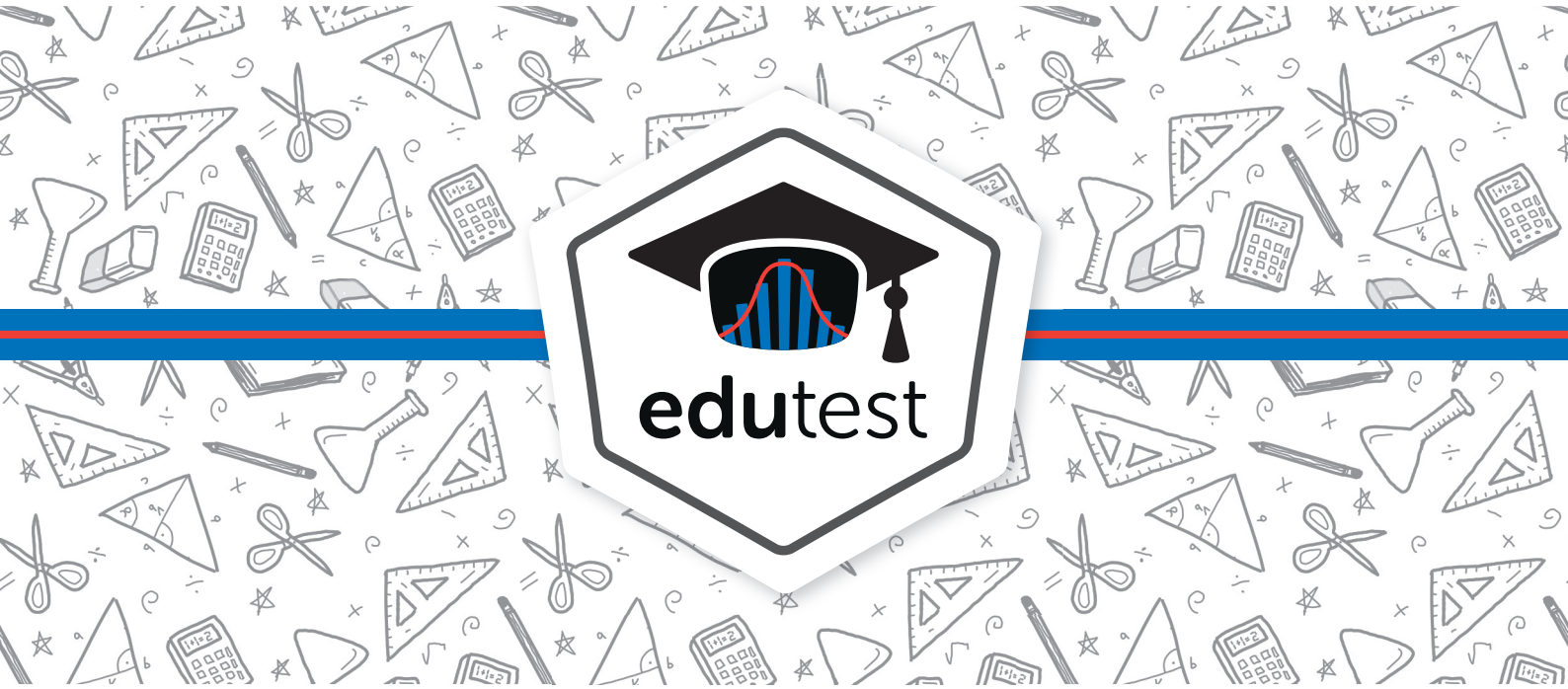


Soubor postupů a nástrojů
pro zkvalitnění tvorby znalostních testů pomocí psychometrických modelů

Patrícia Martinková, Eva Potužníková, Jan Netík a kol.

PEDAGOGICKÁ FAKULTA
Ústav výzkumu a rozvoje vzdělávání
Univerzita Karlova

Ústav informatiky
Akademie věd ČR

T A
Č R

Projekt TL05000008 Výzvy pro hodnocení znalostí: Analytická podpora tvorby znalostních testů byl spolufinancován se státní podporou Technologické agentury ČR v rámci Programu ÉTA 5.

Obsah

Psychometrická analýza maturitních a jiných testů pomocí interaktivní aplikace ShinyItemAnalysis a modulu EduTest Item Analysis

Patrícia Martinková, Jan Netík, Adéla Hladká

1 Úvod

Se zavedením jednotné maturitní zkoušky a jednotných přijímacích zkoušek na střední školy roste v České republice význam standardizovaných znalostních testů zadávaných jednotně velkým počtům respondentů. Rovněž vysoké školy mohou zařazovat znalostní testy do přijímacího řízení, vlastní testy zadává Česká školní inspekce a řada škol využívá pro svoji potřebu testy soukromých společností. Ruku v ruce s rostoucím dopadem znalostních testů na rozhodování na různých úrovních vzdělávacího systému roste také potřeba komplexní analýzy znalostních testů v rámci jejich vývoje.

Tradiční metody analýzy vícepoložkových testů vycházející z klasické testové teorie poskytnou dobrý prvotní náhled na fungování jednotlivých položek a mohou napomoci odhalit nesprávně formulovanou položku. Mají však několik omezení, např. neumožňují stanovit náročnost testu nezávisle na testované populaci, tedy ani porovnat obtížnost různých variant testu zadávaných například v různých termínech nebo letech, předpokládají také stejnou velikost chyby měření pro všechny úrovně latentní znalosti.

Velký potenciál v tomto ohledu poskytují metody založené na regresních modelech a jejich složitější varianty, modely teorie odpovědi na položku (*Item Response Theory*, IRT), viz např. Martinková a Hladká (2023). Další rozšíření těchto modelů, skupinově specifické modely, pak nabízí možnost zkoumat podrobněji meziskupinové rozdíly, a to až na úroveň položek, pomocí analýzy tzv. odlišného fungování položek (*Differential Item Functioning*, DIF), viz např. Martinková et al. (2017); Martinková a Hladká (2023).

Zatímco tradiční položkovou analýzu lze provést snadno i v běžně dostupném tabulkovém procesoru, implementace regresních modelů, IRT modelů a DIF analýzy již často vyžaduje specializovaný software.

2 ShinyItemAnalysis a modul EduTest Item Analysis

Interaktivní aplikace `ShinyItemAnalysis` (Martinková, Drabinová, & Houdek, 2017; Martinková & Drabinová, 2018; Martinková & Hladká, 2023) si klade za cíl zpřístupnit pokročilé psychometrické metody, včetně IRT modelů a metod pro analýzu odlišného fungování položek, vzdělávacím institucím, testovým společnostem a široké odborné veřejnosti a podpořit jejich integraci do procesu testování znalostí. Aplikace umožňuje rychlé zpracování testových dat a včasnou identifikaci potenciálně problematických položek.

Aplikace je přístupná online na odkazu

https://shiny.cs.cas.cz/ShinyItemAnalysisEduTest

Lze s ní pracovat také lokálně ve volně šiřitelném statistickém prostředí `R`. Nejdříve je

potřeba nainstalovat balíky `ShinyItemAnalysis` a `EduTestItemAnalysis` pomocí příkazů:

```
install.packages("ShinyItemAnalysis", dependencies = TRUE)
install.packages("EduTestItemAnalysis", repos =
    c("https://applstat.github.io/SIArepo/", "https://cloud.r-project.org"))
```

Aplikaci potom spustíme pomocí následujícího příkazu:

```
ShinyItemAnalysis::run_app()
```

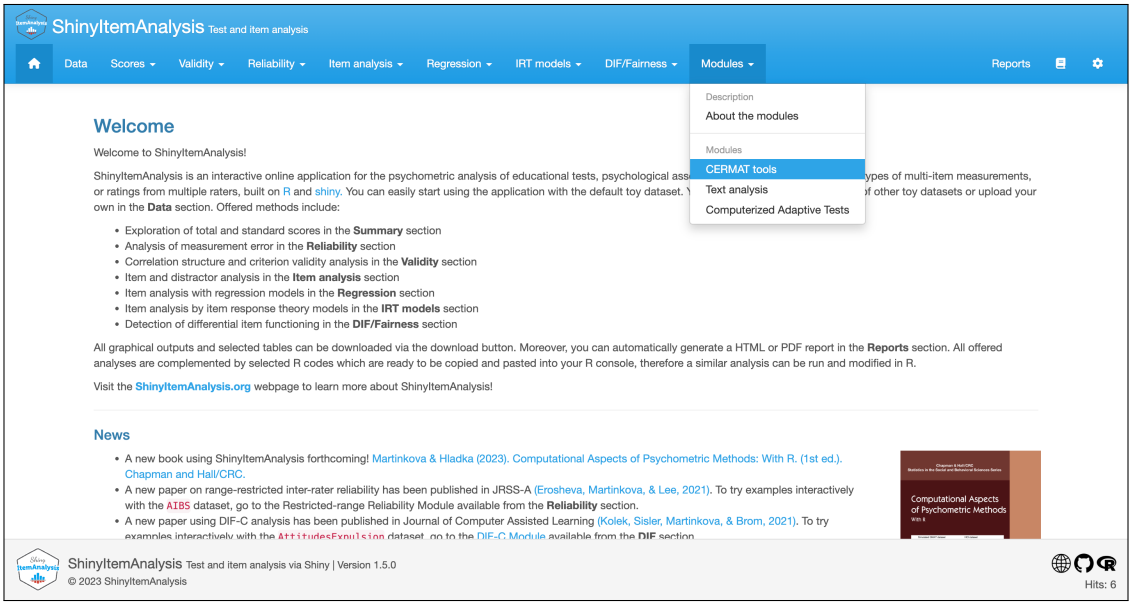
Úvodní stránka interaktivní aplikace uživatele seznamuje s typy analýz, které aplikace nabízí. Aplikace je dále rozdělena do několika záložek (viz obrázek 1). V rámci první záložky je možné zvolit si vybraná cvičná data, nebo si nahrát svá vlastní data ve vhodném formátu. Další záložky pak zpřístupňují analýzu celkových skóre (záložka Scores), analýzy pro získání empirických důkazů o validitě testu (záložka Validity), nebo spolehlivosti testu (záložka Reliability). V následujících záložkách je k dispozici analýza jednotlivých položek, a to jak tradiční metody (záložka Item analysis), tak i metody založené na regresních modelech (záložka Regression) a IRT modelech (záložka IRT models), a nakonec také metody pro analýzu odlišného fungování položek (záložka DIF/Fairness).

2.1 Přídavný modul EduTest Item Analysis

Aplikace `ShinyItemAnalysis` nabízí od verze 1.5.0 také možnost vytvořit a připojit přídavné moduly. Ty jsou poté zařazeny buď v záložce Modules (obrázek 1), nebo v jiné záložce, pokud tam obsahově lépe zapadají. Ve stávající verzi aplikace je k dispozici několik různých modulů, mj. modul pro demonstrace odhadu shody mezi hodnotiteli (inter-rater reliability) v případě omezených vzorků (v záložce Reliability; viz Erosheva, Martinková, & Lee, 2021), modul pro demonstraci analýzy odlišného fungování položek ve změně (DIF-C v záložce DIF/Fairness; viz Martinková, Hladká, & Potužníková, 2020), modul pro analýzu obtížnosti položek na základě textové analýzy zadání (Štěpánek, Dlouhá, & Martinková, 2023), nebo modul pro demonstraci počítačového adaptivního testování v záložce Modules, viz obrázek 1.

