HANDBOOK OF ITEM RESPONSE THEORY
Recenze třídílného sborníku (van der Linden, 2016)

HYNEK CÍGLER
Katedra psychologie, Fakulta sociálních studií Masarykovy univerzity[1]

Abstrakt: *Teorie odpovědi na položku je jednou z dominantních psychometrických teorií měření v psychologii, pedagogickém výzkumu a sociálních vědách vůbec. Zatímco však v případě klasické testové teorie, teorie zobecnitelnosti a dalších „tradičních" teorií máme k dispozici řadu přehledových, autoritativních zdrojů, v případě teorie odpovědi na položku dosud podobná kniha chyběla. Nedávno vydaný třídílný* Handbook of Item Response Theory *(van der Linden, 2016) však více než zaplnil tuto mezeru. Na zhruba 1.500 stranách předkládá vynikající a hlavně komplexní shrnutí aktuálního stavu poznání, a to od jednotlivých modelů přes statistické postupy a estimační nástroje až po aplikaci a konkrétní příklady dobré praxe. Domnívám se, že přinejmenším první a třetí díl sborníku by neměl chybět v příruční knihovně žádného psychometrika či kvantitativně orientovaného sociálního vědce; své si v něm však nečekaně najdou i statistikové, informatici a další.*

titul	Handbook of Item Response Theory
podtitul	Volume 1: Models Volume 2: Statistical Tools Volume 3: Applications
editor	Wim J. van der Linden
rok	2016
vydavatel	Chapman & Hall / CRC
ISBN	978-0-3672-2120-1
rozsah	cca 1.500 stran ve třech dílech
cena	k 9. 7. 2020 cca £98 (2.900 Kč)

[1] Katedra psychologie, Fakulta sociálních studií, Masarykova univerzita. Joštova 10, 602 00 Brno.

Psychometrika je dynamicky rozvíjející se vědní disciplínou, a nejinak je tomu v případě široké palety postupů a statistických nástrojů, které lze souhrnně pojmenovat jako teorii odpovědi na položku (IRT; Item Response Theory). Zdroje staré více než deset let často bývají přinejmenším zčásti zastaralé, řada podstatných postupů v nich chybí úplně. Navíc v případě IRT dosud zcela chyběl nějaký komplexní přehled – na trhu byly dostupné spíše základní publikace a učebnice, které pokrývaly buď jen úvodní témata určená spíše začátečníkům (Bond & Fox, 2009; De Ayala, 2009; DeMars, 2010; Embretson & Reise, 2000), úzce zaměřené knihy pokrývající několik málo specializovaných témat (Leighton & Gierl, 2007; Reise & Revicki, 2015), nebo obecné učebnice psychometriky, které IRT zmiňovaly jen na okraj vedle jiných teorií a postupů. Tato mezera v odborné literatuře vyniká zejména při srovnání s jinými psychometrickými postupy disponujícími řadou skvělých publikací, jako třeba klasickou testovou teorií (Lord & Novick, 1968; Nunnally, 1978) či teorií zobecnitelnosti (Brennan, 2001).

Handbook of Item Response Theory (van der Linden, 2016) je relativně nový, třídílný sborník pokrývající většinu statistických modelů, nástrojů, postupů i praktických aplikací, které jsou dnes řazeny do souhrnného paradigmatu teorie odpovědi na položku. Cílem této recenze není popsat obsah celé publikace, což ostatně není vzhledem více než 1.500stránkovému rozsahu možné, ale spíše navnadit českého čtenáře popisem, proč by mohl a hlavně měl být sborník součástí jeho příruční knihovny. Následující recenze je rozdělena do tří kapitol podle jednotlivých dílů sborníku, na které navazuje závěrečný souhrn.

Volume 1: Models

Osobně považuji první a nejrozsáhlejší díl sborníku za stěžejní, a to jak díky jeho rozsahu a komplexitě, tak i díky přispěvatelům – mezi jeho 20 autory nechybí prakticky žádné z významných jmen současné psychometriky. Navíc všude, kde to bylo možné, napsali danou kapitolu sami původní autoři popisovaného IRT modelu (Samejima, Masters či Muraki), což skýtá možnost historického srovnání a určitého nadhledu.

