Oponentský posudek habilitační práce RnDr. Patrície Martinkové, PhD. *Computational aspects of psychometric methods with R*

doc. Stanislav Ježek, Ph.D.

Předložená práce má podobu monografie určené všem odborníkům, kteří potřebují hodnotit psychometrické kvality měřících nástrojů, ať již v kontextu výzkumu či aplikované praxe. Monografie vyšla v anglickém jazyce v renomovaném nakladatelství a je koncipována tak, že může být užitečným zdrojem pro odborníky po celém světě. Monografie vznikla ve spoluautorství s Mgr. Adélou Hladkou, jejíž autorský podíl je v práci jasně deklarován jako minoritní, a tak podle mého mínění nic z této perspektivy nebrání akceptovat monografii jako habilitační práci.

Monografie svým zaměřením vhodně doplňuje existující psychometrickou literaturu. Zaměřuje se na výpočetní a praktický aspekt psychometriky, čímž vychází vstříc všem, kdo chtějí či potřebují hodnotit psychometrické vlastnosti měřících nástrojů, které znají z konceptuálně zaměřených textů, a hledají návod, jak analýzy realizovat. V tomto smyslu autorka nabízí čtenářům postupy analýzy ve prostředí jazyka R, které je zcela volně dostupné a má další výhodu v masivní komunitní podpoře. Práce dobře využívá toho, že veškeré analýzy se v R odehrávají prostřednictvím zadávání textových příkazů – čtenář tak má v doplňujících materiálech online k dispozici skripty ke všem analýzám v učebnici (a ještě k nějakým navíc). Pozitivně hodnotím také různorodost datových souborů, které autorka zvolila pro demonstraci analýz. Úlohy k procvičování (exercises) jsou určitě dobrý nápad, který by ale možná ještě stál za další rozpracování. V současné podobě se patrně počítá s přítomností vyučujícího, který může poskytnout k řešením zpětnou vazbu.

Zásadní součástí monografie je aplikace ShinyItemAnalysis a s ní spojená knihovna jazyka R. Ta umožňuje čtenáři ať už na svém počítači nebo na serveru AV ČR realizovat analýzy ještě jednodušeji, pouhým klikáním. Vedle toho poskytuje užitečné funkce, které při praktických analýzách výrazně usnadňují práci. Tento prvek práce je vlastně vynikající podobou psychometrické osvěty. Nejen v psychologii, ale i ve vzdělávání, sociální práci a zdravotnických oborech jsou každodenně administrována množství měřítek, která stojí na psychometrických základech. Často je tato praxe opřena pouze o závěry metodických psychometrických studií, které nemusí přesně odrážet konkrétní kontext nasazení nástroje – může se lišit populace, jazyk, doba apod. Byť uživatelé mohou mít pochyby ohledně fungování nástroje, často neumí své pochyby prověřit, i když k tomu třeba mají vhodná data. Podobně jsou na tom učitelé tvořící nejrůznější testy, od „opakovacích" až po poměrně high-stakes testy s nezanedbatelným vlivem na další život studenta. S pomocí ShinyItemAnalysis si mohou „nízkoprahově" vyzkoušet psychometrické analýzy svých testů a v případě podpory podezření vyhledat odborníka – psychometrika.

Dohromady je soubor monografie, online materiálů a aplikace ShinyItemAnalysis vynikajícím podkladem pro realizaci výpočetně zaměřených psychometrických kurzů. Soubor je vhodný i pro samouky, ale při takovém použití by bylo dobré mít vedle toho i konceptuálně zaměřenou učebnici psychometriky.

V této souvislosti vidím v aplikaci i monografii jednu rezervu, a tou je rozšíření interpretačních vodítek k jednotlivým modelům. Jsem si vědom, že to je zrádné, a autorka zjevně také, protože když nějaká vodítka uvádí, tak vždy neopomene uvést, že klasifikace hodnot toho kterého koeficientu na nízké, střední a vysoké je ryze arbitrární. S tím nelze než souhlasit. Uvažovaný „nízkoprahový" uživatel ale jiná vodítka nemá a mnohdy si ani nemusí uvědomit souvislosti mezi různými psychometrickými ukazateli navzájem a souvislosti s kontextem a cíli analýzy. V praxi se pak setkáváme s jevy jako je unáhlené vyřazování položek na základě toho, že některá z psychometrických vlastností je v arbitrární kategorizaci „špatná", nebo třeba naopak k paušálně pozitivnímu hodnocení nástroje podle ukazatelů fitu modelu, aniž by bylo přihlíženo k tomu, jestli je model a jeho parametry v souladu s teoretickými očekáváními.

