SHINYITEMANALYSIS: ANALÝZA PŘIJÍMACÍCH A JINÝCH ZNALOSTNÍCH ČI PSYCHOLOGICKÝCH TESTŮ

PATRÍCIA MARTINKOVÁ[1], ADÉLA DRABINOVÁ[1,2], JAKUB HOUDEK[1,3]
[1] Oddělení statistického modelování, Ústav informatiky AV ČR
[2] Katedra pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulta Univerzity Karlovy
[3] Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze

Abstrakt: *Tento článek představuje aplikaci ShinyItemAnalysis pro psychometrickou analýzu testů a jejich položek. ShinyItemAnalysis nabízí psychometrické modely v rámci grafického rozhraní pro volně šiřitelné statistické prostředí R a zpřístupňuje tak jeho funkcionalitu širší veřejnosti. Aplikace pokrývá širokou škálu psychometrických metod, od tradiční položkové analýzy až po složitější latentní modely, nabízí cvičné datové soubory, uvádí rovnice modelů, odhady parametrů a jejich interpretaci, jakož i vybraný zdrojový kód, a je tak vhodným nástrojem pro výuku psychometrických konceptů a jejich implementace v R. Aplikace však také nabízí možnost analýzy vlastních dat a generování reportů a aspiruje tak na to být jednoduchým nástrojem pro rutinní analýzu testů a jejich položek. Závěr článku ukazuje, že ShinyItemAnalysis je dostupným, flexibilním a uživatelsky příjemným nástrojem, který může pomoci tomu, aby se statistická analýza přijímacích i jiných znalostních či psychologických testů stala v praxi samozřejmou záležitostí.*

Klíčová slova: *analýza testů, položková analýza, teorie odpovědi na položku, odlišné fungování položek, R, shiny*

[1] Oddělení statistického modelování, Ústav informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8
[2] Katedra pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulta Univerzity Karlovy, Sokolovská 83, 186 75 Praha 8
[3] Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze, nám. W. Churchilla 4, 130 67 Praha 3.

Došlo do redakce: 14. 3. 2017.
Schváleno k publikaci: 23. 7. 2017.

Podstatným krokem vývoje znalostních a psychologických testů je získání důkazů o jejich spolehlivosti, validitě a dobrém fungování jednotlivých položek (AERA, APA, NCME, 2014). V pedagogické praxi jsou dobré psychometrické vlastnosti esenciální pro selektivní testování: funkční přijímací testy jsou základem dobrého výběru uchazečů (Salvatori, 2001). Jsou ale důležité také pro formativní testování v rámci běžné výuky: nepřiměřené, nefunkční nebo neférové testy mohou zeslabit studentův zájem o daný obor, nebo jej odradit od studia (Legewie & DiPrete, 2014).

Rutinní pretestování znalostních testů je běžnou praxí v západní Evropě či v USA. Např. přijímací testy jsou připravovány a ověřovány v mnohastupňovém procesu, položky procházejí vícenásobnou revizí, jak obsahovou, tak pomocí statistických metod (Zwick, 2006). Ve výuce jsou zde také ve větší míře využívané validizované znalostní škály (tzv concept inventories, McFarland et al., 2017; SABER, n.d.). V České republice nejsou validizační studie tak rozšířené, fakulty však postupně věnují svým přijímacím zkouškám i běžným formativním testům větší pozornost (Anděl & Zvára, 2005; Höschl & Kožený, 1997; Kožený, Tišanská, & Höschl, 2001; Rubešová, 2009; Štuka, Martinková, Zvára, & Zvárová, 2012; Zvára & Anděl, 2001).

Psychometrickou analýzu dat lze provést v obecných statistických programech, z nichž mnohé jsou komerční, jako např. SAS (SAS Institute, 2013), SPSS (IBM Corp, 2015), STATA (StataCorp, 2015), aj. K dispozici je také spektrum komerčních programů zaměřených přímo na analýzu položek a psychometrické modely, např. Winsteps (Linacre, 2005), IRTPRO (Cai, Thissen, & du Toit, 2011) nebo ConQuest (Wu, Adams, & Wilson, 1998), viz také van der Linden (2017).

V této práci si představíme analýzu znalostních testů pomocí volně šiřitelného statistického prostředí R (R Development Core Team, 2016, viz také Revelle, 2009; Rusch, Mair, & Hatzinger, 2013) a webové aplikace ShinyItemAnalysis (Martinková, Drabinová, Leder, & Houdek, 2017). ShinyItemAnalysis postupně pokrývá stěžejní psychometrické nástroje a modely, od tradiční analýzy položek přes regresní analýzy až po komplexnější latentní IRT (Item Response Theory) modely. Cílem tohoto příspěvku je představit funkcionality aplikace a její ovládání čtenářům, aby ji mohli snadno využívat ve své praxi. Metody jsou demonstrovány na několika cvičných datových souborech, aplikace také nabízí možnost nahrát a analyzovat vlastní data. Základ zdrojového kódu představovaných funkcí byl připraven a využit v rámci výuky kurzu Item Response Theory Models of Tests pro doktorské studenty oboru Measurement and Statistics na College of Education, University of Washington. Nový balík ShinyItemAnalysis a jeho online verze – aplikace ShinyItemAnalysis – byla také využita v rámci workshopů pro pedagogy lékařských a jiných fakult (Štuka, Vejražka, Martinková, Komenda, & Štěpánek, 2016). Jeví se jako vhodný interaktivní nástroj pro výuku metod analýzy testů. Aspiruje však také na to být jednoduchou volně šiřitelnou aplikací pro rutinní analýzu přijímacích a dalších testů a pro běžné použití pedagogy, kteří své testy připravují.

Spuštění ShinyItemAnalysis

ShinyItemAnalysis (Martinková, Drabinová, Leder, et al., 2017) byla vytvořena jako jedna z knihoven volně šiřitelného programu R (R Development Core Team, 2016). Lze ji tedy instalovat lokálně běžným způsobem v rámci R. Její velkou výhodou je možnost využít webového rozhraní shiny[4] (Chang, Cheng, Allaire, Xie, & McPherson, 2017) a analyzovat data bez nutnosti lokálního spouštění statistického programu R. Je tak přístupná i uživatelům, kteří nechtějí R instalovat a dále využívat.

Webové rozhraní

Online verze aplikace[5] nabízí všechny funkcionality verze instalovatelné lokálně, včetně možnosti načtení svých vlastních dat a možnosti generování reportů ve formátu PDF nebo HTML a exportu grafů ve formátu JPG. Omezením může být rychlost připojení, jakož i rychlost serveru. Ta se může projevit obzvláště v případě, kdy aplikaci najednou využívá větší množství uživatelů, jednotlivé úkoly jsou totiž prováděné postupně v takovém pořadí, jak byly uživateli vyžádány. Časová náročnost pak může být delší než v případě lokálního spuštění.

