Alternativní přístupy k tvorbě a interpretaci (nejen) psychologických testů

Item Response Theory (IRT)
Kowledge Space Theory (KST)

PhDr. Denisa Denglerová, Ph.D.

Historie IRT

Dvě odlišné linie

Evropa – dánský matematik George Rasch, pracoval pro dánskou armádu, modely zabývající se schopností čtení, šifrování

1960 Probabilistic models for some intelligence and attainment test

Inspiroval dva psychometriky, Gerhard Fischer z Vídeňské Univerzity, který Raschův matematický model spojil více s psychologickým uvažováním.

USA - za začátek IRT považuje vydání knihy „Statistical Theories of Mental Test Scores“ (Lord a Novick, 1968), v rámci níž se objevily čtyři kapitoly o IRT napsané Allanem Birnbaumem.

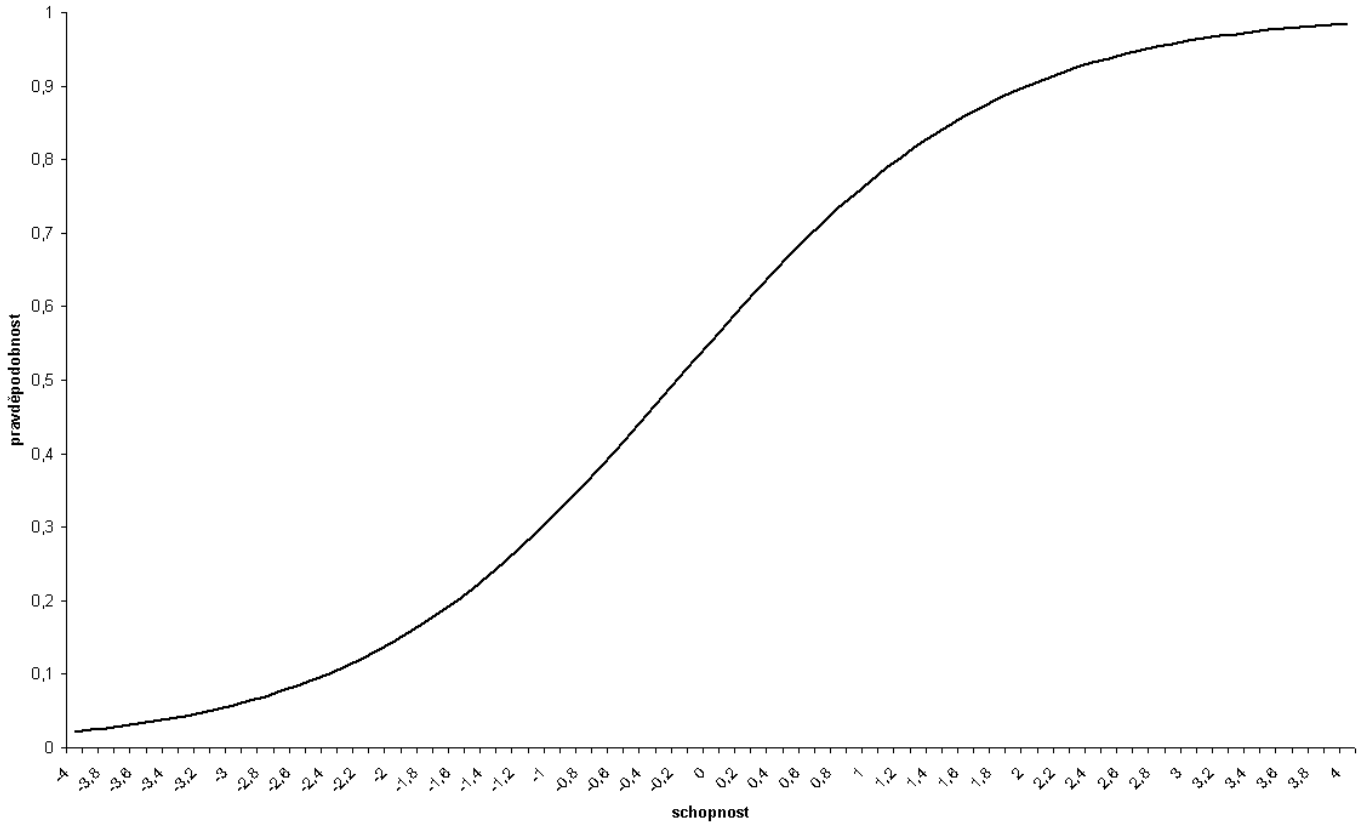
V 70. letech Rasch navštívil University of Chicago, aby tam přednesl sérii přednášek, inspiroval profesora Benjamina Wrighta, množství doktorandských prací současných klasiků IRT – Daves, Weiss, Humbleton…

Dva základní postuláty IRT

Výkon respondenta na testové položce je predikovatelný (vysvětlitelný) množinou faktorů, nazývané rysy, latentní rysy nebo schopnosti.

Vztah mezi výkonem respondenta na položce a množinou rysů, jež tento výkon zapříčiňují, může být popsán monotónní rostoucí funkcí nazývanou charakteristická funkce položky (item characteristic function). Tato křivka má tvar normální ogivy.

Charakteristická křivka položky



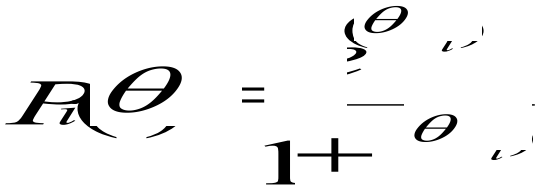
Předpoklad jednodimenzionality a lokální nezávislosti

Společným předpokladem IRT modelů je to, že množina položek (tedy celý test nebo subtest) měří pouze jednu schopnost. Tato podmínka samozřejmě není v reálu nikdy zcela splněna, jde spíše o ideál, k němuž se při výzkumech i jiných aplikacích snažíme co nejvíce přiblížit.

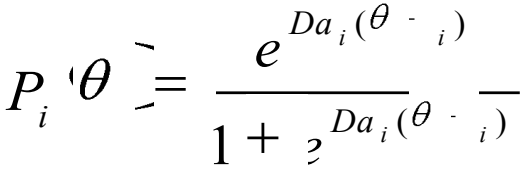
Odpovědi zkoušeného na každé dvě položky jsou statisticky nezávislé, což znamená, že neexistuje žádný vztah mezi odpověďmi respondenta na různé položky. Tento předpoklad částečně nahrazuje požadavek jednodimenzionality, jehož absolutní splnění je nemožné.

Předpoklad lokální nezávislosti nám pomáhá při tvorbě modelu uvažovat právě nad těmi schopnostmi, které opravdu ovlivňují odpovědi na položky.