Tento díl je rozdělen do sekcí podle jednotlivých druhů modelů. Před nimi je zařazen ještě krátký a vynikající úvod (van der Linden, vol. 1, 13–30)[2], kde je představen univerzální modelovací framework IRT a hlavně epistemologická východiska a předchůdci teorie, jako byli Alfred Bined či Louis Thurstone. Následující kapitoly jsou zpravidla velmi přehledné a srozumitelné, a to nejen díky shodné struktuře (historie, východiska, relevance a použití daného modelu, formální/statistická definice modelu, odhad parametrů a identifikace, hodnocení shody s daty a konečně i příklad z praxe), ale i díky oddělení „uživatelských“ pasáží psaných běžným jazykem a „technických“ pasáží zapsaných formálně v jazyku matematické statistiky.

[2] Odkazy na konkrétní kapitoly nejsou uváděny v seznamu literatury; namísto roku obsahují identifikátor dílu (např. vol. 1) a rozsah stran.

První dvě sekce pokrývají zdánlivě běžné binární, respektive nominální a ordinální IRT modely, přesto však stojí za zmínku. V první řadě nepoužívají tradiční dělení na jedno-, dvou-, tří-, atp. -parametrové modely (1PL–4PL), ale nabízejí univerzální definici charakteristické funkce 3PL modelu, ze které lze jednotlivé modely odvodit. Navíc zvažují povahu parametrů položek i osob, které lze považovat za náhodné (random) nebo pevné (fixed) efekty. Do kontrastu k nim je pak postaven Raschův model s radikálně odlišnými epistemologickými východisky. Kromě toho se sekce zaměřují i na méně známé modely, jako například Tutzův sekvenční model (Tutz, vol. 1, 139–152), nebo nahlížejí novou perspektivou na některé dobře zavedené – příkladem je faktorová analýza jako případ IRT modelu pro intervalové spojité proměnné (Mellenbergh, vol. 1, 153–166).

Třetí sekce knihy se zaměřuje na definici multidimenzionálních a multikomponentových modelů s různými východisky a předpoklady, zatímco v dalších sekcích jsou již představeny méně rozšířené či spíše neznámé modely. Chci upozornit zejména na celkem čtyři kapitoly věnované IRT modelům pro popis času a délky odpovídání, případně souběžné multivariační modelování času i správnosti odpovědi (van der Linden, vol. 1, 481–502).

Inspirativní pátá sekce se zaměřuje na neparametrické modely; vyjma tradičního Mokkenova škálování (Sijtsma a Molenaar, vol. 1, 303–322) nabízí rovněž i Bayesovské přístupy (Karabatsos, vol. 1, 323–336) či Ramsayovy křivky (Ramsay, vol. 1, 337–352), zatímco sekce šest představuje dva různé modely pro tzv. non-monotónní (unfolding) položky. V této značně nedoceněné rodině IRT modelů nejen vysoká, ale zároveň i nízká úroveň rysu vede ke spíše souhlasné odpovědi, zatímco ke spíše nesouhlasné odpovědi vede naopak průměrná úroveň rysu.

Sedmou sekci ocení zejména výzkumníci v oblasti pedagogiky, protože se věnuje hierarchickým a multilevel modelům – a to nejen z hlediska struktury respondentů, ale i položek v podobě item-family modelu (Glas a kol., vol. 1, 437–448) či modelů pro hierarchickou parametrizaci skórování volných výpovědí vícero hodnotiteli (Casabianca a kol., vol. 1, 449–466).

Poslední, osmá sekce prvního dílu pak nabízí několik generalizovaných a univerzálních modelovacích rámců, díky kterým získává sociální vědec plnou kontrolu nad definicí modelu, charakteristických funkcí položek či hierarchickou strukturou dat, a může si definovat model na míru svým potřebám. Patří sem např. univerzální framework Muthéna a Asparouhova (vol. 1, 527–540) tak, jak je implementovaný v programu Mplus, či explanační modely De Boecka (De Boeck a Wilson, vol. 1, 565–580) blízce příbuzné multikomponentovým a LLTM modelům, prezentovaným dříve (Janssen, vol. 1, 211–224).