Lokální spuštění v rámci R

Pro lokální spuštění je potřeba mít nainstalováno prostředí R, jehož nejnovější verze jsou dostupné zdarma online.[6] ShinyItemAnalysis lze stáhnout jako běžný balík a nainstalovat v rámci R pomocí následujících příkazů:

```
install.packages("ShinyItemAnalysis")
library(ShinyItemAnalysis)
```

Aplikace si může vyžádat také instalaci dalších balíků, které využívá, nebo jiných balíků, na kterých tyto balíky závisí.[7]

Samotná aplikace se pak spustí pomocí příkazu

```
startShinyItemAnalysis()
```

Načtení dat

Po spuštění ShinyItemAnalysis si uživatel může vyzkoušet funkcionalitu aplikace na čtyřech cvičných datových souborech. Jejich výběr lze provést v záložce *Data*. Prvním z cvičných souborů je 20 položkový soubor „GMAT“ (Martinková, Drabinová, Liaw, et al., 2017) obsahující data simulovaná na základě reálných parametrů z Graduate Admission Management Test (Kingston, Leary, & Wightman, 1985). První dvě položky v tomto souboru byly vygenerovány tak, aby fungovaly odlišně pro muže a ženy (odlišným

[4] Viz https://shiny.rstudio.com/
[5] Viz https://shiny.cs.cas.cz/ShinyItemAnalysis/
[6] Viz https://cran.r-project.org/
[7] Jejich instalace se provede pomocí příkazu `install.packages("Jméno chybějícího balíku")`

fungováním položek se zabývá sekce DIF a DDF analýza, viz dále), přičemž rozložení celkových skóre je u obou skupin naprosto stejné. Dataset obsahuje také generované spojité kritérium pro analýzu predikční validity (analýzou predikční validity se zabývá sekce *Validita*). Dalším souborem je „GMAT 2“ (Drabinová, Martinková, & Zvára, 2017), který se od „GMAT“ liší především v tom, že znalosti žen a mužů se mírně liší. Třetím souborem je „Medical 20 DIF“ (Drabinová et al., 2017). Jedná se o 20 vybraných položek z reálného přijímacího testu na lékařskou fakultu, kde první položka funguje odlišně pro muže a ženy. Posledním souborem je „Medical 100“ (Štuka et al., 2016), což je 100položkový reálný přijímací test na lékařskou fakultu. Pro přijaté studenty dataset obsahuje také vektor pro analýzu predikční validity, v tomto případě indikátor, zda student řádně studuje i rok a půl po přijetí. Všechny zmíněné datové soubory jsou tvořeny výhradně položkami s více odpověďmi (multiple-choice), poslední dva zmíněné připouštějí více správných odpovědí (multiple true-false).

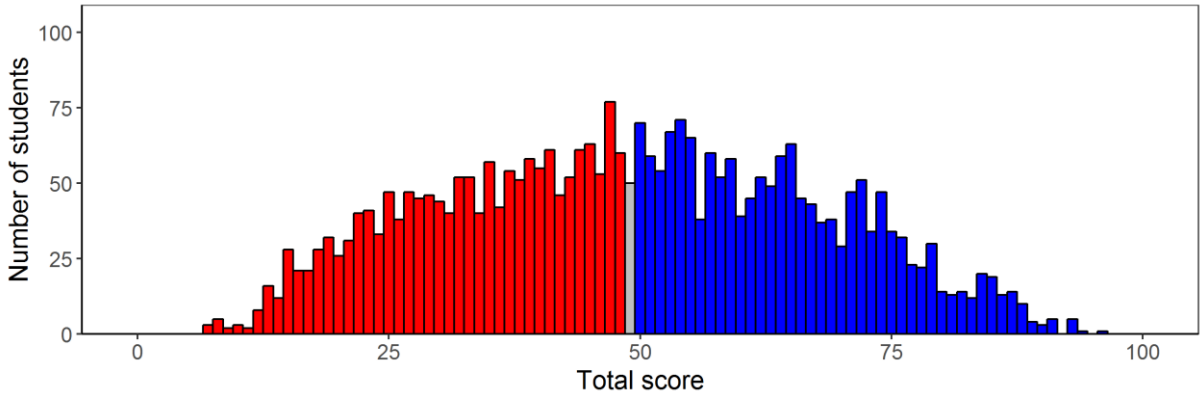
Aplikace na listu *Data* nabízí také možnost nahrát svá vlastní data ve formátu CSV. Je možné pracovat jak s neskórovanými daty pro otázky s více odpověďmi (tj. ve formátu ABCD, kde je možné i více správných odpovědí), tak s daty skórovanými (např. ve formátu 0–1, kde 0 znamená špatnou odpověď a 1 správnou). K datovému souboru je vždy nutné dodat klíč správných odpovědí v samotném CSV souboru; u skórovaných dat (např. na Likertově škále) doporučujeme uvést vektor maximálních možných skóre. Pro analýzu férovosti a odlišného fungování položek je také potřeba vložit soubor poskytující informaci o příslušnosti ke skupině (např. pohlaví, národnost, atd.) s kódováním 0–1, kde 0 značí referenční skupinu, většinou majorita, a 1 skupinu fokální. Pro analýzu predikční validity je dále potřeba vložit soubor poskytující hodnotu kritéria (pro přijímací zkoušky to může být např. průměrný prospěch na VŠ nebo dokončení VŠ studia).

V dalších sekcích budeme funkcionalitu aplikace a jejích jednotlivých částí (záložek) demonstrovat na datech „Medical 100“ a na datech „GMAT 2“.

Souhrnné statistiky

V záložce *Summary* lze najít popisné statistiky celkového skóre (total scores). V souhrnné tabulce jsou např. k nalezení průměr, minimum, maximum, medián a směrodatná odchylka.

Histogram celkových skóre studentů poskytuje bližší informace o jejich rozdělení, viz Obrázek 1.



Obrázek 1: Histogram celkových skóre v datovém souboru "Medical 100", barevně jsou odlišení studenti nad a pod zvolenou hranicí úspěšnosti (tzv. cut-score).

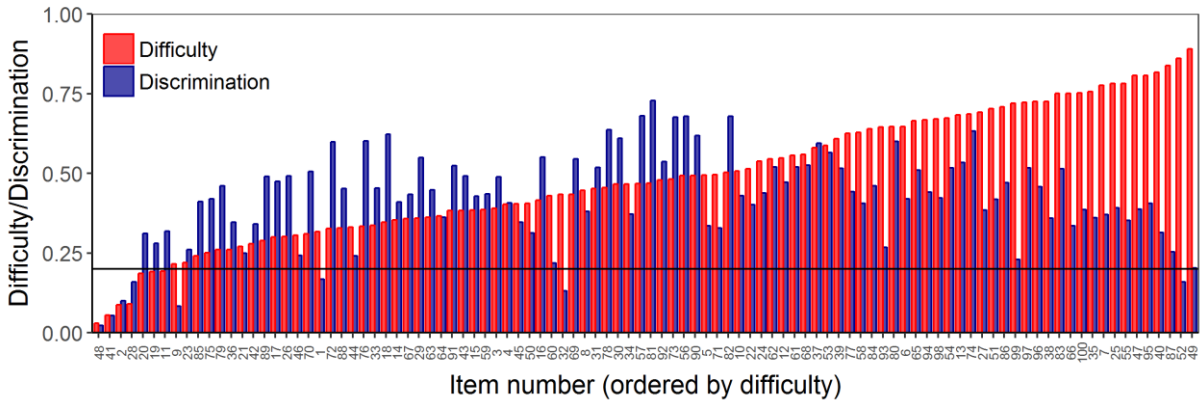
Pro hodnocení studentů může být užitečná tabulka standardizovaných hodnot, kterou lze nalézt v podzáložce *Standard scores*. Zde lze pro dané celkové skóre zjistit, v jakém percentilu se studenti s daným skóre umístili a jaká byla jejich míra úspěchu (tzv. success-rate, tedy poměr získaného skóre k maximálnímu možnému skóre). Najdeme zde také hodnoty tzv. z-skóre a T-skóre (ČŠI, 2015).

Validita

V záložce *Validity* lze prozkoumat korelační strukturu položek pomocí grafického znázornění korelační matice. Dále je nabízen tzv. scree plot, který zobrazuje vlastní čísla varianční matice a může poskytnout představu o dimenzionalitě dat – tedy zda data měří jeden latentní znak, či více. Podzáložka *Predictive validity*, v případě, že data obsahují kritérium (např. úspěšnost v dalším studiu), nabízí analýzu schopnosti testu a jeho položek předpovědět toto kritérium. Pomocí Spearmanova testu lze ověřit, zda celkový skór a kritérium korelují, a tedy zda test měří, co je jeho záměrem.