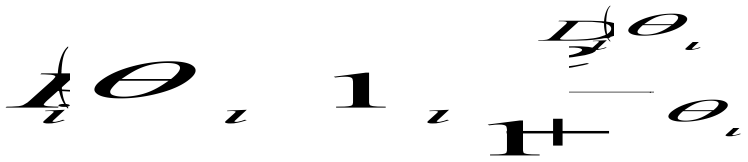
Tři klasické modely v rámci IRT



Raschův model, 1PL



Birnbaumův model, 2PL



Model s uhádnutelností, 3PL

Jednoparametrový logistický model

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

P(θ)… pravděpodobnost, že náhodně vybraný respondent se schopností θ odpoví na položku správně

b… parametr obtížnosti položky

e…Eulerovo číslo

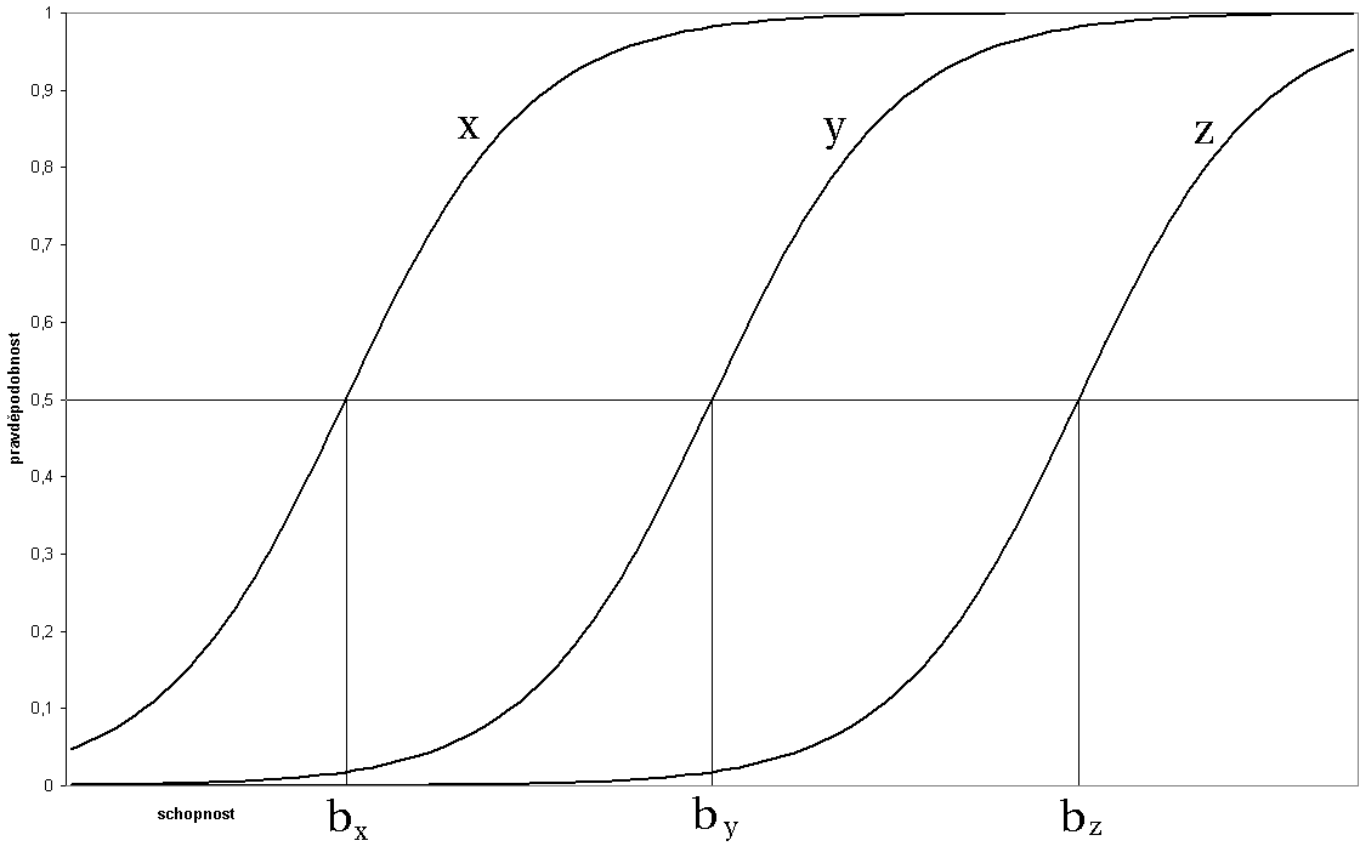
Jednoparametrový logistický model

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

Parametr položky b je bod na škále schopností, v němž je pravděpodobnost správné odpovědi rovna 0,5. Čím vyšší je hodnota b, tím větší schopnost je požadována po respondentovi, aby pravděpodobnost jeho správné odpovědi byla 50%, a tím je tedy položka těžší (obtížnější).

Parametr b…(-4, 4)