Jedním z nabízených modulů v záložce Modules je také EduTest Item Analysis. Tento modul nabízí možnost načtení a následné analýzy neagregovaných položkových dat z maturitních a jednotných přijímacích testů, které jsou k dispozici na portálu Centra pro zjišťování výsledků vzdělávání (CZVV), dostupném na adrese

https://vysledky.cermat.cz/statistika/



Obrázek 1: Úvodní stránka aplikace `ShinyItemAnalysis` s rozbalenou nabídkou jinde nezařazených modulů

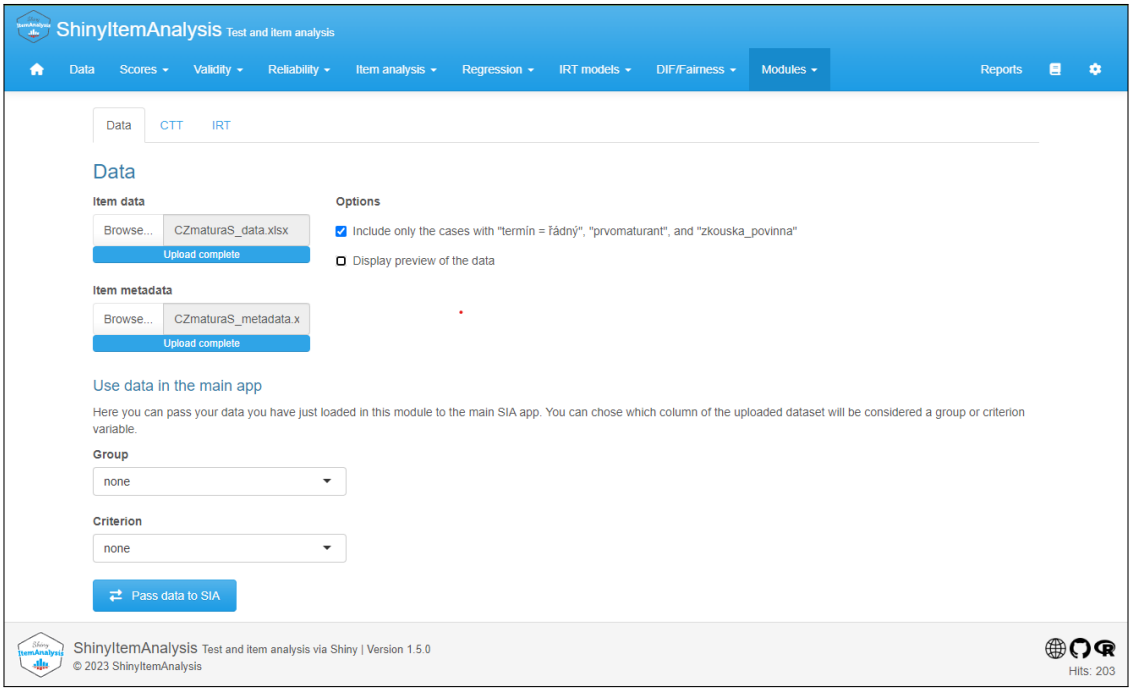
Do modulu je potřeba načíst také metadata popisující položky testu (typ položky, počet možných odpovědí, informace o správné odpovědi), která si musí uživatel připravit z veřejně dostupných testových zadání a klíčů správných řešení.

Modul předpokládá načtení položkových dat ve formátu, v němž jsou zveřejňována data CZVV (např. specifické označování názvů položek či kódování žákovských odpovědí). Je tedy určen především k podpoře uživatelů z řad CZVV i širší odborné veřejnosti při provádění základních a zejména pokročilých analýz dat z jednotných přijímacích a maturitních testů. Zároveň však může sloužit k demonstraci analytických postupů při práci s testovými daty obecně či jako inspirace tvůrcům podobných podpůrných nástrojů.

Demonstraci práce s modulem zde provádíme na datech z maturitního testu z matematiky zadávaného v jarním termínu 2019. Pro snížení výpočetní náročnosti nepracujeme s úplným souborem respondentů, ale používáme náhodný vzorek 2 000 žáků (obrázek 2). Používaná položková data i metadata jsou ke stažení na adrese:

https://edutest.cs.cas.cz/

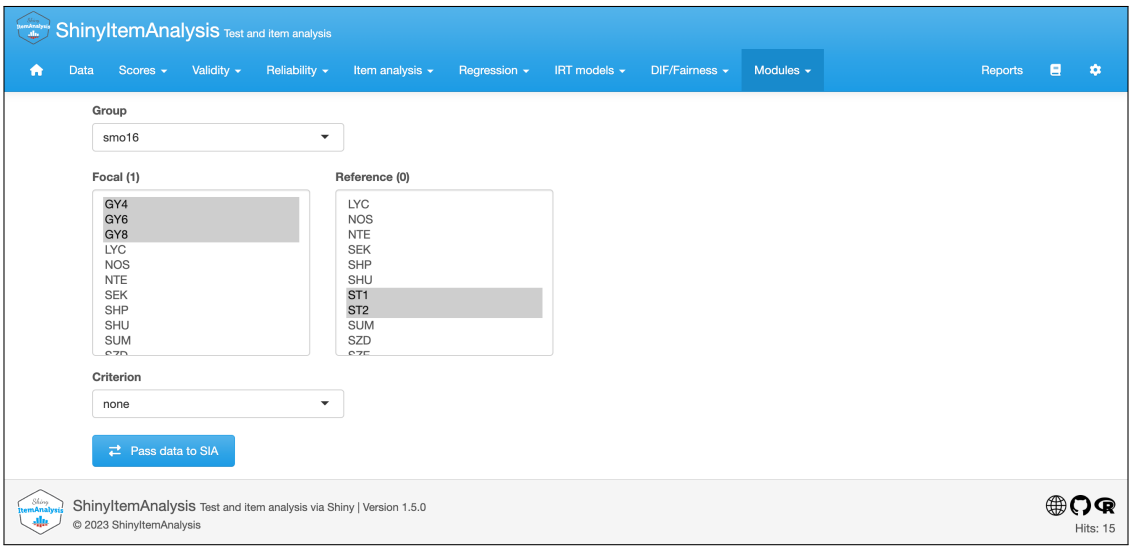
Z nahraných položkových dat jsou automaticky detekovány sloupce bodovaných odpovědí, které jsou dále využity v rámci položkové analýzy. Lze také zvolit skupinovou proměnnou a kritérium, které se využijí v některých typech podrobnějších analýz. Skupinovou proměnnou může být například pohlaví nebo typ školy, kritériem například výsledek v jiném testu, pokud jsou tyto proměnné součástí nahraného datového souboru. V naší demonstraci používáme proměnnou „smo16“, a vytváříme z ní binarizo-



Obrázek 2: Nahrání dat maturitního testu z matematiky z jarního termínu 2019 v modulu EduTest Item Analysis

vanou skupinovou proměnnou, která indikuje typ navštěvované školy (viz obrázek 3). V tomto konkrétním příkladu tvoří první skupinu respondenti studující na gymnáziu (4-, 6- či 8letém; hodnota proměnné = 1) a druhou skupinu respondenti studující na střední odborné škole s technickým nebo technologickým zaměřením (obory ST1 a ST2; hodnota proměnné = 0). Pro žáky z ostatních typů škol má proměnná hodnotu NA (z angl. „not available“) a tito žáci do podrobnějších analýz porovnávajících rozdíly mezi skupinami nevstupují. Jsou však zařazeni do všech ostatních analýz, které nepracují se skupinovou proměnnou.

Modul pak v jednotlivých podzáložkách nabízí analýzy vycházející z klasické testové teorie (*Classical Test Theory*, CTT), i z teorie odpovědi na položku (IRT), které si představíme v dalších částech. Kromě toho modul nabízí i možnost data automaticky upravit do vhodného formátu, který je využíván v ostatních částech interaktivní aplikace `ShinyItemAnalysis`, a analyzovat je pomocí zde dostupných metod (tlačítko Pass data to SIA, viz obrázky 2 a 3).



Obrázek 3: Binarizace proměnné „smo16“

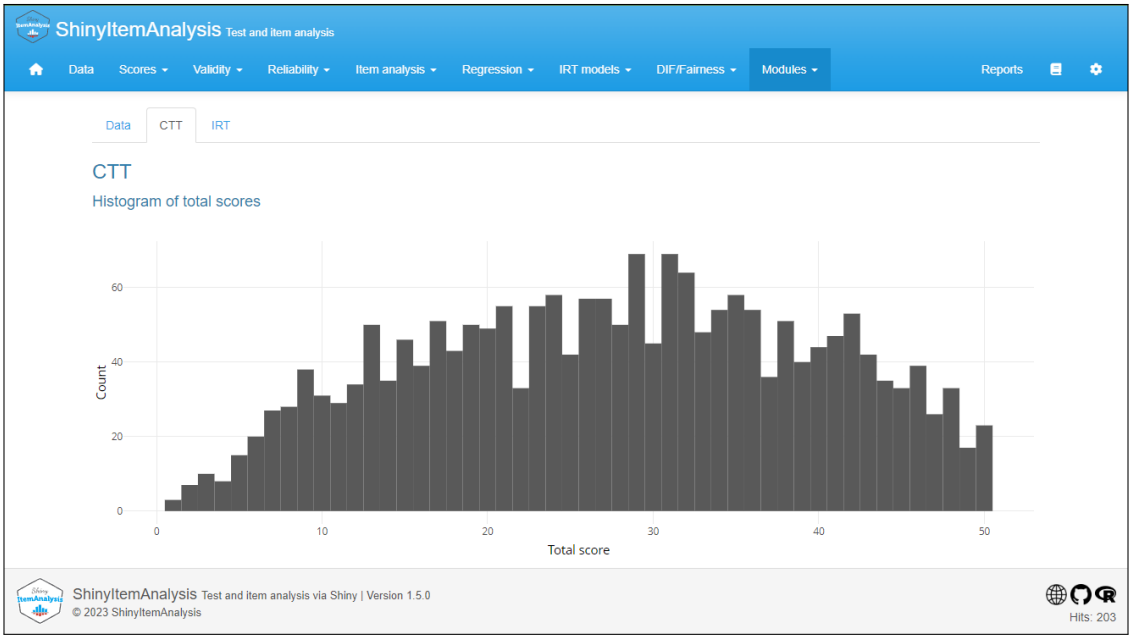
3 Metody klasické testové teorie

V podzáložce CTT jsou k dispozici metody klasické testové teorie a tradiční položkové analýzy. Nabízené grafy a tabulky vycházejí z nástrojů využívaných analytickým oddělením CZVV. Metody CTT využívají k popisu fungování testu a položek podíly, procenta a korelace.

Nejprve je zobrazen histogram celkového počtu bodů (obrázek 4). Podobný graf, avšak s možností interaktivního barevného zobrazení, je nabízen také v hlavní aplikaci `ShinyItemAnalysis` v sekci Scores.

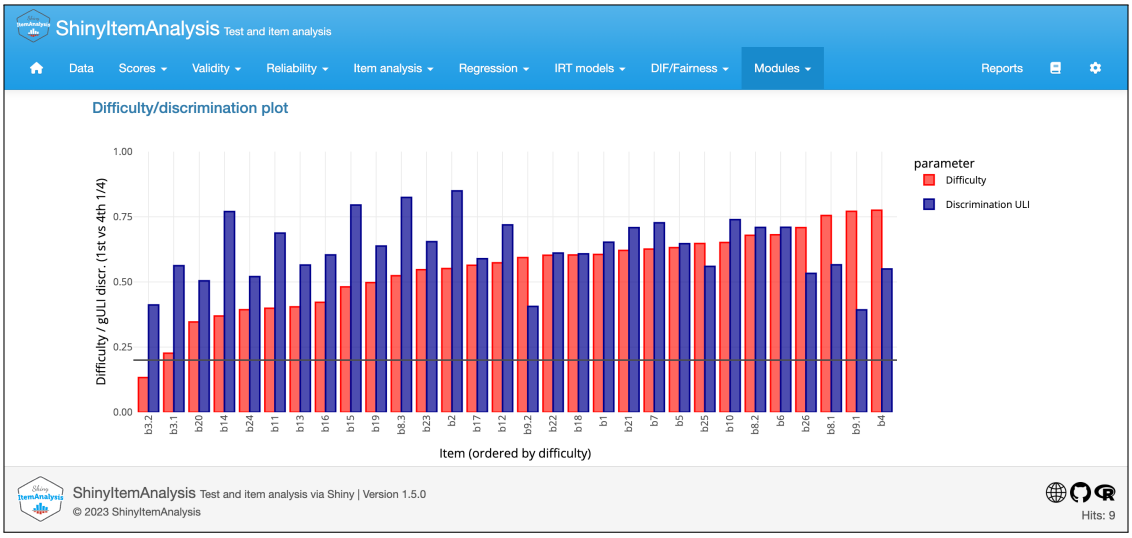
Dále je zobrazen graf obtížnosti (procentuální podíl správných odpovědí na danou položku, červené sloupce v obrázku 5) a diskriminačního indexu položek, který je založený na porovnání nejslabší a nejsilnější čtvrtiny respondentů (tzv. Upper-Lower Index, ULI, modré sloupce). Položky jsou v tomto grafu seřazeny od nejtěžších položek (v tomto případě položky b3.2 a b3.1 vlevo, na které správně zodpovědělo pouze 13 %, resp. 23 % respondentů) po nejlehčí položku (položka b4 vpravo, u které jsme zaznamenali přes 77 % správných odpovědí). Podezřelé by byly ty položky, u kterých je diskriminační index ULI nižší než předem stanovená mez (zpravidla se volí hodnota 0,2, zvýrazněno černou vodorovnou čarou). To zde nenastává, všechny položky se zdají mít dobrou diskriminační schopnost. Při interpretaci výsledků položkové analýzy je však také potřeba mít na paměti, že v případě velmi snadných nebo velmi těžkých položek je nízká diskriminace očekávatelná.

Další varianty tohoto grafu jsou přístupné v hlavní aplikaci `ShinyItemAnalysis` v záložce Item Analysis. Ty umožňují interaktivně zvolit i jiný počet skupin pro definici



Obrázek 4: Histogram celkového počtu bodů

indexu ULI (např. definovat ULI pomocí 1. a 3. třetiny), nebo místo ULI využít pro popis diskriminace korelaci mezi položkovým a celkovým skóre (RIT) nebo korelaci mezi položkovým skóre a skóre v testu bez dané položky (RIR).



