funkce $P_j(\theta)$ se nazývá funkcí odpovědi na položku (IRF); býval používán i název charakteristická křivka položky (OCC). Protože pravděpodobnost omezuje své hodnoty na škále od nuly do jedné, $P_j(\theta)$ nemůže být lineární funkcí argumentu θ. Zřejmě musí monotonně růst v θ. To určuje, jakého typu musí být funkce $P_j(\theta)$.

Historicky první IRT model využíval distribuční funkci standardizovaného normálního rozdělení

$$P_j(\theta) = \int_{-\infty}^{a_j(\theta - b_j)} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) dz \quad . \qquad (3)$$

Z rozboru (3) je zřejmé, že parametr b_j obtížnosti položky j je zároveň bodem na škále veličiny θ, ve kterém subjekt dosahuje pravděpodobnosti úspěchu 0,5. Hodnota a_j je úměrná strmosti křivky odpovědi $P_j(\theta)$ v tomto bodě b_j. Obě tyto vlastnosti modelu vysvětlují interpretaci naznačenou v názvech parametrů: stoupá-li

b_i, posouvá se křivka odpovědi $P_i(\theta)$ doprava a k tomu, aby subjekt dosáhl správné odpovědi s danou pravděpodobností, musí mít vyšší schopnosti (znalosti). Zároveň platí, že čím vyšší je hodnota a_i, tím lépe diskriminuje (rozlišuje) položka mezi pravděpodobnostmi úspěchu subjektů se schopnostmi (znalostmi) pod a nad hodnotou $\theta = b_i$.

Vývoj modelů tohoto typu úzce souvisel s představami a praxí psychofyzikálních měření. Příkladem může být studium závislosti percepce na fyzikálním podnětu, např. hlasitosti zvuku, kdy se měřilo, při jaké intenzitě podnětu začal už subjekt podnět vnímat. Zásadní rozdíl psychofyzikálních měření a měření psychometrických je ovšem v tom, že v psychofyzikálním experimentu bylo možno předem intenzitu podnětu specifikovat, což při měření znalostí či schopnosti možné není. θ je latentní proměnná, což je poměrně nový koncept. Dlouho se také držela představa, že měřená schopnost θ má v referenční populaci normální rozdělení. Napomáhalo se jí transformací, kdy se škála schopnosti dělila na (obvykle 7) intervalů vymezených podle zásady stejného násobku směrodatné odchylky. Středy těchto intervalů se považovaly za diskrétní skóry schopnosti a pomocí nich se „připasovávala" křivka normálního rozdělení.

Před érou počítačů bylo odhadování parametrů velice pracné. Pro model založený na představě normálního rozdělení měřené schopnosti předložili Lord a Novick (1968) vzorce vystihující souvislost parametrů b_i a a_i s klasickým podílem π a biseriálním korelačním koeficientem ρ_i položky j a testového skóre

$$a_j = \frac{\rho_j}{\sqrt{1-\rho_j^2}} \qquad (4)$$

$$b_j = \frac{-\gamma_j}{\rho_j}\ , \qquad (5)$$

kde γ_j se definuje vztahem

$$\gamma_j = \phi^{-1}(\pi_j) \qquad (6)$$

a $\phi(\cdot)$ je distribuční funkce standardizovaného rozdělení. Tyto odhady dlouho sloužily jako počáteční hodnoty iterativních postupů zavedených v počítačové éře.

Jednoparametrový logistický model podle Rasche

George Rasch se zabýval problematikou pedagogického a psychologického měření od konce 40. let. V 50. letech vyvinul dva Poissonovské modely pro testy čtení a model testování inteligence a znalostí, nazvaný jeho jménem. Z formálního hlediska jde o speciální případ modelu zavedeného Birnbaumem.

Rasch byl veden snahou objektivizovat test, tj. vyhnout se při analýze testu závislosti na populacích testovaných subjektů. Šlo mu o to, aby se analýza soustředila na jedince, a oddělila parametry položek a subjektu. Ve svém úsilí vyhnout se statistikám dovolávajícím se populací se Rasch podobal Skinnerovi, který experimentoval na jedincích.

Tento Raschův přístup se odrazil v přechodu od klasické teorie testování založené na populaci a zdůrazňující standardizaci a znáhodnění, k Item Response Theory, v níž se modeluje pravděpodobnost správného řešení položky jako interakce mezi jednotlivou položkou a jednotlivým subjektem.