Volume 2: Statistical Tools

Druhý díl sborníku je pro běžného sociálního vědce zřejmě nejvíce náročný, a zároveň nejméně užitečný pro běžnou výzkumnou praxi; ocení jej naopak programátoři či výzkumníci, kteří se zaměřují na technický vývoj statistických postupů a jejich softwarovou implementaci. Zaměřuje se totiž spíše na postupy odhadu parametrů položek a identifikace modelů, způsoby modelování chybějících dat, různé druhy statistických rozložení či zhodnocení shody modelů s daty. Autory kapitol jsou spíše statistici, informatici a celý díl je psaný technicky náročnějším jazykem. Zároveň však jde o nesmírně cenný přehled statistických nástrojů, který běžně nebývá takto souhrnně popsán. Navíc bych rád upozornil na několik konkrétních pasáží, které jsem shledal skutečně podnětnými.

Hned první sekce v první kapitole nabízí základní přehled odpověďových („link“) funkcí, které propojují neparametrickou nominální odpověď respondenta s parametrickou, latentní úrovní latentního rysu – tedy lineární, logistické či gaussovské link funkce (Albert, vol. 2, 3–22). Další kapitoly pak předkládají přehled diskrétních a multivariačních distribucí, exponenciálních distribucí, loglineárních modelů, vlastnosti součtů náhodných, nenormálně rozložených proměnných a v neposlední řadě pak i informační teorii (s využitím nejen v adaptivním počítačovém testování), což jsou oblasti, ve kterých běžných psycholog či psychometrik nemívá dostatečné vzdělání. Užitečná může být i kap. 2 (Casabianca a Junker, vol. 2, 23–34), která jakoby na okraj předkládá vztahy apriorních a posteriorních distribucí, což skýtá mnohá využití nejen v tzv. bayesovském modelování a výrazně zjednoduší život každému, kdo s ním právě začíná.

Druhá sekce knihy nabízí univerzální a generalizovaný pohled na identifikaci modelu a s tím spojené potíže, stejně jako způsoby práce s chybějícími daty. Ve třetí sekci se pak autoři zaměřují na problematiku odhadů parametrů, a to pomocí tradiční metody maximální věrohodnosti (maximum-likelihood), tak i modernějších postupů založených na EM algoritmu („expectation–maximization“), Monte-Carlo či bayesovských postupech. Za zmínku stojí zejm. teorie optimálního designu popsaná v 16. kapitole (Holling a Schwabe, vol. 2, 313–342), kterou využije každý, kdo se zabývá pilotáží a standardizací psychologických či didaktických testů. S její pomocí lze totiž navrhnout optimální výzkumný vzorek tak, aby byl optimalizovaný odhad zvolených parametrů položek či modelu (na úkor jiných, méně důležitých parametrů). To je pak rozvedeno praktickou ukázkou v dalším díle sborníku (van der Linden, vol. 3, 197–228).

Čtvrtá a poslední sekce se zaměřuje na shodu modelu s daty či vzájemné srovnání modelů, a to s použitím tradiční frekventistické statistiky, informačních kritérií, bayesovského přístupu či postupů založených na odhadu reziduí.

V této poslední sekci mi nicméně chyběl přístup s tzv. limitovanou informací („limited information approach“), založený na kolapsovaných univariačních či bivariačních reziduálních momentech a reprezentovaný zejména rodinou koeficientů M_2, M_2^* či C_2. Ty jsou v případě IRT modelů jednak vhodnější než tradiční přístup založený na plné

matici reziduí (χ^2 či G^2 statistiky), jednak jsou přímo srovnatelné s běžnými ukazateli dobré shody modelů faktorové analýzy. Čtenáře proto v tomto ohledu odkazuji na původní zdroje (Cai & Hansen, 2013; Albert Maydeu-Olivares & Joe, 2006; Alberto Maydeu-Olivares et al., 2011).

Dále mi pak chybělo pojetí tzv. ordinální faktorové analýzy jako gaussovské varianty multidimenzionálního graded-response modelu (GRM). Estimátor WLSMV nad maticí polychorických korelací, který tvoří základ tohoto přístupu, je jen okrajově zmíněn v prvním díle (Muthén a Asparouhov, vol. 1, 527–540) v souvislosti s programem Mplus, ačkoli je dnes velmi rozšířený. V tomto ohledu proto čtenáře rovněž odkazuji na externí literaturu (Asparouhov & Muthén, 2012, 2020; Forero & Maydeu-Olivares, 2009).

Volume 3: Applications

Třetí a poslední díl sborníku (který vyšel až v roce 2018) skýtá přehled možných využití, příkladů dobré praxe a konkrétních aplikací IRT. Své si v něm tak naleznou výzkumníci s rozličným zaměřením a specializací, a to včetně marketingových výzkumníků, pedagogických výzkumníků, kognitivních vědců a dalších.