Charakteristické křivky položek lišící se parametrem obtížnosti

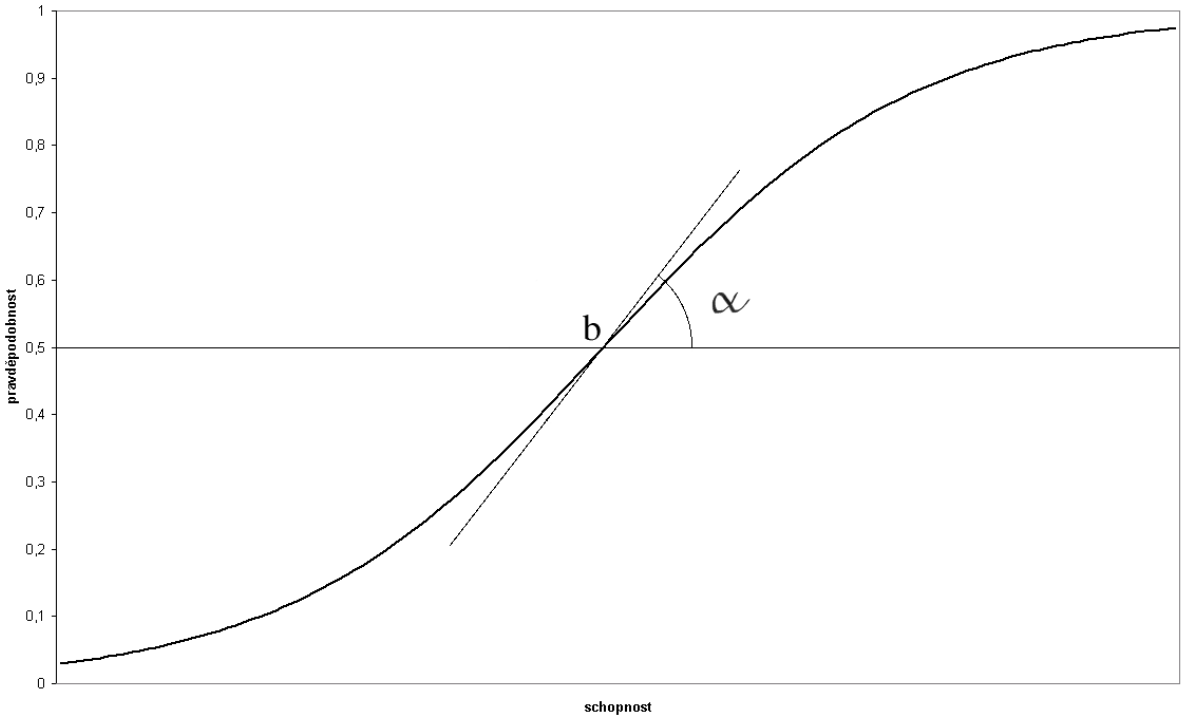


Birnbaumův model

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

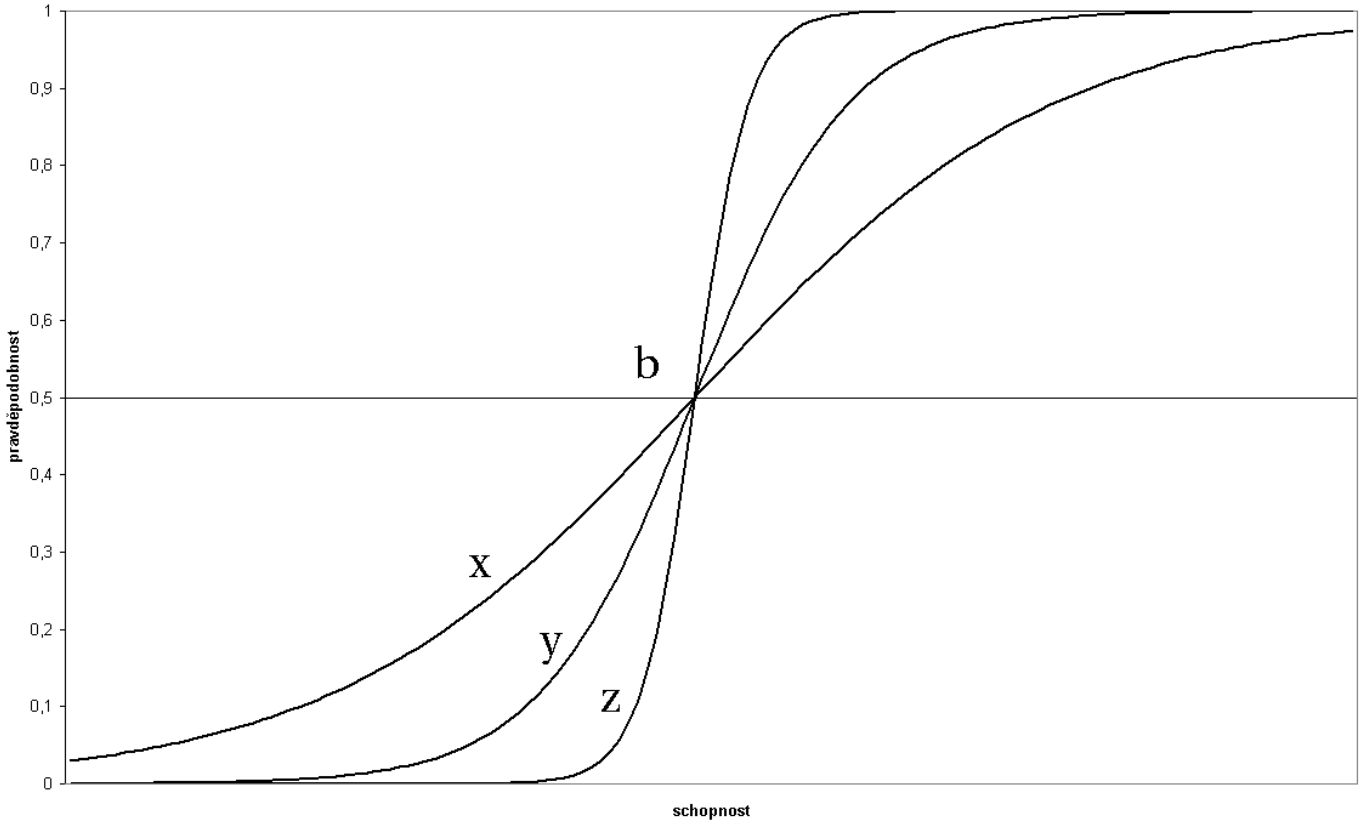
D…konstanta, která formátuje logistickou funkci, aby byla tvarově co nejvíce podobná normální ogivě, má hodnotu 1,7
a…diskriminační parametr, vyjadřuje velikost naklonění charakteristické křivky položky v bodě b
Položky, které jsou v bodě b strmější (a tedy je pro ně parametr a vyšší), mají větší rozlišovací potenciál,vhodnější pro třídění respondentů podle odlišných úrovní schopnosti θ, než položky pozvolnější
Parametr a…teoreticky definován v (-∞, +∞), v psychometrické praxi obvykle (0, 2)

Charakteristická křivka 2PL modelu

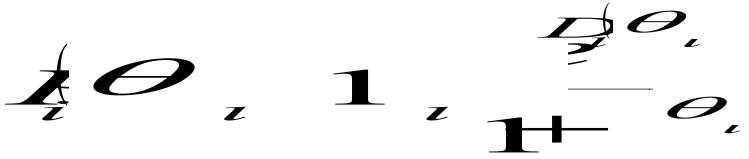


úhel α, který svírá tečna charakteristické křivky položky v bodě s přímkou proloženou úrovní 50% pravděpodobnosti

Charakteristické křivky položek se stejným parametrem obtížnosti, ale lišící se v diskriminačním parametru