Obrázek 5: Graf obtížnosti a diskriminačního indexu položek

Modul dále nabízí jiné možné zobrazení obtížnosti a diskriminace jednotlivých položek, které je více využíváno v rámci CZVV (obrázek 6). Zde jsou obtížnost a diskriminace jednotlivých položek zobrazeny v bodovém grafu. Podezřelé jsou opět položky, jejichž ULI je menší než stanovená mez (body v dolní části grafu), poněkud méně přísně

se opět nahlíží na velmi snadné (vlevo) nebo velmi těžké položky (vpravo).

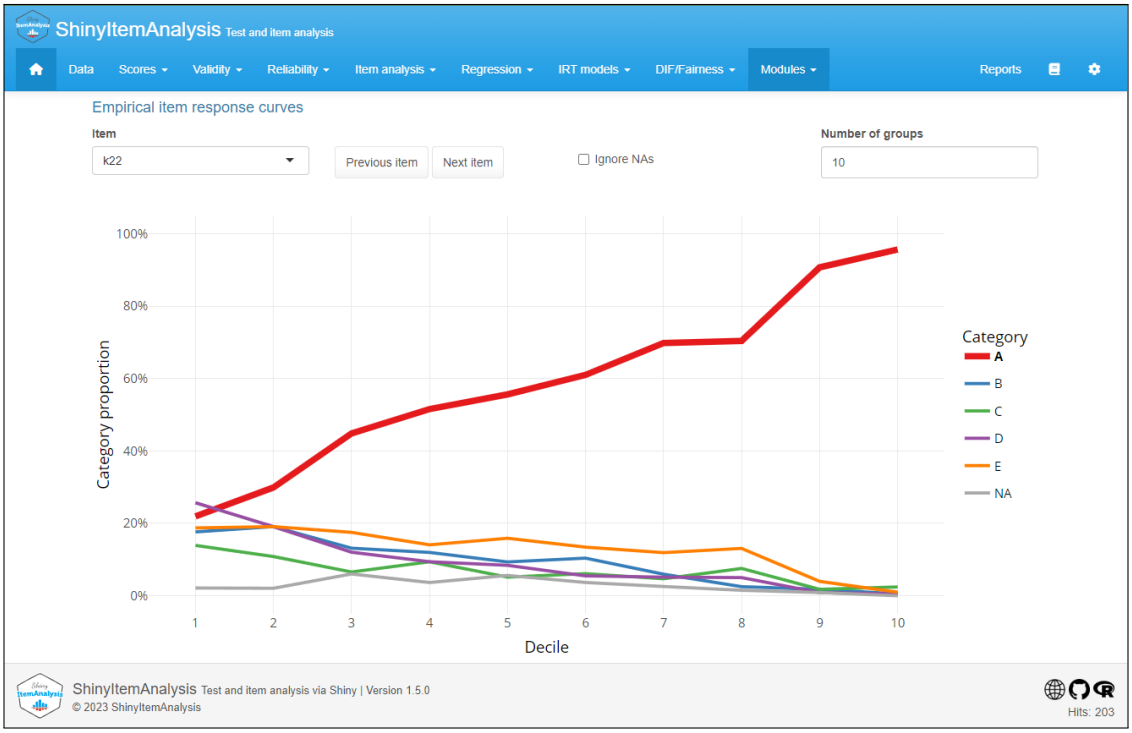
Obrázek 6: Graf závislosti diskriminačního indexu na obtížnosti položek

Podrobnější náhled na fungování jednotlivých položek poskytují grafy tzv. empirických charakteristických křivek (obrázek 7). Správná odpověď (nebo plný počet bodů v případě ordinálně hodnocených položek) je zde zvýrazněna tučně. U správně fungující položky je tato křivka rostoucí, tedy respondenti z vyššího decilu (dle celkového skóre) mají vyšší procentuální podíl správné odpovědi na položku. Naopak empirické charakteristické funkce nesprávných odpovědí jsou typicky klesající, tedy čím zdatnější skupina studentů, tím menší má podíl nesprávných nebo chybějících odpovědí.

U binárních položek, v nichž se rozlišuje pouze správná a nesprávná odpověď, vykresluje graf jednu křivku pro správnou odpověď (tučná červená čára), jednu křivku pro nesprávnou odpověď (slabší barevná čára) a jednu křivku pro chybějící odpověď (slabší šedá čára). Chybějící odpovědi je možné z analýzy vyloučit zaškrtnutím možnosti „Ignore NAs“. U ordinálních položek je navíc vykreslena i křivka pro částečně správnou odpověď, u položek s výběrem odpovědi (*multiple-choice*) jsou vykresleny také křivky pro jednotlivé nesprávné odpovědi (tzv. distraktory). Jejich průběh může pomoci identifikovat nedostatky ve formulaci testové položky.

U položky b22 (obrázek 7) vidíme, že správná odpověď A má rostoucí tendenci, tedy žáci s celkovým skóre z vyššího decilu volili správnou odpověď častěji než žáci z nižšího decilu. Naopak distraktory mají, dle očekávání, klesající tendenci, stejně jako chybějící odpovědi.

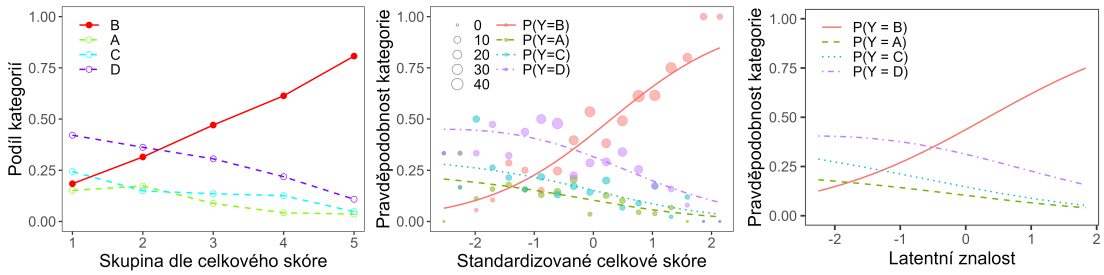
Více možností pro zobrazení empirických charakteristických funkcí položek (např. zobrazení správné kombinace správných odpovědí v případě tzv. multiple true-false položek) poskytuje hlavní aplikace `ShinyItemAnalysis` v záložce Item Analysis.



Obrázek 7: Relativní četnosti jednotlivých odpovědí na položku 22 v rámci decilů celkového počtu bodů a základní statistiky položky

4 Modelování fungování položek

Než si představíme IRT modely, podívejme se nejprve detailněji na možnosti modelování fungování položek pomocí regresních modelů. V předchozí sekci jsme popsali metody **tradiční položkové analýzy**, které využívají k popisu fungování položky podíly a procenta (obrázek 8a.).



a. Tradiční položková analýza b. Regresní model c. IRT model

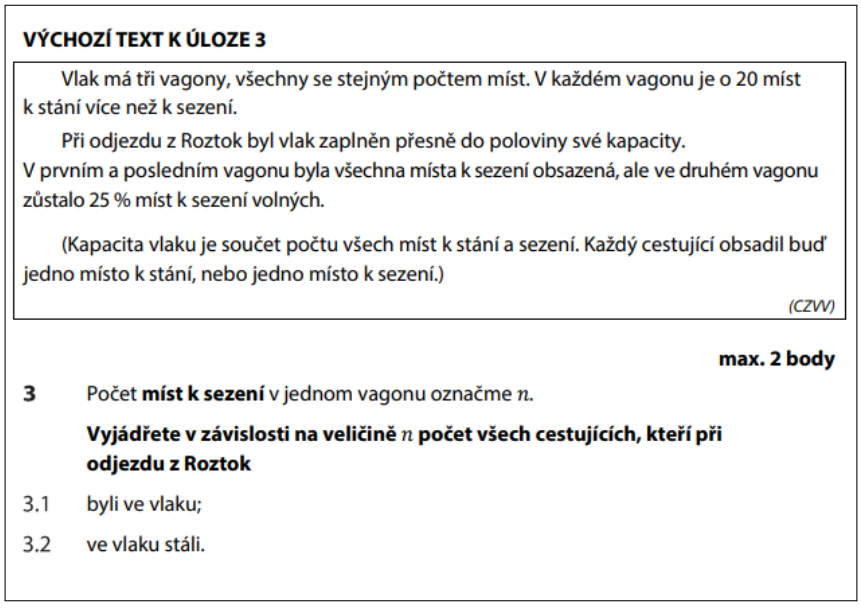
Obrázek 8: Ilustrace různých přístupů k položkové analýze na příkladu stejné položky.

Naproti tomu **regresní modely** (obrázek 8b.) modelují pravděpodobnost dané (např. správné) odpovědi na položku hladkou křivkou, která prokládá empirické hodnoty. Tyto hladké křivky, často zvané charakteristické křivky položek, poskytují infor-

maci o pravděpodobnosti jednotlivých kategorií odpovědi (např. správná vs. nesprávná odpověď v binárních položkách, volba jednotlivých variant v nominálních položkách s výběrem odpovědi) nikoliv pro skupiny respondentů, které jsou zpravidla definovány poměrně hrubě, ale pro všechny úrovně celkového skóre. Charakteristiky položky (její obtížnost, diskriminace, ale také uhádnutelnost nebo nepozornost, viz Drabinová & Martinková, 2017; Hladká & Martinková, 2020) jsou reprezentovány parametry její charakteristické křivky (poloha inflexního bodu, sklon v inflexním bodu, dolní a horní asymptota). Tyto parametry mohou být regresními modely odhadovány pro každou položku zvlášť.

Regresní modely nabízí širokou škálu nástrojů k popisu položek, ať už binárních, ordinálních nebo nominálních. Jelikož jsou parametry odhadovány pro každou položku zvlášť, umožňují regresní modely kombinovat různé typy položek. Jednotlivé regresní modely jsou v rámci `ShinyItemAnalysis` k dispozici v záložce Regression. Níže si představíme příklady možných analýz.

Položku 3.1 (obrázek 9) analyzujeme pomocí modelu logistické regrese na základě standardizovaného celkového skóre. Tato metoda modeluje pravděpodobnost správné odpovědi na danou položku na základě standardizovaného skóre, které má průměr 0 a směrodatnou odchylku 1. Využíváme při tom záložku Regression (obrázek 10) v aplikaci `ShinyItemAnalysis`, do které jsme pomocí tlačítka Pass data to SIA poslali data načtená v modulu EduTest Item Analysis .



VÝCHOZÍ TEXT K ÚLOZE 3

Vlak má tři vagony, všechny se stejným počtem míst. V každém vagonu je o 20 míst k stání více než k sezení.

Při odjezdu z Roztok byl vlak zaplněn přesně do poloviny své kapacity. V prvním a posledním vagonu byla všechna místa k sezení obsazená, ale ve druhém vagonu zůstalo 25 % míst k sezení volných.

(Kapacita vlaku je součet počtu všech míst k stání a sezení. Každý cestující obsadil buď jedno místo k stání, nebo jedno místo k sezení.)

(CZVV)

max. 2 body

3 Počet **míst k sezení** v jednom vagonu označme n.

Vyjádřete v závislosti na veličině n počet všech cestujících, kteří při odjezdu z Roztok

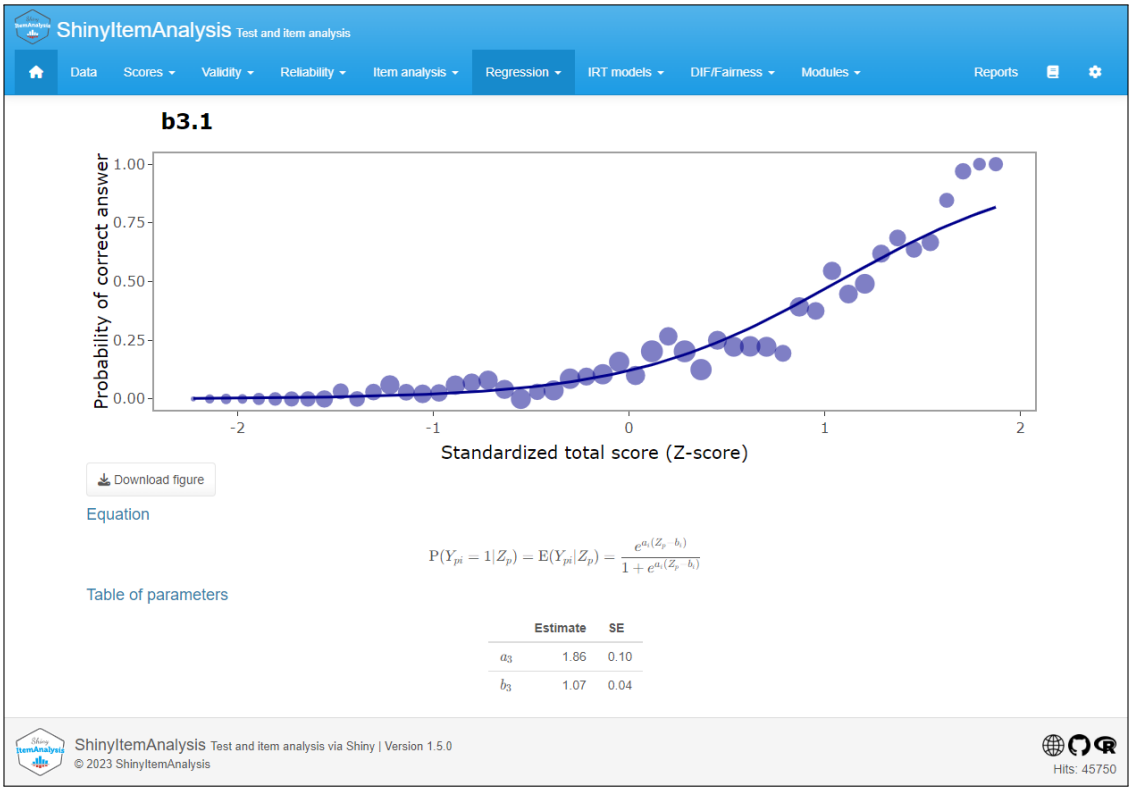
3.1 byli ve vlaku;

3.2 ve vlaku stáli.