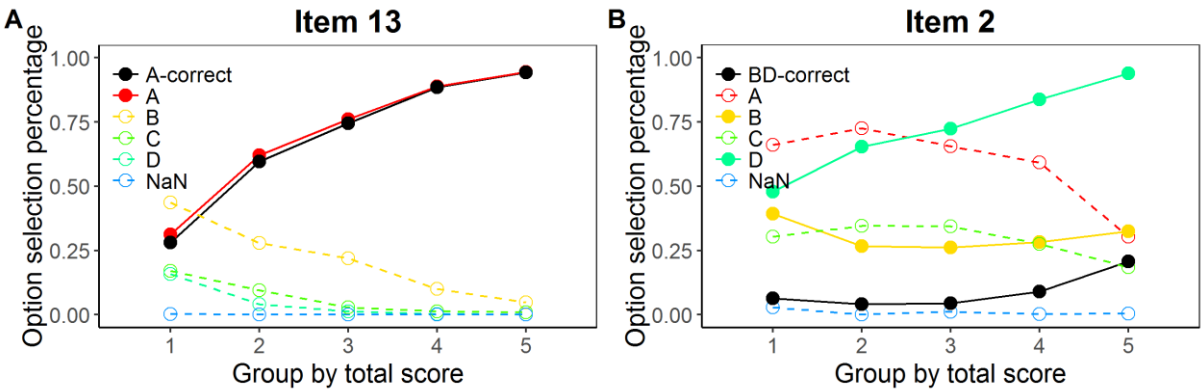
Tradiční analýza položek

Záložka *Item analysis* nabízí analýzy položek v rámci klasické testové teorie. V tabulce položkové analýzy je vyčíslena obtížnost (difficulty) počítaná jako procento správných odpovědí. Dále jsou k dispozici odhady citlivosti (discrimination) dle indexu ULI (z angl. Upper-Lower Index, počítající rozdíl mezi procentem správných odpovědí u nejsilnější a nejslabší třetiny respondentů), indexu RIT (korelační koeficient mezi položkou a celkovým skóre) a indexu RIR (korelační koeficient mezi položkou a celkovým skóre bez položky), podrobněji např. Štuka, Martinková, Vejražka, Trnka, & Komenda (2013). Vyčísleno je také Cronbachovo alfa (Cronbach, 1951) pro celý test a Cronbachovo alfa bez jednotlivé položky. Obtížnost a citlivost pomocí indexu ULI jsou vyobrazeny také graficky (viz Obrázek 2).



Obrázek 2: Grafické vyobrazení obtížnosti a citlivosti položek pro datový soubor "Medical 100", položky jsou řazené dle vzrůstající obtížnosti. Položky s příliš velkou nebo příliš malou obtížností obvykle rozlišují hůř. Položky s citlivostí menší než 0,2 (modrý sloupec pod vodorovnou čarou, zde např. položky 1, 2, 32,…) jsou vnímané jako podezřelé a hodné dalšího prozkoumání (Byčkovský & Zvára, 2007). Pro více možností k tomuto obrázku, viz Martinková et al (2017).

Podrobnější pohled na jednotlivé položky nabízí analýza distraktorů v podzáložce *Distractors.* Zde jsou vyobrazena procenta volby jednotlivých odpovědí (nebo jejich kombinací, v případě, že je více možností správných) dle úrovně celkového skóre respondentů. Vhodnost položky a jednotlivých nabízených odpovědí lze pak posoudit pomocí vertikálního umístění a sklonu příslušné křivky (Štuka et al., 2013). Optimální položka dobře rozlišuje mezi slabšími a silnějšími studenty (Obrázek 3A). U nevhodné položky graf pomůže identifikovat nevhodné distraktory či jiné příčiny, proč položka dobře nefunguje (Obrázek 3B).



Obrázek 3: Analýza distraktorů pro položky č. 13 a č. 2 pro datový soubor „Medical 100": vyobrazení procenta studentů s danou úrovní celkového skóre (tj. 1 = spodních 20 %, 5 = horních 20 %), kteří zvolili jednotlivé nabízené odpovědi (příklad s více možnými správnými odpověďmi, černě vyobrazena kombinace správných odpovědí). A. Správná odpověď (zobrazena plnou čarou) dobře rozlišuje mezi slabšími a silnějšími studenty. B. Správná odpověď (kombinace BD – zobrazena plnou černou čarou) je volena velmi zřídka, položka je velmi obtížná a kvůli tomu špatně diskriminuje. Jak je patrné, problematická je odpověď B (zobrazena žlutou plnou čarou), která je volena v nejvyšší míře nejslabší skupinou.

Pro analýzu distraktorů lze zvolit, do kolika skupin mají být studenti dle celkového skóre zařazeni. Pokud bychom uvažovali skupiny sdružující vždy studenty s daným jednotlivým celkovým skóre, můžeme se snažit procento správných odpovědí modelovat hladkou křivkou. Výše uvedený Obrázek 3 je tak předstupněm regresních modelů, o nichž pojednává následující sekce.

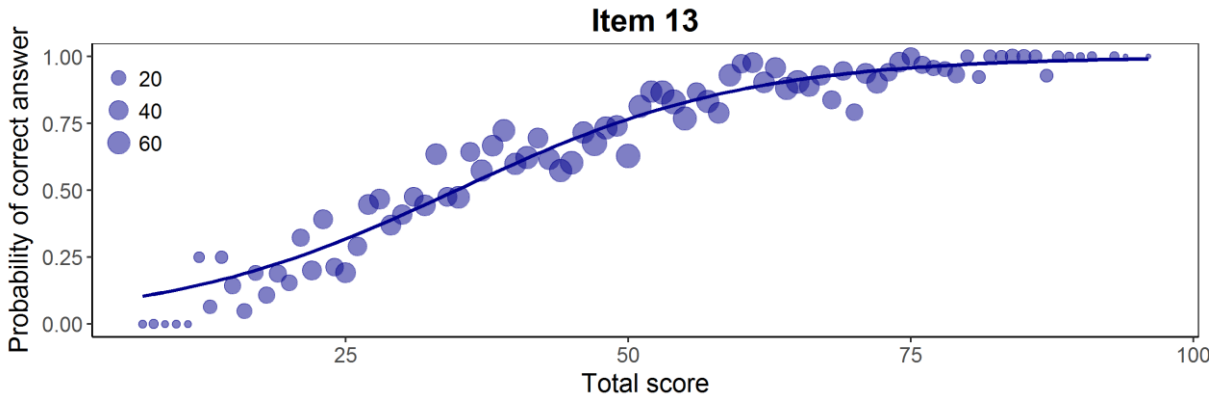
Regresní modely

Regresní analýza nám pomáhá prozkoumat položky a jejich parametry podrobněji. Pravděpodobnost správné odpovědi na položku *j* pro skórovaná data (tj. 1 je správná odpověď, 0 je nesprávná) v závislosti na celkovém skóre X_i lze modelovat pomocí logistické regrese:

$$P(Y_{ij} = 1 | X_i) = \frac{1}{1 + e^{-(b_{0j} + b_{1j} X_i)}} \qquad (1)$$

kde parametr b_0 je absolutní člen a parametr b_1 pak udává efekt celkového skóre X_i.

Místo lomené čáry pro správnou odpověď vyobrazené pomocí analýzy distraktorů je body proložena hladká křivka popsatelná pouhými dvěma parametry (Obrázek 4).



Obrázek 4: Logistická regrese pro položku č. 13 z datového souboru „Medical 100“ znázorňuje pravděpodobnost správné odpovědi v závislosti na celkovém skóre. Body značí empirické hodnoty pravděpodobností a jejich velikost je určena počtem respondentů (20, 40 a 60).