Model s uhádnutelností



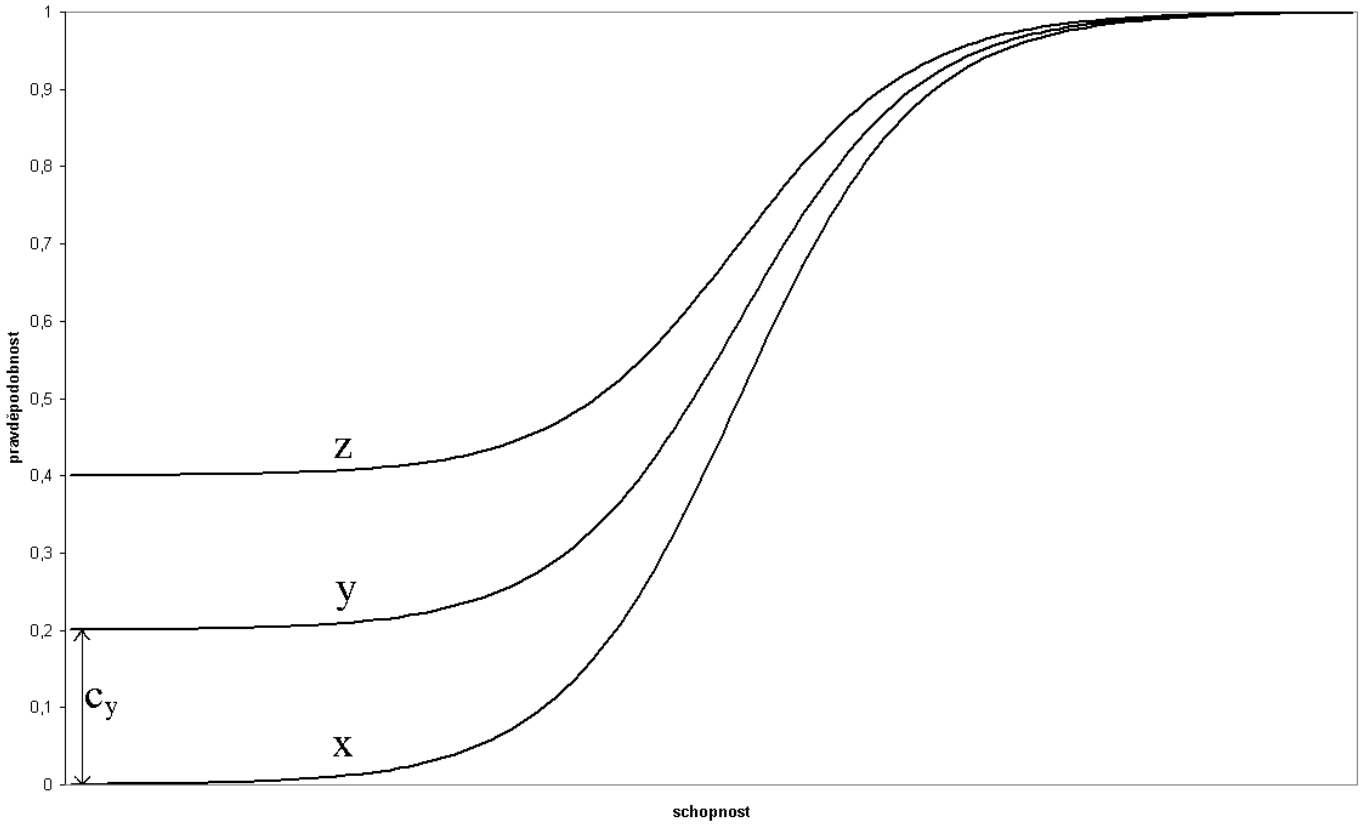
c…hodnota pravděpodobnosti, s jakou i respondent s nulovou měřenou schopností „vyřeší“ položku správně

objevuje se ve formátech položek s vícenásobnou volbou

snaha o minimalizaci parametru uhádnutelnosti

psychologické testy ve výkonové oblasti (např. testy inteligence) a pro pedagogické testování

Charakteristické křivky položek tříparametrového modelu



Specifická objektivita (=vlastnost invariance)

Vlastnost invariance položky a úrovně latentního rysu je základním kamenem IRT, a také hlavním rozdílem oproti klasické testové teorii. Znamená to, že parametry, které charakterizují položku, nezávisí na rozložení schopnosti respondentů a zároveň úroveň schopnosti θ, která charakterizuje respondenta, nezávisí na množině položek.

Důsledek:

překonání omezení CTT

výsledky všech testů založených na CTT mohou být interpretovány a srovnávány pouze v rámci populace, na níž byl test standardizován

porovnávat výsledky různých testů, které však měří stejnou schopnost, není možné

Limity představovaných modelů

- Nejjednodušší modely, pro pochopení principů IRT
- Striktně jednodimenzionální
- Dichotomní

- V současnosti několik desítek různých modelů pro různé úrovně měření (např. Samejimin model pro škálové proměnné)
- Multidimenzionální modely
- Dobrý přehled nabízí Handbook of Modern Item Response Theory (Linden, Hambleton, 1996)

Odhad parametrů

- Odhad parametrů položky
- Odhad schopnostního parametru
- Nejčastěji – společný odhad parametrů položky i probandovy schopnosti

Pravděpodobnostní funkce N respondentů odpovídajících na n položek za předpokladu lokální nezávislosti vypadá následovně

$$L(u_1, u_2, \ldots, u_n / \theta, a, b, c) = \prod_{i=1}^{N} \prod_{j=1} \; {}^{u_{ij}} Q_{ij}^{1-} \; {}^{ij}$$

Informační funkce položky

- umožňuje popsat přínos konkrétní položky
- velký význam při konstrukci testů, neboť právě velikost informace, kterou daná položka přináší, může být vhodným kritériem pro rozhodování, zda položku ponechat nebo ji vyřadit z testu

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

I(θ) je velikost informace, kterou poskytuje položka i při úrovni schopnosti θ.
P(θ) je pravděpodobnost správné odpovědi.

Informační funkce a její souvislosti s parametry položky

- Větší množství informace poskytují položky s vyšší obtížností.
- Diskriminační parametr položky podstatně ovlivňuje velikost informace, kterou daná položka poskytuje. Čím je diskriminační parametr vyšší, tím větší má položka informační hodnotu. Položky s nízkým diskriminačním potenciálem jsou v rámci testu statisticky zbytečné. Mohou však mít význam například na začátku testu jako zácvičné položky, neboť je vhodné, aby je zvládla většina respondentů a neztratila tak motivaci pro další práci s testem.
- Se zvyšující se hodnotou parametru uhádnutelnosti položky samozřejmě informační hodnota klesá, neboť i ti respondenti, kteří nedisponují danou schopností, mají jistou pravděpodobnost (dle velikosti parametru c), že na položku správně odpoví.
- Informační hodnota položky se různí dle úrovně schopností. Položka s relativně velkou obtížností má tedy velkou informační hodnotu mezi respondenty s vysokou mírou dané schopnosti, ale ve střední oblasti schopnosti nám tolik informace neposkytne.

Informační funkce testu a standardní chyba měření

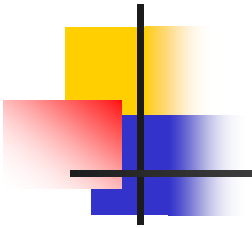
Jako charakteristiku celého testu můžeme používat informační funkci testu, což je součet informačních funkcí všech položek, které test obsahuje.

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

Z informačního přínosu testu můžeme odvodit standardní chybu měření podmíněnou danou úrovní latentního rysu. I(θ) je informace, kterou poskytuje konkrétní test pro respondenta s odhadem schopnosti θ.

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

Informační funkce testu se používá také při srovnávání a zhodnocování dvou či více testů.



Srovnání IRT s klasickou testovou teorií

Nezkreslený odhad vlastností položek

Položka v CTT je obvykle charakterizována p-hodnotou, což je podíl osob, které v rámci standardizačního souboru zodpověděly danou položku správně nebo kladně. Další popisnou statistikou položky je její rozlišovací účinnost. Je to míra, která určuje, jak se respondenti liší ve svých odpovědích na tuto položku. Obě tyto charakteristiky podstatně závisí na úrovni měřené schopnosti a jejím rozložením v populaci. V CTT je tedy **nezkreslený odhad vlastností položky závislý na reprezentativnosti souboru respondentů.**

Z definice charakteristik položek v IRT a způsobu odhadu parametrů IRT modelů vyplývá, že parametry položky jsou stanovovány nezávisle na rozložení schopnosti u konkrétního respondenta. **Nezkreslený odhad vlastností položek tedy můžeme získat z nereprezentativního vzorku.**

Reliabilita a délka testu

V CTT platí: delší test je reliabilnější než kratší test. V rámci IRT toto pravidlo neplatí, reliabilita závisí na jiných aspektech než je délka testu.