Obrázek 9: Zadání položek 3.1 a 3.2

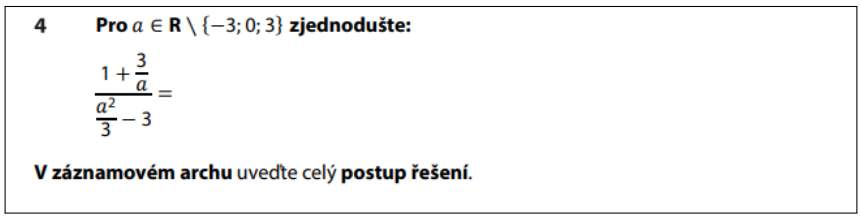
Jedná se o obtížnou položku, která dobře diskriminuje, jak bylo patrné už z tradiční položkové analýzy (obrázek 5). V položkové analýze pomocí regresních modelů

jsou obtížnost a diskriminace položky reprezentovány příslušnými parametry její charakteristické křivky, které aplikace uvádí v tabulce parametrů (ve spodní části obrázku 10). Např. parametr $b = 1{,}07$ u položky b3.1 lze interpretovat následovně: Celkové skóre 1,07 směrodatné odchylky **nad** průměrem je potřeba k tomu, aby žák/žákyně odpověděl/a na tuto položku správně s pravděpodobností 50 %. Jinými slovy, parametr b udává obtížnost položky, polohu tzv. inflexního bodu zobrazené charakteristické křivky položky. Parametr $a = 1{,}86$ pak značí sklon křivky v tomto inflexním bodě a reprezentuje diskriminační schopnost položky.

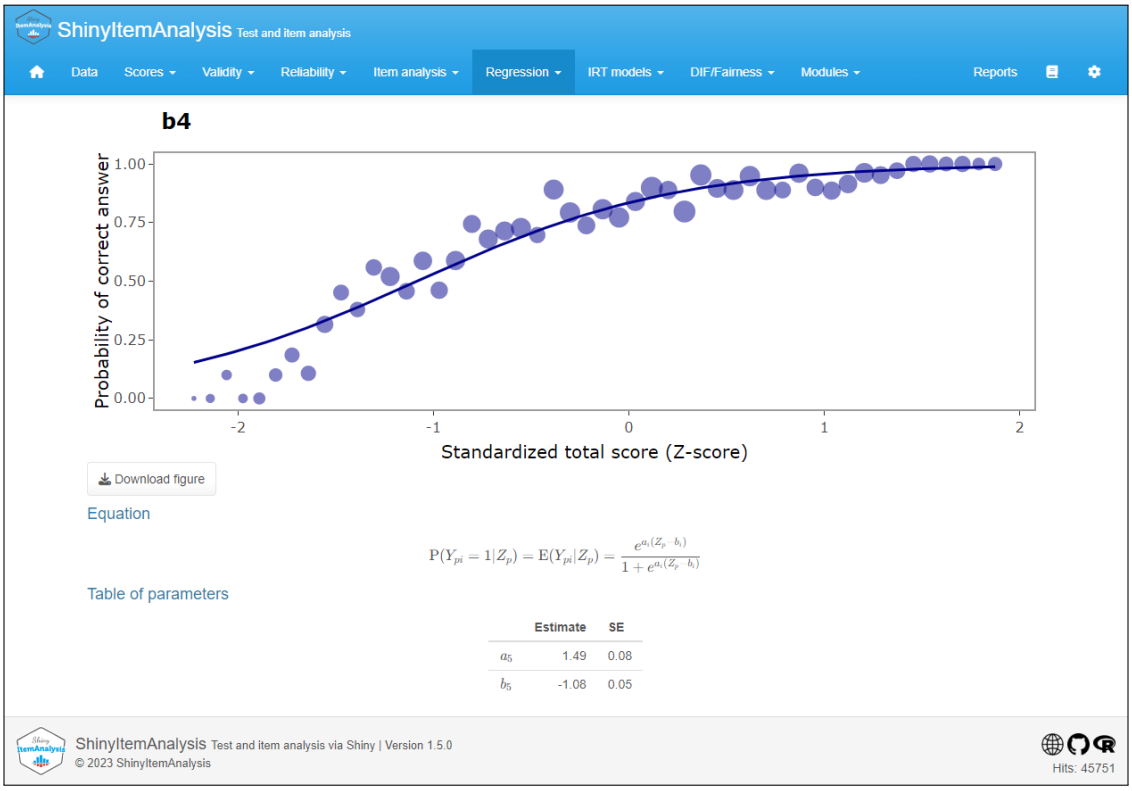


Obrázek 10: Logistický regresní model pro odhad pravděpodobnosti správné odpovědi na položku 3.1 na základě standardizovaného skóre

Položku 4 (obrázek 11) analyzujeme opět pomocí modelu logistické regrese na základě standardizovaného celkového skóre s využitím záložky Regression (obrázek 12). Tato úloha byla naopak jednoduchá. Parametr $b = -1{,}08$ značí, že celkové skóre 1,08 standardní odchylky **pod** průměrem stačí k tomu, aby žák/žákyně odpověděl/a tuto položku správně s pravděpodobností 50 %. Parametr $a = 1{,}49$ značí opět sklon křivky v tomto inflexním bodě, neboli diskriminační schopnost položky, která je i v tomto případě vysoká.



Obrázek 11: Zadání položky 4



Obrázek 12: Logistický regresní model pro odhad pravděpodobnosti správné odpovědi na položku 4 na základě standardizovaného skóre

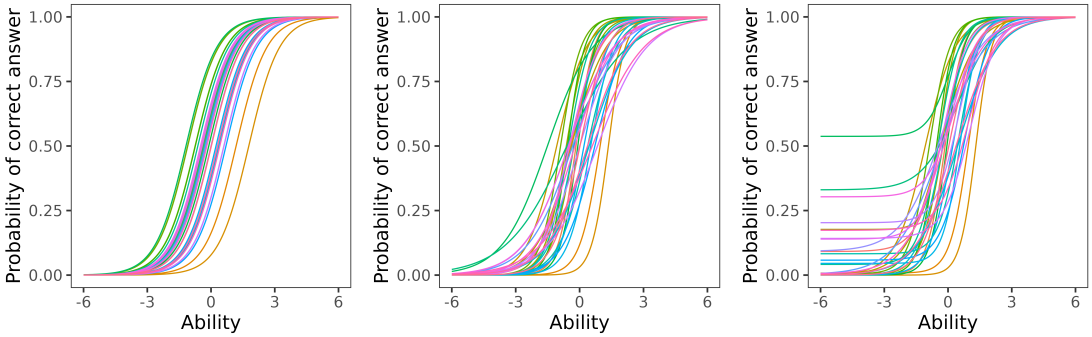
5 Modely teorie odpovědi na položku

Modely teorie odpovědi na položku (IRT; obrázek 8c.) na rozdíl od výše popsaných regresních modelů navíc předpokládají, že úroveň znalosti respondentů je neznámá a je třeba ji odhadnout společně s parametry položek. Parametry položek (obtížnost, diskriminace, uhádnutelnost nebo nepozornost) jsou nyní odhadovány pro všechny položky současně.

Podobně jako výše popsané regresní modely, IRT modely zahrnují širokou škálu nástrojů pro binární, ordinální i nominální položky. V rámci `ShinyItemAnalysis` jsou IRT modely k dispozici v záložce IRT models. Aplikace v současné chvíli nabízí pět modelů pro binární položky (Raschův model, 1–4 parametrické logistické (PL) IRT modely)

a pro nominální (*multiple-choice*) položky tzv. Nominal Response Model (Bock, 1972; Martinková & Hladká, 2023).

Raschův a 1PL model (Rasch, 1960, obrázek 13a.) jsou vhodné v případě, že u jednotlivých testových položek očekáváme rozdílné obtížnosti, ale stejnou diskriminaci. V Raschově modelu je diskriminace fixována na hodnotu 1, v 1PL modelu je odhadována z dat pro všechny položky současně, je však fixován rozptyl abilit respondentů na 1. 2PL model (Birnbaum, 1968, obrázek 13b.) odhaduje hodnotu diskriminace pro každou položku. 3PL model (Birnbaum, 1968, obrázek 13c.) navíc umožňuje odhadnout uhádnutelnost správné odpovědi, tedy pravděpodobnost, že i respondent s velmi nízkou úrovní znalosti odpoví na položku správně. 4PL model (Barton & Lord, 1981) pak navíc počítá s nepozorností respondentů, tedy může omezit pravděpodobnost správné odpovědi pro respondenty s vysokou úrovní znalosti na hodnotu nižší než 1. V rámci hlavní aplikace není možné různé typy položek kombinovat, v rámci modulu EduTest Item Analysis to však možné je, jak si ukážeme v části 5.1.



a. 1PL IRT model b. 2PL IRT model c. 3PL IRT model

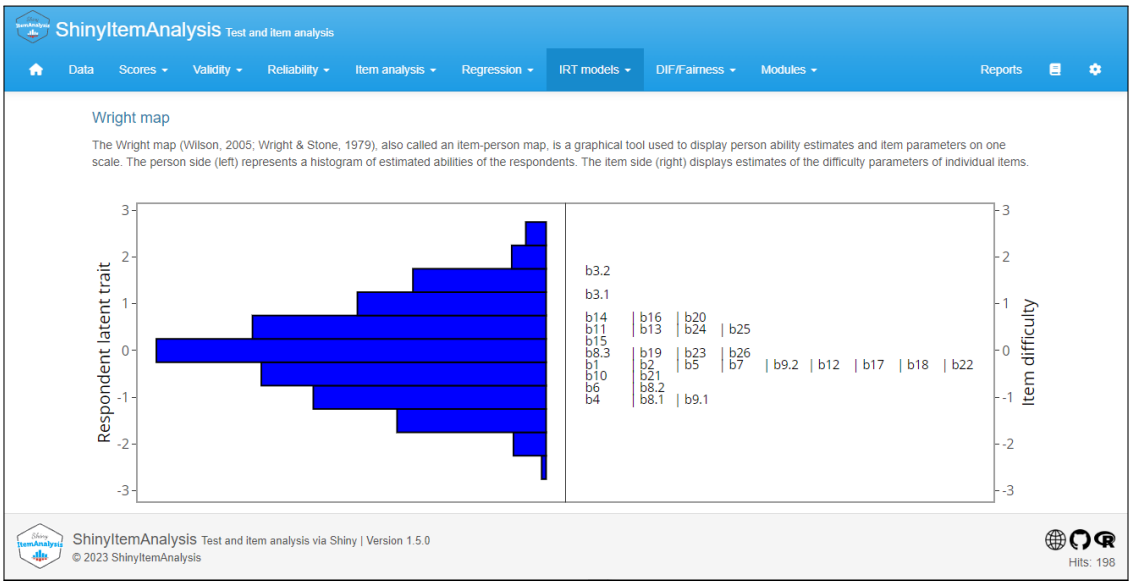
Obrázek 13: Ilustrace různých IRT modelů.

Níže si popíšeme různé funkcionality a příklady využití IRT modelů, které demonstrujeme na maturitním testu z matematiky.

Raschův a 1PL IRT model vyjadřují vztah mezi znalostí respondenta a pravděpodobností určité odpovědi na položku, která závisí také na obtížnosti dané položky. Model pro každou položku odhaduje jeden parametr b, který vyjadřuje její obtížnost. Vztah mezi znalostí respondenta a pravděpodobností odpovědi na položku graficky znázorňují charakteristické křivky položek (obrázek 13a.). Z charakteristických křivek položek lze názorně vyčíst, která položka je nejsnazší (křivka zcela vlevo) a která nejtěžší (křivka zcela vpravo).