Tento model (1), nabízený v podzáložce *Logistic*, může být upravován či rozšiřován. Postupně jsou nabízeny následující regresní modely:

- Podzáložka *Logistic Z* poukazuje na skutečnost, že využijeme-li místo celkových skóre X_i jejich standardizovanou verzi Z_i (tzv. z-skóre, které mají nulovou střední hodnotu a jedničkový rozptyl), tvar křivky se nezmění, pouze se změní odhady parametrů b_0 a b_1.

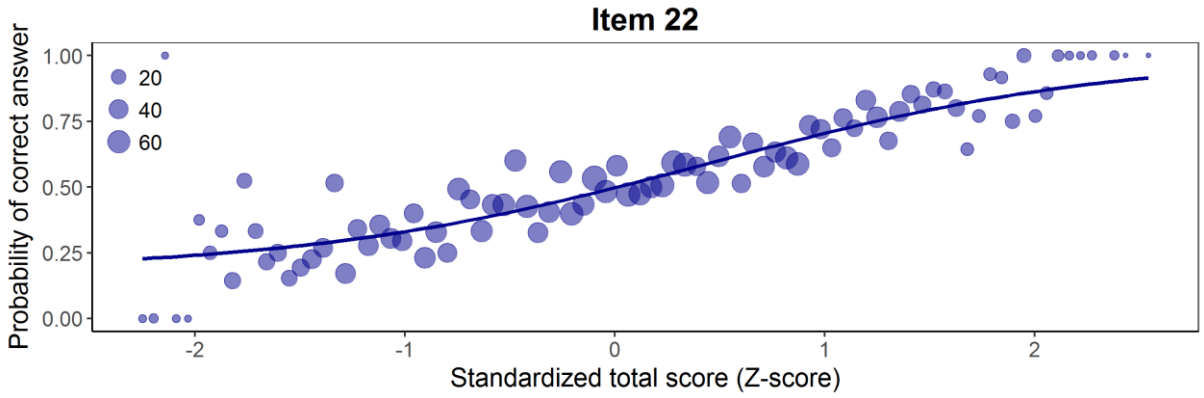
- V podzáložce *Logistic IRT Z* je nabízen logistický model s parametrizací odpovídající tzv. IRT modelům: Místo parametrů b_0 a b_1 jsou modelovány obtížnost b a rozlišovací schopnost (diskriminace) a, přičemž existuje jednoznačný vztah mezi původními a novými parametry

$$P(Y_{ij} = 1|X_i) = \frac{1}{1 + e^{-a_j(Z_i - b_j)}} \quad (2)$$

  Tvar křivky se opět nemění, odlišné jsou pouze odhady parametrů a jejich interpretace.
- Podzáložka *Nonlinear IRT Z* nabízí rozšíření dosud dvouparametrického modelu (2) na model tříparametrický (3PL). Nový 3PL model dovoluje navíc modelovat pravděpodobnost, že položka byla uhádnuta bez potřebné znalosti (c-parametr uhádnutelnosti), což je běžná situace obzvlášť v testech s nabídkou odpovědí:

$$P(Y_{ij} = 1|X_i) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(Z_i - b_j)}} \quad (3)$$

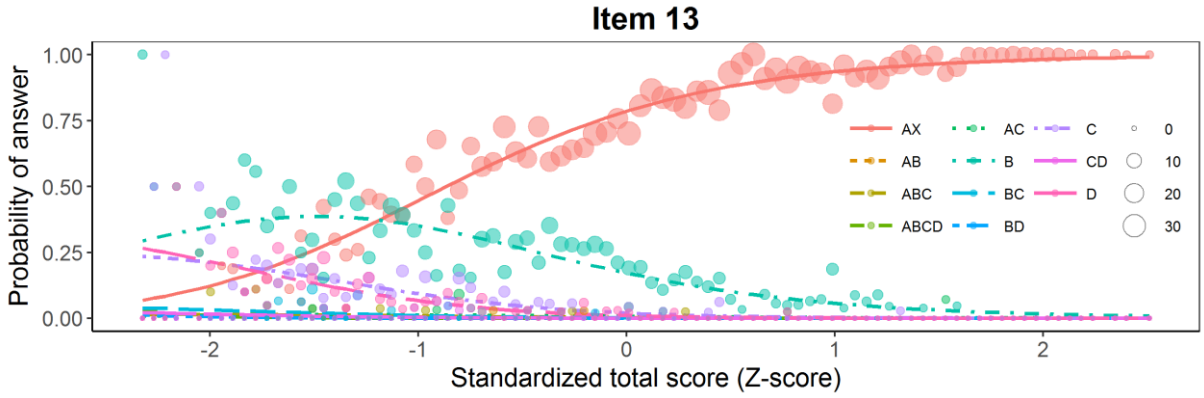
  Nově odhadnutá křivka (3) tedy připouští nenulovou dolní asymptotu (Obrázek 5).



Obrázek 5: Nelineární regrese pro položku č. 22 z datového souboru "Medical 100". Na rozdíl od logistické regrese nelineární (tříparamerický) model navíc povoluje možnost nenulové levé asymptoty, tedy modelování pravděpodobnosti uhádnutí položky bez potřebné znalosti. Body značí empirické hodnoty pravděpodobností a jejich velikost je určena počtem respondentů (20, 40 a 60).

- V podzáložce *Model comparison* aplikace nabízí porovnání dvouparametrických a tříparametrických modelů pro jednotlivé položky pomocí informačních kritérií AIC (Akaike, 1974), BIC (Schwarz, 1978), kde platí, že čím nižší hodnota kritéria, tím lepší model, a pomocí testu poměrem věrohodnosti (Agresti, 2013).
- Podzáložka *Multinomial* umožňuje pomocí multinomické regrese prozkoumat také pravděpodobnosti špatných odpovědí (distraktorů), viz Obrázek 6. Tento

model je obecnější variantou ordinálních modelů, lze jej tedy využít pro analýzu položek hodnocených na Likertově či jiné ordinální škále. Jako klíč správných odpovědí doporučujeme uvést nejvyšší možné skóre za položku. Multinomický model je také předstupněm tzv. Bockova nominálního modelu (viz dále).



Obrázek 6. Multinomická regrese pro položku č. 13 z datového souboru „Medical 100“. Červená plná čára znázorňuje pravděpodobnost správné odpovědi, ostatní křivky pak pravděpodobnosti volby jednotlivých distraktorů. Body značí empirické hodnoty pravděpodobností a jejich velikost je určena počtem respondentů (0, 10, 20 a 30).

IRT modely

Tzv. IRT modely (z angl. Item Response Theory, tj. modely teorie odpovědi na položku) vycházejí z výše uvedených regresních modelů. Koncepčně lze o nich uvažovat jako o smíšených regresních modelech, kde znalost studenta není reprezentována celkovým skóre X (resp. ani jeho standardizovanou verzí Z), ale je modelována pomocí náhodných efektů θ. Tyto náhodné efekty je možno interpretovat jako skutečnou (avšak latentní, tedy nepozorovanou) znalost studentů. O znalosti studentů se předpokládá, že se řídí daným rozdělením, obvykle se uvažuje normální rozdělení s nulovou střední hodnotou a kladným rozptylem.

Aplikace nabízí celkem čtyři typy IRT modelů pro skórovaná data (Urbánek, Denglerová, & Širůček, 2011): Raschův model a dále jednoparametrický (1PL), dvouparametrický (2PL) a tříparametrický (3PL) model, jehož rovnici zde uvádíme:

$$P(Y_{ij} = 1|X_i) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta_i - b_j)}} \quad (4)$$

Jednodušší 2PL model je obdobně jako u předchozích regresních modelů speciálním případem 3PL modelu (4), kde uvažujeme, že parametr uhádnutelnosti c je nulový. Ještě jednodušší 1PL model získáme tak, že navíc parametr diskriminace a fixujeme pro všechny položky. U nejjednoduššího IRT modelu, tzv. Raschova modelu, je tento parametr roven jedné, $a = 1$.