Struktura monografie je dobře koncipovaná. Provádí čtenáře psychometrickými analýzami od těch nejjednodušších až po pokročilé aplikace jako je adaptivní testování. Zvlášť se mi líbí kapitola 6 představující regresní modely souvislosti mezi rysem/kritériem a odpověďmi na položky. Přijde mi to jako didakticky velmi efektivní most k IRT modelům. Čtenář se tak nejprve seznamuje s paletou možných parametrizací těchto souvislostí a až potom s tím, že prediktorem v této souvislosti může být latentní proměnná reprezentující měřenou charakteristiku a že tyto souvislosti mohou být modelovány všechny najednou v jednom modelu.

Naopak kapitolou, která by podle mého názoru stála za rozšíření, je kapitola s poněkud zavádějícím názvem „Validita". Po stránkové rekapitulaci tří různých druhů dokladů o validitě tvoří zbytek kapitoly zhruba dvacetistránkový přehled základů statistiky spolu s jejich realizací v jazyce R. I když akceptuji zaměření monografie na hodnocení měřících nástrojů, a ne na jejich tvorbu, je validita natolik zásadním tématem, že bych čekal její důkladnější propojení se statistickými postupy. Zůstanu-li u jádra této kapitoly, kterým je hledání souvislostí mezi skórem z testu a kritérii (tj. kriteriální validita), líbilo by se mi, kdyby si čtenář odnesl třeba také to, jaké hodnoty korelace mezi testem a kritériem lze považovat za doklad validity a jaké ne (a jak se to za různých okolností liší). Stejně tak by bylo dobré, aby si čtenář odnesl to, že p-hodnota z *t*-testu rozdílu dvou skupinových průměrů není sama o sobě dokladem validity, že dokladem je spíše dostatečná velikost tohoto rozdílu, třeba v metrice Cohenova *d*, a že to „dostatečná" je za různých okolností různé. Obecně je interpretace statistik založena na konfrontaci námi očekávaných hodnot statistik s hodnotami, k nimž jsme v analýze došli. Proto je nutné pomáhat čtenářům s vytvářením a kultivací takových očekávání. Pro matematika mohou být zčásti samozřejmá, ale sociálně-vědný uživatel velmi ocení explicitní instrukce, jak si očekávání vytvořit a jaký.

Dalším příkladem z mé perspektivy nenaplněného potenciálu jsou různé způsoby odhadu parametrů představené v kap. 2.4 (a dalších kapitolách). Jejich zařazení mi udělalo radost, protože toto téma je v sociálně-vědných úvodech do statistiky obvykle přeskakováno a statistiky jsou představovány spíše jako výsledky dosazení do vzorečku, a ne jako odhady čísel, která mají nějaké žádoucí vlastnosti. Autorka ale představuje princip OLS a ML odhadu způsobem, který je přístupnější lidem zběhlým v matematice, a už dál nepokračuje tím, proč je toto téma relevantní, v jakých situacích se s volbou způsobu odhadu bude čtenář setkávat, a co je při této volbě zvažovat. Přijde mi to škoda, protože zvažování různých způsobů odhadu je u psychometrických modelů v praxi velmi častým tématem.

Kapitola o interní struktuře testu představuje standardní analytické nástroje od korelací různých typů položek až po explorační a konfirmační faktorovu analýzu s vloženou shlukovou

analýzou. I zde dostává čtenář funkční mix matematických principů a praktických ukázek analýz v R (či v ShinyItemAnalysis). Také zde, možná více zřetelně než u jiných metod, mi chybí podpora pro rozhodování v té široké paletě voleb, které je zde třeba činit. Jaký způsob odhadu zvolit, jakou rotaci, jaký algoritmus shlukování ... a jak moc na těch volbách záleží? To jsou otázky, které mohou adepta na počátku úplně paralyzovat. A to ještě nejsme u interpretace – jaké korelace/faktorovou strukturu považovat za podporu validity a proč? Vykládám si to tak, že z perspektivy hodnocení již hotových metod je nejčastější otázkou unidimenzionalita, popř. jednoduchá struktura a nežádoucí odchylky od ní, ať již ve smyslu celého testu nebo jednotlivých položek. Složitější faktorové uvažování je možná důležitější v počátečních fázích vzniku testu. Pokud je to tak, asi by bylo vhodné to explicitně reflektovat. Zde by asi bylo jedno z mála míst, kde bych byl rád za korekci v aplikaci ShinyItemAnalysis, a to na dvou konkrétních místech: (1) chybí mi zobrazení korelační matice faktorů v případě použití šikmé rotace a (2) výchozí práh pro skrývání faktorových nábojů bych doporučil snížit z 0,3, popř. nastavit default na to náboje vůbec neskrývat a ponechat to jako možnost.