Model chybného čtení postuluje Poissonovo rozdělení pro počet chyb při čtení textu. Předpoklad je oprávněný, je-li proces čtení stacionární a nemění se např. únavou subjektu nebo posunem obtížnosti textu.

Je-li text tvořen T slovy a X je náhodná veličina označující počet špatně přečtených slov, má zmíněný Poissonovský model tvar

$$P(X=x|T)=e^{-\lambda}\frac{\lambda^x}{x!}, \qquad (7)$$

kde λ je parametr mající význam očekávaného počtu špatně přečtených slov. Z toho plyne, že $\xi = \lambda/T$ je pravděpodobnost, že náhodně zvolené slovo textu bude chybně přečteno.

Rasch postupoval tak, že dále modeloval parametr ξ jako funkci dalších parametrů zachycujících schopnost subjektu a obtížnost textu. Je-li θ_i odrazem schopnosti subjektu i a δ_j odrazem obtížnosti textu j, pak ξ_{ij} by mohl být nový parametr klesající s rostoucím θ_i a rostoucí s rostoucím δ_j. Rasch navrhl, aby

$$\xi_{ij}=\frac{\delta_j}{\theta_i}. \qquad (8)$$

Je-li tento model aplikován na soubor subjektů $i=1,\ldots, N$, které čtou soubor textů $j=1,\ldots, n$, platí, že součet $X_{.j}$ chyb, kterých se při čtení textu j dopustil soubor N subjektů, je postačující statistikou pro odhad parametru δ_j. Je známo, že rozdělení veličiny X_{ij} při daném $X_{.j} = x_{.j}$ je binomické s parametrem úspěchu nezávislým na δ_j. Vliv obtížnosti textu při odhadování schopnosti subjektu se odstraní, založí-li se inference na podmíněné věrohodnostní funkci opírající se o toto binomické rozdělení.

Z představy o struktuře parametru ξ_{ij} se odvozuje hlavní Raschův model chování testové položky j. Rasch navrhuje modelovat pravděpodobnost úspěchu ($U_{ij}=1$) při konfrontaci subjektu i se schopností θ_i s položkou j o obtížnosti δ_j takto

$$P(U_{ij}=1|\beta)=\frac{\frac{\theta_i}{\delta_j}}{1+\frac{\theta_i}{\delta_j}}$$

$$=\frac{\theta_i}{\theta_i+\delta_j} \quad . \qquad (9)$$

Představíme-li si parametry θ_i a δ_j v logaritmickém tvaru jako e^{θ_i} a e^{δ_j}, může být model přeformulován do ekvivalentní podoby (tzv. jednoparametrového logistického modelu)

$$P(U_{ij}=1|\beta)=\frac{1}{1+\exp\{-(\theta-\delta_j)\}}$$

$$=\frac{\exp(\theta-\delta_j)}{1+\exp(\theta-\delta_j)} \qquad (10)$$

Dá se ukázat, že model umožňuje odhadování schopnosti subjektu nezávisle na obtížnosti položek (item-free).

Dvou a tří-parametrový logistický model podle Birnbauma

Birnbaumův hlavní příspěvek k rozvoji IRT byl návrh nahradit distribuční funkci normálního rozdělení ve (3) logistickým modelem

$$P_j(\theta)=\frac{1}{1+\exp\{-a_j\ (\theta-b_j)\}} \quad . \qquad (11)$$

Náhrada byla motivována zjištěním, že distribuční funkce logistického rozdělení s jednotkou škály 1,7, tj. $L(1,7x)$ se věrně podobá distribuční funkci normálního rozdělení, tj. $N(x)$. Platí totiž

$$|N(x)-L(1,7x)|<0,01 \quad \text{pro} \quad -\infty<x<\infty.$$

Výhodou logistické funkce je přitom lepší interpretovatelnost parametrů a_j a b_j položky j.