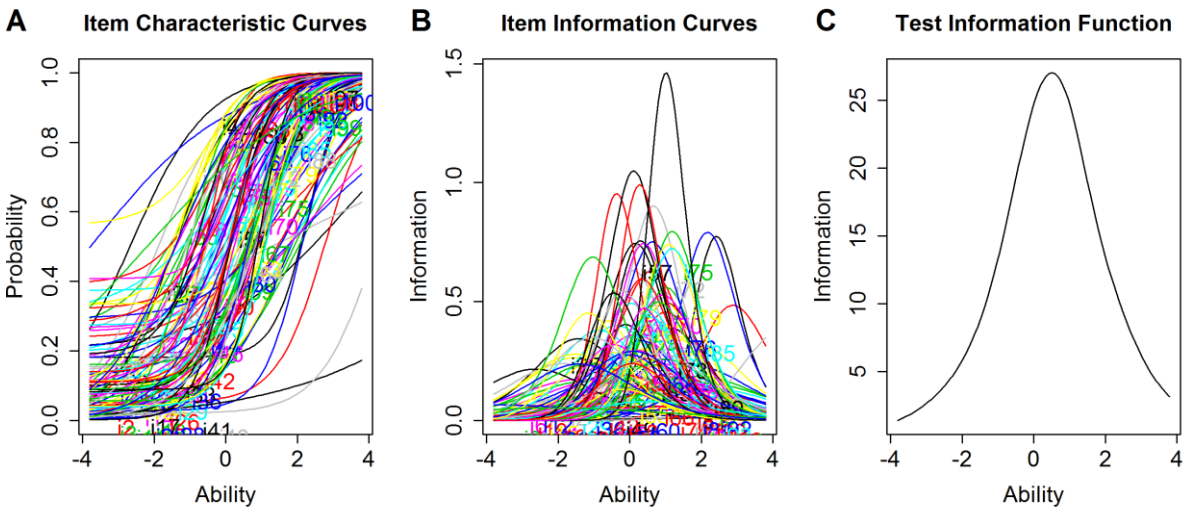
2PL IRT model a 3PL IRT model jsou obdobou regresních modelů (2) a (3), odhadují však studentovu znalost o něco přesněji – místo využití celkového skóre (resp. *z*-skóre) znalost studenta odhadují společně s parametry položek. Cenou za tuto vyšší přesnost bývá u většiny víceparametrických IRT modelů větší výpočetní náročnost, jakož i časté problémy s konvergencí obzvláště pak pro malé rozsahy výběru, či pro případy velkého množství položek. Vysoká výpočetní náročnost je patrná v aplikaci ShinyItemAnalysis obzvlášť pro volbu 2PL a 3PL IRT modelů na 100 položková data „Medical 100“, kde odhadnutí parametrů modelů trvá i několik minut. Pro data s menším počtem položek (např. datový soubor „GMAT“) je výpočetní složitost řádově nižší.

Pro neskórovaná data aplikace navíc nabízí Bockův nominální model (Bock, 1972), který slouží, obdobně jako multinomická regrese, ke zkoumání chování distraktorů. Na rozdíl od multinomické regrese, která využívá celkové skóre testu, Bockův nominální model, stejně jako ostatní IRT modely, odhaduje znalosti studentů pomocí náhodných efektů a uvádí se v IRT parametrizaci.

Obecně lze u IRT modelů vlastnosti jednotlivých položek podrobněji prozkoumat pomocí grafů tzv. charakteristických křivek položek (ICC z angl. Item Characteristic Curves, Obrázek 7A), které znázorňují pravděpodobnost správné odpovědí v závislosti na znalosti studenta.

Dalším důležitým nástrojem pro tvorbu testů jsou grafy tzv. informačních funkcí položek (IIC z angl. Item Information Curves, Obrázek 7B). Tyto jsou přímo odvozeny z charakteristických funkcí (s využitím derivace). Čím vyšší je hodnota IIC pro danou hodnotu latentního rysu, tím lépe položka pro tuto úroveň diskriminuje. IIC se využívají obzvláště v tzv. adaptivních testech (CAT z angl. Computer Adaptive Tests, viz Jelínek, Květon, & Vobořil, 2011), kde informační funkce slouží jako kritérium výběru optimální položky pro studenta s danou odhadnutou znalostí. Jaký tvar mají charakteristické a informační funkce položky pro různé hodnoty parametrů si lze vyzkoušet v podzáložce *Training: Characteristics and information curves.*

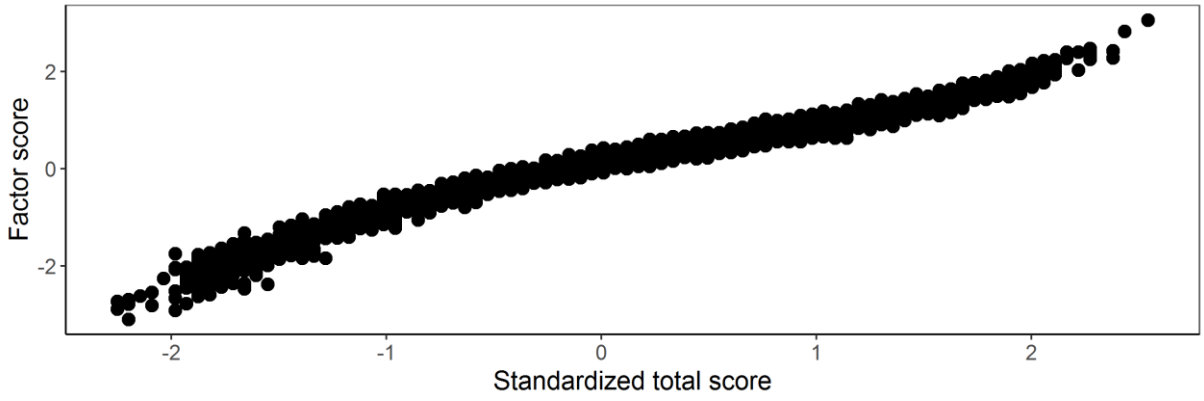
Konečně složením informačních funkcí položek získáme informační funkci (TIF z angl. Test Information Function) celého testu (Obrázek 7C). Ta nám mj. napovídá, pro jakou úroveň latentního rysu test jako celek diskriminuje nejlépe. Lze ji vnímat jako odhad spolehlivosti testu závislý na znalosti studenta (de Ayala, 2009; pro index spolehlivosti v IRT modelech viz také Andrich, 1982 a Martinková & Zvára, 2007).



Obrázek 7. Analýza datového souboru „Medical 100“ pomocí 3PL IRT modelu. A. Charakteristická křivka každé položky je určena odhadnutými parametry položky: obtížnost *b* je charakterizována polohou inflexního bodu (tj. bod, v němž má křivka největší sklon) na x-ové ose, diskriminační schopnost *a* pak určuje sklon křivky v tomto inflexním bodě. Konečně uhádnutelnost *c* je charakterizována levou (dolní) asymptotou. B. Informační funkce jednotlivých položek. C. Informační funkce celého testu.