Kapitola o reliabilitě pěkně předkládá klasické způsoby reliability i jejich zobecnění v odhadech reliability pomocí složek rozptylu a v teorii zobecnitelnosti. Ukazuje, jak pro tento účel využít lineární mixed-effects modely, což umožňuje jednak poměrně snadno škálovat úvahy o reliabilitě a pak relativně snadno použít různé způsoby odhadu, včetně bayesovského, v situaci problémů s odhadem. Jediná věc, která mi zde chybí, je použití odhadů reliability ke stanovení intervalů spolehlivosti pro individuální skóry, protože tam je pro uživatele nedostatečná reliabilita nejzřetelněji viditelná. Celkově jde ale o výborné seznámení s konceptem reliability a jejího odhadování.

Kapitola o klasické položkové analýze představuje základní pojmy a postupy, které pak v rozvíjejí sofistikovanější modely v následujících kapitolách. Čtenář zde najde i zajímavé inovativní prvky, jako je hodnocení distraktorů u multiple-choice položek nikoli prostřednictvím průměrů celkového skóru/kritéria ve skupině lidí, kteří si daný distraktor zvolili, ale prostřednictvím proměňující se popularity distraktoru s rostoucím celkovým skórem/kritériem. Nabízím ke zvážení nahrazení potenciálně zavádějící (a zavedené) zkratky „item reliability" plným „item-deleted scale reliability". Také bych doporučit zdůraznit, že klasická položková analýza předpokládá unidimenzionalitu škály, kterou položky tvoří (u IRT modelů to zmíněno je). Jakkoli se to může zdát jako samozřejmost, viděl jsem řadu analýz, které navrhovaly vyřazování či úpravy položek na základě položkové analýzy explicitně multidimenzionální škály. Navíc, vedle konstatování je v tomto případě vhodné také uvést příklady toho, jak mohou různé podoby vícedimenzionality zkreslovat statistiky v položkové analýze.

Kapitoly 8 a 9 představují IRT modely pro různé typy položek a jsou těžištěm monografie. IRT modely jsou představeny teoreticky spolu s popisem různých způsobů odhadu parametrů i prakticky s využitím různých balíčků v R. Líbí se mi „bezešvé" propojení klasických IRT modelů a jejich parametrizace v rámci zobecněných lineárních mixed modelů. To zvyšuje šanci, že bude čtenář moci ke zvládnutí IRT modelů využít dřívější statistické znalosti a dovednosti. Co by možná ještě stálo za explicitní uchopení, je to, že parametry modelů odhadnuté různými způsoby se liší. Pro nezkušeného čtenáře může být tento fakt nepříjemný a může pro něj být obtížné rozlišit malé nevýznamné rozdíly v hodnotách parametrů od větších rozdílů, které mohou vést k úvaze o tom, zda algoritmus odhadu v něčem neselhal. Také by mi přišlo dobré více komunikovat, jak náročné jsou ty které modely na velikost

vzorku, a jaká jsou rizika spojená s malými vzorky jaké jsou běžné například u lokálně vyvíjených znalostních testů.

Kapitola o DIF analýze ukazuje autorčinu rozsáhnou zkušenost s IRT modely a tímto typem analýzy. Co mě však překvapilo je, že analogická aktivita v rámci konfirmačně faktorového modelu pod zastřešujícím termínem *measurement invariance* není vůbec zmíněn. Líbí se mi opět systematické předložení jednoduchých, regresních a model-based postupů. I zde by ale bylo hezké, kdyby bylo v textu zahrnuto nějaké shrnutí toho, jak si má uživatel v této paletě možností vybírat, popřípadě shrnutí výhod a nevýhod model-based postupů oproti analýzám jednotlivých položek. Z pohledu matematika se to možná může zdát triviální.

Závěrečná kapitola nabízí čtenáři především náhled na fungování adaptivního testování; ostatní témata jsou omezena na několik odstavců. I v tomto tématu je čtenáři k dispozici ShinyItemAnalysis, kde si lze interaktivně „pohrát" se simulací.

Shrnutí: Předložená práce je vlastně souborem monografie, online materiálů a interaktivní psychometrické aplikace a představuje významný příspěvek k rozšíření používání psychometrických analýz u nás i v zahraničí a tím snad i ke kvalitě používaných psychometrických měřících nástrojů. Celé dílo odráží vysokou psychometrickou odbornost i rozsáhlé zkušenosti autorky. Není pouze didaktickým dílem; obsahuje řadu autorských příspěvků k psychometrické praxi. V monografii najdou užitečné informace jak matematici či informatici, kteří mohou být pověřeni psychometrickými analýzami, ale i sociální vědci, kterým schází znalost toho, jak analýzy realizovat. Z perspektivy obou skupin by však bylo přínosné další rozšíření. V případě sociálních vědců by to určitě bylo o interpretační vodítka a vodítka pro rozhodování.

Celkově se domnívám, že autorka předloženou prací prokazuje vysokou odbornou i pedagogickou erudici. Doporučuji práci přijmout v předložené formě a na jejím základě doporučuji udělit titul docent.

V Brně, 6. března 2023

doc. Stanislav Ježek, Ph.D.