Raschův a 1PL model umožňují také interpretovat obtížnost položek vzhledem k odhadnuté úrovni znalostí pomocí tzv. Wrightova grafu (*Wright map*). Ten v levé části

zobrazuje histogram odhadnutých latentních znalostí, v pravé části pak odhadnuté obtížnosti jednotlivých položek. To nám umožňuje ověřit, zda je v testu obsaženo dostatečné množství položek pro ty úrovně latentní znalosti, které jsou v daném kontextu důležité. Pro data z maturitního testu z matematiky analyzovaná pomocí 1PL modelu histogram v levé části Wrightova grafu (obrázek 14) naznačuje nejčastěji odhadnutou latentní znalost matematiky okolo 0, resp. mírně pod nulou. Tomu odpovídá i rozložení obtížností jednotlivých položek testu, které mají obtížnost nejčastěji v tomto rozmezí.

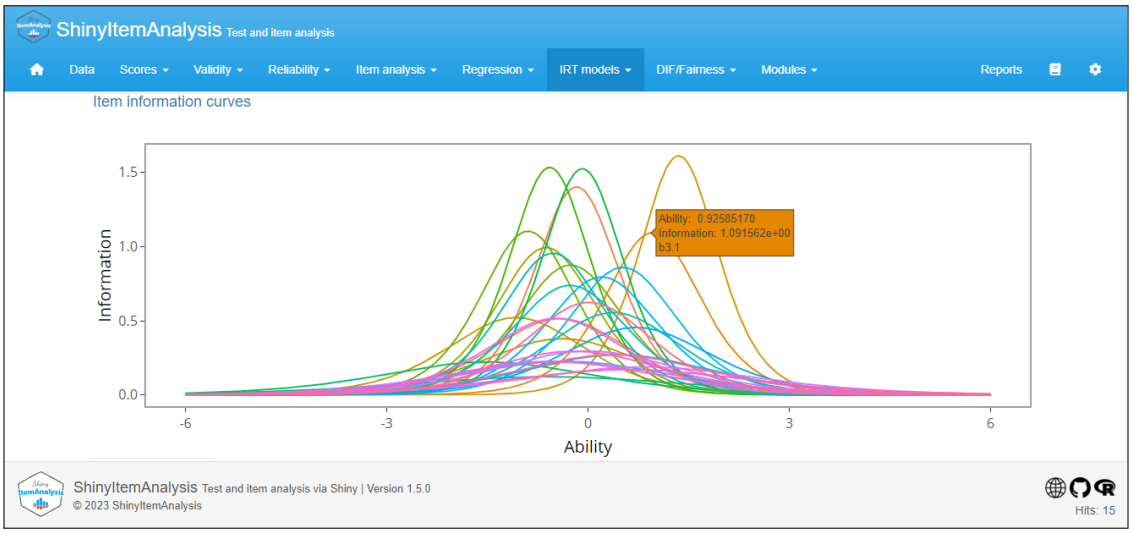


Obrázek 14: Wrightův graf

2PL (dvouparametrický logistický) IRT model se od Raschova a 1PL modelu odlišuje tím, že pro vyjádření vztahu mezi znalostí respondenta a jeho odpovědí na položku zavádí vedle obtížnosti položky také druhý parametr a, který představuje diskriminační schopnost položky. Charakteristické křivky položek mají nejen různé posunutí, ale mohou mít také různý sklon (obrázek 13b.). Čím je křivka strmější, tím lépe daná položka rozlišuje mezi respondenty s nižší a vyšší úrovní znalosti. Pro analýzu dat ze znalostních testů představuje 2PL model obvykle realističtější vyjádření skutečnosti, protože různé položky jsou v praxi zpravidla různě účinné z hlediska své schopnosti rozlišovat mezi méně a více zdatnými respondenty.

Z charakteristických křivek lze matematicky odvodit tzv. informační funkce jednotlivých položek. Informační funkce položek ukazují, jakou míru informace jednotlivé položky nesou pro hodnocení různých úrovní latentní znalosti. Informační funkce mají maximum v blízkosti hodnoty parametru obtížnosti položky a s rostoucí vzdáleností od této hodnoty na obě strany množství informace poskytované položkou klesá. Konkrétní

tvar informační funkce je závislý na parametru diskriminace položky. Při použití 1PL modelu mají informační funkce všech položek stejný tvar a liší se pouze v posunutí na ose x, při použití 2PL modelu mají různé tvary a různá maxima (obrázek 15).



Obrázek 15: Informační křivky položek modelované pomocí 2PL modelu

Pohled na informační funkce položek pak umožňuje posoudit, nakolik jsou jednotlivé položky přínosné pro měření požadovaných znalostí. Na obrázku 15 je například vidět, že položka 3.1 z maturitního testu z matematiky analyzovaného pomocí 2PL modelu je velmi informativní pro hodnocení znalostí nadprůměrných respondentů, ale pro měření znalostí podprůměrných respondentů je její informační přínos prakticky nulový. O znalostech slabších žáků přinášejí více informací jiné položky z testu. Test neobsahuje žádné položky, které by nesly velké množství informací o znalostech velmi slabých žáků (např. s odhadovanou úrovní znalostí menší než -2). Při tvorbě jakéhokoliv znalostního testu je obecně žádoucí, aby obsahoval položky, které jsou dostatečně informativní pro očekávané úrovně znalostí respondentů, resp. pro takové úrovně znalostí, které by měl měřit nejpřesněji. Např. v maturitním testu je důležité, aby dobře měřil znalosti v okolí hranice úspěšnosti a v případě známkování dosažených výsledků také nad ní. Pod touto hranicí už nemusí být příliš informativní, protože účelem testu není detailně zhodnotit znalosti velmi slabých žáků, kteří nedosáhnou hranice úspěšnosti. Informační funkce lze dále využít např. v počítačovém adaptivním testování (podrobněji viz příslušná část v tomto Souboru postupů a nástrojů) či ve vyvažování verzí testů (tzv. *equating*).

Sečtením informačních funkcí jednotlivých položek získáme informační funkci celého testu, z které lze odvodit, jaký je informační přínos testu pro měření různých úrovní znalostí. Maturitní test z matematiky je nejvíce informativní pro znalost těsně

pod 0 (obrázek 16).

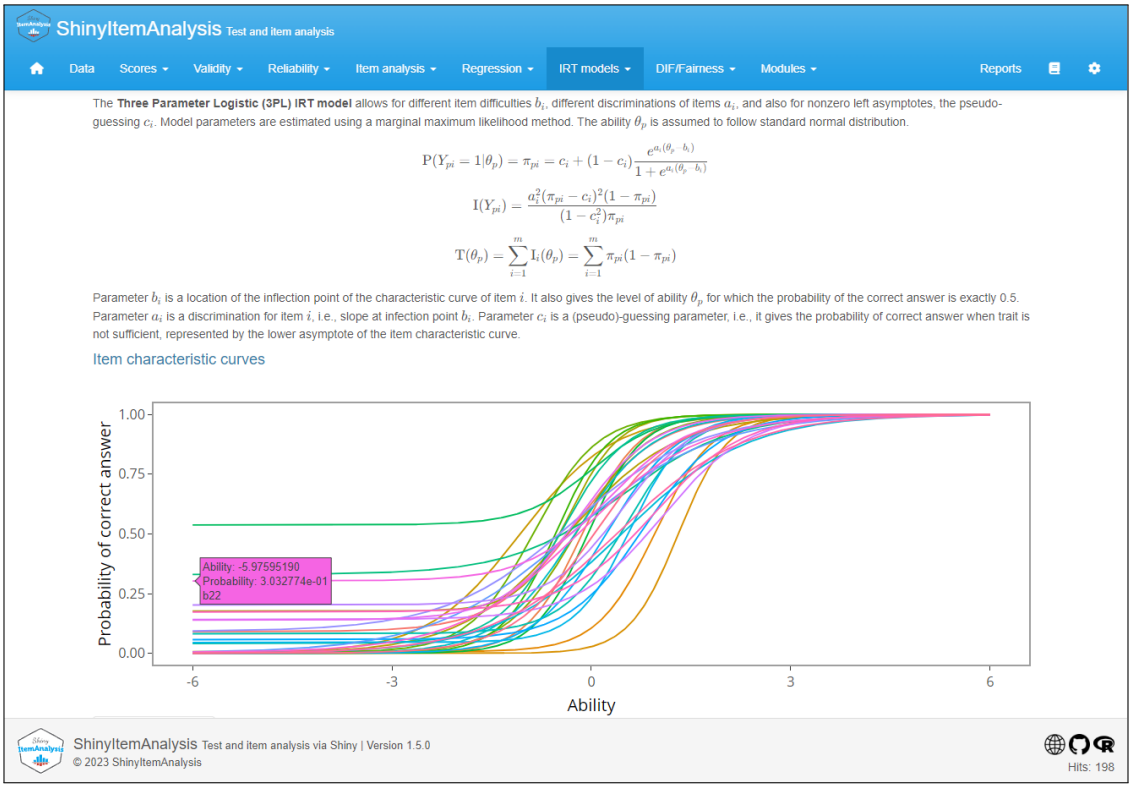
	Total score	Z-score	F-score	SE(F-score)
Respondent 1	31.00	0.28	0.44	0.27
Respondent 2	30.00	0.20	0.37	0.27
Respondent 3	33.00	0.45	0.01	0.25
Respondent 4	28.00	0.03	-0.03	0.25
Respondent 5	26.00	-0.13	-0.10	0.25
Respondent 6	23.00	-0.38	-0.67	0.26

Obrázek 16: Informační křivka testu modelovaná pomocí 2PL IRT modelu

Z informační funkce testu můžeme dále vypočítat standardní chybu měření, která je nepřímo úměrná informaci (čím více informací test přináší, tím menší je chyba měření). Protože informační funkce není pro daný test konstantní, ale nabývá různých hodnot pro různé úrovně znalostí, není ani chyba měření pro všechny úrovně znalostí stejná. Typicky pro latentní znalost okolo průměru je chyba měření nejmenší, kdežto pro okrajové hodnoty je chyba měření největší. Na rozdíl od klasické testové teorie tak IRT modely umožňují určit, s jakou přesností test měří znalosti jednotlivých respondentů, kteří dosáhli různých výsledků. To je důležité kritérium pro posouzení kvality testu a jeho vhodnosti pro daný účel a danou populaci. Aplikace `ShinyItemAnalysis` zobrazuje chybu měření v jednom grafu spolu s informační funkcí testu (obrázek 16), což usnadňuje posouzení kvality testu. Navíc nabízí pro každého respondenta přepočet jeho celkového bodového skóre na standardizované skóre a na latentní skóre odhadnuté pomocí IRT modelu společně s odhadnutou chybou měření. Hodnoty všech respondentů lze z aplikace stáhnout v přehledné tabulce. Např. v maturitním testu z matematiky měl respondent 1 celkové skóre 31 bodů, což odpovídá standardizovanému skóre 0,28 (tj.

0,28 směrodatné odchylky nad průměrem). Jeho latentní znalost odhadnutá pomocí 2PL modelu je pak 0,33 (SE = 0,29). Respondent 4 měl celkové skóre 28 bodů, což odpovídá standardizovanému skóre 0,03 (tj. 0,03 směrodatné odchylky nad průměrem). Odhadnutá latentní znalost je pak 0,00 (SE = 0,28).

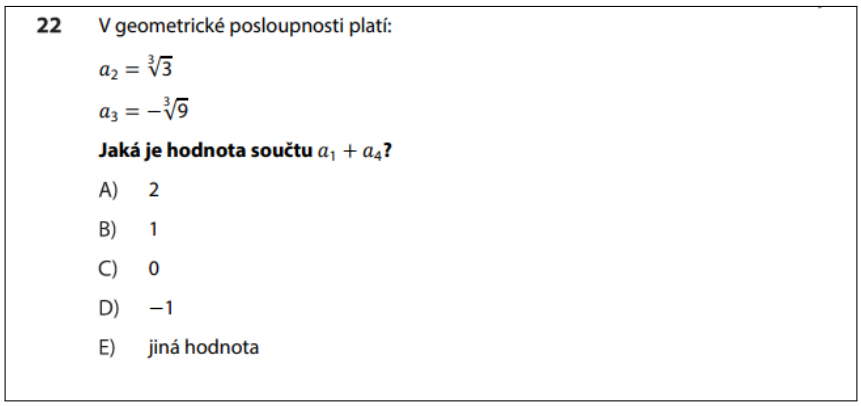
3PL (tříparametrický logistický) IRT model zohledňuje vedle obtížnosti a diskriminace také pravděpodobnost uhádnutí položky (parametr c). Charakteristické křivky položek mají nejen různé posunutí a různý sklon, ale mohou mít také hodnotu dolní asymptoty vyšší než nula (obrázky 13c. a 17), pokud správnou odpověď na danou položku volí i respondenti s velmi nízkými znalostmi. 3PL model je relevantní pro analýzu dat z testů, které reálně umožňují uhodnutí správných odpovědí, tedy typicky testů složených z položek s výběrem odpovědi. Umožňuje posoudit kvalitu položek s ohledem na jejich uhádnutelnost. V praxi však může 3PL model narážet na problémy s výpočetní kapacitou. Zároveň parametr uhádnutelnosti mění definici parametru obtížnosti, což komplikuje jeho interpretaci.