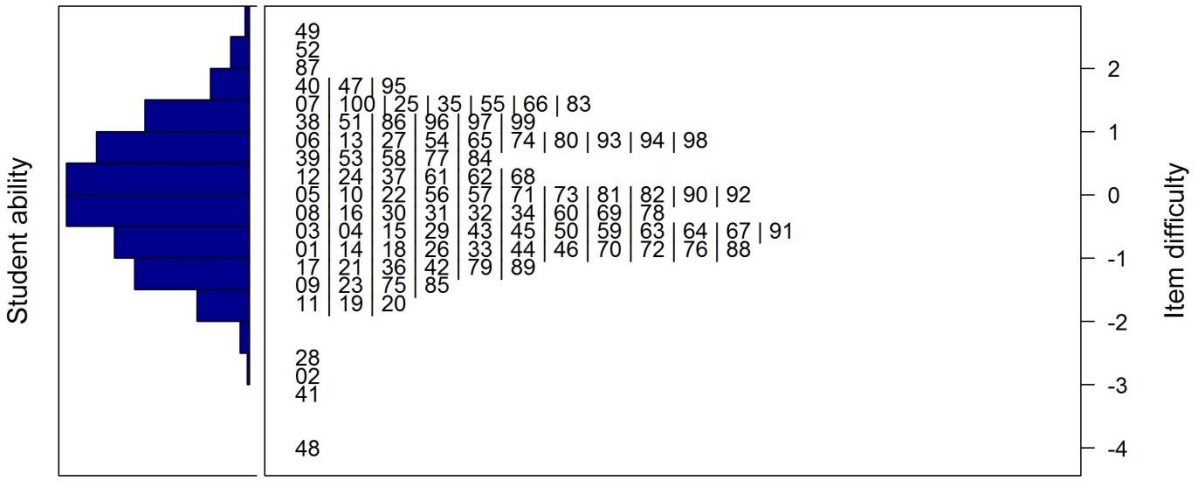
Společně s odhady koeficientů se obvykle uvádějí tzv. fit statistiky, které popisují, nakolik data z jednotlivých položek odpovídají zvolenému modelu. Aplikace nabízí $S - X^2$ statistiku (Ames & Penfield, 2015). Je-li statistika pro danou položku signifikantní (p hodnota menší než 0,05), znamená to, že zvolený model není vhodný pro danou položku. Pokud taková situace nastane u většího množství položek, lze doporučit využití složitějšího, např. vícerozměrného, modelu. Modely lze také, obdobně jako v předchozí sekci, porovnávat pomocí informačních kritérií (AIC, BIC, kde opět platí, že čím nižší hodnota kritéria, tím lepší model) a také pomocí testu poměrem věrohodnosti (de Ayala, 2009).

Jak již bylo naznačeno v úvodu této sekce, IRT modely (s dvěma a více parametry) odhadují znalost studenta přesněji než celkové skóre. O vztahu celkového skóre (či *z*-skóre) s odhadnutou skutečnou znalostí nám dává představu Obrázek 8.



Obrázek 8: Bodový graf (scatter plot) vztahu odhadu znalosti pomocí 3PL IRT modelu a *z*-skóre pro data „Medical 100“. Korelace mezi hodnotami je 0,98.

Další výhodou IRT modelů je, že umožňují kalibraci znalostí studentů a obtížnosti položek na stejnou škálu. Tzv. item-person map (Obrázek 9), někdy zvanou také Wright map (Wilson, 2005; Wright & Stone, 1979), pak lze využít pro posouzení vhodnosti položek pro námi zvolenou populaci. Z grafu lze na první pohled vyčíst, zda jsou pokryty všechny skupiny respondentů. V případě potřeby tak lze doplnit test o položky s chybějící obtížností, zaměřit se lze také na doplnění položek s obtížností okolo dané hranice úspěšnosti (cut-score).

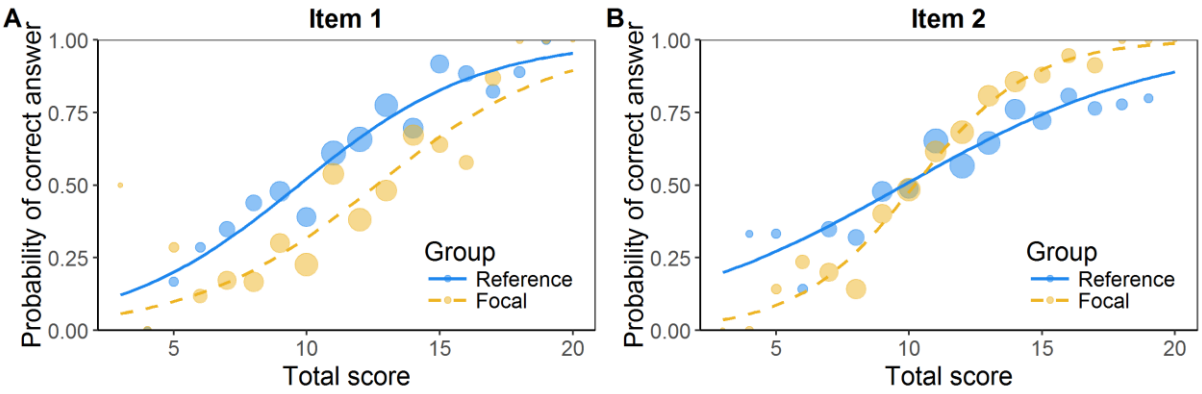


Obrázek 9: Wright map pro datový soubor „Medical 100“. V levé části se nachází histogram odhadnutých latentních znalostí respondentů, v pravé jsou pak vyobrazeny položky dle jejich odhadnuté obtížnosti. Jelikož z parametrů položek tento graf vyobrazuje pouze parametr obtížnosti, k jeho generování je obvykle využíván Raschův nebo 1PL IRT model.

DIF a DDF analýza

O diferenciálním fungování položek (DIF z angl. Differential Item Functioning) mluvíme, když studenti z jedné skupiny dosahují jiné pravděpodobnosti správné odpovědi na položky než studenti z druhé skupiny, jejichž celková úroveň znalosti je shodná (viz Obrázek 10). Položky vykazující DIF vůči příslušným skupinám jsou tak potenciálně neférové.

Přítomnost položek vykazujících DIF přitom nemusí být doprovázena rozdílem v celkovém skóre mezi skupinami: Skupiny se mohou lišit v rozložení celkových skóre a přitom žádná z položek nemusí vykazovat DIF. Naopak, potenciálně neférové položky mohou být přítomné i v případě, kdy se skupiny v celkovém počtu bodů neliší. V aplikaci se lze o této skutečnosti přesvědčit na datech „GMAT 2“, která jsou simulována tak, že první dvě položky vykazují DIF, přestože rozdíl mezi skupinami v celkovém počtu bodů není signifikantní (pro případ, kdy jsou rozdělení celkových skóre dokonce zcela totožné, viz Martinková, Drabinová, Liaw, et al., 2017).



Obrázek 10: Odlišné fungování položek (DIF) v datovém souboru “GMAT 2”. A. uniformní DIF (křivky mají stejnou diskriminaci (sklon), ale liší se v obtížnosti), B. neuniformní DIF (křivky mají odlišnou diskriminaci a kříží se). Body značí empirické hodnoty pravděpodobností a jejich velikost je určena počtem respondentů.

Složka Fairness/DIF nabízí širokou škálu metod, které mohou DIF a tedy i potenciálně neférové položky odhalit. Výpočty jsou založené na knihovnách difR (Magis, Beland, Tuerlinckx, & Boeck, 2010) a difNLR (Drabinová et al., 2017). Obecně lze metody dělit na klasické, tj. založené na celkovém skóre (či jeho standardizované formě), a na IRT metody.

Aplikace nabízí následující metody založené na celkovém skóre:

- Podzáložka *Delta Plot* představuje tradiční metodu detekování DIFu pomocí grafického zobrazení proporcí správných odpovědí, tzv. delta plot (Angoff & Ford, 1973).

- Podzáložka *Mantel-Haenszel* obsahuje tradiční metodu založenou na analýze kontingenčních tabulek včetně výpočtu podílů šancí správné odpovědi pro jednotlivé položky a daná celková skóre (Mantel & Haenszel, 1959).
- Podzáložka *Logistic* využívá již představený logistický model (1), tentokrát s rozšířením o efekt skupiny jako nezávislé proměnné (Swaminathan & Rogers, 1990).
- Podzáložka *Logistic IRT Z* navíc využívá místo celkových skóre jejich standardizovanou verzi *Z* a transformuje odhadnuté parametry (obdobně jako model (2)), aby byly porovnatelné s parametry odhadnutými pomocí 2 PL IRT modelu.