Birnbaum rovněž navrhl třetí parametr, jímž by model dokázal zachytit skutečnost, že u položek typu multiple-choice může i subjekt bez znalosti zkoušeného tématu dosáhnout správného řešení položky j prostě náhodným uhádnutím (guessing). Model má pak podobu

$$P_j(\theta) = c_j + (1-c_j)\frac{1}{1+\exp\{-a_j(\theta-b_j)\}} \quad . \qquad (12)$$

Model (12) je v souladu s následující představou (Komenda a Mazuchová, 1995; Komenda a Zapletalová, 1996):

Látku, na jejíž znalost se odvolává testová položka j, ovládá podíl $(1-\pi)$ 100% referenční populace, zatímco zbytek, daný podílem π 100% , příslušnou látku neovládá: Pokud jde o vztah mezi znalostí látky a chováním subjektu, platí pravidlo, že v případě znalosti látky odpovídá subjekt správně s jistotou, zatímco v případě její neznalosti dosahuje správné odpovědi s pravděpodobností c_j, vymezující možnost náhodného uhádnutí správného řešení položky j.

Podle věty o úplné pravděpodobnosti bude tedy položka j správně vyřešena (subjektem náhodně vybraným z referenční populace) s pravděpodobností

$$(1-\pi)\cdot 1 + \pi\, c_j \quad , \qquad (13)$$

kterou lze upravovat do tvaru

$$c_j + (1-c_j)(1-\pi). \qquad (13a)$$

Porovnáme-li (13a) s (12), je zřejmé, že ve výrazu pro pravděpodobnost $P_j(\theta)$ správné odpovědi na položku j hraje zlomek na pravé straně ve (12) roli pravděpodobnosti, že, subjekt z referenční populace téma položky j skutečně ovládá. V případě $c_j=0$ pak pravděpodobnost, že subjekt položku j vyřeší, je totožná s pravděpodobností, že její téma také skutečně ovládá.

Na grafu funkce $P_j(\theta)$ uváděné ve (12) mají tři uváděné parametry následující interpretaci:

- a_j souvisí přímoúměrně s diskriminační schopností položky j; čím vyšší je a_j, tím strměji probíhá křivka $P_j(\theta)$ vzhledem ke škále θ na vodorovné ose,
- b_j je ukazatelem obtížnosti testové položky j; vyšší hodnoty b_j odpovídají obtížnějším položkám a obráceně; zvýšení b_j posouvá křivku $P_j(\theta)$ doprava,
- c_j stanovuje pravděpodobnost náhodného uhádnutí správného řešení položky j subjektem náhodně vybraným z referenční populace; c_j je dolní asymptota křivky $P_j(\theta)$, tj. její limita pro θ klesající k $-\infty$.

Model (11) je znám pod názvem dvouparametrový logistický model, zatímco (13) definuje tříparametrový logistický model.

Birnbaum rovněž zavedl pro popis informační struktury testu Fisherovu míru množství informace o neznámé schopnosti θ, obsažené v testu o n položkách, jako

$$I(\theta)=\sum_{j=1}^{n} I_j(\theta)$$

$$=\sum_{j=1}^{n}\frac{\left[P_j'(\theta)\right]^2}{P_j(\theta)\left[1-P_j(\theta)\right]}\quad , \qquad (14)$$

kde $I_j(\theta)$ je množství informace o parametru θ obsažené v odpovědi na položku j testu, a $P_j'(\theta)=dP_j(\theta)/d\theta$ je 1. derivace $P_j(\theta)$ podle θ.

Parametry logistických modelů navrhl Birnbaum odhadovat metodou maximální věrohodnosti. Postupy jsou silně podmíněny možnostmi softwareového zabezpečení náročných iterativních postupů. Dnes takový software existuje.

V citovaných pracích Komendy a Mazuchové (1995) či Komendy a Zapletalové (1996) byly navrženy některé aproximace zjednodušující postup odhadování parametrů.

V rozvoji IRT byly podnětné metodologické příspěvky Lazarsfeldovy, který chápe model jako mechanismus spojující latentní parametry teoretických konceptů (schopnosti subjektu, obtížnosti a dalších parametrů testové položky) s manifestními daty, v nichž se projevuje chování subjektu.

Literatura

Komenda S., Mazuchová J.: *Tvorba a testování testu. Olomouc 1995*

Komenda S., Zapletalová J.: *Analýza didaktického testu a její počítačová podpora. Olomouc 1996*

Van der Linden Wim J., Hambleton R.K.(Eds.): *Handbook of Modern Item Response Theory. Springer 1997*

Prof. RNDr. Stanislav Komenda, DrSc.
Ústav sociálního lékařství a zdravotní politiky
Lékařská fakulta UP
Hněvotínská 3
775 15 Olomouc