Obrázek 17: 3PL IRT model se zvýrazněnou položkou b22

Příklad využití 3PL modelu si ukážeme na položce 22 z maturitního testu z matematiky. Úloha 22 (obrázek 18) nabízela pět možných odpovědí. V tomto případě bychom tedy očekávali hodnoty parametru c okolo 0,2. Uhádnutelnost ovšem záleží také na atrak-

tivitě nabízených distraktorů (nesprávných odpovědí). V této úloze byla uhádnutelnost odhadnuta jako $c = 0{,}3$, tedy o něco vyšší (obrázek 17). V případě, kdy modelujeme uhádnutelnost, se mění interpretace parametru obtížnosti b. Nyní odpovídá latentní znalosti, která je potřebná k správnému zodpovězení dané položky s pravděpodobností mezi spodní asymptotou definovanou parametrem c a horní asymptotou, tj. 1, tedy nikoliv 50 % jako u 1PL a 2PL modelů. V této položce se jedná o pravděpodobnost rovnou $^{(0,3+1)}/_{2} = 0{,}65$, tj. 65 %. Tomu pak odpovídá latentní znalost a parametr $b = 0{,}26$. Jedná se opět o inflexní bod charakteristické křivky. Interpretace diskriminačního parametru a zůstává stejná, tj. sklon křivky v inflexním bodě.


22 V geometrické posloupnosti platí:

$a_2 = \sqrt[3]{3}$

$a_3 = -\sqrt[3]{9}$

Jaká je hodnota součtu $a_1 + a_4$?

A) 2

B) 1

C) 0

D) −1

E) jiná hodnota

Obrázek 18: Zadání položky 22

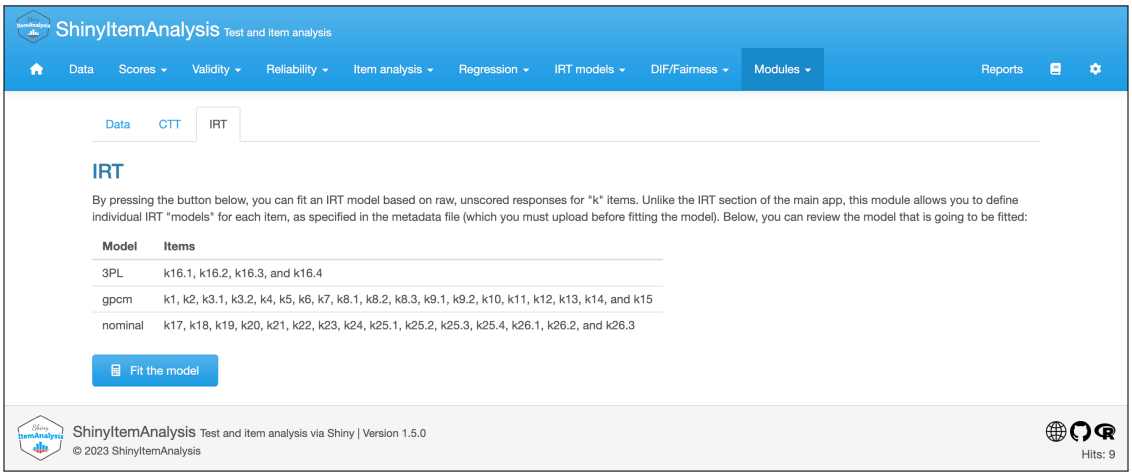
Hlavním rozdílem IRT modelů oproti výše představeným regresním modelům je to, že kromě parametrů jednotlivých položek odhadují také latentní znalost. Zatímco v regresním přístupu je typicky každá položka modelována samostatně, IRT přístup zpracovává všechny položky současně, což spolu s odhadem latentní znalosti může být výpočetně náročné. Pro získání správných hodnot odhadů parametrů také IRT modely vyžadují dostatečné množství pozorování (respondentů), např. více než 500 pro 3PL model. Na druhou stranu však lépe pracují se statistickou nejistotou spojenou s analýzou dat. Díky tomu, že odhadují také latentní znalost, poskytují IRT modely i pro odhad latentní znalosti konfidenční interval, a to i pro každého respondenta. Zároveň umožňují matematicky vyjádřit skutečnost, že reliabilita testu není konstantní, ale je funkcí znalosti respondenta. IRT modely lze navíc využít i k vzájemnému propojení více verzí testu (např. variant A a B nebo testů z různých zkušebních termínů), neboť v případě existence tzv. kotvících položek (položek zadaných ve více verzích testu) umožňují získat odhady parametrů z různých verzí testu na společné škále. Aplikace `ShinyItemAnalysis` usnadňuje použití IRT modelů prostřednictvím jednoduše ovládaného uživatelského rozhraní a zároveň poskytuje základní interpretaci získaných výsledků. Podrobnější seznámení s principy IRT modelování včetně požadavků na data nabízejí např. Martinková a Hladká

(2023).

5.1 Položkově specifický IRT model

Podobně jako výše popsané regresní modely nabízí i IRT přístup možnost volit různé modely pro různé položky. Oproti regresním modelům odhaduje tzv. položkově specifický IRT model parametry všech položek najednou společně s latentní znalostí respondentů. Odhadnutá latentní znalost tedy odráží i to, že položky jsou různého typu, a může být tudíž přesnější než celkové skóre či jeho standardizovaná hodnota, které se používají u regresních modelů.

Možnost kombinovat typy položek pomocí položkově specifického IRT modelu je k dispozici v rámci modulu EduTest Item Analysis, kde se typ položky určí automaticky z nahrávaných metadat (obrázek 2). Využívají se zde nejen základní modely pro binární položky, ale také složitější modely pro ordinální a nominální položky. IRT analýza je dostupná pod podzáložkou IRT. Položkově specifický IRT model pak pro ordinálně bodované položky využije tzv. zobecněný model částečného kreditu (*Generalized Partial Credit Model*), pro položky s výběrem odpovědi model nominálních odpovědí (*Nominal Response Model*) a pro binární položky s možností hádání využije 3PL model (obrázek 19). Pro provedení IRT analýzy je nutno zmáčknout tlačítko „Fit the model“.

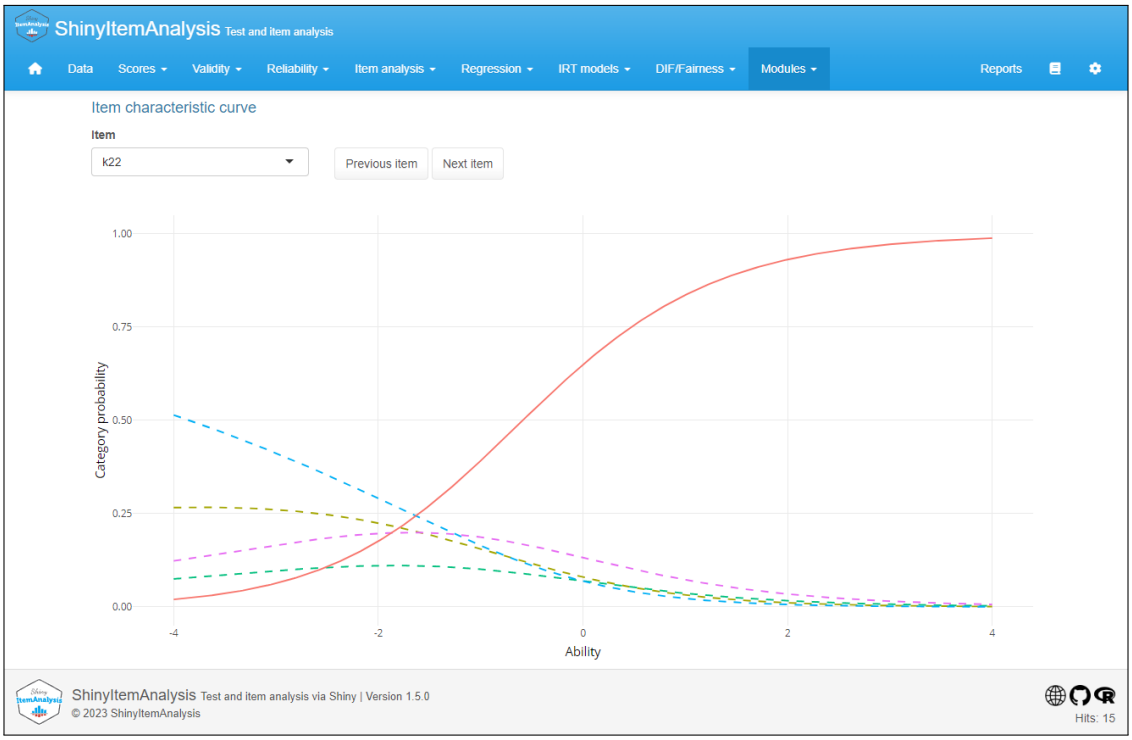


Obrázek 19: Rozdělení položek podle typů IRT modelů

Modul EduTest Item Analysis pro položkově specifický IRT model nabízí graf informační křivky celého testu, graf standardní chyby měření latentní znalosti a reliability testu v závislosti na hodnotě latentní znalosti. Dále je pak možné zobrazit charakteristické a informační křivky jednotlivých položek.

Ukázku analýzy dat z maturitního testu z matematiky pomocí položkově specific-

kého IRT modelu v modulu EduTest Item Analysis představíme opět na položce 22 (obrázek 18), která nabízela pět možných odpovědí. Pro analýzu této položky s výběrem odpovědi je nyní využit *Nominal Response Model* (obrázek 20), který modeluje nejen pravděpodobnost správné odpovědi, ale i pravděpodobnost volby jednotlivých distraktorů. Díky tomu může např. tvůrce testu detailněji vyhodnotit jejich atraktivitu pro respondenty s různou úrovní latentní znalosti. Vidíme, že např. pro průměrného respondenta (odhadnutá latentní znalost blízká 0), byla pravděpodobnost volby správné odpovědi A 65 %. Pravděpodobnosti voleb jednotlivých distraktorů pro respondenty s průměrnou úrovní latentní znalosti pak byly 13 % pro možnost E, 8 % pro možnost B a 7 % pro každou z možností C a D.

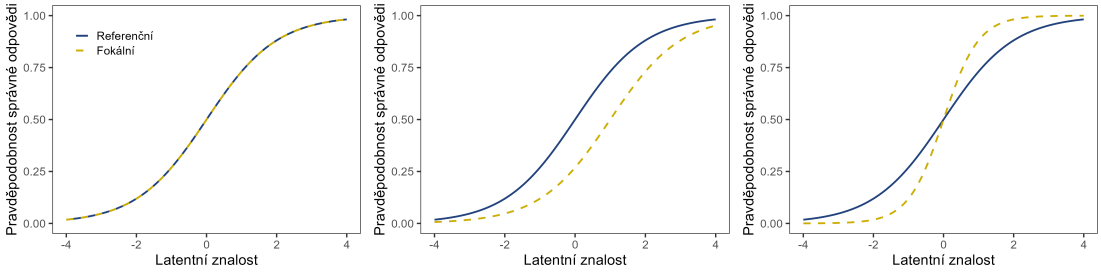


Obrázek 20: Charakteristické křivky správné odpovědi a distraktorů pro položku 22

6 Odlišné fungování položek

Odlišné fungování položek (*Differential Item Functioning*, DIF) je situace, kdy respondenti se stejnou znalostí, ale z jiných skupin mají různou pravděpodobnost správné odpovědi na položku. V nejjednodušším případě rozlišujeme 2 skupiny: referenční (typicky většina) a fokální (typicky menšina nebo znevýhodněná skupina). DIF dále klasifikujeme jako uniformní (obrázek 21b.), pokud má jedna skupina vyšší pravděpodobnost správné

odpovědi než druhá skupina pro všechny úrovně latentní znalosti, a neuniformní (obrázek 21c.), pokud má pro část úrovní latentní znalosti vyšší pravděpodobnost správné odpovědi jedna skupina a pro ostatní úrovně latentní znalosti druhá skupina. Např. na obrázku 21b. mají respondenti s latentní znalostí 0 z fokální skupiny pravděpodobnost správné odpovědi 0,5, kdežto respondenti se stejnou latentní znalostí, ale z referenční skupiny mají pravděpodobnost správné odpovědi přibližně 0,75. V případě, že by položka nefungovala odlišně, byla by pravděpodobnost obou skupin shodná pro všechny úrovně latentní znalosti (obrázek 21a.).



a. Žadný DIF b. Uniformní DIF c. Neuniformní DIF

Obrázek 21: Ilustrace rozdílného fungování položek

Pokud položka funguje odlišně, může být potenciálně neférová. Neférová je taková položka, která zvýhodňuje jednu skupinu vůči druhé kvůli tomu, že její správné zodpovězení usnadňují určité sekundární znalosti, které však daný test nemá hodnotit. Je třeba zdůraznit, že ne všechny odlišně fungující položky jsou neférové. Může se stát, že položka funguje odlišně, ale přesto měří pouze primární (zamýšlenou) latentní znalost. Analýza odlišného fungování může pomoci vytipovat potenciálně neférové položky, a proto by měla být běžnou součástí psychometrické analýzy znalostních testů. Pro posouzení férovosti položek je nicméně vždy nutná interpretace obsahu položky odborníky na danou oblast.