V IRT je přesnost měření zjišťována pomocí informační funkce testu, a ta je sumou informačních funkcí použitých položek. Proto závisí na přispění jednotlivých položek, a ne pouze na jejich počtu (tedy délce testu).

Standardní chyba měření

Standardní chyba měření v CTT je vlastností testu a nezávisí na konkrétní úrovni měřené schopnosti, tudíž je konstantní pro všechny dosažené skóry. Zároveň je však standardní chyba specifická pro danou populaci, na níž byl test standardizován.

Standardní chyba v rámci IRT je definována jako odmocnina převrácené hodnoty celkového informačního přínosu testu, tedy je variabilní a závisí na úrovni měřeného latentního rysu. Také umožňuje zobecnění na různé populace.

Ze znalosti standardní chyby měření můžeme sestrojit intervaly spolehlivosti. Omezení CTT způsobuje, že všichni probandi mají stejný rozsah intervalu spolehlivosti (neboť i standardní chyba je shodná), v IRT je možné sestavit intervaly spolehlivosti pro odhad schopnosti θ pro konkrétního probanda.

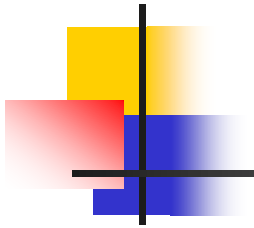
Porovnávání testových forem

V rámci CTT je to velký problém, neboť porovnávat testová skóre je optimální pouze u paralelních forem testu. Schopnost respondenta má smysl brát v úvahu pouze v kontextu daného testu, proto jakékoliv porovnávání odlišných forem testů nepřináší žádnou adekvátní informaci.

IRT však tvrdí, že velikost informace, kterou nám test přináší, závisí na úrovni latentního rysu respondenta. V rámci IRT je tedy vždy lepší odhad úrovně schopnosti získán použitím neparalelních forem testu. Nejpřesnější odhad schopnosti probanda obdržíme na základě adaptivního testování.

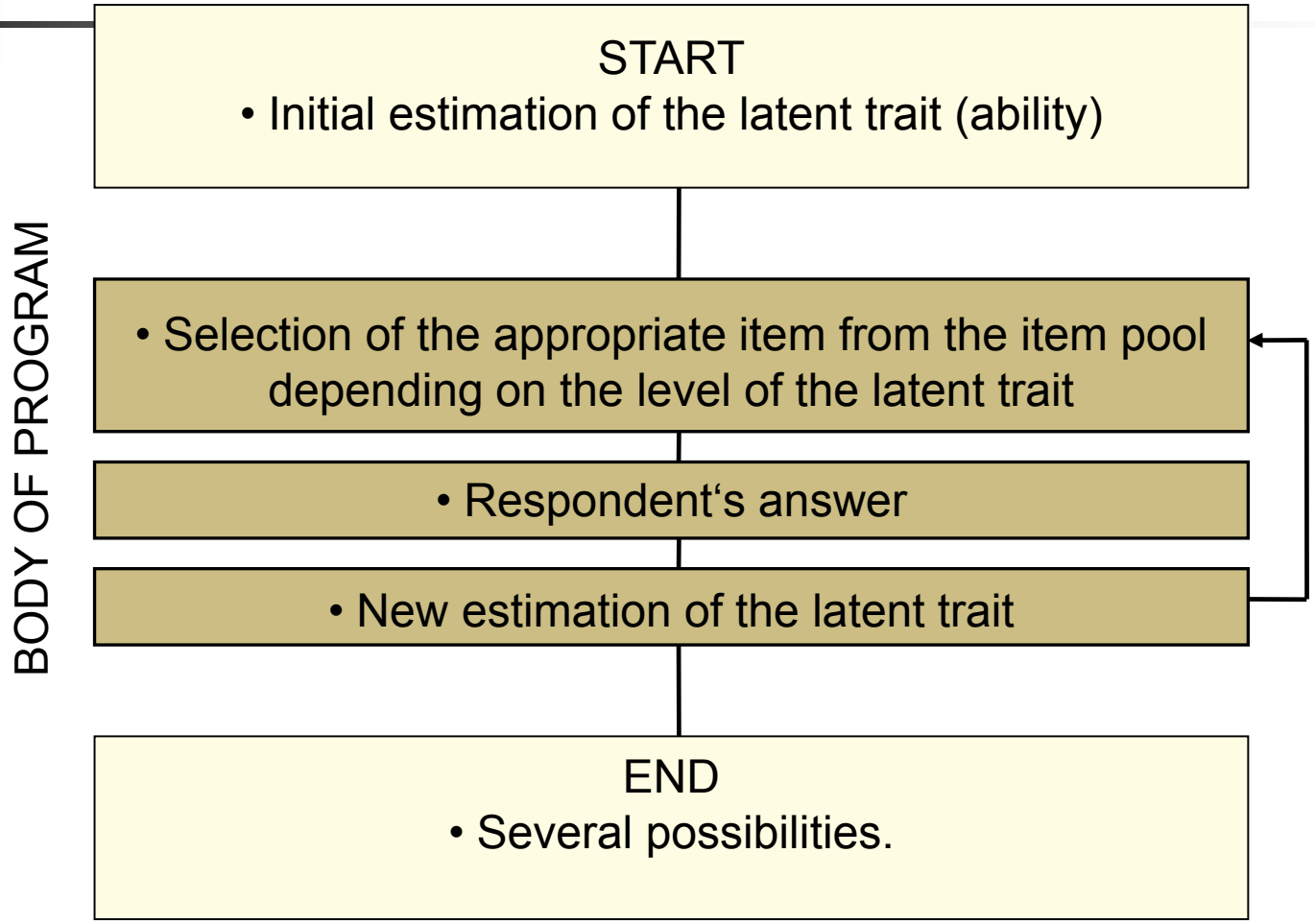
Formát položek v rámci testu

V CTT vedou smíšené formáty položek k nevyváženému dopadu na celkový testový skór. Tuto nepříznivou situaci překonává IRT, neboť zde vliv na celkový skór nemá formát položek, ale jejich parametry a tudíž je možné s použitím smíšených formátů položek vytěžit optimální testový skór.