V první sekci se čtenář dozví vše podstatné o kalibraci parametrů položek, jejich vyvažování napříč různými formami testů a způsoby administrace, analýzách diferenciálního fungování položek (DIF) či analýze (nechtěné) multidimenzionality; nutno však podotknout, že právě DIF analýza a celkově invariance měření je ve sborníku spíše opomenutým tématem; zájemce však snadno nalezne rozšiřující zdroje (např. Millsap, 2011; Millsap & Yun-Tein, 2004; Osterlind & Everson, 2009; Zumbo, 1999). Poměrně zajímavá je pak kap. 5 (Luecht, vol. 3, 87–106), která se soustředí na kalibraci dat získaných s využitím počítačové techniky, automatizovanou identifikaci ne-paralelních a diferenciálně fungujících položek a vůbec postupům automatizace nejen v pedagogickém testování.

Druhá sekce se zaměřuje na respondenty. Skýtá vyčerpávající přehled postupů pro skórování IRT testů, vyhodnocování shody modelu s odpověďmi respondentů (tzv. „person fit“ ukazatelů) za řadou rozličných účelů, jako je diagnostika úpadku pozornosti či koncentrace, podvádění a aberantní chování vůbec (Glas a Khalid, vol. 3, 107–126). Jako velmi přínosnou hodnotím rovněž hned následující sedmou kapitolu, ve které Hambleton a Zenisky (vol. 3, 127–142) předkládají postupy pro reportování a interpretaci IRT skórů a nastolují potřebné standardy. V tomto ohledu je více než zajímavé, že prakticky pomíjejí interpretaci založenou na pořadí – tedy v Česku tolik oblíbené percentily. Pro mě bylo navíc poněkud překvapivé, že zrovna do této sekce byla zařazena kap. 8 o vyvažování paralelních forem testu na základě pozorovaného skóre („observed-score equating“; van der Linden, vol. 3, 143–164). Má zde ovšem samozřejmě své místo, a to díky popisu lokálního vyvažování, korekcím vyvažovacích funkcí na chybu měření a dalším technikám založeným na odhadech latentních rysů respondentů v rámciIRT.

Třetí sekce opět nalézá využití zejm. v didaktickém testování a v tzv. „high-stake large-scale" testování. Kromě již zmíněného optimálního testového designu se totiž zaměřuje na postupy adaptivního testování a zejména pak zajištění standardního settingu, zvažování rychlosti práce a vůbec časové náročnosti testu (nejen) v didaktickém testování znalostí, kde časová omezení mohou snižovat konstruktovou validitu testu, a rovněž i postupy pro zajištění bezpečnosti testů a znění položek. Nejen tato kapitola by tak mohla sloužit jako inspirace pro české společnosti zabývající se testováním.

Poměrně zajímavý přehled, který zaujme zdaleka nejširší spektrum čtenářů, je čtvrtá sekce. Ta totiž předkládá konkrétní příklady dobré praxe použití IRT v nejrůznějších oblastech: popořadě v pedagogickém výzkumu a srovnávání skupin v „large-scale assessment" (Mazzeo, vol. 3, 297–312), psychologickém měření rysů (De Boeck, vol. 3, 313–328), kognitivní diagnostice a klasifikaci nejen v medicíně (Wang a Chang, vol. 3, 329–348), klinické diagnostice ve zdravotnictví (Gershon a kol., vol. 3, 349–364), aplikovaném marketingovém výzkumu (de Jong a Böckenholt, vol. 3, 365–386) a konečně při měření intraindividuální i skupinové změny s využitím specificky Raschova modelu (Fischer, 387–406).