Aplikace `ShinyItemAnalysis` nabízí širokou škálu metod pro detekci odlišného fungování položek. Jsou zde nabízeny klasické metody jako je tzv. Delta metoda, Mantel-Haenszelův test či metoda SIBTEST. Dále jsou k dispozici skupinově specifické regresní modely, popsané v sekci 4, které navíc zahrnují efekt skupinové proměnné a její interakci s odhadnutou znalostí (typicky standardizované testové skóre). Aplikace také nabízí metody založené na IRT modelech. V ilustracích níže jsme se zaměřili na DIF detekci pomocí skupinově specifických regresních modelů.

Typickým příkladem neférové položky je položka obsažená v americkém testu SAT (*Scholastic Aptitude Test*) z 60. let 20. století:

„Běžec je k maratonu jako“
A. vyslanec k velvyslanectví
B. mučedník k masakru
C. veslař k regatě
D. rozhodčí k turnaji
E. koně ke stájím

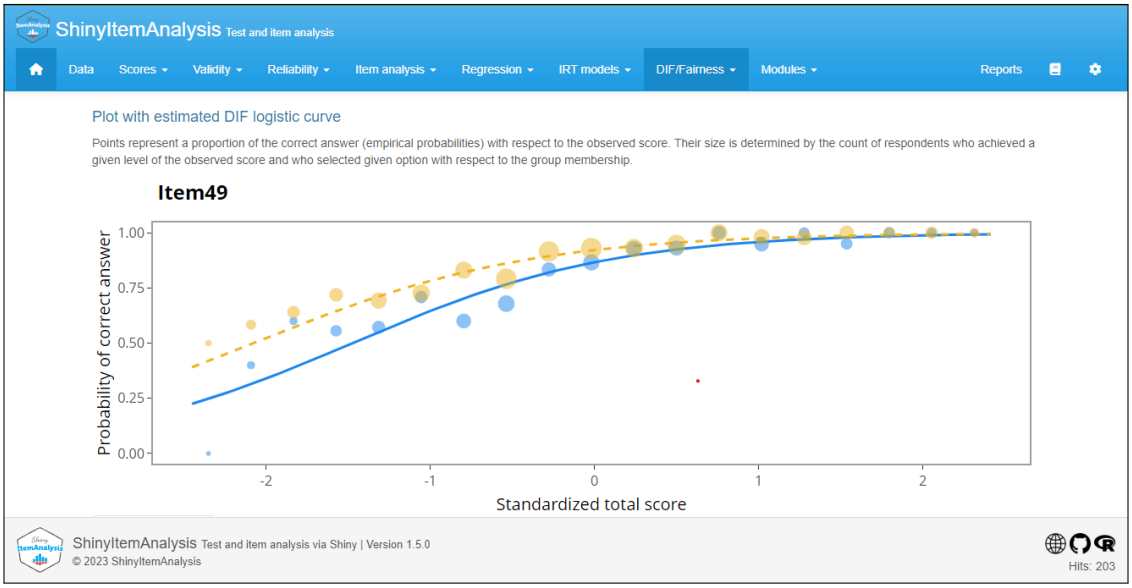
Správná odpověď byla C. veslař k regatě. V tomto případě odpovídali častěji správně bílí než afroameričtí studenti se stejnými celkovými znalostmi (což by odpovídalo ilustrativnímu znázornění na obrázku 21b., kde referenční skupina by byli bílí studenti a fokální skupinu by tvořili afroameričtí studenti). Test měl měřit verbální abstrakci, ale tato položka navíc měřila znalost sportu (regaty), který byl v dané době typickým sportem pro bílé studenty. Proto byla položka označena za neférovou a vyřazena z testu.

Jak jsme již naznačili výše, ne všechny položky, které fungují odlišně, jsou nutně neférové. Může se stát, že určitá položka je detekována pomocí DIF analýzy jako položka s odlišným fungováním, ale stále měří pouze primární (zamýšlený) latentní rys. To se může např. stát, když je položka zaměřena na specifickou oblast či dimenzi primárního latentního rysu.

Takovým příkladem může být položka, která se objevila v přijímacím testu na lékařskou fakultu:

„Nedostatek vitaminu D v dětství může způsobit“
A. křivici
B. kurděje
C. nanismus
D. mentální retardaci

Správná odpověď byla A. křivici. V tomto případě častěji správně odpověděly studentky než studenti se stejnými celkovými znalostmi (obrázek 22). Test měl měřit znalosti biologie člověka, přičemž položka testuje znalost dětských nemocí, které ke znalostem biologie člověka patří. Daná položka, ačkoliv funguje odlišně, je férová.

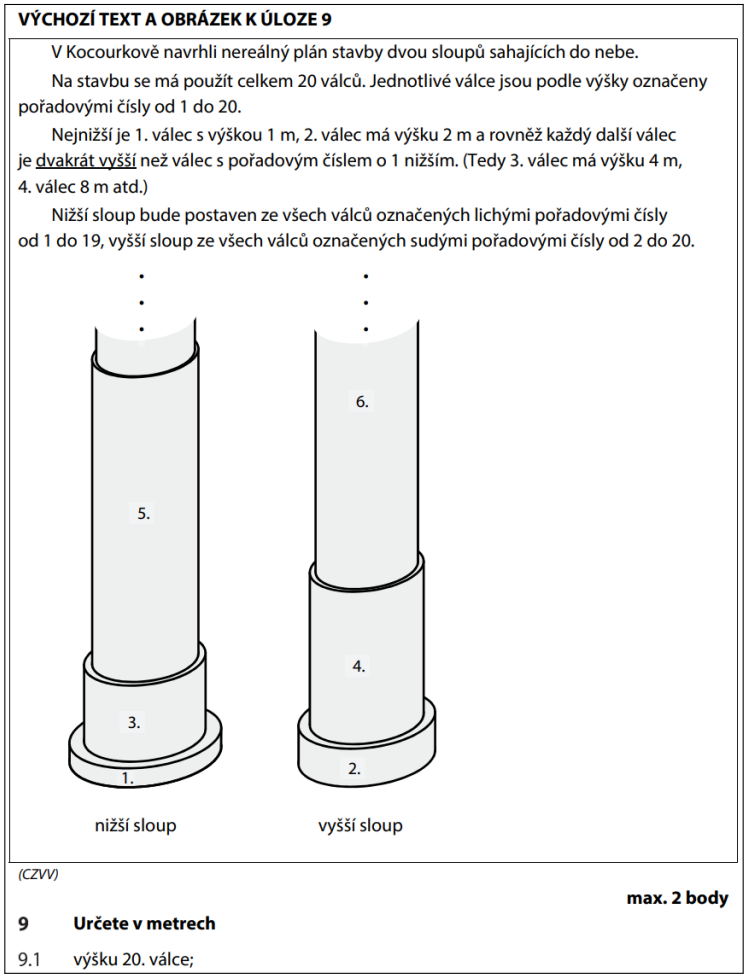


Obrázek 22: Odlišně fungující férová položka v přijímacím testu na lékařskou fakultu

Detekce odlišného fungování položek a DIF analýza nutně neslouží jen k posouzení potenciální neférovosti. Mohou sloužit také k detailnějšímu porozumění vzdělávacím výsledkům nebo k odhalení silných a slabých stránek různých skupin respondentů.

Příkladem takového využití DIF analýzy může být např. porovnávání respondentů z různých typů škol, které zde demonstrujeme na porovnání mezi žáky gymnázií a žáky středních odborných škol s technickým nebo technologickým zaměřením (obory ST1 a ST2 dle klasifikace SMO16). V maturitním testu z matematiky bylo pro tyto dvě skupiny detekováno odlišné fungování položky 9.1 (obrázek 23).

Žáci technických a technologických oborů (plná čára na obrázku 24) při stejné celkové znalosti řešili úlohu o stavbě věže snáze než žáci gymnázií (přerušovaná čára na obrázku 24). Např. žák technického či technologického oboru se standardizovaným skóre -1 měl pravděpodobnost správné odpovědi cca 75 %. Naproti tomu žák gymnázia se stejným standardizovaným skóre měl tuto pravděpodobnost jen cca 60 %. Ačkoliv položka funguje pro tyto dvě skupiny žáků odlišně, testuje pouze znalost matematiky, a je tedy férová. Poukazuje spíše na vyšší zdatnost žáků technických a technologických oborů řešit takovýto typ úloh.



VÝCHOZÍ TEXT A OBRÁZEK K ÚLOZE 9

V Kocourkově navrhli nereálný plán stavby dvou sloupů sahajících do nebe.

Na stavbu se má použít celkem 20 válců. Jednotlivé válce jsou podle výšky označeny pořadovými čísly od 1 do 20.

Nejnižší je 1. válec s výškou 1 m, 2. válec má výšku 2 m a rovněž každý další válec je dvakrát vyšší než válec s pořadovým číslem o 1 nižším. (Tedy 3. válec má výšku 4 m, 4. válec 8 m atd.)

Nižší sloup bude postaven ze všech válců označených lichými pořadovými čísly od 1 do 19, vyšší sloup ze všech válců označených sudými pořadovými čísly od 2 do 20.

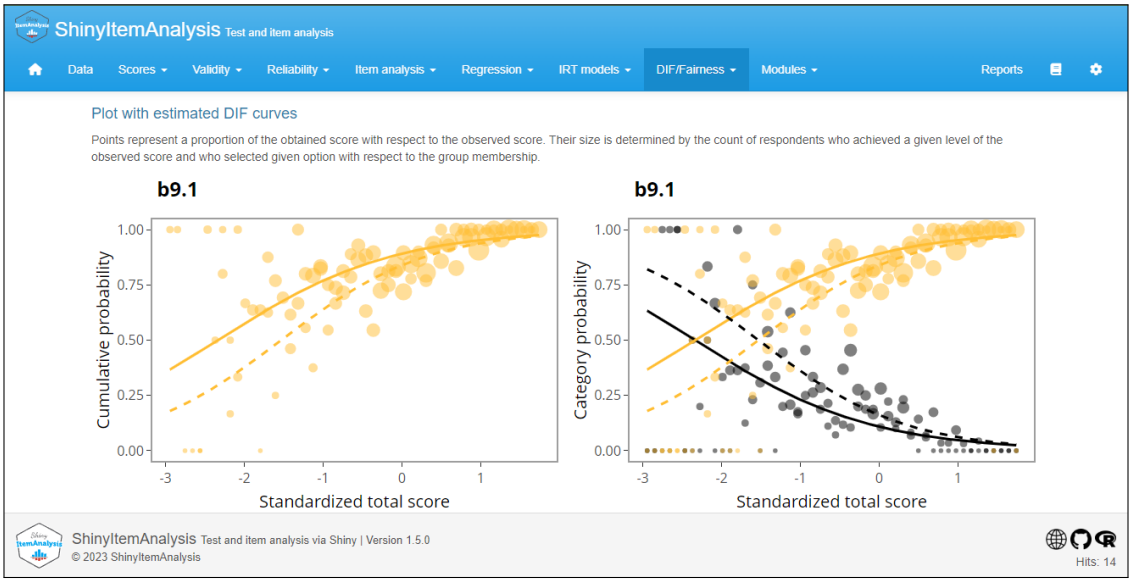
(CZVV)

max. 2 body

9 Určete v metrech

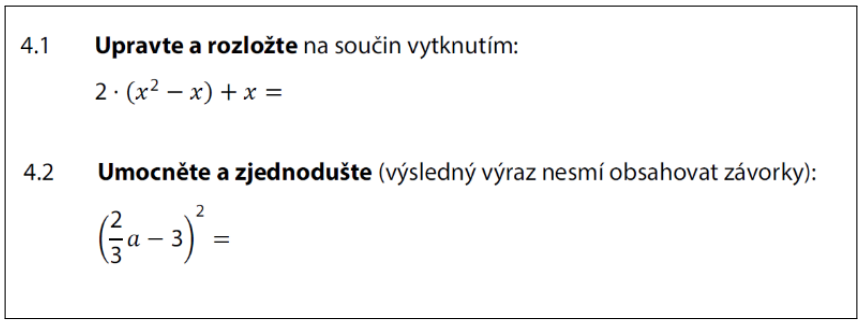
9.1 výšku 20. válce;