Dále aplikace nabízí tyto metody detekce DIFu založené na IRT modelech (1PL, 2PL, 3PL):

- Podzáložka IRT Lord představuje Waldův test (Lord, 1980).
- Podzáložka IRT Raju poskytuje test založený na ploše mezi logistickými křivkami (Raju, 1990).

Konečně, podzáložka *DDF* nabízí možnost zkoumání tzv. odlišného fungování distraktorů (DDF z angl. Differential Distractor Functioning) pomocí multinomické logistické regrese. Odlišné fungování distraktorů je situace, kdy dva studenti se stejnou znalostí ale z jiné skupiny mají různou pravděpodobnost volby (alespoň) jedné z nabízených odpovědí. Více o DIF/DDF viz Martinková, Drabinová, Liaw et al. (2017).

Výstup v HTML/PDF

Shiny aplikace rovněž nabízí možnost automatického vytvoření reportů. V těchto generovaných reportech můžeme nalézt vybrané textové nebo tabulkové výstupy spolu s grafickými interpretacemi analýz.

Jmenovitě se ve verzi 1.2.0 jedná o:

- Souhrnné popisné charakteristiky datového souboru.
- Histogram četností studentů na základě celkového skóre.
- Graf obtížností a diskriminací položek.
- Graf analýzy distraktorů v porovnání s výsledky multinomické regrese.
- Grafické a tabulkové výstupy uživatelem voleného IRT modelu.
- Grafický výstup logistické regrese u položek, u kterých byla zjištěna odlišnost fungování mezi skupinami (DIF).
- Graficky výstup analýzy odlišného fungování distraktorů (DDF).

Reporty je možné generovat ve formátu PDF a HTML. Generování reportů ve formátu HTML je zcela samostatné a nevyžaduje žádná dodatečná opatření ze strany uživatele. Generování ve formátu PDF vyžaduje instalaci aktuální verze systému TeX a příslušných

balíků. V rámci online verze lze vygenerovat PDF výstup bez nutnosti instalace TeXu, generování výstupů může však zabrat několik minut.

Dále lze v rámci generování výstupů nastavit, jak má vypadat obsah reportu. Uživatel si tak může zvolit, zda chce zahrnout korelační strukturu, zda a jaký preferuje IRT model, nebo si vybrat, které části DIF analýzy mají být zahrnuty.

Diskuse a závěr

V tomto článku jsme pojednali o možnostech analýz znalostních a psychologických testů pomocí aplikace ShinyItemAnalysis založené na statistickém programu R. Výhodou R je jeho finanční dostupnost (je k dispozici zcela zdarma) a otevřenost (zdrojový kód většiny funkcí je dohledatelný, lze jej tedy využít pro lepší porozumění i další inspiraci). V současné době je pro psychometrickou analýzu dostupné široké spektrum R knihoven (Rusch et al., 2013)[8].

Jako nevýhoda R může být vnímána nutnost psaní zdrojového kódu. Představená aplikace ShinyItemAnalysis nabízí vhodnou a uživatelsky přívětivou alternativu. Aplikace je dostupná také online a má grafické rozhraní, které nevyžaduje znalost programovacího jazyka R. Přesto je vybraný kód součástí aplikace a může tak sloužit jako učební pomůcka či ilustrace, jak provádět vlastní analýzy pomocí R.

Aplikace současně nabízí možnost načtení a následného analyzování vlastních dat, jakož i vytvoření stručného reportu. Aspiruje tak na to být jednoduchou a volně dostupnou aplikací pro rutinní analýzu znalostních a psychologických testů, využitelnou fakultami pro analýzu jejich přijímacích řízení, či pedagogy v rámci formativního testování během výuky.

Funkčnost aplikace byla demonstrována na datech z přijímacího řízení na lékařskou fakultu (Štuka et al., 2016, viz také Štuka et al., 2012). Možné využití aplikace je však mnohem širší. Validizace, a tudíž analýza testů a jejich položek, je důležitá při konstrukci dalších znalostních testů (McFarland et al., 2017), psychologických testů, ale také např. škál pro hodnocení klinického stavu nebo zdravotního postižení (Řasová, Martinková, Vyskotová, & Šedová, 2012), či dotazníků o kvalitě života.

Aplikace ShinyItemAnalysis je i nadále ve vývoji a je možné ji dále rozšiřovat či vylepšovat. Autoři plánují v dalších verzích zpřístupnit některé polytomní modely pro položky hodnocené na Likertově škále, např. partial credit model nebo ranking scale model (viz de Ayala, 2009). Tyto modely jsou speciálním případem Bockova nominálního modelu (Muraki, 1992), který již nyní v aplikaci nabízený je, avšak bez omezení daných zmíněnými polytomními modely. Podněty pro případná další rozšíření

[8] Viz také https://cran.r-project.org/web/views/Psychometrics.html

aplikace lze vložit také online, či je přímo naprogramovat a nabídnout k zakomponování do stávající aplikace.[9]

ShinyItemAnalysis může také sloužit jako inspirace pro vytvoření nové podobné aplikace, obsahující analýzy šité na míru potřebám dané instituce či daným datům (pro bohatý seznam příkladů viz webové stránky věnované shiny aplikacím[10]).

Věříme, že ShinyItemAnalysis napomůže zvýšit povědomí o psychometrických modelech a metodách pro analýzu znalostních a psychologických testů a jejich položek. Přejeme si, aby se i díky ní v České republice analýza a pretestování znalostních testů staly standardem a samozřejmou záležitostí. Taková praxe může vést ke zkvalitnění znalostních testů používaných ve výuce i v přijímacím řízení.

Poznámky

Tento text vznikl v rámci projektu *Odhad psychometrických vlastností jako součást vývoje přijímacích testů* podpořeném GA ČR (projekt GJ15-15856Y). Autoři děkují Hynkovi Cíglerovi, Čestmíru Štukovi, Martinu Vejražkovi a dvěma anonymním recenzentům za cenné připomínky k dřívější verzi textu.

Patrícia Martinková, Adéla Drabinová & Jakub Houdek: ShinyItemAnalysis: Analyzing admission and other educational and psychological tests.

In this paper we introduce ShinyItemAnalysis application for psychometric analysis of educational and psychological tests and their items. ShinyItemAnalysis provides graphical interface and web framework to open source statistical software R and thus opens up its functionality to wide audience. Application covers broad range of methods and offers data examples, model equations, parameter estimates, interpretation of results, together with selected R code, and is thus suitable for teaching psychometric concepts with R. The application also aspires to be a simple tool for routine analysis by allowing the users to upload and analyze their own data and by generating analysis report. We conclude by arguing that psychometric analysis should be a routine part of test development in order to gather proofs of reliability and validity of the measurement. With example of admission test to medical faculty, we demonstrate how ShinyItemAnalysis may provide a simple and free tool to routinely analyze tests.

[9] Viz https://github.com/patriciamar/ShinyItemAnalysis/issues
[10] Viz https://shiny.rstudio.com/

Literatura

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association.

Agresti, A. (2013). Categorical Data Analysis. Wiley. Retrieved from http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470463635.html

Akaike, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Ames, A. J., & Penfield, R. D. (2015). An NCME Instructional Module on Item-Fit Statistics for Item Response Theory Models. Educational Measurement: Issues and Practice, 34(3), 39–48. https://doi.org/10.1111/emip.12067

Anděl, J., & Zvára, K. (2005). Přijímací zkouška z matematiky na MFF v roce 2004. Pokroky Matematiky, Fyziky a Astronomie, 50(2), 148–161. Retrieved from http://hdl.handle.net/10338.dmlcz/141263%0A

Andrich, D. (1982). An Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response Pattern. Education Research and Perspective, 9(1), 95–104.