Počítačové adaptivní testování

Software pro realizaci CAT



Průběh algoritmu CAT

I. Section – Selection of suitable model (good model-data fit)

the program works with 3 models (1PL, 2PL, 3PL) currently it is still possible to use other models, especially designed for personality testing

II. Section – Selection of procedure for estimation of ability

marginal maximum likelihood estimation

marginal maximum a posteriori estimation (Bayes estimation)

III. Section – The begining of test

several (three) randomly selected items are administred to respondent, the items should have lower parameter of difficulty (estimation of ability - positive motivation of respondent)

the item with the particular parameter of difficulty is chosen

(if the level of ability has been approximately known)

Průběh algoritmu CAT

IV. Section – Selection of the Item to Administer

- the item with the highest information function's value
- random selection from items, the information contribution of which is higher then the a priori set value
- random selection from several items with the highest information
- function's values together with all items with information function's
- values higher than the a priori set value

V. Section – Format of Items

- a respondent chooses an answer from offered choices (different answers)
- open ended statement – the answer is compared with the list of possible answers

Průběh algoritmu CAT

VI. Section – Item pool

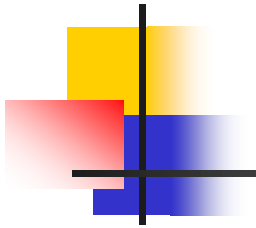
- All items are taken from 1 item pool (unidimensional intelligence test)
- There are more item pools, every pool measures different trait. The item to administer is chosen from the pool with the highest standard error within the estimation of the latent trait. As a result we receive more estimates of different traits (Eysenck test – one item pool for neuroticism, and other pool for extroversion).
- Stratified item pool – items in the pool are distributed to groups which are regularly alternated during the administration of the pool (if there is not a suitable item in the current group, it is skipped). The result is one estimation of one skill (test measures mathematics skill of pupilsand the groups are addition, subtraction, multiplication, division).

Průběh algoritmu CAT

VII. Sections – End of program

Program can be designed to stop when:

- the maximum (a priori set number of items) test length is reached
- the pool has run out of the suitable items (in the case of small item
- pool)
- the estimation of a latent trait exceeds the pass-fail criterion
- the level of the trait is estimated with the sufficient precision
  - the standard error is lower than the set value
  - the difference of standard errors between two last items is sufficiently small



Teorie vědomostního prostoru

Teorie vědomostního prostoru - KST

Teorie vědomostního prostoru umožňuje uspořádat vědomosti jedince do přehledné struktury, z které vyplývá, která vědomost je nutná pro vybudování vědomostí dalších. Zároveň nabízí nástroj pro detekování toho, které vědomosti jsou na sobě zcela nezávislé a které spolu naopak určitým způsobem souvisí. Základy KST formulovali **Doignon a Falmagne v roce 1985.**

Vědomostní doména Q je konečná množina problémů mapující nějakou vědomostní oblast.

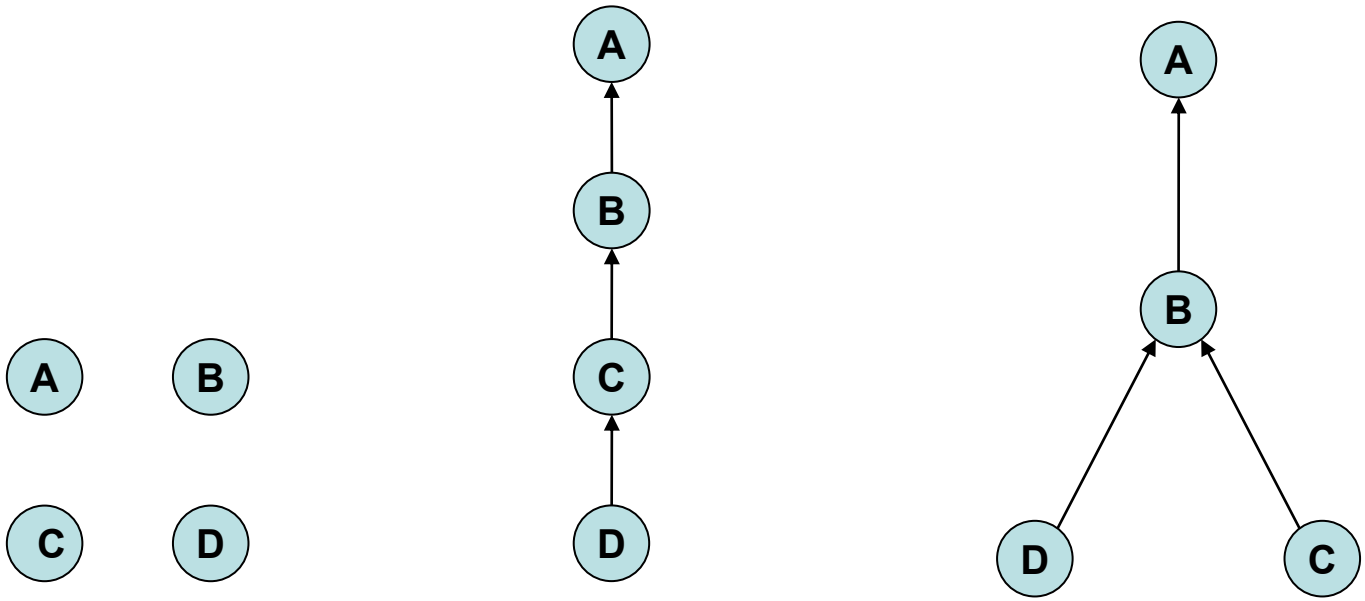
Vědomostní stav K je podmnožina všech problémů z vědomostní domény, které je určitý jedinec schopen úspěšně vyřešit.

Vědomostní struktura *K* je množina všech vědomostních stavů, které pro určitou doménu Q mohou být pozorovány v nějaké populaci. Vědomostním prostorem se struktura stává pokud je uzavřena na sjednocení i na průnik.

Prerekvizitní relace

Prerekvizitní relace je binární relace, která udává, že ze správného vyřešení položky A můžeme usuzovat na správné vyřešení položky B. Jinými slovy: pokud respondent správně zodpoví (resp. zodpoví v indikovaném směru v případě nevýkonových testů) položku A a z toho můžeme usuzovat na správné zodpovězení položky B, potom dvojice (B, A) je v prerekvizitní relaci.

Prerekvizitní relace



Položky na prvním obrázku mezi sebou nemají žádný vztah, v druhém případě se jedná o lineární uspořádání vůči jedné vlastnosti (známé z Guttmanova škálování). Třetí příklad znázorňuje typickou sekvenci ve vědomostní struktuře (položky C a D mezi sebou nemají žádný vztah, obě jsou však prerekvizitami pro položku B).