Obrázek 23: Zadání položky 9.1

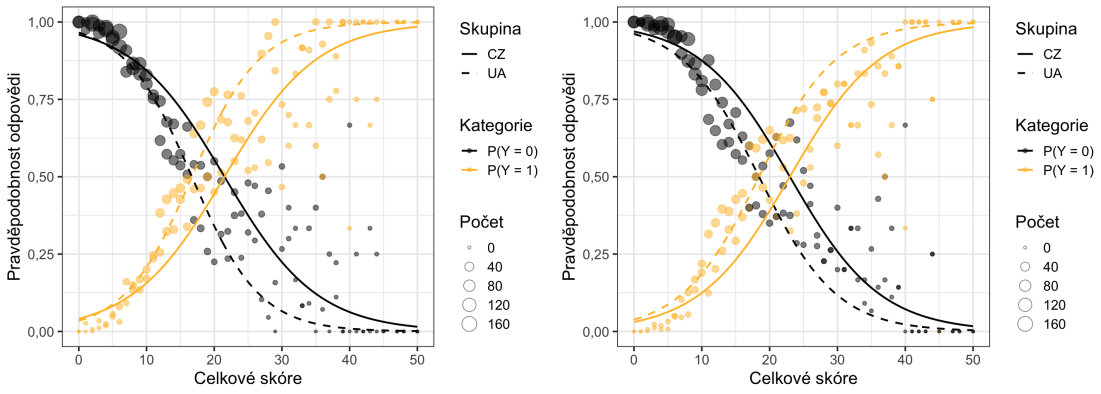


Obrázek 24: Detekce odlišného fungování pro položku 9.1

Dalším příkladem DIF detekce je analýza dat z jednotné přijímací zkoušky z matematiky z roku 2023 pro skupinu žáků, kteří ji skládali v českém jazyce (referenční skupina) a v ukrajinském jazyce (fokální skupina). V demonstraci jsou využita data z 1. řádného termínu jednotné přijímací zkoušky pro čtyřleté obory středních škol. Pro lepší srovnatelnost obou skupin jsme z referenční skupiny vybrali pouze žáky, kteří se hlásili na stejné kombinace škol jako žáci z fokální skupiny. Odlišné fungování bylo detekováno např. u úloh 4.1 a 4.2 (obrázek 25) a úlohy 9 (obrázek 27).



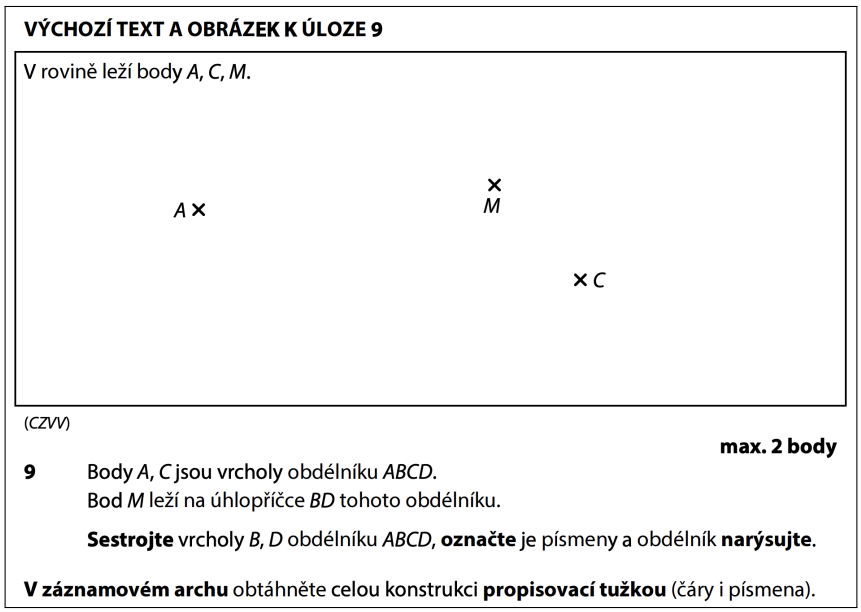
Obrázek 25: Zadání úloh 4.1 a 4.2 v jednotné přijímací zkoušce z matematiky, 2023



a. Položka 4.1 b. Položka 4.2

Obrázek 26: Příklady rozdílně fungujících položek, v nichž měli ukrajinští žáci vyšší pravděpodobnost zisku plného počtu bodů než ostatní žáci se stejným celkovým skóre

U položek 4.1 a 4.2 (obrázek 25) měli nově příchozí žáci z Ukrajiny **větší** pravděpodobnost správné odpovědi při stejném celkovém skóre než ostatní žáci (obrázek 26). Např. ukrajinský žák s celkovým skóre 30 měl pravděpodobnost správné odpovědi v úloze 4.1 cca 94 %, kdežto český žák se stejným skóre měl pravděpodobnost jen cca 80 % (obrázek 26a.). Důvodem může být např. rozdílné kurikulum matematiky na základních školách v ČR a na Ukrajině nebo minimum textu. Ačkoliv položky fungují odlišně pro české a ukrajinské žáky, obě testují pouze znalost matematiky, a jsou tedy férové.

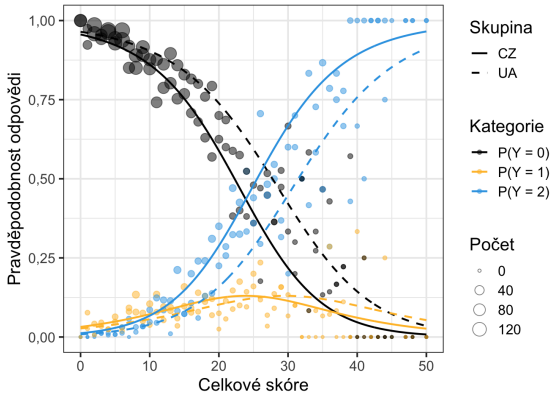


Obrázek 27: Zadání úlohy 9 v jednotné přijímací zkoušce z matematiky

Odlišné fungování bylo dále detekováno např. v geometrické konstrukční úloze 9 (obrázek 27), za kterou bylo možné získat maximálně 2 body. V této úloze měli nově příchozí ukrajinští žáci **menší** pravděpodobnost získání plného počtu bodů při stejném celkovém skóre než ostatní žáci (obrázek 28). Např. ukrajinský žák s celkovým skóre 40 měl pravděpodobnost získání 2 bodů cca 75 %, kdežto český žák se stejným skóre měl pravděpodobnost cca 90 % (obrázek 28). Důvodem může být opět rozdílné kurikulum, větší množství textu s obtížnější terminologií,[1] odlišnosti ve zvyklostech při rýsování nebo rozdíly v hodnocení[2] atp. Ačkoliv položka funguje odlišně pro české a ukrajinské žáky, testuje pouze znalost matematiky, a je tedy férová.

[1] Nově příchozí ukrajinští žáci psali test v ukrajinštině, položka ale mohla být obtížnější pro žáky, kteří před vypuknutím války používali jako primární jazyk ruštinu.

[2] Každé řešení sice hodnotí vždy dva hodnotitelé, ale hodnotitelé testů zadávaných v ukrajinštině se teoreticky mohou systematicky lišit od hodnotitelů testů zadávaných v češtině.



Obrázek 28: Rozdílně fungující položka 9, v níž měli ukrajinští žáci nižší pravděpodobnost zisku plného počtu bodů než ostatní žáci se stejným celkovým skóre

7 Závěr

Interaktivní aplikace `ShinyItemAnalysis` a přídavný modul EduTest Item Analysis zpřístupňují psychometrickou analýzu maturitních a dalších testů a jejich položek. Modul umožňuje nahrát volně dostupná neagregovaná maturitní data bez nutnosti jejich dalších úprav. V podzáložce CTT zpřístupňuje analýzu podobnou té, kterou běžně provádí CZVV v rámci tvorby podkladů pro validační komisi. Modul dále nabízí možnost upravit data do formátu vhodného pro aplikaci `ShinyItemAnalysis`, a využít tak dalších analýz, které tato aplikace nabízí, včetně dalších možností tradiční položkové analýzy, analýzy fungování položek pomocí regresních modelů, IRT modelů a skupinově specifických modelů vhodných pro testování odlišného fungování položek. V podzáložce IRT pak modul EduTest Item Analysis navíc zprostředkovává tzv. položkově specifický IRT model, což je funkcionalita, která zatím v hlavní aplikaci dostupná není.

Jelikož ve velmi podobném formátu jsou ukládána také data z testů jednotné přijímací zkoušky, téměř okamžitě je možné aplikaci a modul EduTest Item Analysis využít také pro psychometrickou analýzu těchto testů. Pro každý test je však potřeba vždy připravit soubor s metadaty popisující typy a vlastnosti jednotlivých položek.

Představený modul aplikace `ShinyItemAnalysis` je připraven především pro tvůrce testů (v tomto případě CZVV), kterým zpřístupňuje podrobnější psychometrickou analýzu. Výsledky této analýzy lze dále využít např. pro jednání validační komise aj. Nástroj však lze využít i pro analýzu dat CZVV k výzkumným účelům nebo pro výuku na vysokých školách.

Největší přínos aplikace lze spatřovat ve zpřístupnění regresních a IRT modelů a metod pro analýzu odlišného fungování položek. IRT modely poskytují oproti tradiční položkové analýze také hodnotu informační funkce a potažmo chyby měření pro každou

úroveň latentní znalosti, a tedy i pro každého respondenta. Z představených modelů je nejvíce flexibilní položkově specifický model, který umožňuje modelovat položky, kde se očekává hádání, 3PL modelem. Nutno podotknout, že zahraniční testové společnosti často využívají pouze IRT modely tzv. Raschova typu (např. pro binární data 1PL model, pro ordinální data tzv. model částečného kreditu *Partial Credit Model*), které předpokládají pro všechny položky stejnou diskriminační schopnost. Pro tyto modely platí, že celkové skóre je tzv. suficientní statistikou pro odhadované latentní skóre (viz např. Martinková & Hladká, 2023). To jinými slovy znamená, že dva respondenti se stejným celkovým skóre budou mít vždy stejnou hodnotu odhadnuté latentní znalosti, takže dosažené bodové skóre lze jednoznačně převést na IRT skóre. U ostatních typů modelů není vztah mezi hodnotami bodového skóre a IRT skóre jednoznačný, a proto jsou tyto modely méně vhodné pro využívání IRT skóre k hodnocení účastníků testování. Mohou však přinést užitečné informace o kvalitě testových položek i testu jako celku, které metody založené na klasické testové teorii neposkytují.

Analýza odlišného fungování položek (DIF) přidává novou informaci o fungování položek. V zahraničních testových agenturách je DIF analýza běžně používaná pro detekci potenciálně neférových položek, tedy položek, u kterých odlišnému fungování dochází z důvodu existence sekundární latentní ability, kterou však test nezamýšlel testovat. U položek, které vykazovaly DIF v rámci maturitních testů k takové situaci nedocházelo. DIF analýza však poukázala na silné a slabé stránky jednotlivých skupin, ať už to byly dvě skupiny SŠ oborů, vybraná škola vs. ostatní školy stejného typu v ČR, nebo ukrajinští vs. čeští studenti.

Aplikace `ShinyItemAnalysis` i představený přídavný modul EduTest Item Analysis lze do budoucna dále rozšiřovat. Lze např. implementovat další, pokročilejší, regresní i IRT modely. Konkrétně pak ordinální IRT modely, položkově specifické regresní modely, nebo regresní model, který se odhaduje pro všechny položky zároveň. Dále lze také vylepšovat automaticky generovaný PDF report.

Dále je také možné vytvořit obdobné moduly, které funkcionality interaktivní aplikace `ShinyItemAnalysis` zpřístupní pro jiné typy dat, např. testů zadávaných Českou školní inspekcí, a zpřístupní tak pokročilou psychometrickou analýzu již obsaženou v aplikaci, nebo další nadstavby i pro tyto další testy.

Literatura

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), 1–24. doi: 10.1002/j.2333-8504.1981.tb01255.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

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. doi: 10.1007/BF02291411

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement*, *54*(4), 498–517. doi: 10.1111/jedm.12158

Erosheva, E. A., Martinková, P., & Lee, C. J. (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(3), 904—919. doi: 10.1111/rssa.12681

Hladká, A., & Martinková, P. (2020). difNLR: Generalized logistic regression models for DIF and DDF detection. *The R Journal*, *12*(1), 300–323. doi: 10.32614/RJ-2020-014

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503–515. doi: 10.32614/rj-2018-074

Martinková, P., Drabinová, A., & Houdek, J. (2017). ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů [ShinyItemAnalysis: Analyzing admission and other educational and psychological tests]. *TESTFÓRUM*, *6*(9), 16–35. doi: 10.5817/TF2017-9-129

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE-Life Sciences Education*, *16*(2), rm2. doi: 10.1187/cbe.16-10-0307

Martinková, P., & Hladká, A. (2023). *Computational aspects of psychometric methods: With R*. Chapman and Hall/CRC. doi: 10.1201/9781003054313

Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, *66*, 101286. doi: 10.1016/j.learninstruc.2019.101286

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danmarks Paedagogiske Institut.

Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, *11*(19), 4104. doi: 10.3390/math11194104