Angoff, W. H., & Ford, S. F. (1973). Item-Race Interaction on a Test of Scholastic Aptitude. Journal of Educational Measurement, 10(2), 95–106. Retrieved from http://www.jstor.org/stable/1433905

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1), 29–51. https://doi.org/10.1007/BF02291411

Byčkovský, P., & Zvára, K. (2007). Konstrukce a analýza testů pro přijímací řízení. Univerzita Karlova v Praze, Pedagogická fakulta. Retrieved from https://books.google.cz/books?id=mvvjtgAACAAJ

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows. Lincolnwood, IL: Scientific Software International.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

ČŠI. (2015). Honocení výsledků vzdělávání didaktickými testy. Retrieved from http://www.csicr.cz/cz/Aktuality/Hodnoceni-vysledku-vzdelavani-didaktickymi-testy

de Ayala, R. J. (2009). The theory and practice of item response theory. New York, NY: Guilford Press.

Drabinová, A., Martinková, P., & Zvára, K. (2017). difNLR: Detection of Dichotomous Differential Item Functioning (DIF) and Differential Distractor Functioning (DDF) by Non-Linear Regression Models. Retrieved from https://cran.r-project.org/package=difNLR

Höschl, C., & Kožený, J. (1997). Predicting academic performance of medical students: The first three years. The American Journal of Psychiatry, 154(6), 86.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2017). shiny: Web Application Framework for R. Retrieved from https://cran.r-project.org/package=shiny

IBM Corp. Released. (2015). IBM SPSS Statistics for Windows, Version 23.0. 2015.

Jelínek, M., Květoň, P., & Vobořil, D. (2011). Testování v psychologii: Teorie odpovědi na položku a počítačové adaptivní testování. Praha: Grada.

Kingston, N., Leary, L., & Wightman, L. (1985). An Exploratory Study of the Applicability of Item Response Theory Methods to the Graduate Management Admission Test. ETS Research Report Series. https://doi.org/doi.org/10.1002/j.2330-8516.1985.tb00119.x

Kožený, J., Tišanská, L., & Höschl, C. (2001). Akademická úspěšnost na střední škole: prediktor absolvování studia medicíny. Československá Psychologie : Časopis pro Psychologickou Teorii a Praxi, 45(1), 1–6. Retrieved from http://www.medvik.cz/link/bmc01014269

Legewie, J., & DiPrete, T. A. (2014). The High School Environment and the Gender Gap in Science and Engineering. Sociology of Education, 87(4), 259–280. https://doi.org/10.1177/0038040714547770

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. Rasch Measurement Transactions, 19(3), 1032.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 42, 847–862. https://doi.org/10.3758/BRM.42.3.847

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. JNCI: Journal of the National Cancer Institute, 22(4), 719.https://doi.org/10.1093/jnci/22.4.719

Martinková, P., Drabinová, A., Leder, O., Houdek, J. (2017). ShinyItemAnalysis: Test and Item Analysis via Shiny. Version 1.2.0 Retrieved from https://cran.r-project.org/package=ShinyItemAnalysis

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J., & Price, R. M. (2017). Checking equity: Why DIF analysis should be a routine part of developing conceptual assessments. CBE-Life Sciences Education, 16(2), rm2. http://www.lifescied.org/content/16/2/rm2.full.pdf+html?with-ds=yes https://doi.org/10.1187/cbe.16-10-0307

Martinková, P., Štěpánek, L., Drabinová A., Houdek J., Vejražka M., Štuka Č. Semi-real-time analyses of item characteristics for medical school admission tests. In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems. Accepted.

Martinková, P., & Zvára, K. (2007). Reliability in the Rasch Model. Kybernetika, 43(3), 315–326. Retrieved from http://dml.cz/bitstream/handle/10338.dmlcz/135776/Kybernetika_43-2007-3_4.pdf

McFarland, J., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., Modell H., Wright, A. (2017). Development and validation of the Homeostasis Concept Inventory. CBE-Life Sciences Education, 16(2), ar35. http://www.lifescied.org/content/16/2/ar35.full.pdf+html?with-ds=yes. https://doi.org/10.1187/cbe.16-10-0305

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. ETS Research Report Series, 1992(1).

R Development Core Team. (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna Austria, 0, {ISBN} 3-900051-07-0. https://doi.org/10.1038/sj.hdy.6800737

Raju, N. S. (1990). Determining the Significance of Estimated Signed and Unsigned Areas Between Two Item Response Functions.pdf. Applied Psychological Measurement, 14, 197–207.

Revelle, W. (2009). An introduction to psychometric theory with applications in R. Retrieved from http://www.personality-project.org/r/book/

Rubešová, J. (2009). Souvisí úspěšnost studia na vysoké škole se středoškolským prospěchem? Pedagogická Orientace, 19(3), 89–103.

Rusch, T., Mair, P., & Hatzinger, R. (2013). Psychometrics With R: A Review Of CRAN Packages For Item Response Theory. Center for Empirical Research Methods, Discussion Paper Series, (November).

Řasová, K., Martinková, P., Vyskotová, J., & Šedová, M. (2012). Assessment set for evaluation of clinical outcomes in multiple sclerosis - psychometric properties. Patient Related Outcome Measures, 3, 59–70. Retrieved from https://www.dovepress.com/assessment-set-for-evaluation-of-clinical-outcomes-in-multiple-scleros-peer-reviewed-article-PROM

SABER. (n.d.). Biology Concept Inventories and Assessments. Retrieved March 9, 2017, from http://saber-biologyeducationresearch.wikispaces.com/DBER-Concept+Inventories

Salvatori, P. (2001). Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. Advances in Health Sciences Education, 6(2), 159–175. https://doi.org/10.1023/A:1011489618208

SAS Institute Inc. (2013). SAS 9.4 Language Reference: Concepts. Cary, NC, USA: SAS Institute Inc.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6(2), 461–464. https://doi.org/10.2307/2958889

StataCorp. (2015). Stata Statistical Software: Release 14. 2015. https://doi.org/10.2307/2234838

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. Source Journal of Educational Measurement, 27(4), 361–370. Retrieved from http://www.jstor.org/stable/1434855

Štuka, Č., Martinková, P., Vejražka, M., Trnka, J., & Komenda, M. (2013). Testování při výuce medicíny. Konstrukce a analýza testů na lékařských fakultách. (Vyd. 1.). Praha: Karolinum. Retrieved from http://www.wikiskripta.eu/Testy

Štuka, Č., Martinková, P., Zvára, K., & Zvárová, J. (2012). The prediction and probability for successful completion in medical study based on tests and pre-admission grades. The New Educational Review, 28, 138–152. Retrieved from http://www.educationalrev.us.edu.pl/dok/volumes/tner_2_2012.pdf

Štuka, Č., Vejražka, M., Martinková, P., Komenda, M., & Štěpánek, L. (2016). The use of test and item analysis for improvment of tests. In Mefanet. Brno. Retrieved from http://www.mefanet.cz/index.php?pg=konference--prezentace

Urbánek, T., Denglerová, D., & Širůček, J. (2011). Psychometrika: měření v psychologii. Portál.

van der Linden, W. J. (2017). Handbook of Item Response Theory, Three Volume Set. CRC Press.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Taylor & Francis. Retrieved from https://doi.org/10.4324/9781410611697

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: Mesa Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2008). ConQuest: Multi-Aspect Test Software. Camberwell: Australian Council for Educational Research.

Zvára, K., & Anděl, J. (2001). Connections between the results of entrance examinations and successful completion of studies at the Faculty of Mathematics and Physics. Pokroky Mat. Fyz. Astron., 46(4), 304–312. Retrieved from http://dml.cz/dmlcz/141097

Zwick, R. (2006). Higher education admission tests. In Educational Measurement (4th ed.). Westport, CT: American Council on Education/Praeger.