Poslední, pátá sekce je potom přehledem dostupného IRT softwaru, na kterou jsem se osobně velmi těšil a možná i proto jsem byl silně zklamaný. Zdá se mi totiž, že některé kapitoly byly napsané před více roky a publikace se dočkaly se značným zpožděním. Řada informací je zastaralá, jiné důležité informace zcela chybějí. Pozornost je věnována některým minoritním programům či softwaru s velmi úzkým využitím (jako např. Firestar, WinGen aj.), zastaralým a spíše historickým nástrojům (BILOG, PARSCALE), zatímco řada moderních nástrojů zcela chybí. Nepříjemné to je zejm. v kap. 20 (Rusch a kol., vol. 3, 407–420), která představuje dostupné balíčky v prostředí R: navzdory extrémně rychlému vývoji je nejnovější citace z roku 2015, druhá z roku 2014 a další ještě starší, samotné R je citované ve verzi z roku 2012. Zcela tak chybí dnes zřejmě nejrozšířenější balíček mirt (Chalmers, 2012) s první verzí z roku 2011 nebo balíček lavaan (Rosseel, 2012) s podporou ordinálních dat a tedy i IRT od roku 2012. Obdobně chybí i lepší popis specifických bayesovských nástrojů pro Gibsovo vzorkování; zmíněna je pouze rodina programů BUGS, OpenBUGS, WinBUGS a JAGS. Naopak zcela chybí oblíbený program STAN (s první verzí z roku 2013) a rozhraní pro jeho ovládání z prostředí R.

Závěrečné shrnutí

Recenzovaný handbook rozhodně není vhodný pro „psychometrické začátečníky" s žádnými či minimálními zkušenostmi s teorií odpovědi na položku. Ty bych odkázal na základnější učebnice, zejména *Item Response Theory* od de Ayaly či jiné (např. Bond & Fox, 2009; De Ayala, 2009; DeMars, 2010; Embretson & Reise, 2000).

Pokud však základní znalosti máte a chcete si je prohloubit, pokud hledáte komplexní přehled IRT a reprezentativní, dostatečně obsáhlý autoritativní zdroj, či pokud prostě jen

potřebujete IRT pro vaši práci při vývoji testů různého druhu, pak je Handbook of Item Response Theory podle mého názoru tím nejlepším, co si můžete do své knihovny pořídit. Vybrané kapitoly poslouží také jako základní uvedení do problematiky a rozcestník na další zdroje při řešení prakticky všech otázek spojených s IRT, které je možné si představit. Upozorňuji, že použití není omezeno pouze na tradiční psychometriky s psychologickým vzděláním. Na své si přijdou statistici zabývající se sociálně-vědními daty, informatici či třeba kognitivní vědci, kteří se dotýkají témat jako je reakční čas a využívají typický experimentální výzkum s opakováním rozsáhlého množství podnětů různého druhu.

Málokdy se poštěstí mít valnou většinu základů některého rozsáhlejšího vědního oboru či teorie pohromadě v jediné publikaci. Domnívám se, že editorovi van der Lindenovi se to v tomto případě podařilo. Je sice nepravděpodobné, že celý handbook vydrží být aktuální přes 50 let, jako se to v případě klasické testové teorie poštěstilo knize Statistical Theories of Mental Test Scores Lorda a Novicka (1968) – myslím si ovšem, že i přes extrémně rychlý vývoj komputačních a statistických nástrojů kniha nezastará ještě pěkných pár let. Samozřejmě kromě představených softwarových nástrojů.

Reference

Asparouhov, T., & Muthén, B. (2012). *Comparison of computational methods for high dimensional item factor analysis*. https://www.statmodel.com/download/HighDimension11.pdf

Asparouhov, T., & Muthén, B. (2020). *IRT in Mplus. Version 2*. https://www.statmodel.com/download/MplusIRT.pdf

Bond, T. G., & Fox, C. M. (2009). *Applying the Rasch Model : Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates, Inc.

Brennan, R. l. (2001). *Generalizability Theory*. Springer-Verlag.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Chalmers, R. P. (2012). mirt : A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

DeMars, C. (2010). *Item Response Theory*. Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Inc.

Forero, C. G., & Maydeu-Olivares, A. (2009). *Estimation of IRT Graded Response Models: Limited Versus Full Information Methods*. https://doi.org/10.1037/a0015825.supp

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive Diagnostic Assessment for Education*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Maydeu-Olivares, Albert, & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Maydeu-Olivares, Alberto, Cai, L., & Hernández, A. (2011). Comparing the Fit of Item Response Theory and Factor Analysis Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333–356. https://doi.org/10.1080/10705511.2011.581993

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge, Taylor & Francis Group.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research*, *39*(3), 479–515. https://doi.org/10.1207/S15327906MBR3903_4

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage.

Reise, S. P., & Revicki, D. A. (Eds.). (2015). *Handbook of item response theory modeling : applications to typical performance assessment*. Routledge.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

van der Linden, W. J. (Ed.). (2016). *Handbook of Item Response Theory: Three Volume Set*. Chapman and Hall/CRC.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.