Způsoby tvorby vědomostního prostoru

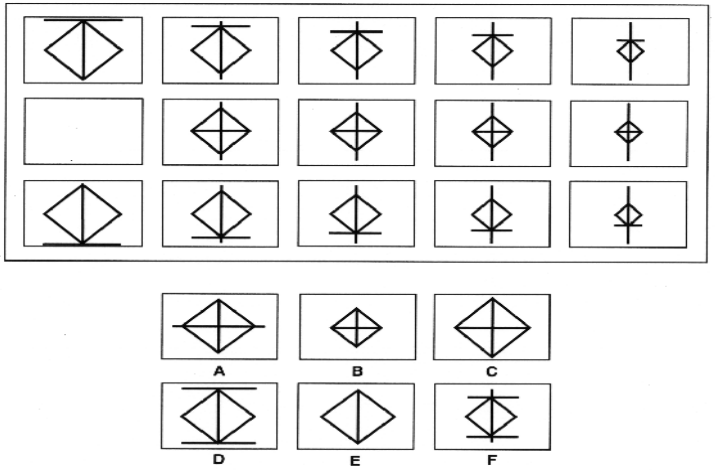
Hierarchie domény Q – doména je již takto strukturována, např. didaktické koncepce

Analýza dat respondentů

Expertní posouzení - několik expertů posuzuje všechny možné neprázdné podmnožiny vybrané z množiny Q vzhledem ke všem dalším možným neprázdným podmnožinám z množiny Q**.**

Pravdivost tvrzení: „Pokud respondent selže na všech položkách z podmnožiny A, bude to mít za následek selhání na všech položkách z podmnožiny B.“

Empirická část
test BOMAT



Inteligenční test, 40 položek

autoři Hossiep, Turck a Hasella, 2002

Analýza v kontextu teorie odpovědi na položku

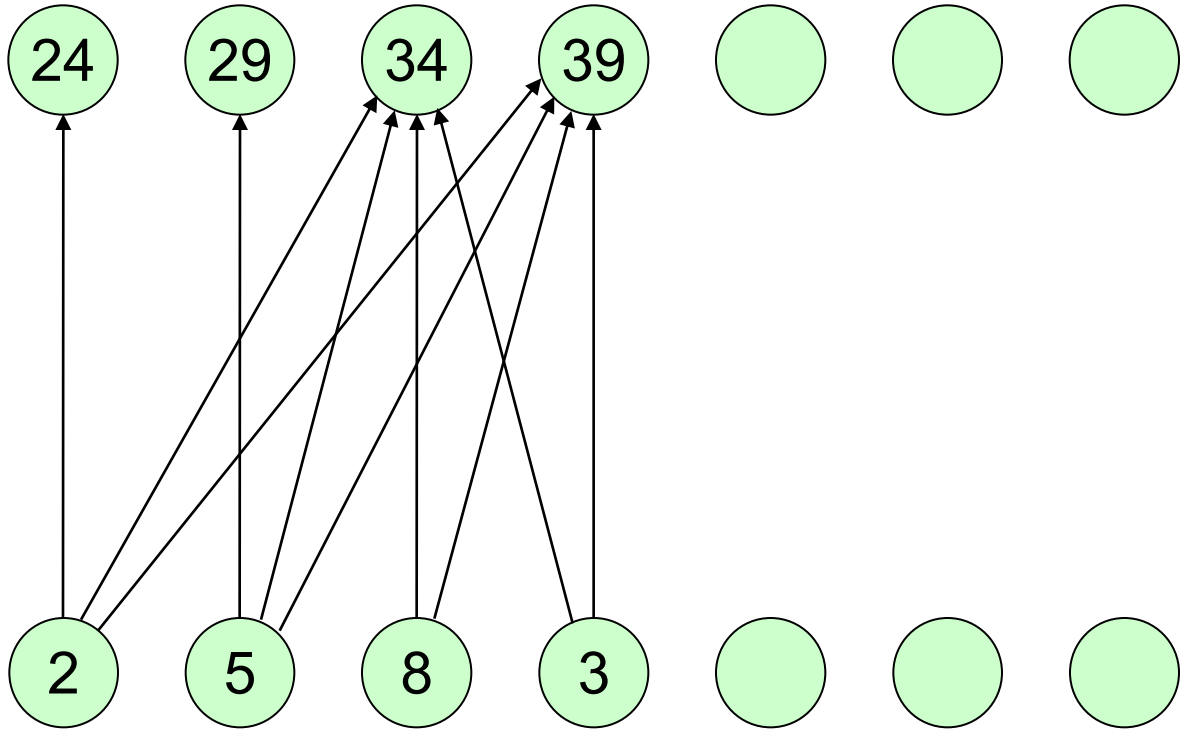
Seřazení položek testu podle parametru obtížnosti (3 PL model)

Analýza v kontextu teorie vědomostního prostoru

Dvoupatrová vědomostní struktura + několik separovaných položek

Vždy několik položek s nižší obtížností tvoří prerekvizity pro položku s vyšší obtížností na základě podobného principu nutného pro vyřešení úlohy

Výsek vědomostního prostoru nad testem BOMAT



Empirická část – NEO FFI

Vědomostní prostory pro každou dimenzi testu zvlášť (detekce vrcholových položek)

Společný vědomostní prostor pro dvě dimenze

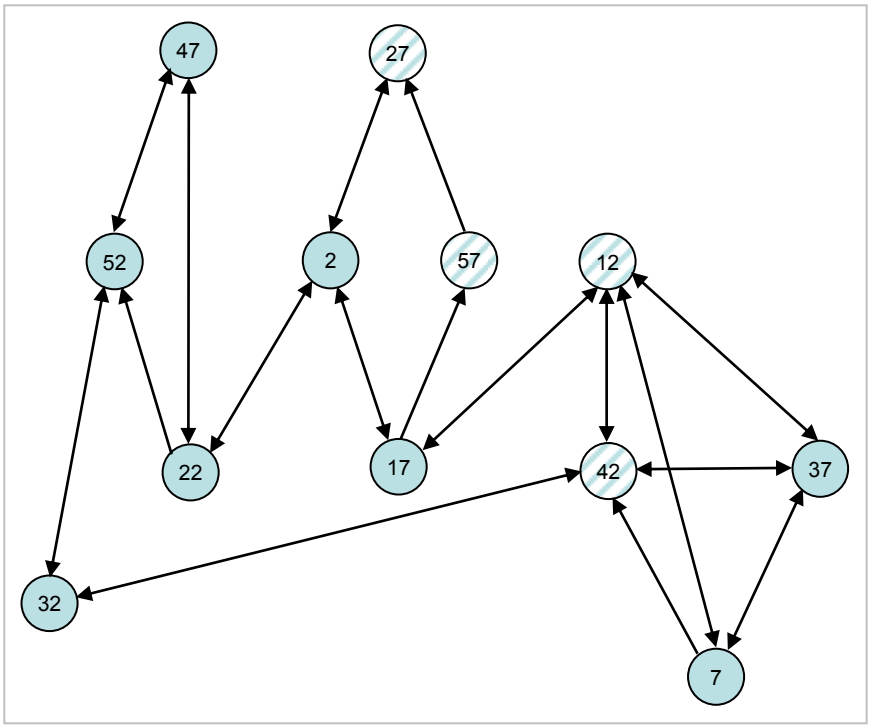
Různé vědomostní prostory ve stejné dimenzi dle sociodemografických charakteristik (muži x ženy, dospělí x adolescenti)

Různé vědomostní prostory ve stejné dimenzi ze souboru náhodně rozděleného na dvě poloviny (vysoká shoda svědčící pro reliabilitu metody)

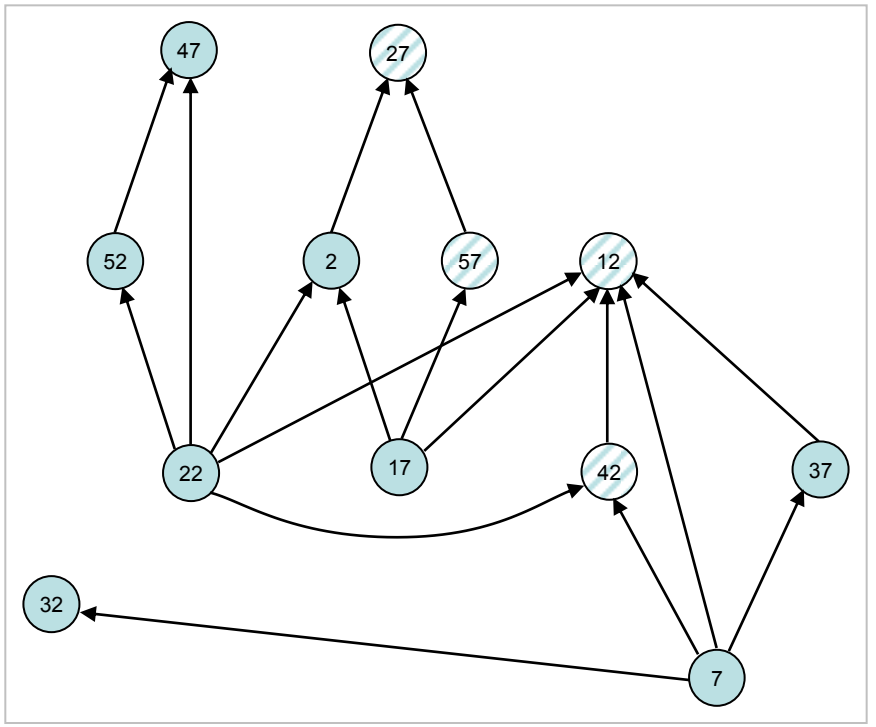
Různé vědomostní prostory ve stejné dimenzi získané odlišnými způsoby (porovnání prostorů vygenerovaných z dat respondentů a prostorů vzniklých expertním posouzením)

Dva různé způsoby tvorby vědomostního prostoru dimenze extroverze

Míra shody vědomostních prostorů je 70%, pomineme-li obousměrnost 85%.

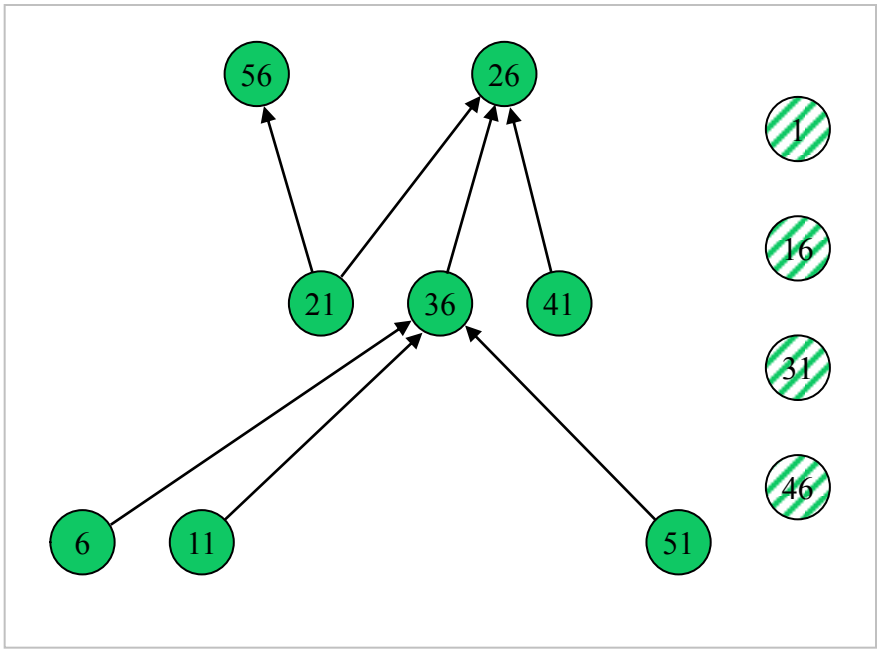


Expertní posouzení

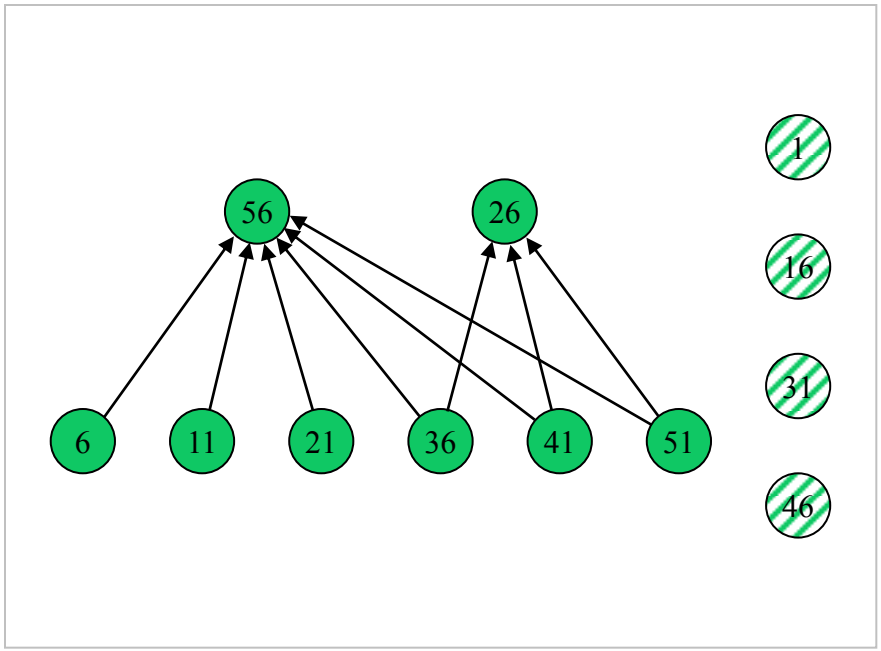


Analýza dat respondentů

Rozdíly v dimenzi neuroticismus podle pohlaví



Vědomostní prostor mužů



Vědomostní prostor žen

Položka 56 „Někdy se stydím, že bych se nejraději neviděl.“

Závěr

Teorie vědomostního prostoru přináší nové možnosti do psychometriky, konstrukce a interpretace testů.

Tři hlavní přínosy:

Oblast adaptivních testů – KST slouží jako algoritmus z něhož může vycházet počítačové adaptivní testování, zefektivňuje dotazovací proceduru výběrem vhodných položek.

Detekce odlišných, nepravděpodobných odpověďových vzorců – východisko pro lži skóre, detekci opisování, ale i kreativní nestandardní řešení…

Oblast sémantická – umožňuje postihnout, jak různí lidé chápou smysl položek či jejich skupin.